Plenary Sessions

Opening address

Some Aspects of the 'Phonetic Sciences', Past and Present

Eli Fischer-Jørgensen
Copenhagen, Denmark

Dames en heren,

Het is voor mij een grote eer en een bijzonder plezier hier in Nederland als eerste te spreken. Ik ben kort na de oorlog een half jaar in Nederland geweest, en die tijd behoort tot mijn beste herinneringen. Ik heb sindsdien een bijzondere sympathie bewaard voor het nederlandse landschap, de nederlandse kunst en de nederlandse mensen.

Mr. President, dear Colleagues,

I first want to thank the Committee for inviting me to give this talk. I feel it as a great honour, in fact as *too* great an honour. I know of various collegues who could have done it better, and I am somewhat ashamed that I accepted it. But, as I just mentioned, I have a soft spot in my heart for Holland. Moreover, that was two years ago, when I had just retired and thought that I would have plenty of time for reading and writing; perhaps I might even become more intelligent – who knows? But that was, of course, a vain hope. – Anyhow there are a few things I should like to say.

This is a sort of jubilee. It is the tenth International Congress of Phonetic Sciences, and it is approximately 50 years (more exactly 51 years) since the first congress took place in 1932, also in Holland.

It is true that on various occasions (1965 and 1982) Eberhard Zwirner has pointed to the fact that the congress in Amsterdam in 1932 was not really the first International Congress of Phonetics: there was one in 1914 (but due to the war no proceedings were ever published), and there was one again in 1930 in Bonn. That is correct, but these were congresses of experimental phonetics, whereas the congress in Amsterdam was the first congress of what was called 'the phonetic sciences', and that makes a difference.

It was not by chance that Holland was chosen as the place for the congress in 1932. Holland has a long and rich tradition in phonetics. One of the most impressive older works is the book by Petrus Montanus van Delft in 1635: 'Bericht van een nieuw konst genaemt de spreeckonst', a remarkable and very original work, which has rarely met with the appreciation it deserves, perhaps because it was written in Dutch and, moreover, used a forbidding terminology. In the first decades of this century, thus in the years before the

congress in Amsterdam, Holland had become an important centre of phonetic research with a number of very competent phoneticians, for instance Zwaardemaker, Eijkman, van Ginneken, and Louise Kaiser. Zwaardemaker and Eijkman had published an excellent textbook – or rather handbook – of phonetics in 1928 with original contributions on many points. The new phonological theories had also been quickly – but not uncritically – accepted in Holland, for instance by De Groot and Van Wijk. A few years later (1932) Van Wijk published an introduction to phonology which was less dogmatic and much easier to read than Trubetzkoy's Grundzüge, and which might have made phonology more popular if it had been written in e.g. English. As early as 1914 a Dutch society for experimental phonetics had been founded, which in 1931 was transformed into a Society for Phonetics. Dutch phoneticians also published a periodical, 'Archives néerlandaises de phonétique expérimentale' (from 1927) which in the first years exclusively, and later to a large extent was based on contributions from Dutch phoneticians, and the University of Amsterdam had a lecturer in phonetics (Louise Kaiser) from 1926.

This brilliant tradition has continued to the present day with phonetic research centers and excellent phoneticians at various universities and at the Institute for Perception Research in Eindhoven. Their contributions are well known. I will therefore only mention that, although several Dutch phoneticians must have been very busy organizing this congress, there are more than forty section papers by Dutch phoneticians. It is thus not simply for sentimental reasons that this tenth congress is also being held in Holland. It is scientifically very well motivated.

The congress in Amsterdam in 1932 was originally – like those in 1914 and 1930 – planned as a congress on experimental phonetics. But the Dutch committee widened its scope on the initiative of its chairman, the psychologist Van Ginneken. Van Ginneken was an impressive personality, and his appearance was impressive too (for instance, he had long hair long before its time); and he was a man of vision. Some of them were rather wild, but some were fruitful. One of them was that all those who were interested in any aspect of speech sounds should meet and work together. Therefore invitations were sent out to a broad spectrum of scholars from different sciences, and the name of the congress changed to 'congress of phonetic sciences'. The topics of the congress were announced to be: physiology of speech and voice, the development of speech and voice in the individual and in mankind, anthropology of speech and voice, phonology, linguistic psychology, pathology of speech and voice, comparative physiology of the sounds of animals, and musicology; and the congress program included a meeting of the so-called 'Internationale phonologische Arbeitsgemeinschaft'. But shortly after the invitations had been sent out, the International Society of Experimental Phonetics which had taken the original initiative gave up participating as a society because its president, E. Scripture, was afraid that the economic crisis would prevent too many members from coming. The committee, however, continued its work with Louise Kaiser as general secretary.

I do not think that the name 'phonetic sciences' is good terminology but it may be viewed as shorthand for 'disciplines' (like phonetics and phonology) which have the speech sound as their main object, plus various sciences which among other objects include some aspects of the speech sound, like physiology, acoustics, psychology, etc. And at least it was clear what the committee intended, and since both title and intention have been kept since then, it was a very important decision. It was also a very good idea to bring various groups of people together just at that time. In the thirties there was not much contact between different sciences interested in speech sounds, and between the more closely related approaches there was even suspicion and antagonism. The adherents of classical phonetics regarded the use of instruments with pronounced scepticism and, on the other hand some experimental phoneticians, like Scripture, rejected everything that was not expressed in figures. He considered non-experimental phonetics an illusion and 'the investigator', he said, 'might be, and preferably should be, congenitally deaf and totally ignorant of any notions concerning sound and speech' (1936). Panconcelli-Calzia had also emphasized that the language spoken by the subject was irrelevant. The phonetician was only interested in their vocal tracts. He considered phonetics as belonging to the natural sciences.

The Prague phonologists accepted this view of phonetics, describing it as a science which investigated sounds, irrespective of their function, whereas phonology described the functional aspect of sounds and belonged to the humanities. By this claim and also by emphasizing that phonology was something quite new they succeeded in offending both the adherents of classical phonetics, who had always, more or less explicitly, taken the communicative function of speech sounds into account, and the more linguistically orientated experimental phoneticians.

The congress in Amsterdam, which, like the next two congresses, had only plenary sessions, managed to bring people together, but you still feel a certain tension in the reports of the discussions. I think it was not until the third congress in Ghent, which was the first congress I attended, that there was a real breakthrough in the understanding between phonologists and phoneticians, owing particularly to the contributions by Zwirner, Roman Jakobson and Van Wijk. Nowadays, these old antagonisms are forgotten. Everybody recognizes that phoneticians must use instruments and that speech sounds must be studied from both a material and a functional point of view (although this mutual recognition does not always include close cooperation). But as late as in the fifties there were still linguistic centers in Europe where phonology (and structural linguistics on the whole) was regarded as a new and dangerous heresy, where you saw smiles fade away and faces getting a very rigid expression of you dared to admit that you found these trends interesting, and where young linguists who were interested in them had to hold clandestine meetings.

In America the development was much more harmonious because it was for many years dominated by Bloomfield, for whom phonetics and phonology were complementary approaches.

It is a good thing that the wide scientific scope of the congresses has been retained. But of course they have changed in character during these fifty years.

In the first place there has been an enormous increase in the number of participants and of papers. At the first congress there were 136 participants. During the following congresses the number increased slowly to almost 300, with a sudden jump up to about 550 at the fifth congress in Prague in 1967, followed by a more steady increase to the approximately 650 members of this congress, five times as many as at the first congress. The number of papers has increased even more: from 40 in 1932 to about 100 at the fourth congress and then growing rapidly to the almost 400 section papers of this congress, apart from symposia and plenary lectures; and the number of authors has grown even more, since now one third of the papers are the result of team work, whereas in 1932 all papers had only a single author.

The large number of members and papers of course causes various inconveniences. You can only attend a small fraction of the meetings you find interesting; and it may be difficult to get into contact with the people you want to meet. On the other hand, I find that these big open congresses serve a useful purpose. It is important to have a forum where people from different fields can meet, and it is important to have congresses that are open to everybody interested. The smaller conferences may give more scientific output, but generally only established scholars are invited. The big open congresses offer the only possibility for young phoneticians from various countries to meet each other and older colleagues.

The enormous increase of papers reflects a general explosive growth in phonetic publications. Thirty years ago it was still possible to read the more important publications in the whole field. Now it is not even possible to keep up with the literature within one's own special field of interest. I think the moment has come where it would be extremely useful to start a journal of abstracts in phonetics and phonology with competent contributors, who could tell what is new and valuable in a paper. And it could also be useful if the phonetic journals would include surveys of specific areas at regular intervals perhaps dividing the work among themselves.

There has, of course, also been a change in emphasis as far as the subjects treated are concerned. A good deal of the change can be ascribed to technological progress. There was from the start an interest in the acoustic and perceptual aspects, but the possibilities of research were modest. At the beginning of the century it could take hours to analyse a single cycle. Nevertheless, there were patient scholars who undertook this work, but not many. At the first congress there were only two papers on acoustics and none on perception. At this congress there is a very large number of papers dealing with both these subjects. I had not expected the increase in papers on acoustic phonetics to have taken place until the first congress that was held after the war, in 1961. As a matter of fact, the increase took place at the congress in 1938 in Ghent, where about 17 percent of the papers dealt with acoustic phonetics compared to 5 percent in 1935.

The explanation is probably that in the mid thirties a number of instruments for acoustic investigation were constructed, mostly by German engineers, and most of them were demonstrated in Ghent. But then the war broke out, and after the war new instruments were built, mostly by Swedish and American engineers, partly according to the same principles, but much handier and easier to use, and one may tend to forget the achievements of the thirties.

The progress in acoustic phonetics, and particularly the possibility of speech synthesis, gave a new impetus to the study of speech perception and a better basis for the study of prosodic phenomena, and this is reflected in the congress papers after the war. At the same time there was an obvious decrease in the study of speech production, reflected in a small number of papers within this field at the first congresses after the war. In the beginning of the seventies this changed again. I do not think this was simply a consequence of the invention of new transducers and a better EMG-technique. It may have been the other way round . It had become possible, particularly due to the work of Fant, Stevens and others, to relate details of production to the acoustic results, and thus production came into focus again as a very important step in the communication chain. The causal relations within this chain are now central topics in phonetic research, including the discussion of models for both production and perception. The brain is still a missing link in this chain, although we know more than we did a few years ago. We may at least hope that neurophonetics may be a central topic at the next congress.

The fact that the proceedings of the first congresses contain a number of papers treating phonetics from a biological point of view probably had a rather specific explanation, namely the interests of the first president of the international council, Van Ginneken. There is, for instance, at the first congress an informative paper by Negus describing the larynx of various species of animals, ending with the human larynx and Van Ginneken himself developed one of his more fantastic theories about the heredity of speech sounds. He believed, and even considered it as proven, that all phonological systems and moreover the relative frequency of speech sounds can be explained by Mendel's laws of heredity, according to the pattern: a man sho has **k** as only consonant marries a woman who has **m** as only consonant, and each of their children will then inherit one of the sounds **k,m,p,ŋ** distributed according to Mendel's laws, and learn the others from their sisters and brothers. This theory was not pursued, and biological considerations did not play any role at later congresses. They have come up again at this congress, but in a quite different form.

Other changes during the 50 years were rather conditioned by the shift in dominating trends in linguistics as part of shifts in the general cultural pattern and philosophical approach of the period. These shifts were, of course, in the first place influential for phonology (and up till the ninth congress about 20 percent of the papers dealt with phonological problems), but also for the relations between phonology and phonetics.

During the first thirty years the dominant linguistic trend was structuralism. In Europe it was mainly represented by Prague phonology with its emphasis on phonological oppositions and phonological systems, aiming at a general typology and involving the demonstration of universal tendencies. Roman Jakobson's distinctive feature theory was a further development of this trend. Prague phonology was dominant on the European continent in the beginning of the period; later the extreme formalism of glossematics had a certain influence but never gained many real adherents. In Great Britain most phoneticians adhered to Daniel Jones' practical approach, or else to Firth's prosodic phonology.

Whereas Prague phonology was accused (by Doroszweski at the first congress) of 'platonism with 2400 centuries' delay', this could not be said of American structuralism, which was deeply rooted in behaviourism and was principally interested in finding waterproof methods for setting up the phonemes of a language and stating their possibility of combination, but not in systems or universal tendencies. Transformational grammar including generative phonology was in the first place a reaction against American structuralism, a widening of the perspective by taking account of the cognitive functions of the human mind and attempting to set up an explanatory theory. But the exclusively morphophonemic approach of generative phonology with underlying forms and derivation by explicit, ordered rules and with abolition of a separate phoneme level had a sweeping success, also in Europe.

At the moment there is no dominating school of phonology, but a number of new, partly more concrete and surface oriented trends: natural phonology, metrical phonology, lexical phonology, autosegmental phonology, dependency phonology, etc. Some may find that this is a deplorable disintegration. But it may also be seen as a sign of more independent thinking, and these approaches may all contribute to a deepening of our insight into the function of language. They are, to a large extent, complementary descriptions of the same linguistic data.

A feature common to American structuralism and generative phonology was that the role ascribed to general phonetics was rather modest, its main task being to deliver the phonetic categories used to identify the contrastive segments and features. For this purpose auditory identification was generally considered sufficient. Phonetics was not asked to contribute to the explanation of phonological systems or developments. American structuralism was, on the whole, suspicious of explanations, and the explanatory procedure of generative phonology was extremely abstract, based on notational conventions implying that fewer symbols were used for natural rules. Glossematics accepted only purely formal explanations, whereas the Prague School looked for explanation in an interplay between formal and phonetic factors. But structural explanations were preferred.

Once the phonological structure of the individual language was set up, the primary task of the phonetician was to analyze the phonetic manifestation of the contrastive segments and features, which were supposed to contain invariant properties.

This assumption proved fruitful in giving rise to a whole trend in phonetic research – the search for the invariant. It was clear from the very start of the period, at least after Menzerath's studies of coarticulation, that it could not be found in speech production. Then it was looked for in acoustics, and some still hope to find it there, but at least it was not very obvious. The next hope was the invariant motor command, and this hope contributed to the renewed interest in speech production and particularly in EMG, and gave rise to the motor theory of speech perception. Unfortunately, however, the electromyographic recordings generally showed different innervations for different variants. We must look higher up for invariants. Perhaps Martin Joos (1948) was right in assuming that we have stored invariant phonemes in the brain, but in the production of a concrete word the overlapping innervation waves are combined already in the cerebellum or perhaps at a still higher level. We still do not know that. Perhaps we may also store dyads or words. – Anyhow, as emphasized recently by Lindblom (1982), one should not look for invariance, only for what he calls 'perceptual equivalence', since the speaker is aware of the fact that listening is an active process and that the listener does not need all the cues for individual phonemes in order to identify a word. This is also confirmed by various papers on word recognition at this congress.

Other papers point to the enormous variability of speech. Different languages use different production processes to attain almost identical sounds, different individuals use different muscles to produce the same acoustic results, and different perceptual strategies to analyse the acoustic stimuli. Moreover, modern phonological and sociophonetic studies emphasize the heterogeneous character of the speech community and the possibility of individual speakers having different norms. This is an important condition for sound change, which was stressed – in the fifties – by Fónagy and now, combined with the idea of natural selection, by Lindblom.

On the whole, there is at present an increasing reaction to a purely formal approach, a renewed interest in the concrete speech performance, in the biological and social embedding of language, and in language history. The isolationism of structural and transformational grammar was perhaps a necessary step in the development of linguistics, but in the long run it was detrimental to progress.

This sets new tasks for general phonetics, in particular the contribution to a better understanding of the structure of phonological systems and their development. Lindblom, who has emphasized this repeatedly, has taken up the old idea, expressed explicitly by Passy and Jespersen, and in more elaborate terms by Martinet, of an intended balance between articulatory economy and sufficient perceptual contrast. What is new and exciting in his approach is the attempt to obtain a quantitative formulation of this balance, based on extensive research. This will certainly lead to a better understanding of universal tendencies, but I do not believe that it is possible to reach exhaustive causal explanations, not to speak of predictions, of concrete

changes, which are always due to an intricate interplay of physical, physiological, communitive, and social factors.

Phonetics is, according to its subject (the speech sounds, that is: sounds with a communicative function) part of linguistics. The deductive nomological causal explanation as used in natural sciences can, as far as I can see, only be applied to limited areas of phonetic research, for instance the relation between articulation and its acoustic results, not to typology or phonological change. Here we must be content with statistical and teleological explanations.

The task of explanation requires close cooperation between phonetics and phonology. It is therefore deplorable that the participation of phonologists has decreased so drastically at the present congress. The phonetician describing a concrete language does not need to know the subtleties of different phonological theories, but at least the basic principles, and particularly for the description of prosodic facts quite a bit of linguistic insight is required. The phonetician who wants to explain things must also know a good deal about language typology. On the other hand, phonology needs phonetics, not only for identifying sounds but also for the purpose of explanation.

Lastly let me point to a similarity between the first congress and the tenth, a similarity in the conditions for research. Both congresses take place in a time of economic crisis and in a very tense and threatening political situation. The two things may not be unrelated. There is an old English saying: 'When poverty comes in at the door, love flies out at the window'. The economic crisis is oppressive, but it is not yet as bad as in the thirties. In a paper from the first congress it is said, for instance, that no normal phonetics laboratory can afford buying and using an oscillograph. A phonetic crisis may hamper research, - it cannot stop it. I cannot help thinking of Marguerite Durand, who did excellent phonetic research using on old kymograph which would only start moving when you had thrown a pail of water on the rope connecting it to the motor. We can do with poverty, but we cannot do without love.

The political situation is still more threatening than it was in the thirties, and I think some of us now and then ask ourselves if it really makes sense doing phonetic research if our whole civilisation is doomed, - whether it is not a more urgent task to try to improve mutual understanding and confidence among people. Perhaps it is. However: Man is certainly the most destructive of animals, and perhaps he does not deserve to survive. On the other hand, he is also the most constructive animal, the most creative; and if we give up creating art and seeking truth, do we not then betray just that which gives us a sort of moral right to survive? That which makes us human?

Therefore, let us leave these gloomy thoughts and start our discussions. And an international congress has, after all, not only the purpose of promoting science, but also the purpose of promoting mutual understanding. I wish the tenth congress of phonetic sciences much success in both respects!

Acknowledgements:

I am grateful to Jørgen Rischel for valuable suggestions concerning the content of this talk, and to Elizabeth Uldall for improving my English style.

References

(References are only given for a few more concrete points in the talk)

Proceedings of the International Congress of Phonetic Sciences, 3-8 July 1932. *Archives néerlandaises de phonétique expérimentale* VIII-IX 1933, 92-312.

Delft, Petrus Montanus v. (1635). *Bericht van een nieuwe konst genaemt de spreeckonst.*

Ginneken, J. v. (1932). La tendance labiale de la race méditerraneé et la tendance laryngale de la race alpinée. *Proceedings of the international congress of phonetic sciences, 3-8 July 1932. Archives néerlandaises de phonétique expérimentale* VIII-IX, 1933, 167-83.

Joos, M. (1948). Acoustic Phonetics. *Language monograph,* **23** 109ff.

Lindblom, B. (1982). The interdisciplinary challenge of speech motor control. *Wenner-Gren Center International Symposium Series,* **36**: Speech Motor Control, 1-18.

Scripture, E. (1936). In: Bulletin of the internation society of experimental phonetics III. *Archives néerlandaises de phonétique expérimentale* **XII**, 135.

Wijk, N. v. (1939). *Phonologie*: een hoofdstuk uit de structurele taalwetenschap. The Hague.

Zwaardemaker, H. and Eijkman, L.P.H. (1928). *Leerboek der Phonetiek.* Haarlem.

Zwirner, E. (1965). Eröffnungsansprache. *Proceedings of the Fifth International Congress of Phonetic Sciences*, Munster 16-22 August 1964.

Zwirner E. (1982). 50 Jahre 'Permanent Council', *Phonetica* **39**, 374-378.

Opening address

Some Aspects of the 'Phonetic Sciences', Past and Present

Eli Fischer-Jørgensen
Copenhagen, Denmark

Dames en heren,

Het is voor mij een grote eer en een bijzonder plezier hier in Nederland als eerste te spreken. Ik ben kort na de oorlog een half jaar in Nederland geweest, en die tijd behoort tot mijn beste herinneringen. Ik heb sindsdien een bijzondere sympathie bewaard voor het nederlandse landschap, de nederlandse kunst en de nederlandse mensen.

Mr. President, dear Colleagues,

I first want to thank the Committee for inviting me to give this talk. I feel it as a great honour, in fact as *too* great an honour. I know of various collegues who could have done it better, and I am somewhat ashamed that I accepted it. But, as I just mentioned, I have a soft spot in my heart for Holland. Moreover, that was two years ago, when I had just retired and thought that I would have plenty of time for reading and writing; perhaps I might even become more intelligent – who knows? But that was, of course, a vain hope. – Anyhow there are a few things I should like to say.

This is a sort of jubilee. It is the tenth International Congress of Phonetic Sciences, and it is approximately 50 years (more exactly 51 years) since the first congress took place in 1932, also in Holland.

It is true that on various occasions (1965 and 1982) Eberhard Zwirner has pointed to the fact that the congress in Amsterdam in 1932 was not really the first International Congress of Phonetics: there was one in 1914 (but due to the war no proceedings were ever published), and there was one again in 1930 in Bonn. That is correct, but these were congresses of experimental phonetics, whereas the congress in Amsterdam was the first congress of what was called 'the phonetic sciences', and that makes a difference.

It was not by chance that Holland was chosen as the place for the congress in 1932. Holland has a long and rich tradition in phonetics. One of the most impressive older works is the book by Petrus Montanus van Delft in 1635: 'Bericht van een nieuw konst genaemt de spreeckonst', a remarkable and very original work, which has rarely met with the appreciation it deserves, perhaps because it was written in Dutch and, moreover, used a forbidding terminology. In the first decades of this century, thus in the years before the

congress in Amsterdam, Holland had become an important centre of phonetic research with a number of very competent phoneticians, for instance Zwaardemaker, Eijkman, van Ginneken, and Louise Kaiser. Zwaardemaker and Eijkman had published an excellent textbook – or rather handbook – of phonetics in 1928 with original contributions on many points. The new phonological theories had also been quickly – but not uncritically – accepted in Holland, for instance by De Groot and Van Wijk. A few years later (1932) Van Wijk published an introduction to phonology which was less dogmatic and much easier to read than Trubetzkoy's Grundzüge, and which might have made phonology more popular if it had been written in e.g. English. As early as 1914 a Dutch society for experimental phonetics had been founded, which in 1931 was transformed into a Society for Phonetics. Dutch phoneticians also published a periodical, 'Archives néerlandaises de phonétique expérimentale' (from 1927) which in the first years exclusively, and later to a large extent was based on contributions from Dutch phoneticians, and the University of Amsterdam had a lecturer in phonetics (Louise Kaiser) from 1926.

This brilliant tradition has continued to the present day with phonetic research centers and excellent phoneticians at various universities and at the Institute for Perception Research in Eindhoven. Their contributions are well known. I will therefore only mention that, although several Dutch phoneticians must have been very busy organizing this congress, there are more than forty section papers by Dutch phoneticians. It is thus not simply for sentimental reasons that this tenth congress is also being held in Holland. It is scientifically very well motivated.

The congress in Amsterdam in 1932 was originally – like those in 1914 and 1930 – planned as a congress on experimental phonetics. But the Dutch committee widened its scope on the initiative of its chairman, the psychologist Van Ginneken. Van Ginneken was an impressive personality, and his appearance was impressive too (for instance, he had long hair long before its time); and he was a man of vision. Some of them were rather wild, but some were fruitful. One of them was that all those who were interested in any aspect of speech sounds should meet and work together. Therefore invitations were sent out to a broad spectrum of scholars from different sciences, and the name of the congress changed to 'congress of phonetic sciences'. The topics of the congress were announced to be: physiology of speech and voice, the development of speech and voice in the individual and in mankind, anthropology of speech and voice, phonology, linguistic psychology, pathology of speech and voice, comparative physiology of the sounds of animals, and musicology; and the congress program included a meeting of the so-called 'Internationale phonologische Arbeitsgemeinschaft'. But shortly after the invitations had been sent out, the International Society of Experimental Phonetics which had taken the original initiative gave up participating as a society because its president, E. Scripture, was afraid that the economic crisis would prevent too many members from coming. The committee, however, continued its work with Louise Kaiser as general secretary.

I do not think that the name 'phonetic sciences' is good terminology but it may be viewed as shorthand for 'disciplines' (like phonetics and phonology) which have the speech sound as their main object, plus various sciences which among other objects include some aspects of the speech sound, like physiology, acoustics, psychology, etc. And at least it was clear what the committee intended, and since both title and intention have been kept since then, it was a very important decision. It was also a very good idea to bring various groups of people together just at that time. In the thirties there was not much contact between different sciences interested in speech sounds, and between the more closely related approaches there was even suspicion and antagonism. The adherents of classical phonetics regarded the use of instruments with pronounced scepticism and, on the other hand some experimental phoneticians, like Scripture, rejected everything that was not expressed in figures. He considered non-experimental phonetics an illusion and 'the investigator', he said, 'might be, and preferably should be, congenitally deaf and totally ignorant of any notions concerning sound and speech' (1936). Panconcelli-Calzia had also emphasized that the language spoken by the subject was irrelevant. The phonetician was only interested in their vocal tracts. He considered phonetics as belonging to the natural sciences.

The Prague phonologists accepted this view of phonetics, describing it as a science which investigated sounds, irrespective of their function, whereas phonology described the functional aspect of sounds and belonged to the humanities. By this claim and also by emphasizing that phonology was something quite new they succeeded in offending both the adherents of classical phonetics, who had always, more or less explicitly, taken the communicative function of speech sounds into account, and the more linguistically orientated experimental phoneticians.

The congress in Amsterdam, which, like the next two congresses, had only plenary sessions, managed to bring people together, but you still feel a certain tension in the reports of the discussions. I think it was not until the third congress in Ghent, which was the first congress I attended, that there was a real breakthrough in the understanding between phonologists and phoneticians, owing particularly to the contributions by Zwirner, Roman Jakobson and Van Wijk. Nowadays, these old antagonisms are forgotten. Everybody recognizes that phoneticians must use instruments and that speech sounds must be studied from both a material and a functional point of view (although this mutual recognition does not always include close cooperation). But as late as in the fifties there were still linguistic centers in Europe where phonology (and structural linguistics on the whole) was regarded as a new and dangerous heresy, where you saw smiles fade away and faces getting a very rigid expression of you dared to admit that you found these trends interesting, and where young linguists who were interested in them had to hold clandestine meetings.

In America the development was much more harmonious because it was for many years dominated by Bloomfield, for whom phonetics and phonology were complementary approaches.

It is a good thing that the wide scientific scope of the congresses has been retained. But of course they have changed in character during these fifty years.

In the first place there has been an enormous increase in the number of participants and of papers. At the first congress there were 136 participants. During the following congresses the number increased slowly to almost 300, with a sudden jump up to about 550 at the fifth congress in Prague in 1967, followed by a more steady increase to the approximately 650 members of this congress, five times as many as at the first congress. The number of papers has increased even more: from 40 in 1932 to about 100 at the fourth congress and then growing rapidly to the almost 400 section papers of this congress, apart from symposia and plenary lectures; and the number of authors has grown even more, since now one third of the papers are the result of team work, whereas in 1932 all papers had only a single author.

The large number of members and papers of course causes various inconveniences. You can only attend a small fraction of the meetings you find interesting; and it may be difficult to get into contact with the people you want to meet. On the other hand, I find that these big open congresses serve a useful purpose. It is important to have a forum where people from different fields can meet, and it is important to have congresses that are open to everybody interested. The smaller conferences may give more scientific output, but generally only established scholars are invited. The big open congresses offer the only possibility for young phoneticians from various countries to meet each other and older colleagues.

The enormous increase of papers reflects a general explosive growth in phonetic publications. Thirty years ago it was still possible to read the more important publications in the whole field. Now it is not even possible to keep up with the literature within one's own special field of interest. I think the moment has come where it would be extremely useful to start a journal of abstracts in phonetics and phonology with competent contributors, who could tell what is new and valuable in a paper. And it could also be useful if the phonetic journals would include surveys of specific areas at regular intervals perhaps dividing the work among themselves.

There has, of course, also been a change in emphasis as far as the subjects treated are concerned. A good deal of the change can be ascribed to technological progress. There was from the start an interest in the acoustic and perceptual aspects, but the possibilities of research were modest. At the beginning of the century it could take hours to analyse a single cycle. Nevertheless, there were patient scholars who undertook this work, but not many. At the first congress there were only two papers on acoustics and none on perception. At this congress there is a very large number of papers dealing with both these subjects. I had not expected the increase in papers on acoustic phonetics to have taken place until the first congress that was held after the war, in 1961. As a matter of fact, the increase took place at the congress in 1938 in Ghent, where about 17 percent of the papers dealt with acoustic phonetics compared to 5 percent in 1935.

The explanation is probably that in the mid thirties a number of instruments for acoustic investigation were constructed, mostly by German engineers, and most of them were demonstrated in Ghent. But then the war broke out, and after the war new instruments were built, mostly by Swedish and American engineers, partly according to the same principles, but much handier and easier to use, and one may tend to forget the achievements of the thirties.

The progress in acoustic phonetics, and particularly the possibility of speech synthesis, gave a new impetus to the study of speech perception and a better basis for the study of prosodic phenomena, and this is reflected in the congress papers after the war. At the same time there was an obvious decrease in the study of speech production, reflected in a small number of papers within this field at the first congresses after the war. In the beginning of the seventies this changed again. I do not think this was simply a consequence of the invention of new transducers and a better EMG-technique. It may have been the other way round . It had become possible, particularly due to the work of Fant, Stevens and others, to relate details of production to the acoustic results, and thus production came into focus again as a very important step in the communication chain. The causal relations within this chain are now central topics in phonetic research, including the discussion of models for both production and perception. The brain is still a missing link in this chain, although we know more than we did a few years ago. We may at least hope that neurophonetics may be a central topic at the next congress.

The fact that the proceedings of the first congresses contain a number of papers treating phonetics from a biological point of view probably had a rather specific explanation, namely the interests of the first president of the international council, Van Ginneken. There is, for instance, at the first congress an informative paper by Negus describing the larynx of various species of animals, ending with the human larynx and Van Ginneken himself developed one of his more fantastic theories about the heredity of speech sounds. He believed, and even considered it as proven, that all phonological systems and moreover the relative frequency of speech sounds can be explained by Mendel's laws of heredity, according to the pattern: a man sho has **k** as only consonant marries a woman who has **m** as only consonant, and each of their children will then inherit one of the sounds **k,m,p,ŋ** distributed according to Mendel's laws, and learn the others from their sisters and brothers. This theory was not pursued, and biological considerations did not play any role at later congresses. They have come up again at this congress, but in a quite different form.

Other changes during the 50 years were rather conditioned by the shift in dominating trends in linguistics as part of shifts in the general cultural pattern and philosophical approach of the period. These shifts were, of course, in the first place influential for phonology (and up till the ninth congress about 20 percent of the papers dealt with phonological problems), but also for the relations between phonology and phonetics.

During the first thirty years the dominant linguistic trend was structuralism. In Europe it was mainly represented by Prague phonology with its emphasis on phonological oppositions and phonological systems, aiming at a general typology and involving the demonstration of universal tendencies. Roman Jakobson's distinctive feature theory was a further development of this trend. Prague phonology was dominant on the European continent in the beginning of the period; later the extreme formalism of glossematics had a certain influence but never gained many real adherents. In Great Britain most phoneticians adhered to Daniel Jones' practical approach, or else to Firth's prosodic phonology.

Whereas Prague phonology was accused (by Doroszweski at the first congress) of 'platonism with 2400 centuries' delay', this could not be said of American structuralism, which was deeply rooted in behaviourism and was principally interested in finding waterproof methods for setting up the phonemes of a language and stating their possibility of combination, but not in systems or universal tendencies. Transformational grammar including generative phonology was in the first place a reaction against American structuralism, a widening of the perspective by taking account of the cognitive functions of the human mind and attempting to set up an explanatory theory. But the exclusively morphophonemic approach of generative phonology with underlying forms and derivation by explicit, ordered rules and with abolition of a separate phoneme level had a sweeping success, also in Europe.

At the moment there is no dominating school of phonology, but a number of new, partly more concrete and surface oriented trends: natural phonology, metrical phonology, lexical phonology, autosegmental phonology, dependency phonology, etc. Some may find that this is a deplorable disintegration. But it may also be seen as a sign of more independent thinking, and these approaches may all contribute to a deepening of our insight into the function of language. They are, to a large extent, complementary descriptions of the same linguistic data.

A feature common to American structuralism and generative phonology was that the role ascribed to general phonetics was rather modest, its main task being to deliver the phonetic categories used to identify the contrastive segments and features. For this purpose auditory identification was generally considered sufficient. Phonetics was not asked to contribute to the explanation of phonological systems or developments. American structuralism was, on the whole, suspicious of explanations, and the explanatory procedure of generative phonology was extremely abstract, based on notational conventions implying that fewer symbols were used for natural rules. Glossematics accepted only purely formal explanations, whereas the Prague School looked for explanation in an interplay between formal and phonetic factors. But structural explanations were preferred.

Once the phonological structure of the individual language was set up, the primary task of the phonetician was to analyze the phonetic manifestation of the contrastive segments and features, which were supposed to contain invariant properties.

This assumption proved fruitful in giving rise to a whole trend in phonetic research – the search for the invariant. It was clear from the very start of the period, at least after Menzerath's studies of coarticulation, that it could not be found in speech production. Then it was looked for in acoustics, and some still hope to find it there, but at least it was not very obvious. The next hope was the invariant motor command, and this hope contributed to the renewed interest in speech production and particularly in EMG, and gave rise to the motor theory of speech perception. Unfortunately, however, the electromyographic recordings generally showed different innervations for different variants. We must look higher up for invariants. Perhaps Martin Joos (1948) was right in assuming that we have stored invariant phonemes in the brain, but in the production of a concrete word the overlapping innervation waves are combined already in the cerebellum or perhaps at a still higher level. We still do not know that. Perhaps we may also store dyads or words. – Anyhow, as emphasized recently by Lindblom (1982), one should not look for invariance, only for what he calls 'perceptual equivalence', since the speaker is aware of the fact that listening is an active process and that the listener does not need all the cues for individual phonemes in order to identify a word. This is also confirmed by various papers on word recognition at this congress.

Other papers point to the enormous variability of speech. Different languages use different production processes to attain almost identical sounds, different individuals use different muscles to produce the same acoustic results, and different perceptual strategies to analyse the acoustic stimuli. Moreover, modern phonological and sociophonetic studies emphasize the heterogeneous character of the speech community and the possibility of individual speakers having different norms. This is an important condition for sound change, which was stressed – in the fifties – by Fónagy and now, combined with the idea of natural selection, by Lindblom.

On the whole, there is at present an increasing reaction to a purely formal approach, a renewed interest in the concrete speech performance, in the biological and social embedding of language, and in language history. The isolationism of structural and transformational grammar was perhaps a necessary step in the development of linguistics, but in the long run it was detrimental to progress.

This sets new tasks for general phonetics, in particular the contribution to a better understanding of the structure of phonological systems and their development. Lindblom, who has emphasized this repeatedly, has taken up the old idea, expressed explicitly by Passy and Jespersen, and in more elaborate terms by Martinet, of an intended balance between articulatory economy and sufficient perceptual contrast. What is new and exciting in his approach is the attempt to obtain a quantitative formulation of this balance, based on extensive research. This will certainly lead to a better understanding of universal tendencies, but I do not believe that it is possible to reach exhaustive causal explanations, not to speak of predictions, of concrete

changes, which are always due to an intricate interplay of physical, physiological, communitive, and social factors.

Phonetics is, according to its subject (the speech sounds, that is: sounds with a communicative function) part of linguistics. The deductive nomological causal explanation as used in natural sciences can, as far as I can see, only be applied to limited areas of phonetic research, for instance the relation between articulation and its acoustic results, not to typology or phonological change. Here we must be content with statistical and teleological explanations.

The task of explanation requires close cooperation between phonetics and phonology. It is therefore deplorable that the participation of phonologists has decreased so drastically at the present congress. The phonetician describing a concrete language does not need to know the subtleties of different phonological theories, but at least the basic principles, and particularly for the description of prosodic facts quite a bit of linguistic insight is required. The phonetician who wants to explain things must also know a good deal about language typology. On the other hand, phonology needs phonetics, not only for identifying sounds but also for the purpose of explanation.

Lastly let me point to a similarity between the first congress and the tenth, a similarity in the conditions for research. Both congresses take place in a time of economic crisis and in a very tense and threatening political situation. The two things may not be unrelated. There is an old English saying: 'When poverty comes in at the door, love flies out at the window'. The economic crisis is oppressive, but it is not yet as bad as in the thirties. In a paper from the first congress it is said, for instance, that no normal phonetics laboratory can afford buying and using an oscillograph. A phonetic crisis may hamper research, - it cannot stop it. I cannot help thinking of Marguerite Durand, who did excellent phonetic research using on old kymograph which would only start moving when you had thrown a pail of water on the rope connecting it to the motor. We can do with poverty, but we cannot do without love.

The political situation is still more threatening than it was in the thirties, and I think some of us now and then ask ourselves if it really makes sense doing phonetic research if our whole civilisation is doomed, - whether it is not a more urgent task to try to improve mutual understanding and confidence among people. Perhaps it is. However: Man is certainly the most destructive of animals, and perhaps he does not deserve to survive. On the other hand, he is also the most constructive animal, the most creative; and if we give up creating art and seeking truth, do we not then betray just that which gives us a sort of moral right to survive? That which makes us human?

Therefore, let us leave these gloomy thoughts and start our discussions. And an international congress has, after all, not only the purpose of promoting science, but also the purpose of promoting mutual understanding. I wish the tenth congress of phonetic sciences much success in both respects!

Acknowledgements:

I am grateful to Jørgen Rischel for valuable suggestions concerning the content of this talk, and to Elizabeth Uldall for improving my English style.

References

(References are only given for a few more concrete points in the talk)

Proceedings of the International Congress of Phonetic Sciences, 3-8 July 1932. *Archives néerlandaises de phonétique expérimentale* VIII-IX 1933, 92-312.

Delft, Petrus Montanus v. (1635). *Bericht van een nieuwe konst genaemt de spreeckonst.*

Ginneken, J. v. (1932). La tendance labiale de la race méditerraneé et la tendance laryngale de la race alpinée. *Proceedings of the international congress of phonetic sciences, 3-8 July 1932. Archives néerlandaises de phonétique expérimentale* VIII-IX, 1933, 167-83.

Joos, M. (1948). Acoustic Phonetics. *Language monograph,* **23** 109ff.

Lindblom, B. (1982). The interdisciplinary challenge of speech motor control. *Wenner-Gren Center International Symposium Series,* **36**: Speech Motor Control, 1-18.

Scripture, E. (1936). In: Bulletin of the internation society of experimental phonetics III. *Archives néerlandaises de phonétique expérimentale* **XII**, 135.

Wijk, N. v. (1939). *Phonologie*: een hoofdstuk uit de structurele taalwetenschap. The Hague.

Zwaardemaker, H. and Eijkman, L.P.H. (1928). *Leerboek der Phonetiek.* Haarlem.

Zwirner, E. (1965). Eröffnungsansprache. *Proceedings of the Fifth International Congress of Phonetic Sciences*, Munster 16-22 August 1964.

Zwirner E. (1982). 50 Jahre 'Permanent Council', *Phonetica* **39**, 374-378.

Opening address

Some Aspects of the 'Phonetic Sciences', Past and Present

Eli Fischer-Jørgensen
Copenhagen, Denmark

Dames en heren,

Het is voor mij een grote eer en een bijzonder plezier hier in Nederland als eerste te spreken. Ik ben kort na de oorlog een half jaar in Nederland geweest, en die tijd behoort tot mijn beste herinneringen. Ik heb sindsdien een bijzondere sympathie bewaard voor het nederlandse landschap, de nederlandse kunst en de nederlandse mensen.

Mr. President, dear Colleagues,

I first want to thank the Committee for inviting me to give this talk. I feel it as a great honour, in fact as *too* great an honour. I know of various collegues who could have done it better, and I am somewhat ashamed that I accepted it. But, as I just mentioned, I have a soft spot in my heart for Holland. Moreover, that was two years ago, when I had just retired and thought that I would have plenty of time for reading and writing; perhaps I might even become more intelligent – who knows? But that was, of course, a vain hope. – Anyhow there are a few things I should like to say.

This is a sort of jubilee. It is the tenth International Congress of Phonetic Sciences, and it is approximately 50 years (more exactly 51 years) since the first congress took place in 1932, also in Holland.

It is true that on various occasions (1965 and 1982) Eberhard Zwirner has pointed to the fact that the congress in Amsterdam in 1932 was not really the first International Congress of Phonetics: there was one in 1914 (but due to the war no proceedings were ever published), and there was one again in 1930 in Bonn. That is correct, but these were congresses of experimental phonetics, whereas the congress in Amsterdam was the first congress of what was called 'the phonetic sciences', and that makes a difference.

It was not by chance that Holland was chosen as the place for the congress in 1932. Holland has a long and rich tradition in phonetics. One of the most impressive older works is the book by Petrus Montanus van Delft in 1635: 'Bericht van een nieuw konst genaemt de spreeckonst', a remarkable and very original work, which has rarely met with the appreciation it deserves, perhaps because it was written in Dutch and, moreover, used a forbidding terminology. In the first decades of this century, thus in the years before the

congress in Amsterdam, Holland had become an important centre of phonetic research with a number of very competent phoneticians, for instance Zwaardemaker, Eijkman, van Ginneken, and Louise Kaiser. Zwaardemaker and Eijkman had published an excellent textbook – or rather handbook – of phonetics in 1928 with original contributions on many points. The new phonological theories had also been quickly – but not uncritically – accepted in Holland, for instance by De Groot and Van Wijk. A few years later (1932) Van Wijk published an introduction to phonology which was less dogmatic and much easier to read than Trubetzkoy's Grundzüge, and which might have made phonology more popular if it had been written in e.g. English. As early as 1914 a Dutch society for experimental phonetics had been founded, which in 1931 was transformed into a Society for Phonetics. Dutch phoneticians also published a periodical, 'Archives néerlandaises de phonétique expérimentale' (from 1927) which in the first years exclusively, and later to a large extent was based on contributions from Dutch phoneticians, and the University of Amsterdam had a lecturer in phonetics (Louise Kaiser) from 1926.

This brilliant tradition has continued to the present day with phonetic research centers and excellent phoneticians at various universities and at the Institute for Perception Research in Eindhoven. Their contributions are well known. I will therefore only mention that, although several Dutch phoneticians must have been very busy organizing this congress, there are more than forty section papers by Dutch phoneticians. It is thus not simply for sentimental reasons that this tenth congress is also being held in Holland. It is scientifically very well motivated.

The congress in Amsterdam in 1932 was originally – like those in 1914 and 1930 – planned as a congress on experimental phonetics. But the Dutch committee widened its scope on the initiative of its chairman, the psychologist Van Ginneken. Van Ginneken was an impressive personality, and his appearance was impressive too (for instance, he had long hair long before its time); and he was a man of vision. Some of them were rather wild, but some were fruitful. One of them was that all those who were interested in any aspect of speech sounds should meet and work together. Therefore invitations were sent out to a broad spectrum of scholars from different sciences, and the name of the congress changed to 'congress of phonetic sciences'. The topics of the congress were announced to be: physiology of speech and voice, the development of speech and voice in the individual and in mankind, anthropology of speech and voice, phonology, linguistic psychology, pathology of speech and voice, comparative physiology of the sounds of animals, and musicology; and the congress program included a meeting of the so-called 'Internationale phonologische Arbeitsgemeinschaft'. But shortly after the invitations had been sent out, the International Society of Experimental Phonetics which had taken the original initiative gave up participating as a society because its president, E. Scripture, was afraid that the economic crisis would prevent too many members from coming. The committee, however, continued its work with Louise Kaiser as general secretary.

I do not think that the name 'phonetic sciences' is good terminology but it may be viewed as shorthand for 'disciplines' (like phonetics and phonology) which have the speech sound as their main object, plus various sciences which among other objects include some aspects of the speech sound, like physiology, acoustics, psychology, etc. And at least it was clear what the committee intended, and since both title and intention have been kept since then, it was a very important decision. It was also a very good idea to bring various groups of people together just at that time. In the thirties there was not much contact between different sciences interested in speech sounds, and between the more closely related approaches there was even suspicion and antagonism. The adherents of classical phonetics regarded the use of instruments with pronounced scepticism and, on the other hand some experimental phoneticians, like Scripture, rejected everything that was not expressed in figures. He considered non-experimental phonetics an illusion and 'the investigator', he said, 'might be, and preferably should be, congenitally deaf and totally ignorant of any notions concerning sound and speech' (1936). Panconcelli-Calzia had also emphasized that the language spoken by the subject was irrelevant. The phonetician was only interested in their vocal tracts. He considered phonetics as belonging to the natural sciences.

The Prague phonologists accepted this view of phonetics, describing it as a science which investigated sounds, irrespective of their function, whereas phonology described the functional aspect of sounds and belonged to the humanities. By this claim and also by emphasizing that phonology was something quite new they succeeded in offending both the adherents of classical phonetics, who had always, more or less explicitly, taken the communicative function of speech sounds into account, and the more linguistically orientated experimental phoneticians.

The congress in Amsterdam, which, like the next two congresses, had only plenary sessions, managed to bring people together, but you still feel a certain tension in the reports of the discussions. I think it was not until the third congress in Ghent, which was the first congress I attended, that there was a real breakthrough in the understanding between phonologists and phoneticians, owing particularly to the contributions by Zwirner, Roman Jakobson and Van Wijk. Nowadays, these old antagonisms are forgotten. Everybody recognizes that phoneticians must use instruments and that speech sounds must be studied from both a material and a functional point of view (although this mutual recognition does not always include close cooperation). But as late as in the fifties there were still linguistic centers in Europe where phonology (and structural linguistics on the whole) was regarded as a new and dangerous heresy, where you saw smiles fade away and faces getting a very rigid expression of you dared to admit that you found these trends interesting, and where young linguists who were interested in them had to hold clandestine meetings.

In America the development was much more harmonious because it was for many years dominated by Bloomfield, for whom phonetics and phonology were complementary approaches.

It is a good thing that the wide scientific scope of the congresses has been retained. But of course they have changed in character during these fifty years.

In the first place there has been an enormous increase in the number of participants and of papers. At the first congress there were 136 participants. During the following congresses the number increased slowly to almost 300, with a sudden jump up to about 550 at the fifth congress in Prague in 1967, followed by a more steady increase to the approximately 650 members of this congress, five times as many as at the first congress. The number of papers has increased even more: from 40 in 1932 to about 100 at the fourth congress and then growing rapidly to the almost 400 section papers of this congress, apart from symposia and plenary lectures; and the number of authors has grown even more, since now one third of the papers are the result of team work, whereas in 1932 all papers had only a single author.

The large number of members and papers of course causes various inconveniences. You can only attend a small fraction of the meetings you find interesting; and it may be difficult to get into contact with the people you want to meet. On the other hand, I find that these big open congresses serve a useful purpose. It is important to have a forum where people from different fields can meet, and it is important to have congresses that are open to everybody interested. The smaller conferences may give more scientific output, but generally only established scholars are invited. The big open congresses offer the only possibility for young phoneticians from various countries to meet each other and older colleagues.

The enormous increase of papers reflects a general explosive growth in phonetic publications. Thirty years ago it was still possible to read the more important publications in the whole field. Now it is not even possible to keep up with the literature within one's own special field of interest. I think the moment has come where it would be extremely useful to start a journal of abstracts in phonetics and phonology with competent contributors, who could tell what is new and valuable in a paper. And it could also be useful if the phonetic journals would include surveys of specific areas at regular intervals perhaps dividing the work among themselves.

There has, of course, also been a change in emphasis as far as the subjects treated are concerned. A good deal of the change can be ascribed to technological progress. There was from the start an interest in the acoustic and perceptual aspects, but the possibilities of research were modest. At the beginning of the century it could take hours to analyse a single cycle. Nevertheless, there were patient scholars who undertook this work, but not many. At the first congress there were only two papers on acoustics and none on perception. At this congress there is a very large number of papers dealing with both these subjects. I had not expected the increase in papers on acoustic phonetics to have taken place until the first congress that was held after the war, in 1961. As a matter of fact, the increase took place at the congress in 1938 in Ghent, where about 17 percent of the papers dealt with acoustic phonetics compared to 5 percent in 1935.

The explanation is probably that in the mid thirties a number of instruments for acoustic investigation were constructed, mostly by German engineers, and most of them were demonstrated in Ghent. But then the war broke out, and after the war new instruments were built, mostly by Swedish and American engineers, partly according to the same principles, but much handier and easier to use, and one may tend to forget the achievements of the thirties.

The progress in acoustic phonetics, and particularly the possibility of speech synthesis, gave a new impetus to the study of speech perception and a better basis for the study of prosodic phenomena, and this is reflected in the congress papers after the war. At the same time there was an obvious decrease in the study of speech production, reflected in a small number of papers within this field at the first congresses after the war. In the beginning of the seventies this changed again. I do not think this was simply a consequence of the invention of new transducers and a better EMG-technique. It may have been the other way round . It had become possible, particularly due to the work of Fant, Stevens and others, to relate details of production to the acoustic results, and thus production came into focus again as a very important step in the communication chain. The causal relations within this chain are now central topics in phonetic research, including the discussion of models for both production and perception. The brain is still a missing link in this chain, although we know more than we did a few years ago. We may at least hope that neurophonetics may be a central topic at the next congress.

The fact that the proceedings of the first congresses contain a number of papers treating phonetics from a biological point of view probably had a rather specific explanation, namely the interests of the first president of the international council, Van Ginneken. There is, for instance, at the first congress an informative paper by Negus describing the larynx of various species of animals, ending with the human larynx and Van Ginneken himself developed one of his more fantastic theories about the heredity of speech sounds. He believed, and even considered it as proven, that all phonological systems and moreover the relative frequency of speech sounds can be explained by Mendel's laws of heredity, according to the pattern: a man sho has **k** as only consonant marries a woman who has **m** as only consonant, and each of their children will then inherit one of the sounds **k,m,p,ŋ** distributed according to Mendel's laws, and learn the others from their sisters and brothers. This theory was not pursued, and biological considerations did not play any role at later congresses. They have come up again at this congress, but in a quite different form.

Other changes during the 50 years were rather conditioned by the shift in dominating trends in linguistics as part of shifts in the general cultural pattern and philosophical approach of the period. These shifts were, of course, in the first place influential for phonology (and up till the ninth congress about 20 percent of the papers dealt with phonological problems), but also for the relations between phonology and phonetics.

During the first thirty years the dominant linguistic trend was structuralism. In Europe it was mainly represented by Prague phonology with its emphasis on phonological oppositions and phonological systems, aiming at a general typology and involving the demonstration of universal tendencies. Roman Jakobson's distinctive feature theory was a further development of this trend. Prague phonology was dominant on the European continent in the beginning of the period; later the extreme formalism of glossematics had a certain influence but never gained many real adherents. In Great Britain most phoneticians adhered to Daniel Jones' practical approach, or else to Firth's prosodic phonology.

Whereas Prague phonology was accused (by Doroszweski at the first congress) of 'platonism with 2400 centuries' delay', this could not be said of American structuralism, which was deeply rooted in behaviourism and was principally interested in finding waterproof methods for setting up the phonemes of a language and stating their possibility of combination, but not in systems or universal tendencies. Transformational grammar including generative phonology was in the first place a reaction against American structuralism, a widening of the perspective by taking account of the cognitive functions of the human mind and attempting to set up an explanatory theory. But the exclusively morphophonemic approach of generative phonology with underlying forms and derivation by explicit, ordered rules and with abolition of a separate phoneme level had a sweeping success, also in Europe.

At the moment there is no dominating school of phonology, but a number of new, partly more concrete and surface oriented trends: natural phonology, metrical phonology, lexical phonology, autosegmental phonology, dependency phonology, etc. Some may find that this is a deplorable disintegration. But it may also be seen as a sign of more independent thinking, and these approaches may all contribute to a deepening of our insight into the function of language. They are, to a large extent, complementary descriptions of the same linguistic data.

A feature common to American structuralism and generative phonology was that the role ascribed to general phonetics was rather modest, its main task being to deliver the phonetic categories used to identify the contrastive segments and features. For this purpose auditory identification was generally considered sufficient. Phonetics was not asked to contribute to the explanation of phonological systems or developments. American structuralism was, on the whole, suspicious of explanations, and the explanatory procedure of generative phonology was extremely abstract, based on notational conventions implying that fewer symbols were used for natural rules. Glossematics accepted only purely formal explanations, whereas the Prague School looked for explanation in an interplay between formal and phonetic factors. But structural explanations were preferred.

Once the phonological structure of the individual language was set up, the primary task of the phonetician was to analyze the phonetic manifestation of the contrastive segments and features, which were supposed to contain invariant properties.

This assumption proved fruitful in giving rise to a whole trend in phonetic research – the search for the invariant. It was clear from the very start of the period, at least after Menzerath's studies of coarticulation, that it could not be found in speech production. Then it was looked for in acoustics, and some still hope to find it there, but at least it was not very obvious. The next hope was the invariant motor command, and this hope contributed to the renewed interest in speech production and particularly in EMG, and gave rise to the motor theory of speech perception. Unfortunately, however, the electromyographic recordings generally showed different innervations for different variants. We must look higher up for invariants. Perhaps Martin Joos (1948) was right in assuming that we have stored invariant phonemes in the brain, but in the production of a concrete word the overlapping innervation waves are combined already in the cerebellum or perhaps at a still higher level. We still do not know that. Perhaps we may also store dyads or words. – Anyhow, as emphasized recently by Lindblom (1982), one should not look for invariance, only for what he calls 'perceptual equivalence', since the speaker is aware of the fact that listening is an active process and that the listener does not need all the cues for individual phonemes in order to identify a word. This is also confirmed by various papers on word recognition at this congress.

Other papers point to the enormous variability of speech. Different languages use different production processes to attain almost identical sounds, different individuals use different muscles to produce the same acoustic results, and different perceptual strategies to analyse the acoustic stimuli. Moreover, modern phonological and sociophonetic studies emphasize the heterogeneous character of the speech community and the possibility of individual speakers having different norms. This is an important condition for sound change, which was stressed – in the fifties – by Fónagy and now, combined with the idea of natural selection, by Lindblom.

On the whole, there is at present an increasing reaction to a purely formal approach, a renewed interest in the concrete speech performance, in the biological and social embedding of language, and in language history. The isolationism of structural and transformational grammar was perhaps a necessary step in the development of linguistics, but in the long run it was detrimental to progress.

This sets new tasks for general phonetics, in particular the contribution to a better understanding of the structure of phonological systems and their development. Lindblom, who has emphasized this repeatedly, has taken up the old idea, expressed explicitly by Passy and Jespersen, and in more elaborate terms by Martinet, of an intended balance between articulatory economy and sufficient perceptual contrast. What is new and exciting in his approach is the attempt to obtain a quantitative formulation of this balance, based on extensive research. This will certainly lead to a better understanding of universal tendencies, but I do not believe that it is possible to reach exhaustive causal explanations, not to speak of predictions, of concrete

changes, which are always due to an intricate interplay of physical, physiological, communitive, and social factors.

Phonetics is, according to its subject (the speech sounds, that is: sounds with a communicative function) part of linguistics. The deductive nomological causal explanation as used in natural sciences can, as far as I can see, only be applied to limited areas of phonetic research, for instance the relation between articulation and its acoustic results, not to typology or phonological change. Here we must be content with statistical and teleological explanations.

The task of explanation requires close cooperation between phonetics and phonology. It is therefore deplorable that the participation of phonologists has decreased so drastically at the present congress. The phonetician describing a concrete language does not need to know the subtleties of different phonological theories, but at least the basic principles, and particularly for the description of prosodic facts quite a bit of linguistic insight is required. The phonetician who wants to explain things must also know a good deal about language typology. On the other hand, phonology needs phonetics, not only for identifying sounds but also for the purpose of explanation.

Lastly let me point to a similarity between the first congress and the tenth, a similarity in the conditions for research. Both congresses take place in a time of economic crisis and in a very tense and threatening political situation. The two things may not be unrelated. There is an old English saying: 'When poverty comes in at the door, love flies out at the window'. The economic crisis is oppressive, but it is not yet as bad as in the thirties. In a paper from the first congress it is said, for instance, that no normal phonetics laboratory can afford buying and using an oscillograph. A phonetic crisis may hamper research, - it cannot stop it. I cannot help thinking of Marguerite Durand, who did excellent phonetic research using on old kymograph which would only start moving when you had thrown a pail of water on the rope connecting it to the motor. We can do with poverty, but we cannot do without love.

The political situation is still more threatening than it was in the thirties, and I think some of us now and then ask ourselves if it really makes sense doing phonetic research if our whole civilisation is doomed, - whether it is not a more urgent task to try to improve mutual understanding and confidence among people. Perhaps it is. However: Man is certainly the most destructive of animals, and perhaps he does not deserve to survive. On the other hand, he is also the most constructive animal, the most creative; and if we give up creating art and seeking truth, do we not then betray just that which gives us a sort of moral right to survive? That which makes us human?

Therefore, let us leave these gloomy thoughts and start our discussions. And an international congress has, after all, not only the purpose of promoting science, but also the purpose of promoting mutual understanding. I wish the tenth congress of phonetic sciences much success in both respects!

Acknowledgements:

I am grateful to Jørgen Rischel for valuable suggestions concerning the content of this talk, and to Elizabeth Uldall for improving my English style.

References

(References are only given for a few more concrete points in the talk)

Proceedings of the International Congress of Phonetic Sciences, 3-8 July 1932. *Archives néerlandaises de phonétique expérimentale* VIII-IX 1933, 92-312.

Delft, Petrus Montanus v. (1635). *Bericht van een nieuwe konst genaemt de spreeckonst*.

Ginneken, J. v. (1932). La tendance labiale de la race méditerraneé et la tendance laryngale de la race alpinée. *Proceedings of the international congress of phonetic sciences, 3-8 July 1932. Archives néerlandaises de phonétique expérimentale* VIII-IX, 1933, 167-83.

Joos, M. (1948). Acoustic Phonetics. *Language monograph,* **23** 109ff.

Lindblom, B. (1982). The interdisciplinary challenge of speech motor control. *Wenner-Gren Center International Symposium Series,* **36**: Speech Motor Control, 1-18.

Scripture, E. (1936). In: Bulletin of the internation society of experimental phonetics III. *Archives néerlandaises de phonétique expérimentale* **XII**, 135.

Wijk, N. v. (1939). *Phonologie*: een hoofdstuk uit de structurele taalwetenschap. The Hague.

Zwaardemaker, H. and Eijkman, L.P.H. (1928). *Leerboek der Phonetiek*. Haarlem.

Zwirner, E. (1965). Eröffnungsansprache. *Proceedings of the Fifth International Congress of Phonetic Sciences*, Munster 16-22 August 1964.

Zwirner E. (1982). 50 Jahre 'Permanent Council', *Phonetica* **39**, 374-378.

Keynote address

Phonetics and Speech Technology

Gunnar Fant
Stockholm, Sweden

1. Introduction

It is my privilege to address to you on a subject fundamental to our congress - phonetics and speech technology. The close ties and mutual dependencies inherent in the history of speech research and in the last decades of intense developments are apparent: Phonetics has attained a technical profile and speech technology has to rely on phonetics to achieve its advanced goals. This is, of course, an interdisciplinary venture also involving the entire field of speech research indepent of faculty. Instead of speaking about phonetics and speech technology, we could make a distinction between theory and applications and point to the development of handicap aids and new methods of clinical diagnosis and rehabilitation, teaching aids, etc. which add to the specialities represented at this congress. I shall make some general comments about this symbiosis and how it affects speech technology and phonetics. I shall also give my view on the general outlooks for the field, and on some of our problems and current research issues.

In the last decade we have experienced a revolution in computer technology and microelectronics that has paved the way for speech technology. There has been a breakthrough in the data handling capacity allowing very complex processing to be performed in small chips that can be produced at a low price in large quantities. There have also been reasonable advances in speech synthesis and speech recognition techniques which have opened new markets. This has created a boom of industrial expectations, a feeling of surfing on a high wave of technological developments towards the fully automated society where we may converse with computers as freely as with human beings. One expression for this optimistic trend is the Japanese national effort in computing and artificial intelligence which they refer to as the development of the 'Fifth generation of computers' which shall include language translation and speech input and output.

Electronic industry has promoted several large-scale marketing reports with prospects for billion dollars sales at the end of the century.

Will all these expectations come through? I am not the one to judge but there is certainly room for some scepticism or at least caution. The rate of increase of the world market has not progressed at the expected rate. So - the surf on the tidal wave of expectations may end in a brake when we are confronted with the reefs of the knowledge barrier, Fig. 1. I am referring to our still meager insight in speech as a language code. We need a fifth generation of speech scientists rather than a fifth generation of computers.

A stagnation of advanced speech technology products and the marketing of cheap, lower performance products may discredit the field. You frequently hear comments such as: 'Speech synthesis by rule has now existed for several years but the quality is still questionable and the rate of improvement is low. Will it ever reach an acceptability for public use?' To make speech recognition really useful we must first learn to handle connected speech with relatively large vocabularies in a speaker-independent mode. Indeed, we are far off from such advanced levels of recognition techniques whilst there appear to exist potentialities for reaching a substantial improvement in the quality of synthetic speech within the next few years. The latter optimistic opinion is shared by the pioneer in speech synthesis, John Holmes, in his report to this Congress and he also expects significant advances in the handling of connected speech to appear fairly soon.

Figure 1. Speech technology and the knowledge barriers.

To the optimistic view we could also add that text-to-speech synthesis already in the present state of the art has opened up new effective means of communication for handicapped, e.g. text-reading aids for the blind and speech protheses for speech handicapped. Also the performance is quite adequate for many special-purpose applications including computer-aided teaching. The Swedish text-to-speech system developed by Carlson and Granström is implemented with a single chi for the terminal synthesis and has an option for operating in six different languages. A similar text-to-speech system developed by Dennis Klatt at MIT has means for changing the speaker type from male to female to child. A flexible choice of speaker type will be quite important in the marketing of synthesizers but this is an area in which we still have much to learn.

There exists a variety of less advanced and cheaper synthesis systems, generally intended for phonetic symbol input but some also capable of handling a proper orthography text input. These devices provide a lower-quality speech. In general, even our best text-to-speech systems are fatiguing to listen to if used for reading long texts.

A substantial part of the speech output market is talking chips which serve as low data-rate recording and play-back systems. They are now introduced in automobiles, household appliances, watches, calculators, and video games. We might even anticipate a sound pollution problem from synthetic voices guiding every step of our daily life.

At present, toy industry and manufacturers of video games have employed phonetic experts to tailor talking chips to simulate special voice types and speaking manners. In the future I believe we can do this more or less by rules. General purpose text-to-speech systems are expected to improve sufficiently in performance to compete with speech coding and concatenating systems, at least when a certain flexibility is desired.

Computer speech input, i.e. speech recognition systems are expected to develop a greater market than speech output systems, at least in terms of sales value. Although we are far off from very advanced speech recognition systems, we might soon expect applications in office automation, e.g. as voice input for word processing systems. An extension of present techniques to handle connected sequences of words would facilitate this application. A speech synthesis monitoring feature could be included.

2. The Computerized Phonetics

The close ties between phonetics and speech technology are apparent. Phonetics has been computerized and has gained new efficient instrumentation and advanced speech processing methods. Of course, computers would have found their way to phonetics anyway but phonetics has now attained somewhat of a technical profile. The more prominent phonetic departments have a staff of engineering assistants and a research budget which was unheard of in former days' humanity faculties but this development has, of course, not come about without an intense engagement of people involved. Phonetics of today has gained a new respect from its vital role in the ever increasing importance of research into human functions. The technical profilation is also apparent in any speech research laboratory whether it is an outgrowth of linguistics, psychology, or a medical department.

This interdisciplinary venture has opened up new channels between formerly isolated faculties. We find young people from a humanities faculty engaged in mathematical problems of signal processing. Conversely, students in electrical engineering and computer science departments make significant contributions to phonetics and linguistics research. Phonetics, within its new profile, takes part in clinical projects and receives funding for basic work in speech recognition and synthesis. This is, indeed, a symbiosis or rather a fusion of research profiles. It is a healthy development much needed in quest of our far reaching goals - but does it not have any negative effects?

Some problems have been apparent all since computer technology penetrated our field. Many phoneticians of an older generation miss the direct contact with their instrumentation which they could handle without engineering support and which gave them an immediate and intimate insight in speech patterns. The old kymograph was indeed valuable in this respect. Even the sound spectrograph, which once revolutionized acoustic phonetics, is in the risk zone of being outdated by multi-function computer analysis programs. However, up till now they have not demonstrated the same temporal resolution as the rotating drum print-out from the ordinary spectrograph, which I still would not be without in spite of access to computer spectrograms with additional synchronized parameters.

At the same time as our appetite grows for more advanced computers systems with analysis and synthesis coordinated in interactive programs, we run into the usual problems of reliability and difficulties in accurately documenting and memorizing complex routines and, as you know, computers have a tendency to break down or to be occupied when you need them most.

Also, if we do not know how to rewrite and expand existing programs, we may become limited by software constraints which are not initially apparent. One example is the widely spread ILS system which, for the benefit of a graphically optimized positioning of curves, has a tendency to discard information on relative intensities comparing successive section frames.

The problem is that neither the software designer nor the user are always aware of needs that emerge from the special properties of speech signals or the research needs. One example is routines for spectrum analysis of unvoiced sounds, for instance of fricatives. Standard FFT routines without additional temporal or spectral averaging retain a random fine structure of almost the same amplitude as that of true formants. The result is a fuzzy spectral picture in which it is hard to see what is a formant and what is a random peak. Spectral smoothing can be attained in many ways. Cepstrum analysis or LPC are useful but the smoothness of the LPC curve can be deceptive since the location of the formant peaks may vary somewhat from sample to sample.

We all know that computers are fast in operation but that programming can take a long time. It is also apparent that computer programming is an art which possesses a great inherent fascination which may distract the user from his basic scientific problem. An intense love-hate relation may develop. I have stayed away from programming until recently, when I started using a Japanese programmable calculator which gives me the great satisfaction of access to fairly complex modeling at the price of time demanding debugging.

One can also raise the partially philosophical problem: Who is the boss? The user or the computer? Can we leave it to the computer to learn about speech or shall we insist on developing our own insights in the many domensions of the speech code? This is really a matter of strategical importance in speech research.

3. Speech Recognition and Research Needs

There are basically two different approaches possible in automatic speech recognition. Either we start by running the computer in a learning mode to store a number of templates of speech patterns from a single or a few subjects, recognition then simply becomes a best match selection. We learn very little about speech this way and we are generally not aware of why the matching incidentally fails.

The other approach needed for large vocabularies and connected speech is phonetically orientated in the sense that it is based on recognition of minimal units that can range from distinctive feature phonemes, diphones, syllables, and words and which require some kind of segmentation.We now approach the general problem of speech research in quest of the speech code and the relation between message units and their phonetic realization with all the variability induced by contextual factors including language, dialect, speaker specific situational and stylistic variations.

It would be a wishful dream to extract all this knowledge merely by computerized statistics, i.e. to collect a very large material of speech, give the computer some help for segmenting transcription, and look up and then just wait for the results to drop out.

Many institutions are now developing such data banks for their research. This is a necessary involvement to make but satisfies a partial need only. We cannot store all possible patterns with table look-ups. To organize the data bank efficiently, we must rely on a continuing development of a model of speech production and generative rules on all levels up to the linguistic frame and down to an advanced vocal tract model which should include all what we know of aerodynamics and source filter interaction. Flanagan in his paper to this Congress describes this process as letting the vocal tract model mimick the speech to be analyzed. This is a dynamic realization of analysis by synthesis, which we will be able to handle once we have gained a sufficient understanding of the speech production process.

It has already been proposed by some people to integrate a text-to-speech synthesis system as a part of the top-down arsenal of speech recognition. As pointed out by John Holmes, this general approach of perturbing synthesis parameters for a best match to a natural utterance is also an effective way of improving synthesis by rule. Here lies perhaps the main advantage of interfacing analysis and synthesis. The basic outcome is that we learn more about speech. Once we have a sufficient insight, we may produce short-cut rules for articulatory interpretations of speech patterns to guide further data collection or for recognition of articulatory events to guide the recognition.

This might be a more realistic approach to attempt a complete match which would require a very advanced adaptability to speaker-specific aspects. Again we are confronted with the constraints of pattern matching procedures.

4. Perception

Now you may ask, why all this emphasis on production? What about models of speech perception and feature theory as a guide for recognition?

First of all, it is apparent that the main drawback of present speech recognition schemes is the handling of bottom-up acoustic data. Either we lose a lot of information-bearing elements contained in rapidly varying temporal events or we perform a maximally detailed sampling in which case substantial information may be lost or diluted by distance calculations, performed without insight in the speech code. Frequency and time-domain adjustments by dynamic programming or by some overall normalization procedure are helpful but do not account for the uneven distribution of information.

Would it not be smarter to base the recognition on models of auditory processing including feature detection? Feature detection is, of course, closely related to the search for articulatory events but with the aid of perception models, we could hope to attain a simpler and more direct specification of the relevant attributes.

Formant frequency tracking is often difficult even for non-nasalized sounds and ambiguities have to be solved with reference to specific spectrum shapes. Models of the peripheral auditory system including Bark scaling, masking, lateral inhibition, and short-time adaptation can provide some improvements in portraying essential characteristics but do not immediately suggest a parametrization. The ultimate constraints are to be found at higher levels of auditory perception but here our insight is more limited and speculative, for instance, in questions of what is a general function and what is a speech mode specific mechanism.

There is now emerging a new duplex view of peripheral auditory analysis on the one hand, the basic concept of short-time spectrum transformed to a spatial discharge rate – on the other hand, the tendency of the outputs from a number of adjacent nerve-endings to be synchronized to a dominant stimulus frequency. The equivalent frequency range over which such synchronization takes place becomes a measure of the relative dominance of a spectral component, and the information about the frequency of the component is, at least for lower frequencies, contained in the neural periodicity patttern. The so-called DOMIN modelling of Carlson and Granström (1982) has its support in the neurophysiological studies of Sachs et al. (1982) and those of Delgutte (1982). A consequence of the DOMIN modelling of Carlson and Granström is that the algorithm, based on a Bark scaled filter bank, detects low-frequency harmonics at high F_0, otherwise formants or formant groups. In the earlier experiments of Carlson and Granström, based on the Bekesy-Flanagan auditory filters which are wider than those of the Bark scale, the system produced something that came close to an F_1 and F_2 detection.

We have already noted that models of the peripheral auditory system do not provide you a complete auditory transform. For a more true representation of the neural transform, we would have to inspect the cortical domain. The psychoacoustic experiments of Ludmilla Chistovich and her colleagues in Leningrad suggest some kind of spatial integration to take place above the level of peripheral hearing. They found that two formants interact to provide a joint contribution to the percept when placed closer than a critical distance of about 3.5 Bark and may then be substituted by a single formant of some weighted mean to provide the same categorical effect. On the other hand, when formants come further apart than the critical distance their relative amplitudes can be varied over a wide range without affecting the identification.

These effects are relevant to the discussion of vowel systems and conform with the early studies of Delattre et al. at Haskins Laboratories who found that back vowels can be simulated by a single formant. I may illustrate the categorical boundary between back vowels and more centrally located vowels by reference to Figure 2 which shows Swedish vowel formants arranged in $F_2' - F_1$ versus $F_2' + F_1$ plot with frequencies transformed to equivalent Bark values. The tendency of fairly equal spacing and regular structure has exceptions which can be related to historical sound changes and a combination of contrast enhancement and reductions. Thus, the Swedish long [u] produced with very high degree of liprounding has advanced articulatorily to a front vowel with a tongue location similar to that of [i:], whilst its short counterpart [ə] resembles a back vowel but for a tongue location sufficiently advanced to transcend the 3.5 Bark F_2-F_1 boundary. Perceptually the long [u] and the short [ə] occupy an extreme low F_1+F_2' 'flatness' feature which they share with their historical origin [u:] and [U] in relation to all other vowels, see further Fant (1973; 1983).

Auditory modelling has now penetrated into the domain of speech dyna-

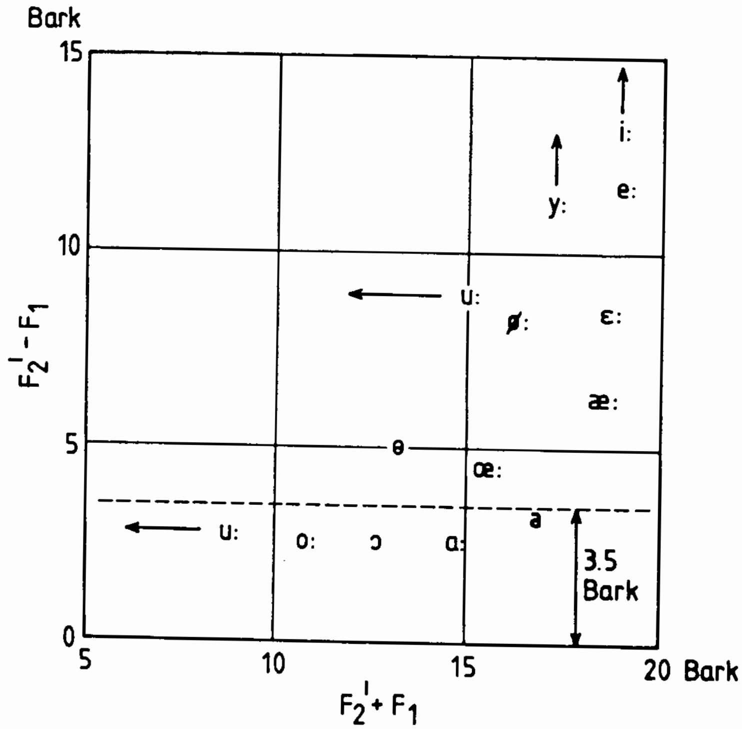


Figure 2. Long and short Swedish vowels ın auditory adjusted F_1' and F_2' scales. Arrows indicate diphthongal extensions. Short vowels of F-patterns close to their long mates are omitted.

mics. There are indications that short-time adaptation effects increase the discriminability of rapid onset patterns and that the frequency resolution is enhanced for timevarying formant patterns. There remains much to be learned about these effects and the role of special feature detectors.

Our lack of understanding of the speech perception mechanism may be illustrated by the two spectrograms of one and the same sentence recorded in an auditorium (see Fig. 3) The upper case spectrogram refers to a microphone close to the speaker and in the lower case it originates from a microphone in the middle of the auditorium. The reverbation distortion does not impede intelligibility much but the spectrographic pattern is blurred to the extent that most of the usual visual cues are lost. How does the auditory system combat noise and reverbation?

5. In Quest of the Speech Code. Variability and Invariance

Although there are shortcuts for special purpose speech recognition and synthesis by rule, it is evident that advanced goals can be reached by intensified fundamental research only. The common knowledge needed, the structure of the speech code, is also the central object of phonetics. Models of production and perception constitute a biological frame within which we can study the speech code. Now even if we possessed perfect general models of production and perception and a maximally effective linguistic framework, we would still have to derive an immense amount of rules and reference data relating message units and speaker categories to observed phonetic sound shapes. Presently available reference data and rules are incomplete and scattered into fractional acoustical phonetic studies. The more complete rule systems are hidden in the software of text-to-speech synthesis systems and are contaminated by elements of ad hoc guess work and by the specific format of the parameter system.

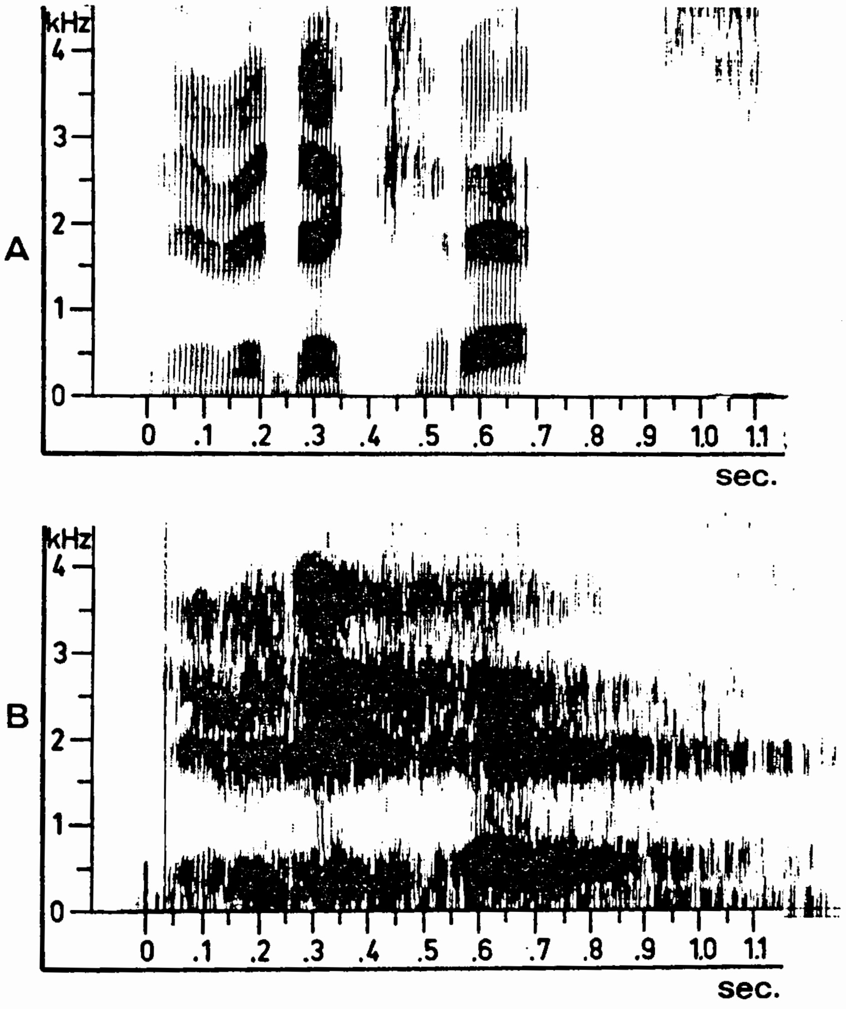


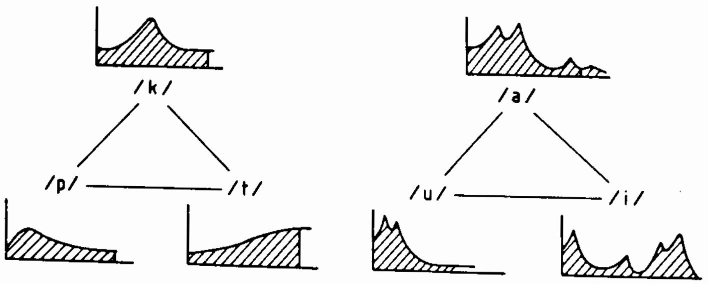
Figure 3. Spectrograms of one and the same utterance from a close talking microphone and in the middle of a reverberant auditorium.

So far, speech technology has relied heavily on linguistic redundancies to ensure an acceptable performance of synthesis as well as recognition, but it is due time to extend fundamental knowledge by large documentary projects around our data banks. When will we have a new version of the book *Visible Speech* with not only illustrations but with reference data and major contextual rules, in other words, the missing links towards the output of generative grammar? When will we have a complete inventory of rules for generating different voice types and stylistic variations?

For applied work it is of no great concern which distinctive feature system we adopt for addressing phonemes as long as we can properly handle their acoustic-phonetic correlates. Prosodic categories should not be defined by single physical parameters. They should be treated the same way as phonological segmentals, that is, as constituents of the message level with rules for their many phonetic realizations.

The study of coarticulation and reduction is of central importance. There is a need to extend the concept of reduction to variations induced by various degrees of stress emphasis and stylistic factors. A typical example is the variation of vowel formant frequencies with the mode of production. We find a more extreme articulation in citation forms than in connected speech and even more extreme in targets in sustained vowels. Emphasis and de-emphasis affect not only target values but in general all speech parameters and their temporal patterning.

The speech code is a theme about variabilities and invariance. Invariance and manifestation rules are closely connected. How do we define invariance?



	Jakobson, Fant, Halle		Chomsky and Halle		
	Compact	Grave	Anterior	Coronal	
/k/	+		–	–	
/p/	–	+	+	–	
/t/	–	–	+	+	
			Back	Low	High
/a/	+	+	+	+	–
/u/	–	+	+	–	+
/i/	–	–	–	–	+

Figure 4. Spectral attributes and DF-specifications of /p,t,k/ and /u,i,a/ according to Jakobson, Fant and Halle; Chomsky and Halle

I feel that we should make a distinction between academical and more pragmatic needs. Roman Jakobson's concept of distinctive feature implies in its most general form a relational invariance. Independent of the sequential context and specific combination of other features in a phoneme, there remains 'ceteris paribus', a vectorial difference along the feature dimension comparing the + alternative and the – alternative.

Obviously, we do a better job in recognition if we make use of all conditional factors affecting the sound shapes of the two candidates. However, a research line adopted by Kenneth Stevens is directed towards, what he calls 'absolute' invariance, which conceptually comes close to the common denominator aspect of the distinctive feature theory. Stevens started out by studying spectrum slope properties of the stop burst and extended his descriptions to temporal contrasts, e.g. the intensity of the burst and that of a following vowel in a certain frequency region. I have suggested an extension of the concept of absolute invariance to employ any description which does not imply a prior phonological identification of the context. In this sense, positional allophones of /k/ and /g/ may be identified by both the degree of spectral concentration and by the location of energy with respect to the format pattern after the release.

Returning to academical issues we find that the use of one and the same feature, such as compactness in both consonant and vowel systems, complicates and dilutes the common denominator whilst there still remains an interesting parallelism, in the Jakobson-Fant-Halle system brought out by the identification of the [k] [p] [t] relations with those within [a] [u] [i]. The Chomsky-Halle system operating with independent consonant place features has its shortcomings in the roundabout labeling of labials as [+anterior [-coronal]. I prefer the output oriented acoustic-perceptual basis. Major spectral attributes are preserved in neurophysiological studies as those of Sachs et al. (1982: 121) see their figure of [i] versus [a] emphasizing the compactness feature.

I am now approaching the more philosophical aspects of phonetics. We are all more or less engaged in studies of the speech code but this is a painstaking slow process. Meanwhile we can make general remarks about the code, e.g. that it has developed with a major concern for the final stage of the speech chain. Roman Jakobson's theme 'we speak to be heard in order to be understood' has had a great impact. This principle is referred to by Bjorn Lindblom as *teleological.* With a slight deletion in this exclusive term, we end up with the word *teology* which has some bearing on issues such as motor theory of speech perception, 'speech is specially handled in perception', the speech code is innate, speech production is a chain process or is preplanned etc.

I am personally in favor of a both-and principle. No single statement is sufficient. Speech is both precise and sloppy. Speech perception involves many parallel processings and may rely on both phonemes, syllables, and words as minimal recognition units. The statement that the truth about segmentation is that you cannot need modification. You both can and cannot. The common denominator of distinctive feature is sometimes easier to describe with reference to articulation than to perception and the reverse is often true. Motor theory of speech perception as well as auditory theory of speech production both have something to contribute to our perspective.

The most absolute statement I can make is that speech research is a remarkable, exciting venture. Most people take speech for granted. A small child can do what 700 wise men and women at this congress do not quite understand. I wish you all an exciting continuation of the congress.

References

Carlson, R. and Granström, B. (1982). Towards an auditory spectrograph. In: *The Representation of Speech in the Peripheral Auditory System*, 109-114. Amsterdam: Elsevier Biomedical Press.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In; *The Representation of Speech in the Peripheral Auditory System*, 131-149. Amsterdam: Elsevier Biomedical Press.

Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: The MIT Press.

Fant, G. (1983). Feature analysis of Swedish vowels – a revisit. *STL-QPSR* **2-3**/1983.

Sachs M.B., Young, E.D., and Miller, M.I. (1982). Encoding of speech features in the auditory nerve. In: *The Representation of Speech in the Peripheral Auditory System*, 115-30. Amsterdam: Elsevier Biomedical Press.

Perception of Speech as a Modulated Signal

Reinier Plomp
Soesterberg and Amsterdam, the Netherlands

1. Introduction

In this contribution on the significance of hearing for speech research I would first like to demarcate what will be discussed. It is not my purpose to present a review of the state of the art of psychoacoustics as far as it is relevant to speech perception. Instead of this the attention will be focused on a comparison of the physical properties of the speech signal with the limits of the ear's capacity to handle this signal. To avoid misunderstandings, the auditory processes to be considered do not include the way in which the signal is decoded in phonetic terms. Therefore, the controversy of whether or not there is a phonetic mode will not bother us (for a recent review, see Repp, 1982).

The auditory system constitutes an important link in the transfer of information from the original speech signal produced by the speaker to the understanding of its message by the listener. Since speech intelligibility in general is the main criterion of the successfulness of this transfer, it will play an important role in our discussion. This means that the present approach does not consider the intelligibility, or any other property, of the individual speech sounds.

As will be shown, it is worthwhile to include in our considerations also the transfer of the speech signal from the mouth of the speaker to the ear of the listener, as far as the reverberation is concerned. This link in the chain is too often neglected in phonetics.

2. Analysis of Speech in Terms of Modulation Frequency

It is common in phonetics to describe the speech signal in terms related to its production, such as formant frequency and place of articulation. This results in a description specific for speech, but not generally applicable to all types of sound relevant in everyday listening. It will be clear that we cannot use such a production- oriented description in studying the transfer of the speech signal on its way from the speaker to the listener; the measures adequate in room acoustics (e.g. reverberation time, sound-absorption coefficient, volume, distance) are general, physical, quantities.

A similar argument holds for the auditory system in the strict sense of the term. Without going into details (see Plomp, 1976), we may compare this system with a set of (overlapping) band-pass filters tuned to different frequencies from low to high. The three main perceptual attributes of sounds: pitch, loudness, and timbre, are strongly related to the physical properties of fundamental frequency, intensity, and frequency spectrum, respectively. (For experimental evidence that sounds with different fundamental frequencies but equal absolute frequency spectra are very similar in timbre, see Plomp and Steeneken, 1971). Together, these three physical quantities given an adequate description of a speech vowel as a periodic vibration (fundamental frequency F_0) with a specific amplitude pattern of the harmonics. This approach also holds reasonably well for the voiced consonants (sonorants); in the nonsonorants the periodic vibration is replaced by a noise-like source. Thus, the speech signal can be interpreted as a carrier (periodic vibration or noise) with a frequency spectrum modulated continuously in time by the vocal tract. Although the temporal variations of F_0 contribute to speech intelligibility, this contribution is relatively small and will be left out of consideration here. This means that we will confine ourselves to the frequency spectrum and how it varies in time.

The significance of this perception-oriented approach can be demonstrated by means of the spectrogram. In a spectrogram the frequency spectrum measured with a set of band-pass filters is plotted as a function of time. It does not represent the fine structure of the signal (instantaneous amplitude) but gives the temporal intensity envelope for the different frequency bands. The important role the spectrogram has played in phonetic research during the last decades may be regarded as a demonstration that the spectrogram is an excellent representation of the information-bearing characteristic of speech.

This conclusion suggests that it makes sense to analyze the speech-signal envelope, reproduced in the spectrogram, in terms of sinusoidal components, as is usual in systems analysis. This analysis should be distinguished from the traditional frequency analysis in terms of audio frequencies. The spectrogram gives the intensity envelope both in time (horizontal) and in frequency (vertical) and it is these two envelopes that should be analyzed. The resulting frequencies are not audio frequencies but modulation frequencies describing the temporal and spectral variations. (In order to avoid confusion between these two types of frequencies, the prefixes 'audio' and 'modulation' will be frequently used.)

As a further illustration of what is important in the transfer of the speech signal, let us consider the spectrogram as an optical image to be transferred. It will be clear that its finer details, quite essential for identifying specific phonemes, are lost if the spectrogram is reproduced on a TV screen or as a newspaper picture with large dots; the medium, including the eye, should be able to preserve the relevant details. In recent years it has become common to quantify the quality of the image transfer by means of the spectral modulation transfer function, representing the faithfulness with which spatial sinusoidal brightness variations are preserved as a function of modulation frequency. Similarly, the transfer of speech as a sound signal can be quantified by means of temporal (in Hz) and spectral (for example, in periods/octave) modulation frequencies. Hence it makes sense to study the speech signal as radiated from the mouth in terms of modulation frequencies.

Since on temporal modulation rather more data are available than on spectral modulation, I will start by discussing what sinusoidal modulations in time are present in speech signals, the effect of reverberation on the transfer of these modulations from the speaker to the listener, and the way in which the limited capacity of the ear to perceive modulations can be expressed, too, by a modulation transfer function. Subsequently, the same points will be considered for the spectral modulations.

3. Temporal Modulation

3.1. The temporal envelope of the speech signal

The intensity of a speech signal as a function of time can be measured with the aid of a squaring circuit followed by a low-pass filter with a cut-off frequency of, say, 50 Hz. In this way a signal is obtained only determined by the fluctuating envelope, not by the fine structure (viz. the audio frequencies) of the speech signal. Figure 1 illustrates such an intensity envelope for a speech fragment of 10 sec; the dashed line represents the average intensity, $\overline{I}$, of this signal.

By means of a frequency analysis of the intensity envelope function of the speech signal the relative importance of different modulation frequencies can be determined. Steeneken and Houtgast (1983) analyzed one-minute speech

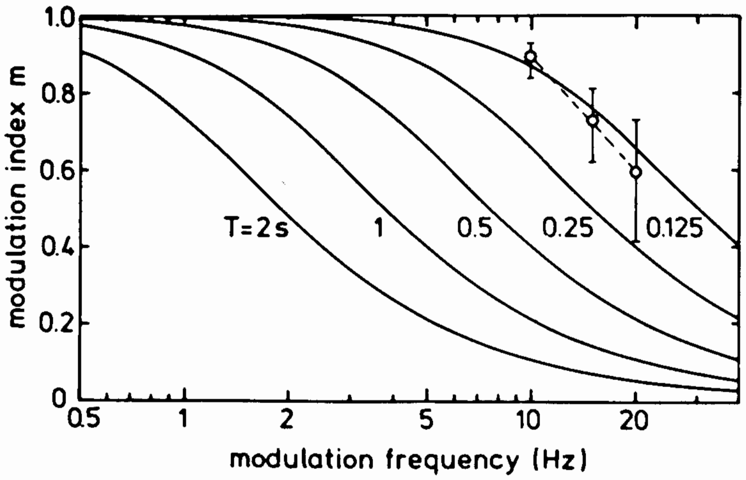


Figure 1 Intensity envelope for a 10-sec speech fragment.

fragments of connected discourse from ten male and ten female speakers who read the same text. By speeding up the envelope function by a factor of 400, they obtained a signal fluctuating in the normal audi-frequency range. This signal was analyzed with a set of one-third octave band-pass filters. Since the modulation index, m, is an appropriate measure for specifying modulation transfer functions, as we will see below, this measure will also be used for the speech signal. It is defined as the average peak value of the filter output amplitude divided by the average value of the unfiltered signal, $\overline{I}$; in Fig. 1 the sinusoids represent the case m=1.

In Fig. 2 the modulation index, averaged over ten male speakers, is plotted as a function of modulation frequency, F (centre frequency of the 1/3-oct band-pass filters). The 1-min speech segments were first analyzed in terms of audio frequencies by means of octave band-pass filters; then for each octave band the temporal intensity envelope was analyzed in terms of modulation frequencies. We see that the resulting curves are rather similar, except for their vertical positions. For all audio frequencies the most important modulation frequencies are 3-4 Hz, related to the number of words/syllables pronounced per sec. With ten female speakers, as well as with other texts, almost the same results were found.

Adopting as a criterion the modulation frequency for which the modulation index is reduced to half its peak value, we see that modulation frequencies are present in speech up to about 15 Hz.

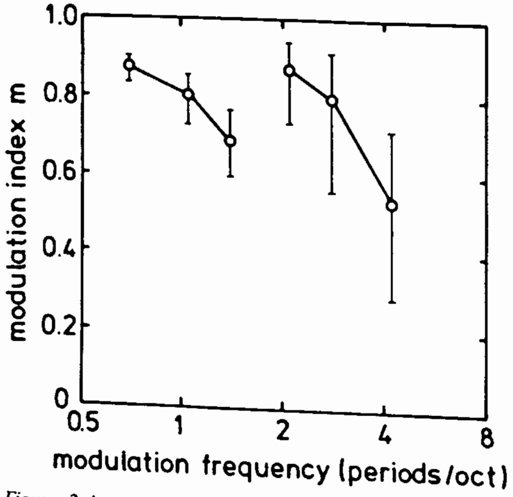


Figure 2 Average temporal envelope spectra in terms of the modulation index for one-min connected discourse from ten male speakers. The number of sentence units, stressed syllables etc. per sec is indicated. The parameter is the centre frequency of the one-oct audio-frequency band.

3.2. The speaker-to-listener modulation transfer function in rooms.

The transfer of the speech signal from the speaker to the listener depends in an enclosed space on the reverberation characteristics of the room. The sound travels from the mouth of the speaker to the ears of the listener via a great many different transmission paths. At a short distance the direct path may be the most important one, but for most positions the sound level will be determined by the indirect paths, each including a number of reflections. Since the time of arrival is given by the total lentgh of the path, the differences in path length result in a blurring of fine temporal details of the speech signal's envelope, comparable with the optical effect of a lens system that is out of focus. In terms of modulation, it means that slow sinusoidal modulations are well preserved, whereas fast modulations are attenuated.

This(modulation)frequency-dependent behaviour of an enclosure can be expressed in the so-called temporal modulation transfer function (TMTF). For an input signal (band of noise) with a 100-% sinusoidally varying intensity, $\overline{I}_i\,(1 + \cos 2\pi Ft)$, at the position of the speaker, the output signal at the listener's ear is given by

$$\overline{I}_0\ [1+m\cos(2F\pi t-\varphi)] \qquad (1)$$

where F= modulation frequency and m as a function of F is defined as the TMTF.

Since the TMTF was first introduced in room acoustics by Houtgast and Steeneken (1973) as a predictor of speech intelligibility, various ways of deriving m(F) from the room parameters have been explored. In the simplest case the sound field is considered as the result of a statistical process of sound reflections without any directional preference, the direct path being excluded. For such a diffuse sound field m(F) is given by

$$m(F) = (1+0.207F^2T^2)^{-1/2} \qquad (2)$$

where T = reverberation time in sec, the time in which a sound decays by 60 dB (Houtgast, Steeneken and Plomp, 1980; Schroeder, 1981. In Fig.3 this equation is plotted as a function of modulation frequency, with T as the parameter. The figure illustrates the blurring effect of high modulation frequencies and the significant role of reverberation time. The reverberation times in everyday life are largely restricted to the range from T= .5 sec (typical living room) to T=2 sec (good concert hall).

For a more accurate determination of the TMTF the statistical approach has to be replaced by a geometrical approach based on the exact dimensions of the room, the positions of the speaker and the listener, and the sound absorption properties of the different boundary surfaces. For a rectangular room the algorithm has been given by Plomp, Steeneken and Houtgast (1980), for a room with oblique walls shaped as a trucated pyramid by

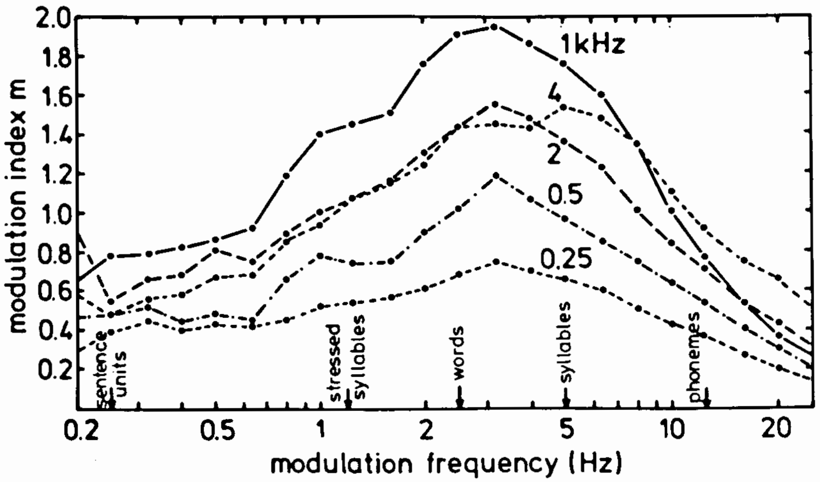


Figure 3 Temporal modulation transfer function in the diffuse sound field with the reverberation time of the room as the parameter. The data points (mean value and standard deviation for 50 subjects) represent the TMTF of the auditory system for 1000 Hz.

Wattel et al. (1982). Whereas those algorithms were based on mirror images, the more general approach by means of ray-tracing has been presented by van Rietschote, Houtgast and Steeneken (1981). In all these models the presence of noise was also taken into account. It has been shown (Houtgast et al. 1980) that from the TMTF over the modulation frequency range 0.4-40 Hz a single measure can be derived, the Speech Transmission Index (STI), which is an excellent predictor of speech intelligibility; as has been verified, this holds generally for Western languages (Houtgast and Steeneken, 1983). This means that the algorithms for computing the TMTF from the room parameters are important tools in designing acoustically good classrooms, conference rooms, auditoria, etc. The TMTF concept has been successfully extended to include band-limiting, nonlinear distortion, and other disturbances of the speech signal which may be present in communication channels (Steeneken and Houtgast, 1980).

4. The temporal modulation transfer function of the ear

In the same way as an enclosure the ear may be regarded as a link in the speech transmission chain that is able to transfer faithfully slow variations of sound intensity but is unable to follow fast modulations. If we could be sure that the ear is linear for modulations, we could derive the TMTF from the just-noticeable intensity modulation as a function of modulation frequency (Viemeister, 1979), but this assumption is not justified. Since we cannot measure directly at the output of the auditory system, we have to use an external test sound to investigate the ear's modulation transfer function. This can be done by measuring the detection thresholds for a very short test sound

coinciding in time with the peaks of the sinusoidally intensity-modulated stimulus and with the valleys of this signal, respectively. The peak-to-valley difference, ΔL, in dB, can be translated in the modulation index m.

$$\Delta L = 10\log\frac{1+m}{1-m}, \text{ or } m = \frac{10^{\Delta L/10}-1}{10^{\Delta L/10}+1}. \qquad (3)$$

A condition for this approach is that detection exclusively depends on signal-to-noise ratio which is true over a large intensity range.

In the literature only few data on peak-to-valley differences are reported (Rodenburg, 1977; Festen et al., 1977; Festen and Plomp, 1981). In the experiment by Festen and Plomp, the sinusoidally modulated sound was white noise low-pass filtered with a cutoff frequency of 4000 Hz. In a two-alternative forced-choice procedure the detection threshold of a 0.4-msec click, octave-filtered around 1000 Hz, was measured for modulation frequencies of 10, 15, and 20 Hz. The results, averaged over 50 normal-hearg subjects, are plotted in Fig. 3.

Taking the frequency for which m=0.5 as a measure of the limit up to which the ear is able to follow temporal modulations, we arrive at a value of about 25 Hz. It is of interest that in the first channel vocoders, already more than 50 years ago, intensity fluctuations up to 25 Hz were considered to be important in speech perception (Flanagan 1965, p. 246).

This experiment leaves unanswered the question of the degree to which the ear's TMTF may depend on audio frequency. The scarce data on the threshold for just-noticeable modulations strongly suggest that for frequencies as low as 250 Hz the TMTF shifts (maybe by as much as a factor of two) to lower frequencies, with the reverse holding for frequencies as high as 4000 Hz (Viemeister, 1979).

Figure 3 allows us to express the ear's sensitivity to modulations in reverberation time, resulting in an estimate of T=0.12-0.15 sec. Since for rooms the reverberation time is almost always at least 0.4 sec, it is clear that in everyday situations the room rather than the ear is the limiting factor in our ability to perceive temporal intensity fluctuations of sounds.

In this derivation of the TMTF it has been taken for granted that the role of phase in the transfer of temporal modulations is negligible. If a room modified the phase relation between the various modulation components, this should affect speech intelligibility. For the diffuse sound field underlying the curves of Fig. 3, the phase shift, relative to F=0Hz, is increased to only 45° at the modulation frequency for which m=0.5, with an asymptote of 90° at high modulation frequencies. Experiments by Viemeister (1977) indicate that for the auditory system, too, the phase shift may be neglected for the range of modulation frequencies relevant in speech perception.

5. Spectral Modulation

5.1. The spectral envelope of the speech signal

Analogously to the case of temporal modulations, we would like to analyze spectral speech envelopes in terms of modulation frequencies. Since no data for a speech fragment of which the frequency spectrum is measured periodically in time are available, we have to estimate the upper limit of spectral modulation frequencies from audio-frequency spectra of individual speech sounds. Both in view of their temporal prominence in speech and their peakedness, vowels are most appropriate for investigating this upper limit.

The spectral envelopes of vowels are characterized by a series of formants, of which the lower three are the most important ones. In addition to their frequency and level, these formants are described by their bandwidth and their interdividual spread in frequency.

Experiments by Dunn (1961) have shown that, in the mid-frequency range (800-2500 Hz), formant bandwidth is, roughly, about 6%. Assuming triangular spectral formant shapes, this implies that two formants have to differ about 15% in order to be separated by a spectral valley of 4.77 dB, corresponding to m=0.5 (equation 3). This peak distance of 15% determines the upper modulation frequency present in speech spectra, equal to about 5 periods/octave.

This value should be considered as an upper estimate, excluding the interindividual spread in formant frequencies. Since speech recognition is based on the absolute rather than the relative position of the spectrum along the frequency scale, it is reasonable to take the interindividual spread into account. For male vowel spectra the standard deviation of formant frequencies is about 10% (Pols, Tromp and Plomp, 1970) which means that 68% of the peaks are within a range of 20% around the average frequency for that particular formant. Interpreting this 20% as a bandwidth to be combined with the 6% of the formant bandwidth, we arrive at a lower estimate of the limit of modulation frequencies present in speech of about 1 period/octave. On the basis of 1/3- octave vowel spectra, the same lower estimate has been obtained (Plomp, 1983).

6. The transfer from the speaker to the listener in a room

The fact that sounds reach the ear via a great many different transmission paths does not only influence the temporal modulations present in the speech signal at the listener's position, but also its spectrum. For steady-state pure tones the sound pressure level at a large distance from the speaker has a theoretical uncertainty with a standard deviation of 5.57 dB (Schroeder, 1954). Measurements at a great many locations in a concert hall have confirmed this value (Plomp and Steeneken, 1973). This uncertainty is a consequence of the vectorial addition of sound waves with random phases; it

is inherent in a diffuse sound field and cannot be reduced by acoustical measures.

In order to get some insight into the effect of this 'noisy' character of the transfer of sound in a room on the speech signal, we can compare it with the interindividual spread in vowel spectra. Using data from Klein, Plomp and Pols (1970) it was found that the spectral variance due to reverberation is about as large as the interindividual spectral variance for male speakers pronouncing the same vowel (Plomp and Steeneken, 1973).

7. The spectral modulation transfer function of the ear

The sum of white noise and its replica by τ sec results in so-called comb-filtered noise with an intensity varying sinusoidally along the frequency scale at a frequency of 1/τ Hz (thus a delay of 5 msec gives noise with peaks at distances of 200 Hz). Similarly as temporally modulated noise, this signal can be used for measuring the peak- to-valley difference for a test tone at a fixed frequency.

With the same group of 50 normal-hearing subjects for which the TMTF was measured (Fig. 3), the spectral modulation transfer function (SMTF) with comb-filtered noise was also investigated. Short 1000-Hz test tones (duration 15 msec) were presented either during or immediately after 500-msec noise bursts. These two conditions, simultaneous masking, were chosen because they result in different values of the ear's SMTF, as the results in Fig. 4 show. This difference is attributed to a sharpening mechanism (lateral suppression, comparable with Mach bands in vision) not effective in simultaneous masking (for more details see Houtgast, 1974; Plomp, 1976).

According to Fig. 4, the ear is able to detect spectral modulations up to a

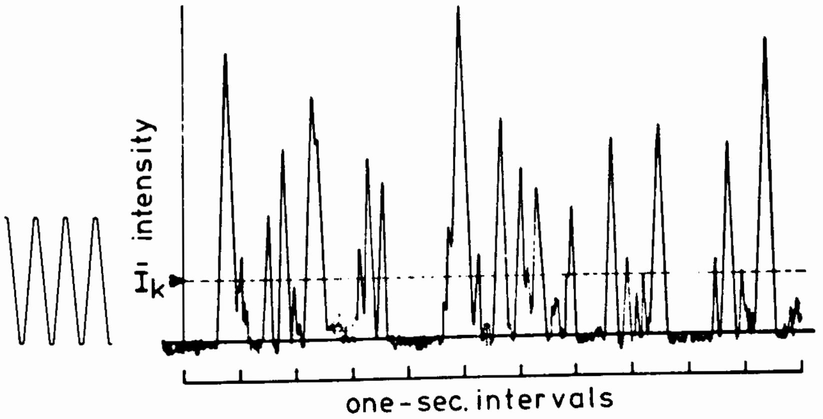


Figure 4. Spectral modulation transfer function (mean value and standard deviation for 50 subjects) of the auditory system for 1000 Hz. The left-hand curve holds for simultaneous masking, the right-hand curve for forward masking.

limit of, on the average, about 4 periods/octave, with lateral suppression included, and of about half this value for the condition without the effect of lateral suppression. From experiments by Houtgast (1974) we may conclude that these figures, measured for 1000 Hz, hold generally over the speech-frequency range. The diagram indicates that for the first-mentioned condition the interindividual differences are rather large; this confirms the finding in other experiments that subjects differ considerably in the effectiveness of the ear's sharpening mechanism.

8. Effect of Hearing Impairment

In the previous sections we have discussed values for the upper limits of temporal and spectral modulation frequencies present in the speech signal as well as the upper limit of the ear's capacity to detect modulations. These data can be summarized as follows:

Type of modulation	Speech signal	Hearing	
		Normal	Impaired
Temporal (Hz)	15	25	10
Spectral (per/oct)	1 to 5	2 to 4	1.4 to 1.6

From this table we may conclude that speech and normal hearing are well matched; the modulations present in speech cover the frequency range over which the ear is able to follow these modulations.

This agreement is disturbed in the case of hearing impairment. The right most column of the table gives average values for a group of 22 hearing-impaired subjects with moderate hearing losses (30 to 60 dB for 1000 Hz) of sensorineural origin. These recent data from Festen and Plomp (1983) were gathered with the same experimental procedures as used in the case of normal hearing.

We see that hearing-impaired subjects are, on the average, unable to hear the fast temporal intensity variations present in speech. According to the TMTF curves of Fig. 3, their hearing handicap in a reflection-free room is , in this respect, comparable with the situation for normal-hearing listeners in a room with a reverberation time of about 0.4 sec.

The table shows that not only the ear's TMTF but also its SMTF is reduced by the hearing impairment. This reduction is much stronger for the nonsimultaneous condition than for the case of simultaneous masking resulting in almost equal values in periods/octave. This suggests that lateral suppression is rather susceptible to hearing loss.

It is well-known that hearing-impaired subjects have special difficulties in understanding speech in noisy and reverberant environments (cf. Plomp and Mimpen, 1979; Plomp and Duquesnoy, 1980; Duquesnoy, 1982). We may conclude that, at least partly, these difficulties are due to the reduction in the ear's upper limits for the detection of temporal and spectral modulations. There are indications that by compensating for the ear's decrease in sensitivity to temporal modulations, speech intelligibility can be enhanced (Lewien, 1982; see Schroeder's contribution to this congress).

References

Dunn, H.K. (1961). Methods of measuring vowel formant bandwidths. *Journal of the Acoustical Society of America,* **33,** , 1737-1746.

Festen, J.M. et al. (1977). Relations between interindividual differences of auditory functions. In: *Psychophysics and Physiology of Hearing.* (E.F. Evans and J.P. Wilson Eds.) London: Academic Press, 311-319.

Festen, J.M. and Plomp, R. (1981). Relations between auditory functions in normal hearing. *Journal of the Acoustical Society of America,* **70,** 356-369.

Festen, J.M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. *Journal of the Acoustical Society of America,* **73,** 652-662.

Flanagan, J.L. (1965). Speech Analysis, Synthesis and Perception. Berlin: Springer Verlag.

Houtgast, T. (1974). *Lateral Suppression in Hearing.* Doctoral Dissertation, Free University Amsterdam.

Houtgast, T. and Steeneken, H.J.M. (1973). The Modulation Transfer Function in room acoustics as a predictor of speech intelligibility. *Acustica,* **28,** 66-73.

Houtgast, T. and Steeneken, H.J.M. (1983). A multi-language evaluation of the RASTI method for estimating speech intelligibility in auditoria. *Acustica,* in press.

Houtgast, T., Steeneken, H.J.M. and Plomp, R. (1980). Predicting speech intelligibility in rooms from the Modulation Transfer Function, I. General room acoustics. *Acustica,* **46,** 60-72.

Klein, W., Plomp, R. and Pols, L.C.W. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America,* **48,** 999-1009.

Lewien, T. (1982). *Filterung von Spracheinhüllenden zur Verständlichkeitsverbesserung bei Innenohr-schwerhörigkeit.* Doctoral Dissertation, Georg-August-University, Göttingen.

Plomp, R. (1976). *Aspects of Tone Sensation.* London: Academic Press.

Plomp, R. (1983). The role of modulation in hearing. In: *Hearing-Physiological Bases and Psychophysics.* (R. Klinke and R. Hartmann, Eds.) Berlin: Springer Verlag, 270-275.

Plomp, R. and Steeneken, H.J.M. (1971). Pitch versus timbre. In: *Proceedings Seventh International Congress on Acoustics, Budapest, Vol.* **3,** 377-380.

Plomp R. and Steeneken, H.J.M. (1973). Place dependence of timbre in reverberant sound fields. *Acustica,* **28,** 50-59.

Plomp, R., Steeneken, H.J.M. and Houtgast, T. (1980). Predicting speech intelligibility in rooms from the Modulation Transfer Function. II. Mirror image computer model applied to rectangular rooms. *Acustica,* **46,** 73-81.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (1970). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America,* **53,** 1093-1101.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin,* **92,** 81-110.

Rietschote, H.F. van, Houtgast, T. and Steeneken, H.J.M. (1981). Predicting speech intelligibility in rooms from the Modulation Transfer Function. IV. A ray-tracing computer model. *Acustica,* **49,** 245-252.

Rodenburg M. (1977). Investigation of temporal effects with amplitude modulated signals. In: *Psychophysics and Physiology of Hearing.* (E.F. Evans and J.P. Wilson, Eds.). London: Academic Press. 429-437.

Schroeder, M. (1954). Die Statistischen Parameter der Frequenzkurven von grossen Räumen. *Acustica,* **4** , 594-600.

Schroeder, M.R. (1981). Modulation transfer functions: Definition and measurement. *Acustica,* **49,** 179-182.

Steeneken, H.J.M. and Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America,* **67,** 318-326.

Steeneken, H.J.M. and Houtgast, T. (1983). The temporal envelope spectrum of speech and its significance in room acoustics. In: *Proceedings of the Eleventh International Congress on Acoustics*, Paris, **7**, 85-88.

Viemeister, N.F. (1977). Temporal factors in audition: A system analysis approach. In: *Psychophysics and Physiology of Hearing*. (E.F. Evans and J.P. Wilson, Eds.). London: Academic Press, 419-427.

Viemeister, N.F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *Journal of the Acoustical Society of America,* **66,** 1364-1380.

Wattel, E., Plomp, R., Rietschote, H.F. van, and Steeneken, H.J.M. (1981). Predicting speech intelligibility in rooms from the Modulation Transfer Function. III. Mirror image computer model applied to pyramidal rooms. *Acustica,* **48,** 320-324.

Speech and Hearing: Some Important Interactions

Manfred R. Schroeder
Göttingen, FRG and Murray Hill, USA

Abstract

I will discuss the following three topics at the interface of speech and hearing:

1. The importance of auditory masking by the speech signal on the perception of speech itself. This 'self-masking' of speech is also important in the synthesis of speech and its digital encoding. In particular, the sizable quantizing noise at the low bit rates desirable for efficient digital transmission and storage of speech signals can be made nearly inaudible by exploiting the masking properties of the human ear. (Collaborators: B.S. Atal and J.L. Hall)
2. Work at Göttingen on processing of speech signals to enhance their intelligibility for the hard-of-hearing, particularly those suffering from sensorineural hearing loss and recruitment. (Collaborators: W.H. Strube, T. Langhans and T. Lewien)
3. Effects of phase on the perception of speech, including the possibility of creating intelligible speech from signals with time-invariant flat power spectra solely by controlling monaural phase. These results point to the importance of temporal cues in the perception of speech beyond presently accepted limits. (Collaborator: S. Mehrgardt).

1. Introduction

I was raised as a physicist and thus, I feel a bit out of place at a Phonetics Congress - but not entirely. In 1953 - 30 years ago - Prof. Werner Meyer-Eppler from the Phonetics Institute in Bonn gave a colloquium talk at the University of Göttingen and I (a young student then) was much impressed by what he had to say about speech and hearing, about linguistics and information theory. My physicist friends pretended that they did not understand what Meyer-Eppler was talking about, but I was so taken with his message that a year later - upon joining Bell Laboratories - I changed my primary research field from microwave physics and physical acoustics to speech.

I remember building the first voice-excited vocoder (VEV) - not from integrated circuits, but from individual components, soldering iron in hand. We discovered that the cepstrum technique - originally suggested to distinguish earthquakes from nuclear explosions - was an ideal tool for the detection of the fundamental frequency of speech.

We manipulated formant frequencies, segmental durations and pitch of natural speech to create the most unnatural speech signals anyone had ever heard. In fact, taking a leave from the great Creator, we tried to change a male voice into a female voice and discovered there is more to the male-female dichotomy than pitch and formant frequencies. (Even after getting formant bandwidths right, our female creations did not sound very inviting - say for sharing an evening at the opera.)

In our work on analyzing and synthesizing speech we were constantly reminded how important hearing - or more generally: auditory perception - is. Let me give you just one small example. At some point in my speech career it occurred to me that the signal-to-noise power ratio (SNR) of voiced speech contaminated by white noise could be enhanced a hundredfold by a pitch-tracking comb-filter. And the physical measurements on a noisy speech signal so processed showed that, indeed, the SNR was improved by more than 20 dB. But when it came to assessing the improvement in speech intelligibility - a subjective measure - the result was nil: the processed speech, although sounding less noisy, was not a bit more intelligible than the unfiltered signal. In other words, whatever my pitch-tracking comb-filter did, the listener's brain could do just as well, using its own biological 'software'.

2. Auditory Masking by Speech Signals

While this early lesson on the interaction of speech and hearing was negative from the point of view of a practical application (enhancement of intelligibility) another interplay between speech and hearing has paid off very handsomely: the exploitation of auditory masking to reduce the audibility of quantizing noise in digital speech. When we first started to apply the principle of linear predictive coding (LPC), everybody was gratified by the resulting high speech quality. Only at very low bit rates some distortion, caused by quantizing noise, could be heard. Then it occurred to me that even this distortion could be eliminated by sweeping the quantizing noise under the 'formant rug', so to speak. Quantizing noise in linear predictive coding typically has a flat spectrum as shown in Fig. 1. But by computing the loudness of the noise in the presence of the speech signal and then minimizing it, resulting in a noise spectrum as shown in Fig. 2, we can make the noise practically inaudible. Even at 1 bit/sample for the prediction residual signal, the quantizing noise is inaudible and input and output speech signals are virtually indistinguishable. Fig. 3 shows some of the steps of this computation. Most of this work was done in collaboration with J.L. Hall and B.S. Atal of Bell Laboratories and P. Calavrytinos at Göttingen (Schroeder et al., 1979*a*, 1979*b*, 1982).

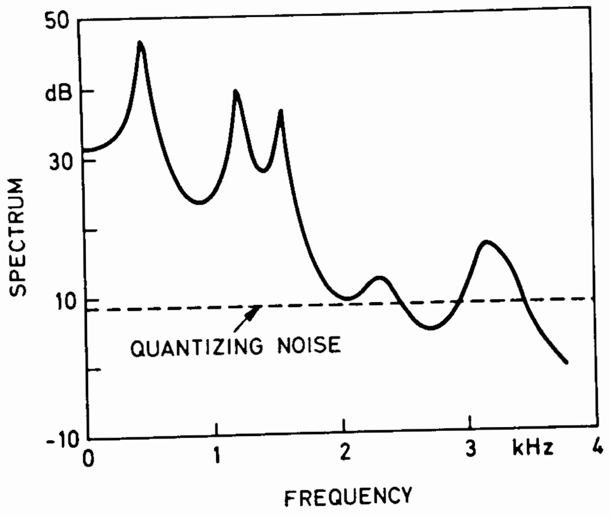


Fig. 1. Quantizing noise in LPC.

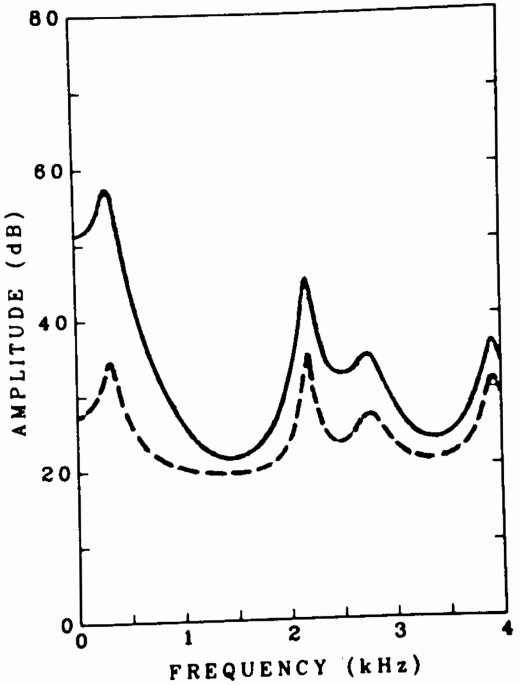


Fig. 2. Minimized loudness of noise in the presence of the speech signal.

SPECTRAL ANALYSIS	$S(t) \longrightarrow \bar{S}(f)$
RELATION BETWEEN FREQUENCY AND TONALNESS	$f = 650 \sinh(\frac{x}{7})$
TRANSFORMATION FROM Hz to Bark	$\bar{S}(f) \longrightarrow S(x)$
NEURAL EXCITATION FOR SPEECH SIGNAL	$E(x) = S(x) * B(x)$
NEURAL EXCITATION FOR QUANTIZING NOISE	$Q(x) = N(x) * B(x)$

Fig. 3. Some of the steps in the computation.

3. Enhancement of Temporal Cues

I mentioned Meyer-Eppler before and will mention him again. When I visited his Institute at the University of Bonn in 1958, he showed me - among other things - an old German patent of his: a kind of vowel-consonant switch that would increase the power level of consonants in connected speech to prevent them from being masked by preceding vowel sounds in a reverberant auditorium. A beautiful idea - except it did not work. The constant switching made the original speech signal rather unintelligible even before it was subjected to reverberation.

Later, when I learned about the modulation transfer function (MTF), I thought I could outwit reverberation by *increasing* the modulation envelope of speech signals - but to no avail. Nevertheless, manipulating the envelope of speech signals has led to success in another application - improving speech intelligibility for the neurally deaf - more specifically, people who suffer from recruitment. Such patients have only a very narrow level range between the thresholds of hearing and discomfort. If speech is to be intelligible to them, its amplitude has to be fitted into the narrow level 'window'. This would call for amplitude compression. However, such compression would decrease an already marginal intelligibility, because much speech articulation is reflected not in stationary amplitudes and spectra, but in *temporal* cues and fast *transitions* between stationary states. Thus, a more sophisticated kind of modulation processing is called for as illustrated in Fig. 4, which shows different modulation transfers (M), versus modulation frequency Ω. Panel 1 corresponds to an unmodified MTF and panel 2 illustrates ordinary dynamic compression, in which all modulation frequencies are equally attenuated (by about 12 dB in the illustration). The third panel shows a reduction of the modulation at low modulation frequencies (below 5 Hz). The fourth panel shows both this reduction *and* an increase in the modulation at high modulation trequencies, thereby both compressing the dynamic range (determined mostly by the low modulation frequencies) and enhancing the transients. (The corresponding operation in vision is called *edge enhancement*.)

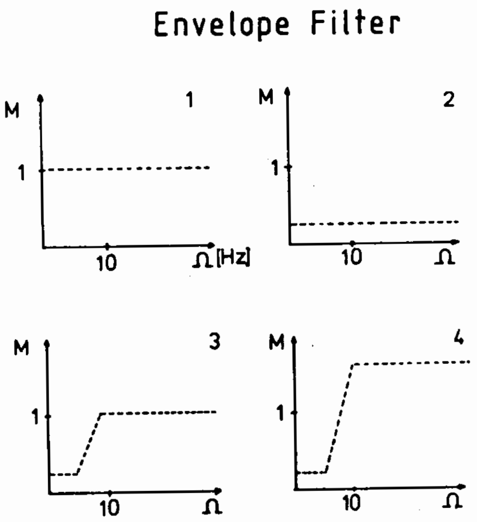


Fig. 4. Different modulation transfers (M) vs. modulation frequency.

The modification of the modulation spectrum of speech is done separately for each critical band of hearing. The success of this method in enhancing speech intelligibility, both in noisy environments and for recruitment patients, was demonstrated by H.W. Strube, T. Lewien, and T. Langhans and described at the 1982 FASE/DAGA Congress in Göttingen.

4. The Importance of the Time Waveform of Speech Signals

In the 1950s, when I first became interested in speech synthesis, I was almost immediately intrigued by the problems of subjective quality of synthetic speech. Vocoders had a reedy, 'electronic' accent and I thought that the excitation waveform, consisting of sharp pulses for voiced sounds, was perhaps to blame. To investigate this question more deeply, I built a generator for 31 coherent harmonics of variable fundamental frequency. The phase of each harmonic could be chosen to be either 0 or π - a total of $2^{30} = 1{,}073{,}741{,}824$ different waveforms, each of which appeared to have its own intrinsic timbre - their identical power spectra notwithstanding. (I wish Seebeck, Ohm and Helmholtz had had a chance to listen to these stimuli!)

For all phase angles set equal to 0, one obtains a periodic cosine-pulse, see Fig. 5. When this waveform is used as an excitation signal for a speech synthesizer, the result is the reedy quality already mentioned. By contrast, if

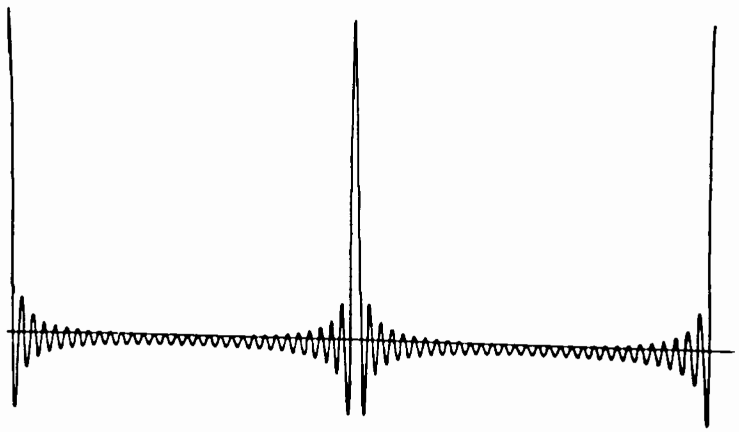


Fig. 5. A periodic cosine-pulse, with phase angles set equal to 0.

one randomizes the phase angles, one gets a less peaky waveform, see Fig. 6, and a mellower sound (Schroeder, 1959). A better-than random choice for the phase angles (one that gives an even less peaky waveform) is given by the formula

$$\varphi_n = \pi n^2/N$$

where n is the harmonic number and N the total number of harmonics in the flat spectrum stimulus. More general formulae, for arbitrary and phase angles restricted to 0 or π, are given in Schroeder, 1970.

Many of the waveforms generated by phase manipulation, although they had smooth spectra without formant structure, had a *vowel*-like quality; I shall return to this astounding observation in a moment. Let me first describe an auditory masking experiment performed by Sönke Mehrgardt at our Institute, which gave an astonishing monaural phase effect.

Mehrgardt took a harmonic tone complex with a speech-like power spectrum. The phase angles of the individual harmonics were either constant or random. The two resulting signals, one reedy sounding, the other with a vowel-like quality, were used to mask a 1200-Hz pure tone. The masked thresholds are shown in Fig. 7. For the random-phase masker, the masked threshold increases with decreasing fundamental frequency of the masker - as expected, because more masker harmonics fall into the critical band around the maskee (the test tone). By contrast for the constant-phase masker (open circles) the masked threshold *decreases* with decreasing fundamental frequency. A paradox, a 40-dB monaural phase effect!

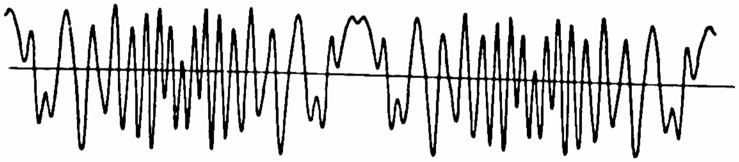


Fig. 6. A periodic cosine-pulse, with randomized phase angles.

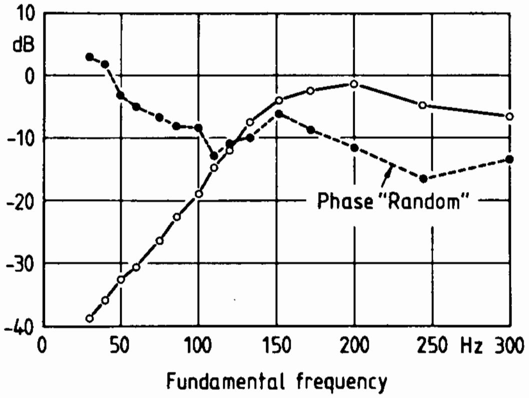


Fig. 7. Direct masking threshold of test tone (1200Hz) masked by a harmonic complex.

How is this possible, considering that G.S. Ohm and H.v. Helmholtz argued about whether there are *any* monaural phase effect at all? A possible explanation is illustrated in Figs. 8 and 9, which show the effects on the time waveforms of doubling the 7th harmonic. While this is clearly visible in the case of the constant-phase spectrum (Fig. 8) no such obvious extra signal at the 7th harmonic is detectable in the case of the 'random'-phase masker (Fig. 9).

But it might be argued that higher centers in the auditory pathway do not 'see' the waveforms shown in these two figures. The inner ear performs a critical-band analysis, resulting in signals as illustrated in Fig. 10.

However, after this bandpass filtering by the ear there remain pronounced waveform differences between constant and 'random'-phase signals. For the constant-phase masker, there are still time gaps in the waveform during which the presence of a test-tone could be be detected, especially at very low fundamental frequency, confirming the result of Fig. 7. If this explanation is to hold water, then a short tone pulse used as a signal should show characteristic variations of the masked threshold with time reflecting the *envelope of the masker*. This is indeed observed as shown in Fig. 11 (Mehrgardt, 1982).

I will not pursue here the possibility of using this experimental paradigm to explore the temporal response characteristics of the inner ear. Rather, I want to stress the importance of temporal detail in speech-like waveforms. These results suggest the existence of short time-windows during which the human ear can 'look' at waveforms and perceive important differences even for given amplitude spectra. If this is true, it should be possible to create speech-like percepts for signals having smooth spectra *without formant structure*.

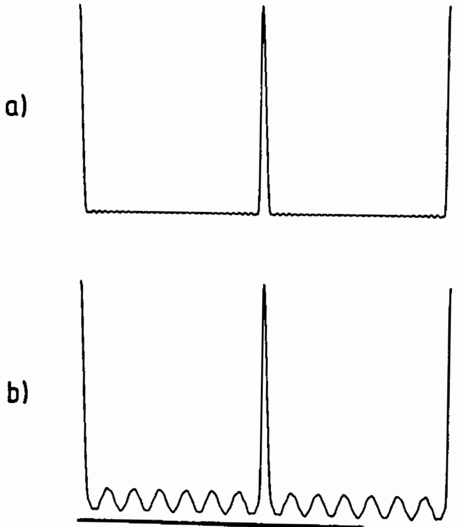


Fig. 8. Waveform of 31 harmonics in zero phase. a) amplitudes Hamming-weighed; b) amplitudes as in a), but amplitude of seventh harmonic doubled.

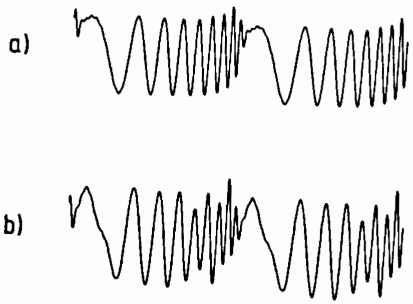


Fig. 9. Waveform of 31 harmonics with "random" phase. a) amplitudes Hamming-weighed; b) amplitudes as in a), but amplitude of seventh harmonic doubled.

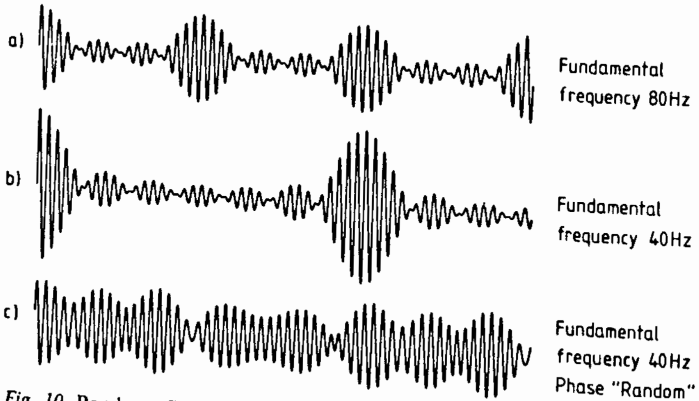


Fig. 10. Bandpass filtered harmonic complex (1/3 octave bandpass, centered at 1500Hz).

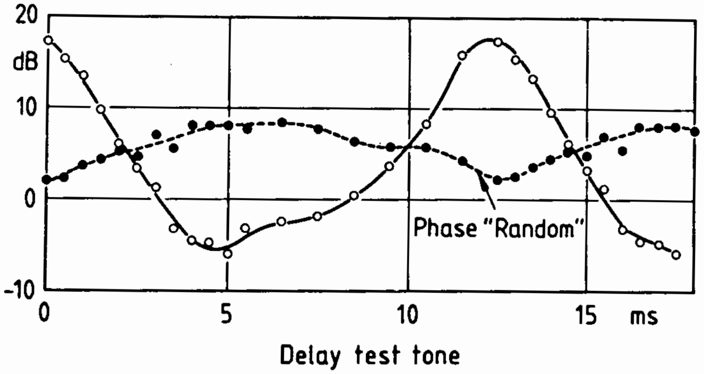


Fig. 11. Direct masking threshold of test tone burst (1600Hz, 5ms) masked by a harmonic complex (fundamental frequency 80Hz).

Such an experiment was performed by Dr. Strube at our institute who synthesized *flat*-spectrum signals corresponding to the three diphthongs /au/, /oI/, and /aI/. This is accomplished by adjusting phase angles below and above the drawn-in vowel spectrum as shown in Fig. 12. All three signals sound clearly speech-like, although there is no formant structure. In fact, they clearly sound like the diphthongs underlying the synthesis. As expected, the diphthongs-like percept is more pronounced at lower fundamental frequencies, consistent with the time-window model.

But is this time-window a complete explanation? More recent results by Mehrgardt hint that it is not. The different maskers illustrated in Fig. 13 all should give the same threshold for the test-tone pulse; but they do not, as shown in Fig. 14. Rather, as the random frequency shifts of the masker components are increased, the threshold increases by about 18 dB although the time gap during which the test tone occurs is clearly defined for all conditions. No such increase of threshold is observed if all frequency components are shifted by the same amount. These results suggest that the ear needs

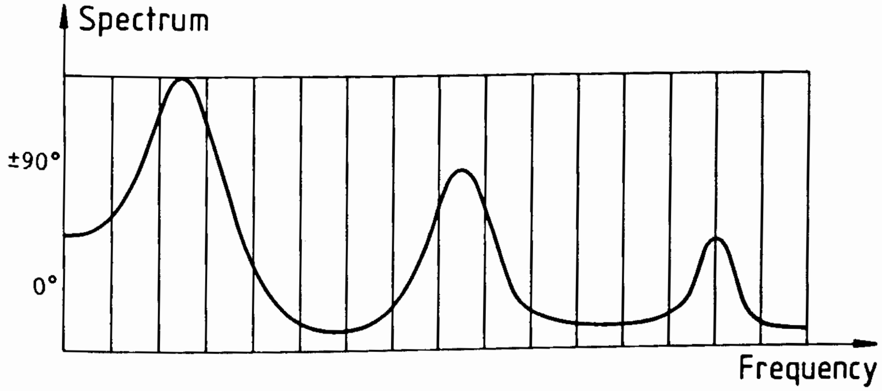


Fig. 12. Flat spectrum signal with adjusted phase angles.

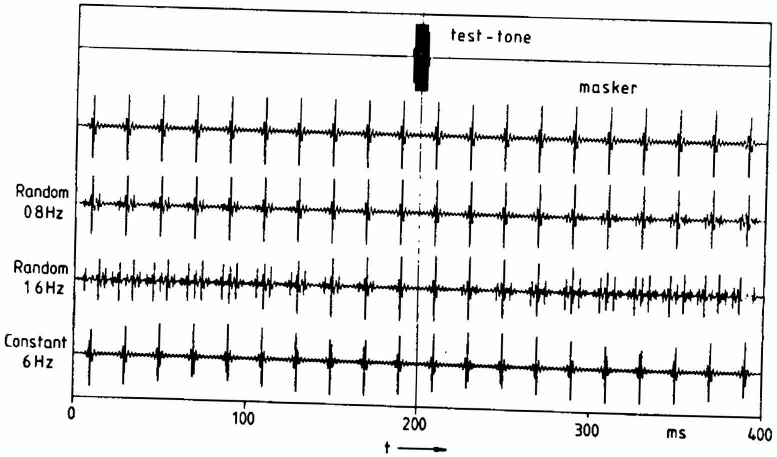


Fig. 13. Different maskers with test tone pulse.

many time gaps (and not just a single one) to optimally detect a single test-tone pulse!

It looks as if *waveform* memory is involved. Is this possible, or even thinkable? We have reached a point where only further experiments can help us. We know quite a bit about the auditory periphery, but woefully little about the higher interpretive functions of hearing. Why do two voices sound like two voices under most conditions? How does our brain integrate the separate harmonic components of a given speech signal into a *single* vowel percept? Precisely under what conditions does such integration take place? Here are some of the most interesting questions for future research.

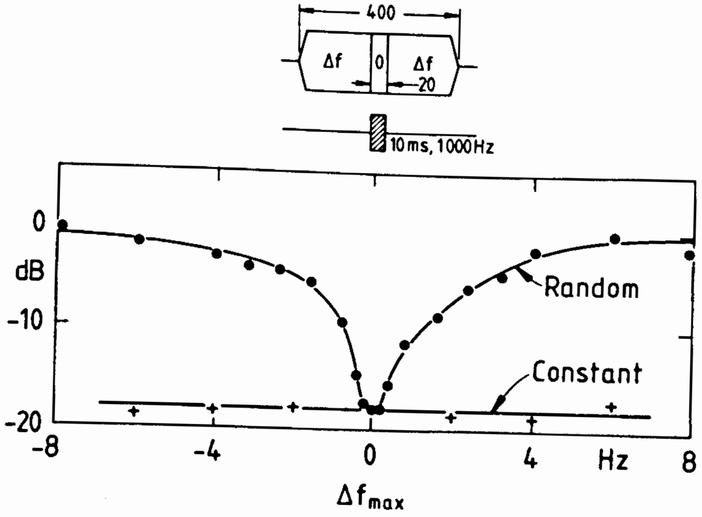


Fig. 14. Increase of threshold with increased random frequency shift of the masker.

What, in speech perception do we hear as 'figure' and what as 'background', to use terms familiar from visual perception. Let me illustrate from an example from language comprehension: There are hundreds of words spelled alike in English and German but having different meanings. I once wrote a little German story using only such words and showed it to a German-speaking friend in the United States and asked him what he thought of it. Answer: 'Nothing, a random collection of words'. Half a year later in *Germany* I showed the same friend the same story and asked him again what he thought of it. The answer this time: 'Quite an interesting story, who wrote it? When I asked him whether he saw anything unusual in the words, the answer was: 'No, I don't see anything.'

As an example of a possibly interesting figure background experiment in speech perception let me sketch Fig. 15. Again we have a flat spectrum with a vowel contour drawn. The harmonic frequency components below and above the vowel line are altered either randomly or more or less coherently along the frequency axis, either by fixed amounts or in a time-varying manner (jitter). At what degree of coherence or jitter do we hear a vowel-like sound? Or will we hear two vowels - or none?

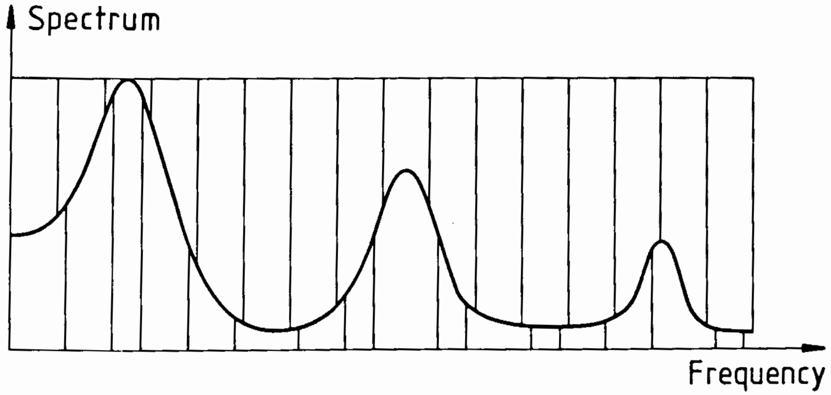


Fig. 15. Flat spectrum signal with random alternations below and above vowel line.

5. Conclusion

What is our main conclusion? We have reaped nice benefits on the basis of our present knowledge. We have made quantizing noise in digital speech practically inaudible even at very low bit rates. We have manipulated the modulation of speech intelligibility in noise and for certain kinds of hearing disorders. And we have shown the importance of temporal cues, beyond prior expectation. But there are still large blank areas on our map of understanding and our most interesting insights into speech and auditory perception are yet to come.

References

Mehrgardt, S. (1982). Kanteneffekte der Mithoerschwelle bei Hochpass- und Tiefpass-Maskierern aus harmonischen Tonkomplexen. In: Fortschritte der Akustik - FASE/DAGA '82, Göttingen: Drittes Physik. Inst., 1223-1234. Göttingen.

Schroeder, M.R. (1959). New results concerning monaural phase sensitivity. *Journal of the Acoustical Society of America,* **31**, 1579(A). See also: Duifhuis, H. (1970). Audibility of high harmonics in a periodic pulse. *Journal of the Acoustical Society of America,* **48**, 888 and Duifhuis, H. (1971). Audibility of high harmonics in a periodic pulse. II. Time Effect. *Journal of the Acoustical Society of America,* **49**, 1155.

Schroeder, M.R. (1970). Synthesis of low-peak-factor signals and binary sequences with low autocorrelation. *IEEE Transactions on Information Theory*, IT-16, No. 1, January 1970.

Schroeder M.R., Atal, B.S., and Hall, J.L. (1979*a*). Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In: *Frontiers of Speech Communication Research.* B. Lindblom and S. Öhman, eds., London: Academic Press, no. 217-229.

Schroeder. M.R.. Atal. B.S.. and Hall, J.L. (1979*b*). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America,* **66**, 1647-1652.

Schroeder, M.R. and Atal, B.S. (1982). Speech coding using efficient block codes. *IEEE Proceedings*, International Congress on Acoustics, Speech and Signal Processing, May 1982, Paris. **3**, 1668-1671.

2. Relation between Speech Production and Speech Perception

Relation between Speech Production and Speech Perception

L.A. Chistovich
Leningrad, USSR

1. Introduction

The problem of relations between auditory representation of speech unit and the 'goal' in the program for this unit production is highly important for both speech production and speech perception theories. Auditory control of timing in execution of motor program is another aspect of the problem. It was supposed that the acoustical events arising at the onset of speech sound production might trigger, after prescribed delay, the execution of the next speech unit (the 'chain' model of production).

Neurophysiological studies of the central auditory system suggest the extraction of two kinds of information from the peripheral auditory pattern - one, most appropriate for timing control (short 'phasic' responses to rapid spectral and amplitude changes) and the other one, more appropriate for specification of the goals (selective responses to specific spectrum shapes, direction of spectral peak transition and so on). It is important to note that auditory neurons seem to have only a primitive memory: they can integrate, with some time constants, the incoming excitations and inhibitions and they can become temporarily blocked after firing. The time window of processing appeared to be different for different neurons but it did not exceed 200 ms.

The aim of this paper is to review some experiments where external speech stimuli were used to control speech production. In speech-by-speech synchronization experiments the subjects produce the prescribed response, only the timing might be controlled by the stimulus. Experiments on mimicking concern the goals formation. Both the goals formation and the timing are involved in shadowing.

2. Speech-by-speech synchronization

The subject can synchronize the production (response) with periodically presented stimuli (clicks, tone pulses) and make stimulus and response overlap in time. It was speculated that if speech stimuli were to be used for synchronization, the speech execution mechanism might mistake the marker of the speech sound onset in the stimulus for the marker of the corresponding onset in the response. It was found: Chistovich et al. (1972) that the interval between V_1 onset in VCV-stimulus and V_2 onset in VCV-response was really more stable than the interval between V_1 and V_2 onsets in the response. The variability of the last interval (between V_1 and V_2 in response) appeared to be much higher in these conditions than in both free and synchronized by clicks VCV productions.

3. Mimicking of isolated sounds

Experiments on mimicking the loudness of fricative sound (Malinnikova, 1968), vowel duration (Chistovich et al., 1966) and tone pitch (Lublinskaja, 1968) were at first aimed to test whether the subject's goal was to match the response with the stimulus. Clear negative results were obtained: the subjects were very good in preserving in responses the orderly relations among stimuli but they did not reproduce the absolute values of the stimuli. (Mimicking of pitch by musically trained subjects was an exception). This points to wired up scales relating the auditory system outputs to motor control parameters. The efforts were aimed at finding out whether these scales are fixed or adaptive and when the last alternative appeared to be true, to study the variables controlling the scale adaptation (Malinnikova, 1971).

Experiments on mimicking synthetic vowel by subjects with different sizes of their vocal tracts (males, females, children of different ages) have shown that the subjects preserve the orderly relations among stimuli in formant frequencies space but they do not match the spectrum of the response with that of the stimulus (Kent, 1978; Kent et al., 1979). There are indications that vowel mimicking is an innate behavior. The important problem is to find out whether continuous or discrete scales relate responses to stimuli in vowel mimicking. Clustering of responses predicted by categorization has been observed (Chistovich et al., 1966; Kent, 1973; Kent, 1978) but far more extensive data are needed for a reliable conclusion.

4. Shadowing

The ability of subjects to rapidly imitate (shadow) natural and synthetic speech is a well documented fact. The data on shadowing stop consonants in VCV stimuli are best suited to discuss the implications of the effect and the problems involved in its analysis. Identification experiments have shown that although some information about consonant identity is conveyed by closure transition and initial part of closure, the subjects rely in phoneme decision on the events following the release of closure. In shadowing VCV the subjects start the consonant production before the release of closure in the stimulus (Kozhevnikov et al., 1965; Porter et al., 1980). That means that auditory information corresponding to closure transition is transformed into motor representation (goal or the set of goals) and could be stored in this form till new auditory data arrive. It was found that consonant response might begin with erroneous articulation, which could be corrected in the course of production. It was tempting to speculate (Kozhevnikov et al., 1965) that response modifications in shadowıng reflect the temporal process of phonetic interpretation. I shall present some arguments against this view and in favour of the idea that the execution of the motor program observed in shadowing and the formation of this program might appear to be two parallel processes controlled by a different kind of auditory information. The latencies of shadowing are equal to simple reaction time. It was found that the actual signal eliciting the response in experiments on simple reaction time to tone is not the tone but the event of onset (presumably on-response). Substitution of the stimulus with the tone of far different frequency results in the same response with just the same latency (Chistovich, 1956). If shadowing resonse to vowel is also triggered by the onset of the stimulus, then by cutting out the late parts of the vowel we might influence the quality but not the latency of the response. The experiments on shadowing the natural whole and truncated vowels confirmed these expectations. The critical stimulus duration determining the initial part of response appeared to be between 50 and 100 ms (Chistovich et al., 1962).

The experiments on shadowing synthetic /ao/, /aæ/ and ai/stimuli with long and variable /a/ duration (Porter et al., 1980) have shown that subjects start correct response to second vowel with a latency of 150 ms from the onset of the formant transitions. The same or a little longer latencies were observed in simple reaction time situation: subjects had to respond by /ao/ to all three kinds of stimuli. This also suggests that the same events trigger the response execution in both tasks.

5. Mimicking of simple sequencies.

Comparison of mimicking response to isolated stimulus with the response to the same stimulus in contest seems to be a good approach to study contextual rules. Pronounced contrast effect has been observed in formant patterns of the second vowel produced in mimicking VV-stimuli with different first vowels (Kent, 1974). It was also observed in vowel durations produced in mimicking VV with different durations of the vowels in the stimulus (Zhukov, 1971). Pitch contrast effect was studied on musically trained subjects, who were instructed to listen to a tone pair and precisely reproduce both stimuli. The subjects followed the instruction when the frequency difference between stimuli was large. When it was small, they made one response higher and the other one lower than the corresponding stimulus. It seems that the subjects tried to preserve the average to the pair pitch and to increase the difference between components of the pair (Lublinskaja, 1970). It is clear that this kind of processing is not compatible with the facts concerning the auditory system. True memory and the ability to read out and modify the previously recorded item are necessary.

6. Concluding remarks

It is obvious that the brain must possess some 'language' to translate the auditory information into information to the motor system. The data on shadowing suggest that the translation occurs with short delay and does not require long auditory memory. The results on mimicking suggest that this audio-motor 'language' is at least partly innate. It is tempting to speculate that several phonetical effects and regularities reflect in fact the structure and the rules of this 'language' and could be found under close examination in various perceptual-motor skills.

References

Chistovich, L.A. (1956). Comparison of conditioned motor reactions based on shock and verbal reinforcements. *Fiziol. zh. USSR,* **43**, 572-580.

Chistovich, L.A., Fant G., de Serpa-Leitao, A., Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quart. Progr. Status Rep. Speech Transm. Lab. Roy. Inst. Technol. Stockholm* **2**, 1-18.

Chistovich, L.A., Klaas Ju.A. (1962). Toward analysis of latency of 'voluntary' reaction to sound. *Fiziol. zh. USSR,* **48**, 899-906.

Chistovich L.A., Lissenko, D.M., Fedorova, N.A. (1972). Speech production control under synchronization with periodically presented speech stimulus. In: *Sensornye sistemy,* **2**, Leningrad, 56-85.

Chistovich, L.A., Zhukova, M.G., Malinnikova T.G., Kozhevnikov, V.A., Borozdin, A.N. (1966). Mimicking and perception of isolated vowels. In: *Mechanizmy recheobrazovania i vospriatia zlozhnych zvukov*. Leningrad, 128-157.

Kent, R.D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica,* **28**, 1-25.

Kent, R.D. (1974). Auditory-motor formant tracking: a study of speech imitation. *J. Speech Hear. Res.,* **17**, 203-222.

Kent R.D. (1978). Imitation of synthesized vowels by preschool children. *J. Acoust. Soc. Amer.,* **63**, 1193-1198.

Kent, R.D., Forner, L.L. (1979). Developmental study of vowel formant frequencies in an imitation task. *J. Acoust. Soc. Amer.,* **65**, 208-217.

Kozhevnikov, V.A., Chistovich, L.A. (1965). Speech: Articulation and perception (Translated from Russian). *Report* **30**, 543. Washington, D.C.: Joint Publication Research Service.

Lublinskaja, V.V. (1968). Mimicking of pitch. *Ztschr. für Phonetik* **21**, 129-134.

Lublinskaja, V.V. (1970). Mimicking of pitch interval in sequences of two tonal stimuli. In: *Upravlenie dvizheniami*. Leningrad, 110-117.

Malinnikova, T.G. (1968). Mimicking of loudness of synthetic fricative consonants. *Ztschr. fur Phonetik,* **21**, 135-139.

Malinnikova, T.G. (1971). Relations between response intensity and the intensity range of stimuli in consonant mimicking. In: *Sensornye sistemy,* **2**, Leningrad, 99-110.

Porter, R.J., Castellanos, F.X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *J. Acoust. Soc. Amer.,* **67**, 1349-1356.

Porter, R.J., Lubker, J.F. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motor linkage in speech. *J. Speech Hear. Res.* **23**, 593-602.

Zhukov, S.J. (1971). About auditory segmentation of /iu/. In: *Sensornye sistemy,* **2**, Leningrad, 71-82.

Relation between Speech Production and Speech Perception

L.A. Chistovich
Leningrad, USSR

1. Introduction

The problem of relations between auditory representation of speech unit and the 'goal' in the program for this unit production is highly important for both speech production and speech perception theories. Auditory control of timing in execution of motor program is another aspect of the problem. It was supposed that the acoustical events arising at the onset of speech sound production might trigger, after prescribed delay, the execution of the next speech unit (the 'chain' model of production).

Neurophysiological studies of the central auditory system suggest the extraction of two kinds of information from the peripheral auditory pattern - one, most appropriate for timing control (short 'phasic' responses to rapid spectral and amplitude changes) and the other one, more appropriate for specification of the goals (selective responses to specific spectrum shapes, direction of spectral peak transition and so on). It is important to note that auditory neurons seem to have only a primitive memory: they can integrate, with some time constants, the incoming excitations and inhibitions and they can become temporarily blocked after firing. The time window of processing appeared to be different for different neurons but it did not exceed 200 ms.

The aim of this paper is to review some experiments where external speech stimuli were used to control speech production. In speech-by-speech synchronization experiments the subjects produce the prescribed response, only the timing might be controlled by the stimulus. Experiments on mimicking concern the goals formation. Both the goals formation and the timing are involved in shadowing.

2. Speech-by-speech synchronization

The subject can synchronize the production (response) with periodically presented stimuli (clicks, tone pulses) and make stimulus and response overlap in time. It was speculated that if speech stimuli were to be used for synchronization, the speech execution mechanism might mistake the marker of the speech sound onset in the stimulus for the marker of the corresponding onset in the response. It was found: Chistovich et al. (1972) that the interval between V_1 onset in VCV-stimulus and V_2 onset in VCV-response was really more stable than the interval between V_1 and V_2 onsets in the response. The variability of the last interval (between V_1 and V_2 in response) appeared to be much higher in these conditions than in both free and synchronized by clicks VCV productions.

3. Mimicking of isolated sounds

Experiments on mimicking the loudness of fricative sound (Malinnikova, 1968), vowel duration (Chistovich et al., 1966) and tone pitch (Lublinskaja, 1968) were at first aimed to test whether the subject's goal was to match the response with the stimulus. Clear negative results were obtained: the subjects were very good in preserving in responses the orderly relations among stimuli but they did not reproduce the absolute values of the stimuli. (Mimicking of pitch by musically trained subjects was an exception). This points to wired up scales relating the auditory system outputs to motor control parameters. The efforts were aimed at finding out whether these scales are fixed or adaptive and when the last alternative appeared to be true, to study the variables controlling the scale adaptation (Malinnikova, 1971).

Experiments on mimicking synthetic vowel by subjects with different sizes of their vocal tracts (males, females, children of different ages) have shown that the subjects preserve the orderly relations among stimuli in formant frequencies space but they do not match the spectrum of the response with that of the stimulus (Kent, 1978; Kent et al., 1979). There are indications that vowel mimicking is an innate behavior. The important problem is to find out whether continuous or discrete scales relate responses to stimuli in vowel mimicking. Clustering of responses predicted by categorization has been observed (Chistovich et al., 1966; Kent, 1973; Kent, 1978) but far more extensive data are needed for a reliable conclusion.

4. Shadowing

The ability of subjects to rapidly imitate (shadow) natural and synthetic speech is a well documented fact. The data on shadowing stop consonants in VCV stimuli are best suited to discuss the implications of the effect and the problems involved in its analysis. Identification experiments have shown that although some information about consonant identity is conveyed by closure transition and initial part of closure, the subjects rely in phoneme decision on the events following the release of closure. In shadowing VCV the subjects start the consonant production before the release of closure in the stimulus (Kozhevnikov et al., 1965; Porter et al., 1980). That means that auditory information corresponding to closure transition is transformed into motor representation (goal or the set of goals) and could be stored in this form till new auditory data arrive. It was found that consonant response might begin with erroneous articulation, which could be corrected in the course of production. It was tempting to speculate (Kozhevnikov et al., 1965) that response modifications in shadowing reflect the temporal process of phonetic interpretation. I shall present some arguments against this view and in favour of the idea that the execution of the motor program observed in shadowing and the formation of this program might appear to be two parallel processes controlled by a different kind of auditory information. The latencies of shadowing are equal to simple reaction time. It was found that the actual signal eliciting the response in experiments on simple reaction time to tone is not the tone but the event of onset (presumably on-response). Substitution of the stimulus with the tone of far different frequency results in the same response with just the same latency (Chistovich, 1956). If shadowing resonse to vowel is also triggered by the onset of the stimulus, then by cutting out the late parts of the vowel we might influence the quality but not the latency of the response. The experiments on shadowing the natural whole and truncated vowels confirmed these expectations. The critical stimulus duration determining the initial part of response appeared to be between 50 and 100 ms (Chistovich et al., 1962).

The experiments on shadowing synthetic /ao/, /aæ/ and ai/stimuli with long and variable /a/ duration (Porter et al., 1980) have shown that subjects start correct response to second vowel with a latency of 150 ms from the onset of the formant transitions. The same or a little longer latencies were observed in simple reaction time situation: subjects had to respond by /ao/ to all three kinds of stimuli. This also suggests that the same events trigger the response execution in both tasks.

5. Mimicking of simple sequencies.

Comparison of mimicking response to isolated stimulus with the response to the same stimulus in contest seems to be a good approach to study contextual rules. Pronounced contrast effect has been observed in formant patterns of the second vowel produced in mimicking VV-stimuli with different first vowels (Kent, 1974). It was also observed in vowel durations produced in mimicking VV with different durations of the vowels in the stimulus (Zhukov, 1971). Pitch contrast effect was studied on musically trained subjects, who were instructed to listen to a tone pair and precisely reproduce both stimuli. The subjects followed the instruction when the frequency difference between stimuli was large. When it was small, they made one response higher and the other one lower than the corresponding stimulus. It seems that the subjects tried to preserve the average to the pair pitch and to increase the difference between components of the pair (Lublinskaja, 1970). It is clear that this kind of processing is not compatible with the facts concerning the auditory system. True memory and the ability to read out and modify the previously recorded item are necessary.

6. Concluding remarks

It is obvious that the brain must possess some 'language' to translate the auditory information into information to the motor system. The data on shadowing suggest that the translation occurs with short delay and does not require long auditory memory. The results on mimicking suggest that this audio-motor 'language' is at least partly innate. It is tempting to speculate that several phonetical effects and regularities reflect in fact the structure and the rules of this 'language' and could be found under close examination in various perceptual-motor skills.

References

Chistovich, L.A. (1956). Comparison of conditioned motor reactions based on shock and verbal reinforcements. *Fiziol. zh. USSR,* **43**, 572-580.

Chistovich, L.A., Fant G., de Serpa-Leitao, A., Tjernlund, P. (1966). Mimicking of synthetic vowels. *Quart. Progr. Status Rep. Speech Transm. Lab. Roy. Inst. Technol. Stockholm* **2**, 1-18.

Chistovich, L.A., Klaas Ju.A. (1962). Toward analysis of latency of 'voluntary' reaction to sound. *Fiziol. zh. USSR,* **48**, 899-906.

Chistovich L.A., Lissenko, D.M., Fedorova, N.A. (1972). Speech production control under synchronization with periodically presented speech stimulus. In: *Sensornye sistemy,* **2**, Leningrad, 56-85.

Chistovich, L.A., Zhukova, M.G., Malinnikova T.G., Kozhevnikov, V.A., Borozdin, A.N. (1966). Mimicking and perception of isolated vowels. In: *Mechanizmy recheobrazovania i vospriatia zlozhnych zvukov.* Leningrad, 128-157.

Kent, R.D. (1973). The imitation of synthetic vowels and some implications for speech memory. *Phonetica,* **28**, 1-25.

Kent, R.D. (1974). Auditory-motor formant tracking: a study of speech imitation. *J. Speech Hear. Res.,* **17**, 203-222.

Kent R.D. (1978). Imitation of synthesized vowels by preschool children. *J. Acoust. Soc. Amer.,* **63**, 1193-1198.

Kent, R.D., Forner, L.L. (1979). Developmental study of vowel formant frequencies in an imitation task. *J. Acoust. Soc. Amer.,* **65**, 208-217.

Kozhevnikov, V.A., Chistovich, L.A. (1965). Speech: Articulation and perception (Translated from Russian). *Report* **30**, 543. Washington, D.C.: Joint Publication Research Service.

Lublinskaja, V.V. (1968). Mimicking of pitch. *Ztschr. für Phonetik* **21**, 129-134.

Lublinskaja, V.V. (1970). Mimicking of pitch interval in sequences of two tonal stimuli. In: *Upravlenie dvizheniami.* Leningrad, 110-117.

Malinnikova, T.G. (1968). Mimicking of loudness of synthetic fricative consonants. *Ztschr. fur Phonetik,* **21**, 135-139.

Malinnikova, T.G. (1971). Relations between response intensity and the intensity range of stimuli in consonant mimicking. In: *Sensornye sistemy,* **2**, Leningrad, 99-110.

Porter, R.J., Castellanos, F.X. (1980). Speech production measures of speech perception: Rapid shadowing of VCV syllables. *J. Acoust. Soc. Amer.,* **67**, 1349-1356.

Porter, R.J., Lubker, J.F. (1980). Rapid reproduction of vowel-vowel sequences: Evidence for a fast and direct acoustic-motor linkage in speech. *J. Speech Hear. Res.* **23**, 593-602.

Zhukov, S.J. (1971). About auditory segmentation of /iu/. In: *Sensornye sistemy,* **2**, Leningrad, 71-82.

Relation between Speech Production and Speech Perception

Hiroya Fujisaki
Tokyo, Japan

1. Introduction

It is apparent that the mechanisms and processes of speech production and those of speech perception are essentially different. Functionally, speech is produced by successive execution of prestored motor programs, while speech is perceived by categorization of successive linguistic units and by recognition of their hierarchical structure. The acoustic characteristics of speech perceived by a speaker are not necessarily identical to those of speech which he/she produces, because of various contextual, idiosyncratic, and dialectical variations. On the other hand, it is quite natural to assume that the original message to be uttered by a speaker and the ultimate message to be received by the same person as listener would have the same form of linguistic representation in his mind. Moreover, many people tend to assume the existence of a further link between speech production and speech perception. In fact, several theories or models have been presented on their possible relations, such as the articulatory reference theory, the analysis-by-synthesis model (Halle and Stevens 1959), the motor theory of speech perception (Liberman et al. 1962), the auditory pattern model (Fant, 1967), the auditory-motor theory of speech production (Ladefoged et al., 1972), etc. Experimental evidences in support of these theories or models, however, have been rather scarce and indirect.

In this paper I will not try to review nor to criticize these theories or models, but will try to review some recent findings that will lead to a more concrete understanding of the possible links between speech production and speech perception. The works to be reviewed here can be classified into the following four areas:

1. Influence of speech production upon speech perception,
2. Influence of speech perception upon speech production,
3. Relation between production and perception in language learning and
4. Speech productions and perception by nonhuman vertebrates.

2. Influence of Speech Production upon Speech Perception

The human process of speech production is constrained by the physiological and physical properties of the mechanisms involved. Thus considerable variations can be found in the realization of the same linguistic unit due to context, changes in speech rate, and other factors such as speaker idiosyncrasy. The process of speech perception has to separate these extraneous factors and extract the relevant linguistic information.

The use of contextual information in speech perception has been well known, but has recently been explicitly observed and measured quantitatively. For example, the presence of a vocalic context was found to modify the identification of the neighboring voiceless fricative consonant in such a way that the coarticulatory influence is compensated for in perception (Kunisaki and Fujisaki, 1977; Mann and Repp, 1980; Whalen, 1981). In the case of Japanese, anticipatory coarticulation by the vowel immediately following a fricative consonant is much stronger than perseveratory coarticulation by the preceding vowel, and the magnitude of perceptual compensation also reflects their difference.

Speech rate is another factor exerting influence on the acoustic characteristics of segments, especially on segmental duration. The existence of perceptual compensation against changes in speech rate has also been studied (Fujisaki, Nakamura and Imoto, 1973, Nooteboom, 1974, 1978: Johnson and Strange, 1982). The range of context used in perception has been found to extend beyond the immediately neighboring phonemes and syllables (Fujisaki, Nakamura and Imoto; 1973, Nooteboom and Cohen, 1975; Martin and Bunnell, 1981).

Another constraint imposed by the production mechanism is that of speaker idiosyncrasy or individual differences in the size and properties of speech organs due to age, sex and other factors. In a study using 3 male and 3 female speakers of American English, Fox (1982) reported a consistent perceptual structure difference of vowels among these subjects as listeners, and also found a high degree of correspondence between perceptual differences and articulatory differences among the subject. In view of the fact that speakers differing widely in the physical size and shape of their vocal tract (e.g. adults and children) can communicate reliably in ordinary situations, however, the above-mentioned influence from speech production to speech perception could not be of any significant magnitude. In fact, Paliwal et al. (1983), in a separate study using 10 speakers of British English, did not find any significant correlation between production and perception of vowels.

3. Influence of Speech Perception upon Speech Production

A syllable-timed language like Japanese imposes perception-based constraints on the temporal organization of speech production. As suggested by Lehiste (1977), isochrony is found to be a perceptual phenomenon, and is manifested as an approximate uniformity of perceived duration of syllables at least in the case of Japanese (Fujisaki and Higuchi, 1979). In a study of production and perception of dissyllabic words consisting of only two vowels, it was demonstrated that the perceived durations of the first and the second vowels tended to be almost equal regardless of vowel combinations but that the articulatory onset of the second vowel, estimated by using a quantitative model of coarticulation varies rather widely depending on specific vowel combinations. There was a clear tendency that a slower articulatory transition is initiated earlier and *vice versa*, indicating that the apparent variability of onset of vowel articulation is the consequence of pre-programming to maintain the uniformity of perceived syllable duration. Even though the results could be explained both in terms of the chain model (i.e. assuming a closed-loop control) and in terms of the comb model (i.e. assuming an open-loop control, Kozhevnikov and Chistovich, 1965) it is more likely that such a perceptual isochrony should be achieved by an open-loop control based on well-prepared program for syllabic timing control (Allen and Tsukahara, 1974).

While the above-mentioned requirement on perceptual isochrony represents a form of perceptual constraint imposed on speech production, the existence of an immediate link from speech perception to speech production has also been suggested by an interesting experiment on perceptuo-motor adaptation (Cooper, 1974; Cooper and Nager, 1975). The result, however, was not replicated in a recent study by Summerfield et al. (1980). As I pointed out elsewhere (Fujisaki 1980), the selective adaptation paradigm is a useful tool, but the results have to be interpreted with caution since it creates a situation never to be encountered in natural speech communication.

4. Speech Production and Perception in Language Learning

The process of second-language learning provides another area of interest where the relation between production and perception of speech can be investigated. In a study on 72 native Spanish-speaking children learning English, both production and perception of voicing distinction in stop consonants (in terms of VOT) have been analyzed from the point of view of a) length of period of learning English, and b) the starting age (Williams, 1979). The results indicated that significant changes occur both in perception and production of voicing toward the pattern of monolingual English-speaking children selected as the control, and that the starting age had a significant effect in production, but not in perception.

The relation between production and perception of word accent was also investigated on 38 students of junior high school in Fukui learning the Tokyo dialect (Sugito and Fujisaki, 1980). The study revealed a strong correclation between the abibility of perception, as indicated by the accuracy of identification, and the ability of production, as indicated by the relative frequency of using correct accent types.

5. What do we learn from Experiments on Nonhuman Vertebrates?

A series of experiments have been reported on the perception of synthetic speech sounds by non-human animals (chinchillas and macaques), and have indicated that these animals show human-like discriminability along certain consonantal continua (Kuhl and Miller, 1978; Kuhl, 1981, 1983), and also can be trained to identify certain vowels (Burdick and Miller, 1975). These results suggest that certain phonetic categories of human speech sounds are based on some psycho-acoustic properties that can be detected also by non-human animals, and provide a basis for an interesting hypothesis concerning the evolution of the human language. One might even say that speech perception takes place in the total absence of the ability of speech production. Similar arguments could also be made on speech production on the basis of the ability of mynahs and parrots which can produce excellent approximations to human sounds. The validity of these arguments will depend, not so much on the facts themselves, but rather on one's definition of speech production and speech perception.

In my opinion, the data on speech perception by animals and prelinguistic infants simply tell us that certain sounds of human languages are selected in such a way as to take advantage of some basic psychoacoustic properties of the sound continuum which can be generated by the human vocal apparatus. It is to be noted, however, that these experiments do not explain, at least up to the present, the origin of all existing categories of speech sounds, especially the categorization of the vowel continuum which is so much language-specific and can hardly be related to any basic psycho-acoustic properties.

References

Allen, G.L. and Tsukahara, N. (1974). Cerebro-cerebellar communication system. *Physiological Review*, **54**, 957-1006.

Bell-Berti, F., Raphael, L.J., Pisoni, D.B. and Sawush, J.R. (1979). Some relationships between speech production and perception. *Phonetica*, **36**, 373-383.

Burdick, C.K. and Miller, J.D. (1975). Speech perception by the chinchilla: Discrimination of sustained /a/ and /i/. *Journal of the Acoustical Society of America* **58**, 415-427.

Cooper, W.E. (1974). Perceptuomotor adaptation to a speech feature. *Perception and Psychophysics*, **16**, 229-234.

Cooper, W.E. and Nager, R.M. (1975). Perceptuo-motor adaptation to speech: an analysis of bisyllabic utterances and a neural model. *Journal of the Acoustical Society of America*, **58**, 256-265.

Fox, R.A. (1982). Individual variation in the perception of vowels: Implications for a perception-production link. *Phonetica*, **39**, 1-22.

Fant, G. (1967). Auditory patterns of speech. In: *Models for the Perception of Speech and Visual Form*. (Wathen-Dunn, W., Ed.). 111-125, Cambridge, Mass.: The M.I.T. Press.

Fujisaki, H. (1980). Some remarks on recent issues in speech perception research. *Language and Speech*, **23**, 75-80.

Fujisaki, H. and Higuchi, N. (1979). Temporal organization of segmental features in Japanese disyllables. In: *Proceedings of the Ninth International Congress of Phonetic Sciences*, Copenhagen, 1979. Vol. II, 275-282.

Fujisaki, H., Nakamura, K. and Imoto, T. (1973). Auditory perception of duration of speech and non-speech stimuli. In: *Proceedings of the Symposium on Auditory Analysis and Perception of Speech*, Leningrad, 1973.

Halle, M. and Stevens, K.N. (1959). Analysis by Synthesis. In: *Proceedings of Seminar on Speech Compression and Processing, AFCRC, 1959, vol.* **2**.

Johnson, T.L. and Strange, W. (1982). Perceptual constancy of vowels in rapid speech. *Journal of the Acoustical Society of America*, **72**, 1761-1770.

Kozhevnikov, V.A. and Chistovich, L.A. (1965). *Speech: Articulation and Perception*. Moscow: Nauka.

Kuhl, P.K. and Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.

Kuhl, P.K. (1981). Discrimination of speech by nonhuman animals: Basic auditory sensitivities conductive to the perception of speech-sound categories. *Journal of the Acoustical Society of America*, **70**, 340-349.

Kuhl, P.K. and Padden, D.M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *Journal of the Acoustical Society of America*, **73**, 1003-1010.

Kunisaki, O. and Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. In: *Proceedings of The Ninth International Congress on Acoustics*, Madrid, 1977. Vol. I, 476. Also in: *Annual Bulletin, RILP, University of Tokyo*, **11**, 85-91.

Ladefoged, P. DeClark, J.D., Lindau, M. and Papcun, G. (1972). An auditory-motor theory of speech production. *UCLA Working Papers in Phonetics* no. **22**, 48-75.

Lehiste, I. 1977). Isochrony reconsidered. *Journal of Phonetics*, **5**, 253-263.

Liberman, A.M., Cooper, F.S., Harris, K.S. and MacNeilage, P.F. (1962). A motor theory of speech perception. In: *Proceedings of the Speech Communication Seminar*, Stockholm, 1962. vol. **2**.

Mann, V.A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics* **28**, 407-412.

Mann, V.A. and Repp, B.H. (1980). Influence of vocalic context on perception of the [š]-[s] distinction. *Perception and Psychophysics*, **28**, 213-218.

Mann, V.A. and Repp, B.H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, **69**, 548-558.

Martin, J.G. and Bunnell, H.T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, **69**, 559-567.

Nooteboom, S.G. (1974). Some context effects on phonemic categorization of vowel duration. *IPO Annual Progress Report*, **9**, 47-54.

Nooteboom, S.G. (1978). Production and perception of vowel length in spoken sentences. *IPO Manuscript* no. 335.

Nooteboom, S.G. and Cohen, A. (1975). Anticipation in speech production and its implications for perception. In: *Structure and Process in Speech Perception*. (Cohen, A. and Nooteboom, S.G. Eds.). 124-145. Berlin-Heidelberg-New York: Springer-Verlag.

Paliwal, K.K., Lindsay, D. and Ainsworth, W.A. (1983). Correlation between production and perception of English vowels. *Journal of Phonetics*, **11**, 77-83.

Repp, B.H. and Mann, V.A. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustical Society of America*, **69**, 1154-1163.

Repp, B.H. and Mann, V.A. (1982). Fricative-stop coarticulation: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **71**, 1562-1567.

Sugito, M. and Fujisaki, H. (1980). Production and perception of word accent in speakers of the Fukui and the Osaka dialects. *Transactions of the Committee on Speech Research, the Acoustical Society of Japan*, **S80**, 221-228 (in Japanese).

Summerfield, Q., Bailey, P.J. and Erickson, D. (1980). A note on perceptuo-motor adaptation of speech. *Journal of Phonetics*, **8**, 491-499.

Whalen, D.H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, **69**, 275-282.

Williams, L. (1979). The modification of speech perception and production in second-language learning. *Perception and Psychophysics*, **26**, 95-104.

Can the Models of Evolutionary Biology be Applied to Phonetic Problems?

Björn Lindblom
Stockholm, Sweden

1. A Biological Style of Inquiry

By way of introduction I would like to make two points. First, 'How do we characterize a biological style of inquiry'? From among the several possible ways of answering such questions I shall choose a formulation that I found in a paper by one of the leading figures in the development of modern biology, Dobzhansky (1965) who says that, confronting any phenomenon in living organisms, the Darwinian biologist has to ask three kinds of questions: The question of (i) *mechanism*: 'How does it work?'; the question of (ii) *function*: 'What does it do for the organism?', and the question of (iii) *origin*: 'How did it get that way?' (both for ontogeny and for phylogeny).

The point is here that the ideal biologist envisioned by Dobzhansky uses an interactive strategy. He uses the three viewpoints in parallel. He asks both HOW- and WHY-questions to elucidate a given problem.

Suppose we apply this thinking to a phonetic problem. Take the problem of speech units. Let us examine the available experimental evidence on the production of speech and moreover let us suppose that we are unable to find any facts seriously contradicting the idea that speaking involves the conversion of discrete psychological units into continuous physical signals. If we were to apply Dobzhansky's program to that particular problem it would not be sufficient to provide a description – no matter how detailed – of what human speakers actually do when they carry out the transformation from the discrete to the continuous. Our account must *also* address the other two issues, *viz.* the question of purpose and the question of origin of the proposed mechanism.

It appears clear that this three-criterion method is a powerful one in that it imposes rather severe constraints on the class of possible accounts that we might come up with for any given phenomenon. In other words, it could in principle help us choose between competing theories. Clearly, this is a valuable aspect that should contribute towards making a biological approach interesting to us.

For the purpose of our discussion we also need to define what we are going to mean by biological explanation. How do biologists deal with the question of function and origin?. As a second point of introduction let us briefly review some aspects of modern theories of evolution that are essential to our theme. According[1] to one widely accepted school of thought – the so-called neo-darwinian 'modern synthesis' (Mayr 1978) – new species have evolved as a result of *natural selection* operating on the *variation* of existing lifeforms. This variation accumulates continually and arises from dynamic genetic processes such as mutation and recombination. The genes of those individuals who survive sufficiently long to have offspring are transmitted to new generations. Genetic material not compatible with survival and propagation tends to be filtered out. Thus natural selection acts as a sort of editor testing the environmental fitness of new genetic messages.

Using a terminology from our own field we could conceive of evolution as a *source-filter* process in which the properties of the 'source' as well as the characteristics of the 'filter' vary in space and time. This 'modern synthesis' thus teaches us that it is as a result of interaction between an extremely rich source of genetic variation and selective environmental filtering that the morphology and behavior of many organisms have become so remarkably well adapted to their environments and often exhibit great adaptability to changing conditions.

Deliberately simplifying let me summarize our review of evolutionary theory and state the basic formula for biological explanations:

BIOLOGICAL FACTS = f(GENETIC VARIATION * NATURAL SELECTION) **(1)**

It says: Biological facts are explained in terms of an interaction between genetic variation and natural selection.

This is a deceptively simple but enormously powerful principle. Now if biologists have been successful in accounting for the enormous diversity of life forms by developing genetics and the Darwinian idea of selection why not follow in their footsteps and try to account for the likewise enormous

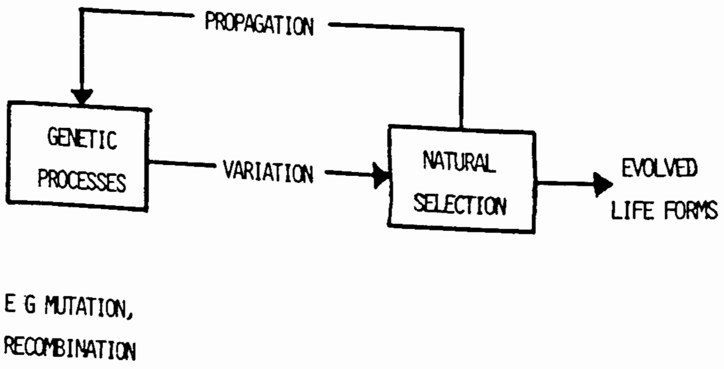


Figure 1. A Model of biological evolution

diversity of speech sounds along similar lines? I will come back in a moment with some specific examples. Let me first invite you all to reflect on the analogous claim for phonetics:

SPEECH SOUNDS = f(PHONETIC VARIATION * SELECTION) **(2)**

It reads: Phonetically as well as phonologically, speech sounds can be explained as arising from an interaction between *phonetic variation* and *selection mechanism.*

I agree with Peter Ladefoged that it is not *immediately* obvious how the data that he represents (Disner 1983, Linker 1982, Nartey 1982) can be explained in any simple way by applying 'biological principles'[2]. However, before we assign to them a secondary role or dismiss them totally we must begin to define them and systematically study their interplay with other factors. That is what I would now like to do by considering, in a preliminary way, the *nature* of *speech units.*

2. Speech units, Self-Organization and System-Generated Structure

In his abstract Ladefoged (1983a) states that the 'units of abstract linguistics – things such as phonemes and features – are of little relevance for speakers and listeners'. He regards such units as social, but not as psychological realities. He returns to such ideas in the proceedings paper (1983b). We can put Ladefoged's claim – a classical topic – in a biological context by introducing the notions of *self-organizing* and *system-generated structure.* The theory of self-organizing systems is a relatively recent paradigm that aims at formulating general laws governing the spontaneous occurrence of order in nature (Jantsch 1981). It can be demonstrated that, wherever there is interaction between subprocesses, this interaction obeys principles of considerable generality and will inevitably impose structuration e.g. on such diverse things as matter, behavior or information.

To convey to you the concept of self-organization more clearly I need to digress for a moment and discuss a distinction which is well known to all of us, viz. the idea of *form and substance.* However, I shall illustrate it with some examples from other disciplines. For instance consider the form of snow flakes (and crystal formation in general), the splash of a drop of milk as displayed in an instantaneous photograph. Or a chemical reaction: The gradual development of so-called spiral waves in a shallow dish. The hexagonal shape of bee honeycomb cells. (For lack of space I omit the pictures shown during the oral presentation and replace them here by referring the reader to my sources: D'Arcy Thompson (1961) and Prigogine (1976, 1980)). I could add many more cases but it is not necessary. They would all exemplify the same thing: the notion of self-organizing system. They also represent phenomena which would be difficult to describe on the basis of an explicit and clear-cut dichotomy into *form* on the one hand and *substance* on the other. Form and substance are inextricably interwoven. And there is no advance specification of FORM. Let me clarify the relevance of these seemingly far-fetched phenomena to linguistics. I shall do so by telling you how termites build their nests. The behavior of these insects has been mathematically analyzed as a self-organizing system by Prigogine (1976) and I am indebted to Michael Turvey (cf. Kugler, Turvey and Shaw, 1982) for bringing this work to my attention.

Termites construct nests that are structured in terms of pillars and arches and that create a sort of 'air-conditioned' environment. The form of these nests appears to arise as a result of a simple local behavioral pattern which is followed by each individual insect: The pillars and arches are formed by deposits of glutinous sand flavored with pheromone, a chemical substance that the animals are sensitive to. Each termite appears to follow a path of increasing pheromone density and deposit when the density starts to decrease. Suppose the termites begin to build on a fairly flat surface. In the beginning the deposits are randomly distributed. A fairly uniform distribution of pheromone is produced. Somewhat later local peaks have begun to appear serving as stimuli for further deposits that gradually grow into pillars and walls by iteration of the same basic stimulus-response process. At points where several such peaks come close, stimulus conditions are particularly likely to generate responses. Deposits made near such maxima of stimulation tend to form arches. As the termites continue their local behavior in this manner the elaborate structure of the nest gradually emerges.

The nest building can be described simply in terms of three rules: To initiate deposit at random! Next time deposit where scent density is maximal! Apply recursively! Note that in this theory there is no explicit mention of the structure of the finished product. No doubt you will agree that we should not attribute to the insects a '*mental target or blue-print*' specifying the final form of the nest. This form is *implicit* in the local behavior of each individual. Consequently the form-substance dichotomy does not apply.

After this digression let us return to Dobzhansky's three questions and the source-filter model in an attempt to apply them to the topic of speech units. Suppose that we try to shed some light on how the mechanism of converting discrete units into continuous signals operates by investigating also the purpose and the origin of this mechanism. If there are such things as phonemes what purposes do they serve and where do they come from? The ontogeny of phonemic coding seems to be a case that clearly calls for a self-organizing model since children proceed from holistic vocalizations to adult segment-based speech as a result of circumstances that they have no direct or conscious control over.

This is work that I am currently doing in collaboration with Peter MacNeilage and Michael Studdert-Kennedy. It will be in one of the chapters of a forthcoming book of ours on *The Biological Bases of Spoken Language*. We explore a hypothesis an early version of which is due to Hockett and which suggests that phonemic coding arose in a 'self-organizing' way from an interaction between vocabulary growth and phonetic constraints. As man's conceptual development was dramatically accelerated a solution to the problem of efficient signal generation and reception seems to have been obtained in parallel.

The basic idea underlying a series of computational experiments is shown in Fig. 2. We begin on the left by specifying a number *k* that represents vocabulary size. We feed this number into a computer program that assigns phonetic shape to these elements in a sequential manner and in the presence of certain performance constraints. The selection of phonetic signals is made from a larger inventory representing universal phonetic possibilities. A phonological analysis of the k phonetic signals is then undertaken. By systematically varying the variables of this 'word game' we hope to be able to investigate whether speech-like units could arise from an interaction between vocabulary development and production/perception constraints. Note that the backbone of the theory is the 'variation-selection model'.

Suppose we attempt to derive the phonetic properties of a small lexicon of *k* words in a manner roughly analogous to the termite story. Replacing deposits by syllables we have:

1. Select first syllable at random!
2. Select next syllable so as to optimize a performance constraints criterion.
3. Apply recursively until k syllabes have been obtained!

We shall develop this analogy in three steps:

1. First we define 'possible vocalization' or 'possible syllable':

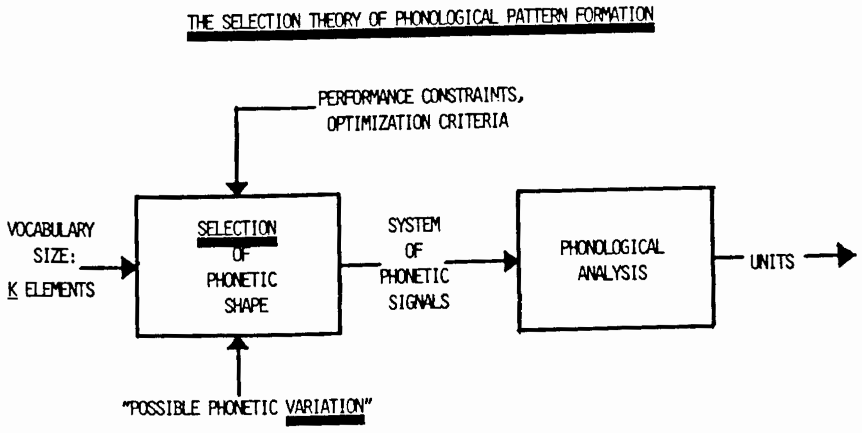

THE SELECTION THEORY OF PHONOLOGICAL PATTERN FORMATION
PERFORMANCE CONSTRAINTS, OPTIMIZATION CRITERIA
VOCABULARY SIZE: K ELEMENTS
SELECTION OF PHONETIC SHAPE
SYSTEM OF PHONETIC SIGNALS
PHONOLOGICAL ANALYSIS
UNITS
"POSSIBLE PHONETIC VARIATION"

Figure 2. Schematic diagram showing components of procedure for deriving systems of phonetic signals as a result of interaction between vocabulary growth and phonetic constraints.

2. Secondly we define the selection process:
3. And finally we define the performance constraints and the criterion of optimization.

We make the assumption that the syllable is an axiomatically given primitive of our theory. It is a gesture starting from articulatory closure and ending in an open configuration.

In principle there is an infinite number of places of closure and open configurations. For computational reasons it is necessary to quantize these possibilities into a certain number of discrete points. We decided that a sufficiently finely graded sampling of the universal phonetic space would be obtained by using the 7 closure onsets and the 19 open configurations shown in Fig. 3. This yielded a total of 7 * 19 = 133 syllables.

By definition each such vocalization is a holistic pattern that would resemble a CV sequence if presented on a spectrogram.Note that this resemblance does not in any way imply that it is analyzed as a sequence of two segments! It should be regarded as a Gestalt trajectory coursing through the atrticulatory/acoustic/perceptual space!

Now let us proceed to the definition of the selection process. The assignment of phonetic shape to *k* distinct meanings can be seen as making *k* choices from a larger inventory of *n* possibilities, that is from the possibilities that the universal phonetic space makes available. For our present purposes we are considering a fragment only of that space *viz.* with *n* equal to 7 x 19 = 133 syllable trajectories.

Given these simplifications we have a combinatorial problem, namely:

SELECT *k* SYLLABLES FROM *n* POSSIBILITIES IN THE PRESENCE OF CERTAIN PERFORMANCE CONSTRAINTS (3)

We chose the performance constraints according to Fig. 4. It is important to note that optimization takes place at two levels: With respect to *individual syllables* as well as with respect to pairs and *systems* of syllables. In the present simulations we explored the following conditions: *Perceptual salience* is qualified as extent of syllable trajectory calibrated in auditorily motivated dimensions: To exemplify, [ɟi] comes out as less salient than [ɟu]. *Extremeness of articulatory gesture* applies both to static configurations and to dynamic events: [ɖ] – closures more extreme than [d̪]'s. [ɟu] represents a more extensive movement than [ɟi].

Articulatory distinctiveness and perceptual distinctiveness are systems parameters. The articulatory dimension is interpreted as sensory discriminability and is computed in terms of 'articulatory distance' as specified by an articulatory model. Perceptual distinctiveness is derived by generalizing results on distance judgements for vowels to holistic syllables. For both of these parameters our metric implies that [d̪dɖ] form a less distinctive system than [bɖG].

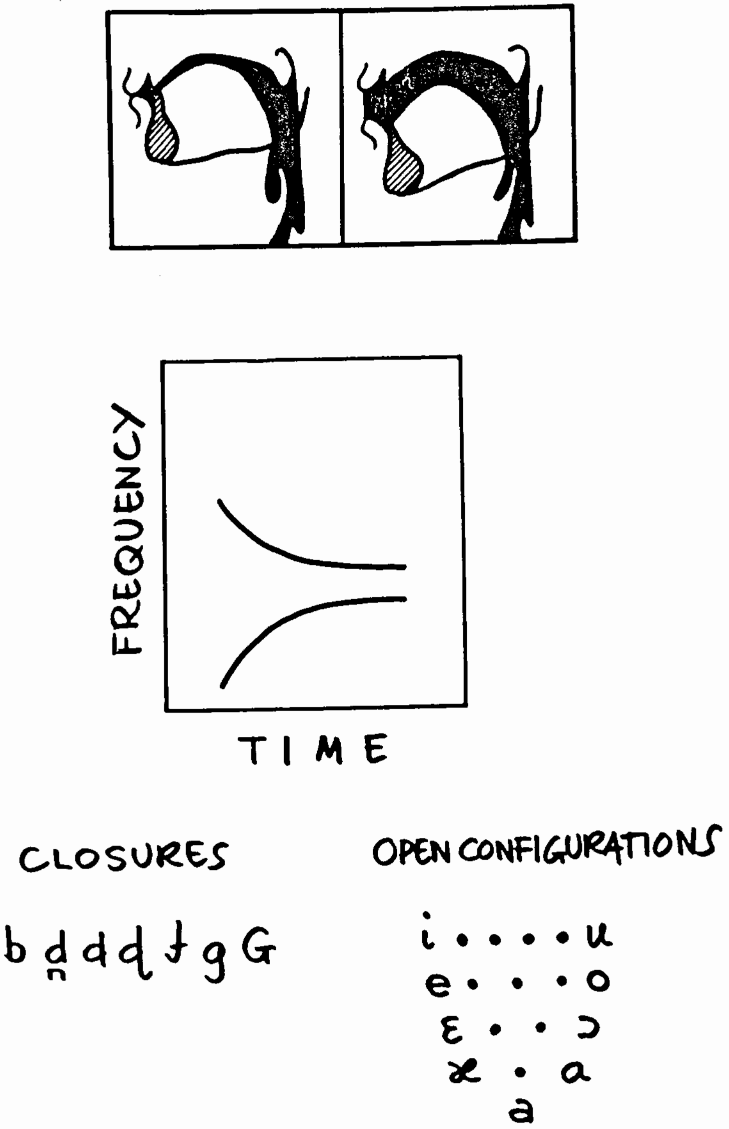
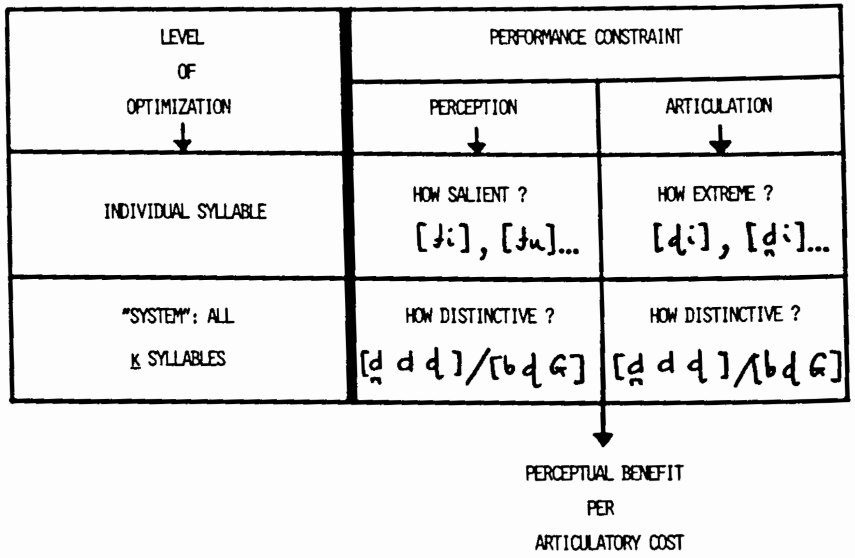


Figure 3. Definition of phonetic primitives of the theory.

For a given arbitrary syllable we obtain numbers in the top row. For a given pair of syllables we generate numbers in all four cells. Those four values are combined into a single number whose meaning can be described verbally as *perceptual benefit per articulatory cost*. The formula is which is applied for each additional syllable is:

$$\sum_{i=1}^{k} \sum_{j=z}^{i-1} 1/(L_{i\,j}/T_{i\,j})^2 < \text{THRESHOLD} \qquad (4)$$



LEVEL OF OPTIMIZATION ↓	PERFORMANCE CONSTRAINT	
	PERCEPTION ↓	ARTICULATION ↓
INDIVIDUAL SYLLABLE	HOW SALIENT ? [ɟi], [ɟu]...	HOW EXTREME ? [ɖi], [d̪i]...
"SYSTEM": ALL K SYLLABLES	HOW DISTINCTIVE ? [d̪ d ɖ]/[b ɖ ɢ]	HOW DISTINCTIVE ? [d̪ d ɖ]/[b ɖ ɢ]

Figure 4. Performance constraints.

Note once more that both individual syllables and the whole system are evaluated. Recall now that to simulate the development of the 'lexicon' we applied the formula repeatedly for each new syllable and continued this procedure until a system of *k* syllables had been obtained. We chose *k* to be 24, a fraction of the total set. Since this method gives results that depend on the initial syllable we repeated the simulations 133 times each time starting from a new syllable. One way of presenting the results is obtained by answering the following question: *For a specific configuration of constraints what is the probability of finding a given syllable in the pooled subsets?*

We pool all the subsets and plot the frequency of each syllable in the pooled set. We find that the results deviate markedly from a pattern of completely uniform preferences which is the result we would have expected had the derivations taken place in a completely unbiased fashion, that is without any performance constraints at all.

In Fig. 5 the results have been arranged in the form of a two-dimensional matrix with rows representing onsets and columns endpoints. Syllables that did occur (at least once in 133 runs) have a black cell. Syllables that did not occur at all have empty cells.

It is immediately apparent that certain rows and columns have more than one entry. This means that syllables suc as [bu, d̪u, gu] etc. contrast. Rows and columns with multiple entries contain syllables that keep one segment constant while varying the other. They identify minimal pairs. Since by definition all syllables have distinct meanings we might conclude that according to standard procedures these minimal pairs contain distinct phonemes. The existence of [bu, d̪u, gu] thus appears to suggest that in these derivations /b, d̪, g/ come out as separate *phonemic segments.*

SYLLABLE END-POINTS

SYLLABLE ONSETS	i	y	ü	ʉ	ɯ	u	e	ø	ə	ɤ	o	ɛ	œ	ʌ	ɔ	æ	ɶ	ɑ	a
b	■	■				■	■	■				■	■			■	■	■	■
d̪						■					■	■			■	■	■	■	■
d																			
ɖ																■			
ɟ															■			■	
g						■					■								

Figure 5. Matrix indicating distribution of derived phonetic syllables in terms of onsets and endpoints.

How is that possible? At no point in the derivations have we analyzed the syllables as a sequence of two segments. We have defined our vocalizations as holistic events. Our theory does not invoke 'segment' as an explicit construct. Neither does it use explicit 'features' although the use of [b, d̪, g] and a certain subset of vowels implies a systematic favoring of certain articulatory properties of 'feature' dimensions.

It should be clear from these considerations that there are neither 'segments' nor 'features' on the generation of these phonetic signals and that it is our preceding linguistic analysis that imputes discrete 'segments' and 'features' to them. Just as 'arches' and 'pillars' are implicit in the behavior of the termites the 'segments' and 'features' represent phonological structure implicitly and non-discretely present in the process of selecting the phonetic system. It is as if the phonetic space becomes 'quantally structured' as phonetic constraints interact with a growing vocabulary.

What do we conclude from these results? Let me extrapolate. Suppose we retain the notion of self-organization and manage to elaborate the theory so that more realistic and language-like phonetic systems can be produced. Will we find that our procedure will eventually derive fully discrete segments and features similar to the ones now postulated by linguistic analysis? Or will our model instead reinforce the notion of implicit psychological realities not in on-line speaker-listener behavior but rather structures indicating that phonemes and features are products of some introspective, metalinguistic ability that we possess as speakers and listeners? Clearly these are questions for future research. For the moment let us conclude that the present results although highly preliminary appear to encourage interest in self-organizing models and further search for biological precursors of phonological and phonetic structure.

3. Role of Socio-Cultural Factors

In response to Ladefoged's remarks on the role of cultural factors and 'the whims of fashion', I would like to make two comments. The description of the phonetic facts we are considering can be seen as an optimization problem. In accordance with that point of view the phonetics of a given language is the result of optimizing a great many dimensions which interact to yield an overall system value:

$$\text{OVERALL SYSTEM VALUE} < \text{THRESHOLD} \tag{5}$$

Assume furthermore that the observed systems need not exclusively represent *optimal* systems but are simply those systems that are *sufficiently optimal* with respect to the threshold criterion. We then realize that there must be many solutions to a given optimization problem. Thus we conclude that also *without* social factors the biologically based conditions would give us diversity and non-uniqueness.

My second point is the following:

Suppose we now postulate that the optimized parameter is *social and perceptual effect per articulatory cost* rather than just perceptual effect per articulatory cost:

$$S_{ij} * L_{i\,j} / T_{i\,j} \tag{6}$$

In Fig. 6 we present this idea schematically. Universal phonetic possibilities are discretely represented and compared among themselves three times in the cells of the matrices the rows and columns of which are labeled identically. For a given cell, or comparison, or contrast if you like, the model provides one social, one perceptual and one production-based value all of which are combined into a single cost-benefit number according to the formula. The cost-benefit matrix on the right serves as the basis of system selections.

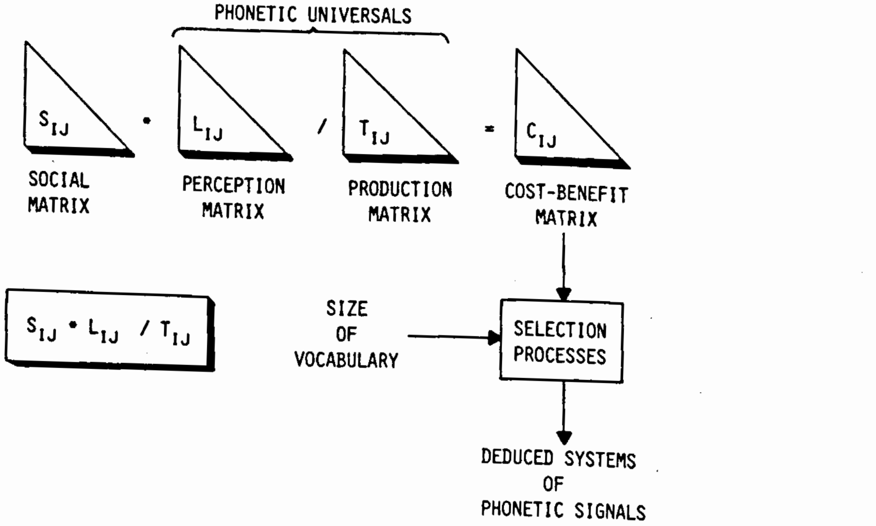


Figure 6. Schematic diagram showing a possible extension of model to accommodate also social factors. The rows and columns of the matrices are labeled identically and refer to universal phonetic possibilities. For any given phonetic contrast – that is given cell – the model provides social, perceptual and production-based coefficients which are combined into a single number and stored in the matrix to the right. This matrix forms the basis of system selections.

Ladefoged emphasizes the important role of the social matrix, and his point is well taken. But however we place our bets on social or biological factors dominating, the problem will be to find ways of determining their relative importance. Fig. 6 illustrates one way of how we might approach such evaluations. Above all I would like to claim that the model described earlier can be extended to accommodate also social factors. Note two things. This extension is still compatible with the source-filter selection model. Secondly, the examples that Ladefoged presents on economic systems, moral codes etc. are discussed in terms of a mechanism applicable also to biological phenomena viz. the principle of self-organization. In view of recent results from research on cultural evolution (Cavalli-Sforza and Feldman, 1981) we may be quite wrong in making too sharp a cut between biology and culture. They seem to share many aspects e.g. self-organization and selection by consequences (Skinner, 1981) [3]

4. Deductive vs. Axiomatic Theories of Phonology

My next comment concerns Ladefoged's pessimism about our being able to formulate a phonetically based, deductive theory of phonology. To be sure we can have no illusions about the magnitude of such a task. But the existence of difficulties does not convince me that there are easy and acceptable short-cuts.

One reason for insisting on a deductive account is based on the fact that the child can be said to derive its phonology deductively. Here is a comment on language development from Rules and Representations by Chomsky (1980, 66-67): ' what we should expect to discover is a system of universal grammar with highly restrictive principles that narrowly constrain the category of attainable grammars, but with parameters that remain open to be fixed by experience. If the system of universal grammar is sufficiently rich, then limited evidence will suffice for the development of rich and complex systems in the mind, and a small change in parameters may lead to what appears to be a radical change in the resulting system. What we should be seeking, then, is *a system* of unifying principles, that is *fairly rich in deductive structure* but with parameters to be fixed by experience. Endowed with this system and exposed to limited experience, the mind develops a grammar that consists of a rich and highly articulated system of rules, not grounded in experience in the sense of inductive justification, but only in that experience has fixed the parameters of a complex schematism with a number of options. The resulting systems, then, may vastly transcend experience in their specific properties but yet be radically different from one another, at least on superficial examination: and they may not be comparable point-by-point in general' (Italics ours).

Personally I see a favored role for phonetics within Chomsky's program of universal grammar but I think it is too early for phoneticians to share Ladefoged's and Chomsky's interest in primarily those aspects that must be described in purely formal terms and are said to make language a unique and specialized structure.

It is true that claiming that language is in part an autonomous biological structure should make perfect sense from the biological point of view. After all specialization is in one sense what evolution is all about.

However, our quarrel with the proponents of uniqueness and autonomy views is a methodological one. Again, let us follow the example of professional biologists who seem to prefer accounts of evolutionary changes that play down 'quantum leaps' as much as possible and that manage to interpret changes in terms of a minimum of de novo developments. This is a parsimonious null hypothesis that can be called the *continuity* or the *tinkering principle* (Jacob, 1977). Applied to our own field its contents would be:

DERIVE SPOKEN LANGUAGE FROM NON-LANGUAGE (7)

Claiming that language is special as Ladefoged and Chomsky do prejudges the issue. For any given phenomenon, it should be preceded by an exhaustive search for preadaptations. Before giving up that search and joining the 'formalist' camp we should make sure that, for example, we have not underestimated the structure-forming power of principles operating in the *self-organizing systems* subserving language. Although clearly untrue (e.g. speciation) the formulation of Linnaeus remains an efficient null hypothesis of biological inquiry: *Natura non facit saltum.*

5. On Explanation

Formal and functional approaches are often regarded as incompatible in current debates among phoneticians and phonologists (Andersen, 1981, Ladefoged, 1983b). In biology, this issue of functional explanation seems to have an analogy in the question: 'Is all evolution adaptive?'.

Functionalism in linguistics if often based on 'utility' arguments. Since many features of both language use and language structure no doubt lack direct utility it appears advisable to take a dim view of such functional arguments. Given the strong emphasis on adaptation by natural selection the reader may at first find such remarks inconsistent.

My point is this: To be able to put linguistic functionalism on a solid basis we need to learn our biology lessons well and avoid caricatures such as viewing 'each bit of morphology, each function of an organ, each behavior as an adaptation, a product of selection leading to a 'better' organism'. Darwin believed in *adaptive* and *nonadaptive change* and pointed to two principles underlying the latter:' (1) organisms are integrated systems and adaptive change in one part can lead to non-adaptive modifications of other features...: (2) an organ built under the influence of selection for a specific role may be able, as a consequence of its structure, to perform many other, unselected functions as well.' The current utility, or inutility, of a structure permits no assumption that selection did, or did not, shape it in a direct way. It may have been selected indirectly as a part of a larger system or through a cumulative action of collective subprocesses. (Quotations from an essay on the human brain (Gould, 1980, 50).

Applying this thinking to our own field a lesson for the phonetician would be that some linguistic phenomena are truly the results of 'adaptive changes' and could thus in principle be explained in functional terms whereas others have arisen nonadaptively and have to be accounted for on a purely formal basis.

In order to arrive at both functional and formal interpretations it would seem that our biologically inspired approach must obey the continuity principle and first lead to exhaustive investigations of all kinds of functional arguments (cf. above). The lesson for the phonologists would in the light of such reasoning be that the very existence of non-adaptive mechanisms in evolution would not a priori make formal, 'non-adaptive', accounts of linguistic observations legitimate until the search for pre-adaptations had been reasonably thorough.

Ladefoged (1983a) states that 'much of our work as phoneticians is simply to provide good descriptions of linguistic events' and that 'phoneticians must be able to document' language 'differences without expecting to explain them'.

His comments are reminiscent of an often quoted remark by Martin Joos (1958; 96) who wrote: 'Trubetzkoy phonology tried to explain everything from articulatory acoustics and a minimum set of phonological laws taken as essentially valid for all languages alike, flatly contradicting the American (Boas) tradition that languages could differ from each other without limit and in unpredictable ways, and offering too much of a phonological *explanation* where a sober *taxonomy* would serve as well.'

To be sure there will be limits to what we may be able to explain but in my opinion we are still far from having reached the end of our resources. We have a long way to go before phonetics ceases to be an interdisciplinary field and achieves a synthesis of its subfields. That development is under way as evident from this conference and will no doubt bring us closer to some of the long-term explanatory goals.

Secondly, as an inhabitant of a sometimes dark and cold country, let me point to an American tradition different from the one that Joos talks about, *viz.* the power of positive thinking. Believing or not believing in long-term explanation is clearly going to make a big difference for how we choose our short-term goals.

Thirdly, the issue of explanation is closely connected with the practical use of phonetics. All over the world to-day scientists including phoneticians feel an increasing pressure of having to produce practically useful results. One way to meet this legitimate challenge is to work for better theories and better explanations as a basis for improved applications. When a speech therapist, an engineer or a language teacher turn their backs on theory we should interpret such behavior more as an indication of the quality of our present explanations than as a confirmation of theory being in principle irrelevant.

The issue of explanation need definitely not be an academic ivory tower pastime.

Acknowledgements

I am particularly indebted to Peter MacNeilage and Michael Studdert-Kennedy. Much of the work reviewed here is the result of our current collaboration.

Special thanks are due also to Jack Fromkin, Bob Harms and Jim Lubker for instructive and inspiring discussions.

1. Current evolutionary theory is a dynamic field full of controversies such as the sociobiology issue (Rose 1982a, 1982b) and the criticism leveled at the 'modern synthesis' by Stephen Jay Gould and others (Gould 1982). For a detailed review of those questions I must refer the reader to the bibliography.

2. The different distributions of vowel qualities in Yoruba and Italian might conceivably be correlated with other factors in the two phonologies e.g. functional load patterns of contrast. An 'uneven distribution' of vowel qualities as in Yoruba does not immediately invalidate all possible 'biological explanations'.

3. For a recent discussion of the role of biological and cultural factors in language change and evolution, see Wang (1982a, 1982b).

References

Anderson, S. (1981). Why phonology isn't "natural". *Linguistic Inquiry*, **12**, 493-589.

Cavalli-Sforza, L.L. and Feldman, M.W. (1981). *Cultural Transmission and Evolution*, Princeton University Press, Princeton, N.Y.

Chomsky, N. (1980). *Rules and Representations*, Columbia University Press, New York.

D'Arcy Thompson. (1961). *On Growth and Form*, abridged edition,Bonner, J.F. (ed.), Cambridge University Press.

Disner, S. (1983). Vowel Quality: the contribution of language particular and language universal factors. *UCLA Working Papers in Phonetics*, **58**

Dobzhansky, T. (1965). *Mendelism, Darwinism, and Evolutionism*. Proceedings of the American Philosophical Society. **109**, 205-215.

Gould, S.J. (1980). Natural Selection and the Human Brain: Darwin vs. Wallace, 47-58 in *The Panda's Thumb*, Norton: New York

Gould, S.J. (1982). Darwinism and the Expansion of Evolutionar Theory. *Science*, **216**, 380-387.

Jacob, F. (1977). Evolution and Tinkering. *Science*, **196**, 1161-1165.

Jantsch, E. (1981). *The Evolutionary Vision:. Toward a Unifying Paradigm of Physical, Biological and Sociocultural Evolution*. Westview, Boulder.

Joos, M. (1958). *Readings in Linguistics*. American Council of Learned Societies. New York.

Kugler, P.N., Turvey, M.T., and Shaw; R. (1982). Is the 'Cognitive Penetrability' criterion Invalidated by Contemporary Physics?. *The Behavioral and Brain Sciences*, **2**, 303-306.

Ladefoged, P. (1983*a*). The Limits of Biological Explanations in Phonetics, In Cohen, A. and M.P.R. van den Broecke (eds.) *Abstracts of the Tenth International Congress of Phonetic Sciences*, 31-37, Foris, Dordrecht, Holland.

Ladefoged, P. (1983*b*), Out of Chaos Comes Order, Present Proceedings.

Lindblom, B., MacNeilage, P. and Studdert-Kennedy, M., (forthcoming). *The Biological Bases of Spoken Language*. Academic Press, San Francisco.

Linker, W. (1982). Articulatory and acoustic correlates of labial activity in vowels: a cross-linguistic study. *UCLA Working Papers in Phonetics*, **56**, 1-154.

Mayr, E. (1978). Evolution. *Sci. Am.* **239** (3), 39-47.

Nartey, J. (1982). On fricative phones and phonemes. *UCLA Working Papers in Phonetics*, **55**, 1-181.

Prigogine, I. (1976). Order through Fluctuation: Self-organization and Social Systems. In Jantsch, E. and Waddington, C.H. (eds.). *Evolution and Consciousness: Human systems in Transition*. Reading, Mass. Addison-Wesley.

Prigogine, I. (1980). *From Being to Becoming*. Freeman, Oxford.

Rose, S. (ed.) (1982*a*). *Against Biological Determinism*. Allison & Busby, London.

Rose, S. (ed.). (1982*b*). *Towards a Liberatory Biology*. Allison & Busby, London.

Skinner, B.F. (1981). Selection by Consequences. *Science* **213**, 501-504.

Wang, W. S.-Y. (1982*a*). Explorations in Language Evolution. *Osmania Papers in Linguistics*, **9**, Department of Linguistics, Osmania University, Hyderabad.

Wang, W. S.-Y, (1982*b*). Variations and Selection in Language Change. In: *The Bulletin of the Institute of History and Philology*, Academia Sinica, Taiwan.

'Out of Chaos Comes Order'; Physical, biological, and structural patterns in phonetics

Peter Ladefoged
Los Angeles, USA

Science is always seeking order; and fortunately for us Nature is always producing patterns out of chaos. In speech this happens at three levels, which we must be careful to keep distinct. At the lower level we can observe the physical patterns, such as the repetitively organized variations in air pressure that arises when sounds are generated. Next there are the patterns that are due to our human biological endowment, such as the tendency for vowels to be dispersed in symmetrical ways. Finally, there are characteristics that are generated by language being a self-organizing social institution. As I will show, these include the appearance of units such as phonemes.

There is not much that need to be said about the physical patterns. We can all appreciate the laws of physics that generate regular movements of the vocal cords when the outgoing breath passes through the suitably adjusted glottis. We can also appreciate the way these same laws generate the complex patterns of air pressure that are produced in the vowel tract as a result of the glottal pulses. The elasticity of the vocal cords, the harmonic structure of formants, and a host of other similar patterned phenomena are now reasonably understood.

At the biological level our knowledge is more meager. But we can explain some of the patterns that occur in language by reference to general principles of human behavior. As a general biological principle, organisms achieve their goal with the least possible effort. We can state this principle equally well the other way round: behaviors that are reinforced and survive are those that use the fewest resources. The notion of intent in achieving a goal is not critical to anything that follows.

Languages exist so that humans can communicate (and so that they can categorize their experiences; this is no doubt important in considering human biological pressures in syntax and semantics, but it is of little relevance in considering the sounds of language). The speaker's goal is to communicate as efficiently as possible. This involves producing a sufficiently distinct sequence of sounds for the listener to be able to get the message in a sufficiently short length of time. Within this aim there are a number of trade offs possible between articulatory effort, auditory distinctiveness, and rate of speech. Usually the speaker is able to take the initiative in setting this balance. Only occasionally does the listener have to interrupt and ask for clarification in some way.

The balance between the conflicting forces is clearly different in different circumstances. Two close friends exchanging information may be able to talk fast, and with a considerable reduction in the degree of precision of the articulatory movements. A speaker addressing an unfamiliar audience on a complex topic may talk more slowly with a more careful articulation. On some occasions, when the listener almost certainly knows the words to expect, virtually no distinctive articulations may be needed. Soldiers on the parade ground can interpret the wordless bellowing of the drill sergeant with great rapidity.

In general, as listeners become more familiar with particular words or phrases, speakers will be able to use more articulatory assimilations. This topic has been well enough covered in the vast literature on historical phonology to need little further documentation here. We might, however, note that as well as obvious assimilations that occur in pronouncing items such as 'handkerchief' and 'in between' as ['hæŋkərtʃif] and [ɪmbətwin], many cases of apparent *dis*similation are actually examples of economy of effort. For example, this is the case for Grassman's law, which states that the first of two aspirated stops in a word will become deaspirated (so that, for example Indo-European *thrikhos becomes classical Greek trikhos 'hair'). This can be interpreted as dissimilation, the consonants in a word becoming more unlike one another. But it is also an example of economy of effort. Aspirated consonants are very distinct from all other sounds (Singh and Black, 1966); but they are also costly in that they use considerable respiratory energy. A word with two such sounds is very costly, and an obvious candidate for pruning in any attempt to reduce the overall effort required for an utterance. Ohala (1981) has given a convincing account of dissimilation in terms of the listener as the source of sound change. I would only add to his account that the conditions are ripe for it to occur because it satisfies the biological drive for economy of communication (not that Ohala would put it that way).

The real cases in which the pressure for auditory distinctiveness affects the sounds of a language are those in which an increase of this kind can be achieved with a relatively low cost in articulatory effort. Many of these affect language paradigmatically, altering the items that can occur at a given place in a phonological structure, rather than syntagmatically, altering items in a sequence as a result of assimilations. The best documented case of the drive for auditory distinctiveness is the dispersion of vowels to fill the vowel space in the most efficient way (Lindblom, 1983). Given that there are a certain number of contrasting vowels in a particular language, it takes little added articulatory effort to increase.the distinctiveness of these vowels by dispersing them appropriately. This principle is far from sufficient to account for all the differences observed in languages with similar numbers of vowels. But it is undoubtedly a powerful underlying force.

The drive for communicative efficiency is not the only source of phonetic effects that may be ascribed to the behavior of individual speakers and listeners. Whenever people speak they identify themselves as belonging to a certain speech community. Every accent of every language is a peculiar set of sounds that is distinguished from all others sets of sounds of all other accents. One cannot speak without an accent of some kind. Part of the function of language is to convey information, part to categorize experience, and part to convey this accentual, sociolinguistic information.

For each of us there is a biological drive to identify oneself as belonging to a particular group, as well as to have one's own idiosyncratic characteristics. My accent is not right for me unless the words have a British RP accent tinged with small Americanisms, and my own particular voice quality. Speakers of every language have to use *exactly* the right vowel and consonant qualities, intonations, rhythms, etc. on pain of being wrongly labeled if they do not. There can be very subtle phonetic differences among languages resulting from this drive to be correctly identified as part of a group; but these phonetic phenomena are important to speakers and listeners. They cannot be ascribed to any general universal principles; they are due to the vagaries of local history and personal desire. But their maintenance can be regarded as ascribable to the behavior of individuals.

We can illustrate the explanatory power of these notions by reference to some recent cross-linguistic studies. In one study of this kind Disner (1983) showed that there were both similarities and differences between the vowel systems of Italian and Yoruba, both of which have seven vowels. The similarities can be ascribed to the fact that both Italian and Yoruba speakers are subjected to the same human drive for greater communicative efficiency. The differences are due in part to the physical laws of sound production affecting speakers of Italian and Yoruba differently, and in part to the biological drive for group identification.

Disner's charts of the formants of a group of 25 Italian speakers and a group of 10 Yoruba speakers are shown in Figure 1. Each vowel is represented by an ellipse (solid lines for Italian and dotted for Yoruba), with its center at the mean for that vowel, its axes along the two principal components of the distribution of the points, and the radii corresponding to two standard deviations. Roughly speaking, therefore, we may say that the ellipses enclose 95% of the points corresponding to each vowel.

Despite some obvious differences which we will discuss in a moment, there is a great overall similarity in the pattern of the vowels in the two languages. In each case the vowels are distributed in a V-shape that takes advantage of the space available. The drive for communicative efficiency results in most of the vowels of each language being fairly distinct from one another.

Some of the differences between the two languages are due to the shapes of the lips of Italian as opposed to Yoruba speakers. The mean first formant frequencies are similar in both sets of vowels, indicating that there are no overall differences in headsize between the two groups of subjects. But, with the exception of /i/ and to a lesser extent /e/, the second formant is lower for the Italian vowels than for the Yoruba vowel. These differences are precisely

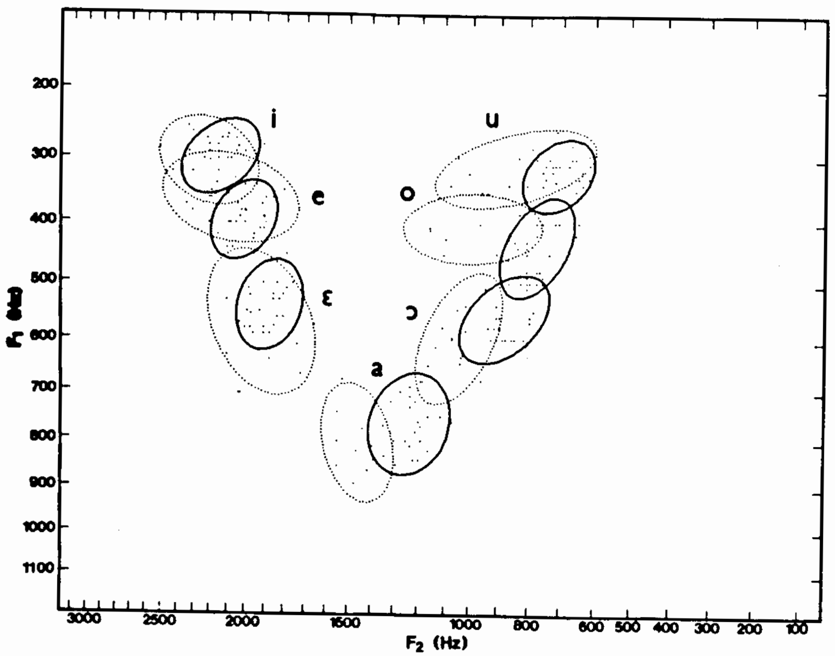


Figure 1. Formant plots of the vowels of 25 speakers of Italian (solid lines) and 10 speakers of Yoruba (dotted lines).

those that one would expect if Yoruba speakers, on the whole, used a larger mouth opening than that used by the Italian. Rounding affects the third formant rather than the second for vowels like [i]; and in other vowels the second (as well as the third) formant is lowered by decreasing the lip aperture. Accordingly, in the case of these two languages, there may be a physical phonetic explanation for at least some of the differences in the formant frequencies between the two groups of speakers. The possibility of overall differences in mouth opening is certainly compatible with the apparent facial differences between speakers of Yoruba and Italian. (This does not of course, imply that a Yoruba could not learn perfect Italian. Any individual speaker could compensate for the overall, statistical, difference in headshape shown in Figure 1, and thus learn perfect Yoruba.)

However as Figure 1 demonstrates, these are far from the only differences between Italian and Yoruba. The most obvious differences are in the F_1 dimension. The vowels of Italian are more evenly distributed than those of Yoruba in which /e/ and /o/ are much closer to /i/ and /u/ than to /ɛ/ and /ɔ/ respectively. The uneven distribution of the Yoruba vowels may be attributed to historical facts concerning the way in which the vowels of the original 9 or 10 vowel system have merged to produce the current 7 vowel Yoruba system. But, nevertheless, this historical explanation does not disguise the fact that present day Yoruba speakers choose to have vowels that are evenly distributed. To some extent they resist the biological pressures for communicative efficiency that undoubtedly exist, and undoubtedly account for phonetic phenomena that are widespread in the languages of the world, preferring instead the uneven distribution of vowel qualities that charactérizes them as Yoruba.

A second illustration demonstrates even more clearly that differences are subject to the whims of fashion. Some articulatory gestures can produce the same sounds - the same acoustic structures - as other quite different movements of the vocal organs; and some languages habitually use one way of producing a given sound, and others another. The evidence on this point is somewhat indirect, but it seems that some languages use one kind of the lip gesture to produce a high back rounded vowel of the [u] type and others produce virtually the same vowel, using a different gesture, with less rounding of the lips being compensated by other vowel gestures such as more lowering of the larynx.

Linker (1982) photographed the lip positions of the vowels in the speech of eight speakers of Cantonese and eight speakers of French. She examined the correlation between formant frequencies and lip position in each of these (and other) languages and showed that there are reliable, statistically significant, differences in the articulatory-acoustic relations. As she says, 'if ...a (mean) speaker of Cantonese wanted to produce an [u] with a given set of formant frequencies, he would have considerably less horizontal opening than a speaker of French producing the same vowel.' She concludes 'These results indicate that languages differ greatly in the lip gestures they use to make the same acoustic distinctions among vowels.'

Presumably children learning to become part of a particular group that speaks a certain language, achieve this goal by watching as well as listening. Children see the lip positions that are typical in their language and learn to make these sounds in that way. There may be a physical explanation for the difference between French and Cantonese associated with the different anatomies of the different racial groups, but it seems unlikely. It is much more likely to be a case of language specific behavior that can be ascribed only to the whims of fashion, and the desire of speakers of each language to maintain their group identity.

There is certainly no anatomical basis for another case in which visual cues during language acquisition are probably the cause of adults having noticeably different articulations when producing acoustically very similar sounds. As has been shown by Ladefoged (1979) a considerably higher proportion of American English speakers use an interdental fricative, as compared with the dental fricative which is more common in British English. In Californian English 75% of the speakers protrude the tongue between the teeth when saying words such as 'thief', whereas virtually no speakers of British English make the fricative in this way. Again, there are no reasons for this other than the desire of British English speakers to do things one way, and American English speakers to behave in another way.

The final illustration concerns coarticulation. We all know that in English, as in many other languages, the place of articulation of a stop is affected by the articulation of the following vowel. Thus the velar stop in 'keep' has a more forward articulation than in 'cop'. But this difference in the stops is only partially explained by claiming that there is coarticulation with the following vowel. We cannot claim that this is a necessary thing that languages have to do. If this were so, it would be nice, not only from the point of view of making biological explanations, but also from the point of view of simplifying the theory of phonology. Life would be easier if there were a set of universal, language independent, rules that took a string of segments specified in terms of a narrow, allophonic, transcription (or a matrix of distinctive features) and converted it into a sequence of continuously varying parameters of the kind required for a complete description (synthesis) of the utterances represented by the string. But there is no force compelling speakers to have a given degree of coarticulation. As indicated in Ladefoged (1972),it seems likely that French and English differ in the ways in which coarticulations occur. The French velar stops in *pique* [pik] and *Pâques* [pak] coarticulate more with the preceding vowel than the corresponding stops in the English words *peak* [pik] and *pock* [pɑk].

Recently Nartey (1982) has given a more rigorous demonstration of the language dependent nature of coarticulation. He recorded (among other data) eight speakers of Amharic and eight speakers of Navaho. Both languages have, among other fricatives, two sibilants which may be transcribed /s,ʃ/. They also have similar vowels that may be transcribed /i,a/. The fricatives in each language were recorded in the context /i-i/ and /a-a/. The acoustic spectra at the midpoint of each fricative was determined using the UCLA WAVES computer analysis system. These spectra were then convolved with an auditory filter as suggested by Bladon and Lindblom (1981) to produce representations of the auditory information.

The upper part of figure 2 shows the mean auditory spectrum of Amharic [ʃ] in the two vocalic contexts; the lower part shows Navaho [ʃ] in similar contexts. There are differences in the sharpness of the spectrum in the two languages. But, over and above this, it may be seen that the Amharic sounds show greater effects of coarticulation with the vowel, the spectrum having a noticeably higher peak in this Bark scale representation when in the context of [i] than when in the context of [a]. In Navaho there is very little coarticulation between this fricative and the adjacent vowel. Again we see that speakers of different languages choose to distinguish themselves by using different phonetic mechanisms that cannot be predicted from the principle of communicative efficiency, but are simply learned behavior that distinguishes one group from another.

The physical laws of speech production, and, at another level, the drives for communicative efficiency and for group identity account for many rule governed phenomena, But many other phonetic facts about languages cannot be explained in terms of these physical and biological principles. This leads to a difference in emphasis between Lindblom and myself. I do not

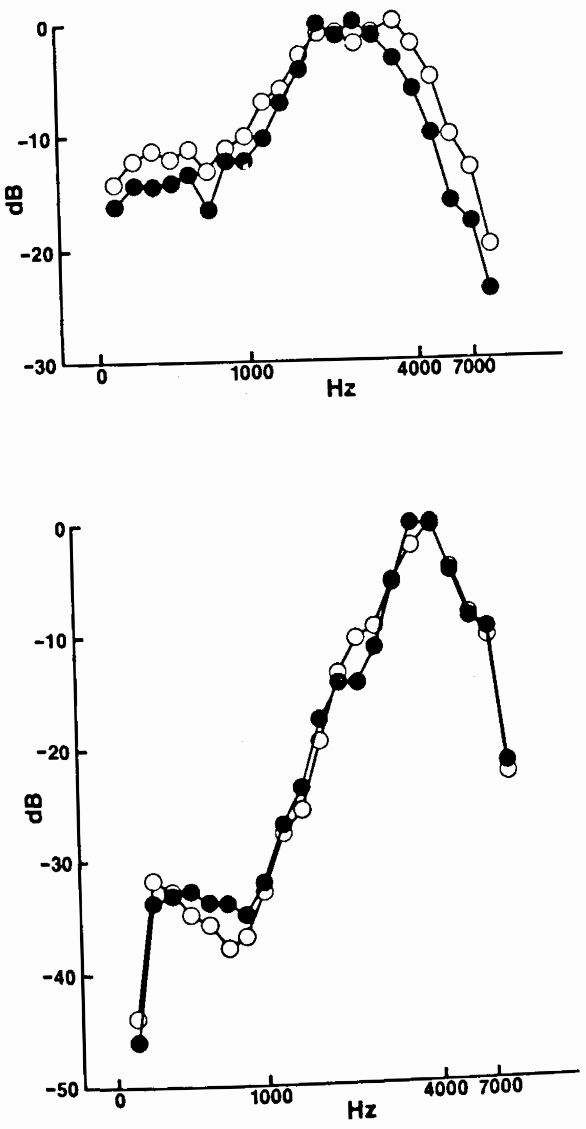


Figure 2. The mean auditory spectra of [ʃ] in the context [i-i] (open circles) and [a-a] (solid points). Data for 8 speakers of Amharic in the upper part of the figure, and for 8 speakers of Navaho in the lower part of the figure.

think that we can directly derive 'the fundamental units and processes of linguistic structure deductively from independent premises anchored in physiological and physical realities' (Lindblom 1983). Instead I agree with Anderson (1981), and other generative phonologists who find that many interesting linguistic observations can be made in formal terms, independent of any other data. I disagree with scholars of the latter group only when they try to relate these linguistic observations to observable mental capacities. It seems to me that the mental nature of language has been somewhat misleadingly presented by Chomsky (1975). His notion that language is an organ of the mind is not very helpful. It is somewhat like saying that digestion is an organ of the body. Digestion is an ability that involves many components, including some things that are normally called organs such as the liver and the pancreas, as well as a number of other things such as saliva, mastication, and bowel movements. Digestion is like language in that it is a system. But neither of them is an organ in the usual sense.

A better way of describing a language is to consider it as an observable social institution, without having to consider what goes on in people's minds. When we consider any social institution we find that it is governed bv different principles from those that govern the behavior of individuals. Principles such as communicative efficiency and identification with a group apply to descriptions of what people do. But a language considered as a system where everything hangs together is to some extent self-organizing.

In order to make this point clear it is worth considering two other examples of self-organizing social institutions. We may begin by comparing a language with a moral code - a system of value judgments applicable in a given community. Any moral code is clearly a product of a society, and is strongly influenced by the surrounding culture. Moral judgements that originally had some utilitarian function rapidly become ritualized. Like pronunciations of words they are as they are because that is the way things are done a a certain society. But morality is also property of an individual, at least to the extent that the individual can choose to perform moral acts. Morality, or at least the capacity for performing moral acts, may even be like language, or the capacity for language, in being innate. Certainly one way to think of morality is as an over-developed herd instinct -- a self-organized, innate drive for the preservation of the herd rather than the individual.

The moral code that we observe (or feel guilty about) is only one example of a social institution. As another, very different, example consider the economic system. There are obvious market forces affecting the price of goods and the cost of labor (the far from inexorable 'laws' of supply and demand). There are also Galbrathian forces such as the conflict between the company management (whose aim is usually growth, which leads to bigger managerial responsibility and salaries) and the company ownership (the shareholders) who want bigger profits, which may well be achieved without growth and with less management). All these forces, and many more (government, international affairs, and perhaps morality) add up to form a social institution, the economic system, which nobody understands and which is certainly not part of anybody's competence. Without people there would be no economic system. It is like language in that it takes at least two to trade. Furthermore, just as people 'know' the rules of their language, in the same sense everyone 'knows' their economic system. We all understand what money can do. But it is obviously ridiculous to take a mentalist approach. Nobody would call economics an organ of the mind.

Many sound patterns are the result of language being a self-organizing social institution, and are not generated at the level of individual behavior. We may start by considering those that result from the filling of a gap in a phonological system. It has often been observed that languages tend to fill holes in the patterns of their segmental inventories. Thus Antilla (1972) shows that Proto-Baltic Finnic had a system 'in which the short vowels had one degree of height more than the long ones, and contained the only front rounded vowel in the whole system ... (Modern) Finnish has filled every single gap and ended up with perfect symmetry.' To take another example, it is not at all surprising for a language such as English, which at one stage had four voiceless fricatives /f, θ, s, ʃ/ and three voiced ones /v, ð, z/, to acquire the missing voiced fricative [ʒ], as we have done recently. But it should be noted that this does not occur because of the biological drive acting on individual speakers and listeners. Filling holes in an abstract phonological system does not increase communicative efficiency for the individual. There is no increase in ease of articulation or auditory distinctiveness for any *existing* possible utterance. Nor, in general, does it help to identify the speaker in any way. (The exception is when the new sound is being brought in by borrowing from some other language or dialect. Using this sound may mark the speaker as wishing to identify with the speakers of the other language or dialect, as, for example, when a speaker of Southern British English starts using a voiceless fricative /χ/ in words of Scottish origin, such as 'loch'.)

The discussion of possible linguistic changes in terms of holes in phonological patterns is sometimes formulated in a slightly different way. There are a number of occasions when linguists talk about the segmental inventories of a language being such that they facilitate or hinder possible sound changes. Thus Maddieson (1984) suggests that if a language does not have /v/ it is more likely to develop a phonemically contrastive /β/ as a result of phonologizing an intervocalic [β] allophone of /b/.

Adding a new sound is like the emergence of a new species in biological evolution. It is possible to claim that it happens because God sees a gap and wants it to be filled because it is easier for the world to be that way. But an equally good claim is that if there is an ecological niche to be filled, events (the random mutation of genes, evolution is not purposive) will conspire to fill it. In the same way a new phoneme is more likely to occur (to be borrowed or to be phonologized from an existing allophone), if it fits nicely into an existing pattern. Note that this cannot be explained in terms of the behavior of individual speakers and listeners, just as the development of a new species is not due to the action of individuals. The communicative efficiency principle does not apply to how languages organise their sounds. When we discuss phonological systems we have moved to considering language not as part of an individual's behavior, but as a self organizing institution.

Many of the patterns currently described by linguists are patterns that occur simply in language considered as a social institution. In addition to the hole in the pattern phenomena, there are what Kisseberth (1970) calls phonological conspiracies. Dauer (1983) has assembled an excellent case for regarding stress timing in English in this way. Many people have observed an apparent tendency in English for stresses to recur at regular intervals of time. But it seems that this may be due to a fortuitous combination of circumstances. The fact that English words have a somewhat regular stress pattern, the possibility of alternative stress patterns in some words, and of dropping stresses in some sequences of words, the reduction of weak syllables, and the clitic-like nature of many grammatical formations, all these things and more combine to lead to the occurrence of stresses at appropriate intervals.

Perhaps the most startling conspiracy - one that seems to have deceived by far the majority of linguists - is the appearance of phonemes. Accounts of human behavior in terms of phonemes are nearly always examples of what has been called the psychologist's fallacy - the notion that because an act can be described in a given way that it is necessarily structured in that way. As far as I can see, phoneme size units play only a minor role in human behavioral acts such as normal speaking and listening. I have argued this point elsewhere (Ladefoged 1980) and will not consider it in detail here. All that is necessary is to examine some apparent counter arguments.

The first concerns children babbling, playing with newly discovered sounds. Typically they make sounds such as [babababa....dadadada] but sometimes they will produce sequences such as [bibibebe]. Does this mean that they have discovered the notion of a segment, and are trying out different vowels? I doubt it. I think it is much more likely to be random, uncontrolled behavior, and it is only the linguist who hears it as a set of different vowels. But in any case I do not want to deny that there is a strong commonality among all syllables beginning with the same consonant. After all the Aramaic scholars and others who invented syllabaries spotted this and wrote syllables such as [bi, ba, bu] with the same symbol. But the important point is that no child ever babbles sequences such as [badaga badaga]. And for hundreds of years the Aramaic scholars never noticed that there was something in common between syllables such as [ba da ga]. According to Gelb (1952), writing has been invented many times, and there have been many independent developments of syllabaries. At least in some languages the syllable is a clear, intuitively apprehended, concept. But only once in the whole history of writing has it occurred to people to break syllables into separate segments that can be recombined to form other syllables; and that occurred because of fortuitous circumstances. It took the happy chance of the coming together of speakers of Semitic languages, who had symbols for syllables beginning with pharyngeal and other non-IndoEuropean consonants, and Greeks, who did not need these symbols with their original values and chose to use them for vowels. The Greek writing system, the only original alphabet, was produced out of the spare symbols of a syllabary. If phonemes are behavioral concepts, it is difficult to see why the alphabet was invented only once; and, for that matter, why children do not alternate phonemes and babble [badaga badaga].

Speech errors are sometimes cited as another piece of evidence against the claim that phoneme size units play only a minor role in speech production and perception. One of the commonest errors in rapid speech is transposing two elements as in a spoonerism, saying, for example, 'fast pew' instead of 'past few' (Fromkin 1973). In such cases it is often apparently segments that are transposed. But note that it is not just any two segments that get out of order. The previously cited phrase could never have appeared as 'paf stew' with the final element in one syllable becoming initial in another. The segments that move always have to be in the same positions in different syllables. I do not know how speech errors involving transpositions occur. But it is no way obvious that they are simply transpositions of segments.

A final piece of seeming counter-evidence is the existence of a number of language games and secret languages that can be described in terms of phonemic interchanges. For example, Cantonese speakers who are unfamiliar with alphabetic notions can nevertheless produce a form of secret speech in which vowels and consonants are reversed (Wu, personal communication). In this speech words such as [ma] are produced as [am], but this is still not a truly segmental approach; words such as [tʃaŋ] come out as [aŋtʃ]. Speakers do not produce [ŋtʃ] (the reverse of the phonemes), and do not even recognize [ŋaʃt] as the true reversal of the phonetic segmental order. Again what is even more important is that in all the language communities I know of that have secret languages or games of this kind, there are always some speakers who find it very difficult, or even virtually impossible, to follow rules involving segments. (Fluent speakers of Pig Latin often fail to understand how difficult it is for those of us without an equivalent mis-spent youth to talk in this way.) Notions concerning phonemes size units have to be carefully learned - they are in no way naturally available.

Lindblom (1983) has suggested a nice analogy that can be extended to make this point clear. He has pointed out that termite nests appear to the outside observer to have a most intricate structure. There are great pillars and arches that rival those of medieval cathedrals. But it does not follow from this that individual termites know about arches. In fact they are simply following a very straightforward pattern of behavior, governed (in nest building) by a single rule: deposit grains of earth near other grains of earth that are scented with a termite secretion. At first this leads to random depositing of earth. But very soon the deposits are on top of other recent deposits and the pillars grow. As two pillars grow taller the scent is strongest on the sides closest to each other; and so those two sides grow together and form an arch. All from a single, simple, rule. Phonemes may be like arches in termite nests, visible to outside observers, but having no meaningful role in the activity of the individuals producing them. Speech *appears* to be composed of sequences of segments because of the interactions of the different systems of which it is composed. The complex gestures involved in producing syllables have diverse parts that look as if they are categorically distinct. We call these diverse parts vowels and consonants, but we must always remember that these are just names for readily distinguishable aspects of the stream of speech. Those of us who have been exposed to an alphabetic tradition may be influenced so that we are very conscious of the possibility of describing speech in terms of units of this kind. But illiterates may have little or no concept of speech segments (Morais et al, 1979). Similarly those involved in adult literacy campaigns report that the concept of the segment is far from self evident. Intelligent adults who have been taught to write a few words cannot perform tasks such as naming other words that begin with the same segment (Jackson, 1982). A language consultant who has been working extensively with a linguist will be able to learn the phonemic principle (Sapir, 1949), just as a child can learn to read and write. But this is hardly evidence for phonemic units in the normal process of speaking and listening.

We can carry the termite analogy a step further still. Just because the individual termite cannot be considered responsible for the design of the arches and pillars in a termite nest, it does not follow that it is not interesting to describe these pillars and arches. They are a necessary part of the termite nest; when a pillar or an arch is needed to support the edifice which the community requires, if the nest is to be one which survives, then the requisite structure will be present. In that sense, a termite nest is self-organized. In a similar way a language gets the sounds that it needs. The segments and phonemes are present in the structure of that abstract entity, the language, considered as a social institution. Indeed, as Halle commented several years ago, 'Almost every insight gained by modern linguistics from Grimm's law to Jakobson's distinctive features depends crucially on the assumption that speech [or, in my terms, language] is a sequence of discrete entities.' (Halle, 1964)

But despite the value of segments as descriptive units, it seems almost certain that the phonemic principle is not part of our genetic endowment (as it surely must be for those who view it as an innate ability). The manipulation of phonemes is an acquired ability. Evolutionists teach us that such things are properties of a culture, and not of an individual's physiology. The invention (not, for me, the discovery) of the alphabet occurred far too recently for it to have become part of our DNA. Indeed, as Gould (1981) puts it: '*Homo sapiens* arose at least 50,000 years ago and we have not a shred of evidence for any genetic improvement since then..... All that we have accomplished, for better or worse, is a result of cultural evolution.'

Our endeavors include building (like termites) social institutions such as language, morality, and economic systems. Each has, to a great extent, become its own thing, so that it is no longer entirely explicable in terms of outside forces. The evolution of language has involved its feeding upon itself, so that it must be described partly in terms of unique principles. The current accounts may not be correct. Descriptions of languages in terms of the fashionable metrical phonology (Halle and Vergnaud, 1980) are, after all, at least superficially very different from those of the older generative phonology (Chomsky and Halle, 1968). But the thread of the uniqueness of language

that runs through them is still valid. And they are all, despite their authors' claims, descriptions of social institutions and not explanations of mental activities. Like termites who do not know how to build an arch, ordinary speakers and listeners do not know the sound pattern of English.

References

Anderson S. (1981). Why phonology isn't 'natural'. *Linguistic Inquiry* **12**, 493-589.

Antilla, R. (1972). *An Introduction to Historical and Comparative Linguistics.* New York: Macmillan.

Bladon, A. and Lindblom, B. (1981). Modeling the judgment of vowel quality differences. *J. Acoust. Soc. Amer.* **69.5**, 1414-1422.

Chomsky, N. (1975). *Reflections on Language.* New York: Parthenon Press.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* New York: Harper and Row.

Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. *J. Phonetics* **11**, 51-62.

Disner, S. (1983). Vowel quality: the contribution of language particular and language universal factors. *UCLA Working Papers in Phonetics* **58**.

Fromkin, V. (1973). *Speech Errors as Linguistic Evidence.* The Hague: Mouton.

Gelb, I. (1952). *A Study of Writing.* Chicago: University of Chicago Press.

Gould, S.J. (1981). *The Panda's Thumb.* New York: Norton.

Halle, M. (1964). On the bases of phonology. In: *The Structure of Language* (Fodor and Katz, eds), Englewood Cliffs, New Jersey: Prentice Hall, 604-612.

Halle, M. and Vergnaud, J-R. (1980). Three dimensional phonology. *J. Linguistic Research* **1.1**, 83-105.

Jackson, E. (1982). Personal communication.

Kisseberth, C. (1970). On the functional unity of phonological rules. *Linguistic Inquiry* **1**, 291-306.

Ladefoged, P. (1972). Phonetic prerequisites for a distinctive feature theory. In: *Papers in linguistics and phonetics to the memory of Pierre Delattre.* (Valdman, ed), The Hague: Mouton, 273-285.

Ladefoged, P. (1979). Review of J.C. Catford: Fundamental Problems in Phonetics. *Language* **55.4**, 904-907.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language* **56.3**, 485-502.

Lindblom, B. (1983). Economy of speech gestures. In: *The Production of speech* (MacNeilage, ed): New York Springer Verlag.

Linker, W. 1982). Articulatory and acoustic correlates of labial activity in vowels: a cross-linguistic study. *UCLA Working Papers in Phonetics* **56**, 1-154.

Maddieson, I. (1984). *Patterns of Sounds.* Cambridge: Cambridge University Press.

Morais, J. et al. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition* **7**, 323-331.

Nartey, J. (1982). On fricative phones and phonemes. *UCLA Working Papers in Phonetics* **55**, 1-181.

Ohala, J. (1981). The listener as a source of sound change. *Chicago Linguistic Society.* Papers from the parasession on language and behavior. 178-203.

Sapir, E. (1949). The psychological reality of phonemes. In: *Selected Writings of Edward Sapir.* (ed. D.G. Mandelbaum) Berkeley: University of California Press. 46-50.

Singh, S. and Black, J.W. (1966). Study of twenty six intervocalic consonants as spoken and recognized by four language groups. *J. Acoust. Soc. Amer.* **39**, 639-656.

Smith, C. and Tager-Flusberg, H. (1980). The relationship between language comprehension and the development of metalinguistic awareness. Paper presented at the Fifth Annual Boston University Conference on Language Development.

4. Psycholinguistic Contributions to Phonetics

Perceiving Speech and Perceiving Words

W.D. Marslen-Wilson
Nijmegen, the Netherlands

1.

Psycholinguistic research into spoken language comprehension, and phonetic research into the processes of acoustic-phonetic analysis, are both, in principle, part of the same general domain of inquiry. Both disciplines are concerned with aspects of the process whereby human listeners map from sound onto meaning. This implies, therefore, a close dependence between them.

In the past, however, there has been surprisingly little direct contact between the two disciplines. Research in phonetics – as, for example, Nooteboom (1979) has documented – tends to pay little attention to the wider functional context within which the processes of acoustic-phonetic analysis presumably operate. Conversely, psycholinguists – even those working on spoken word-recognition – tend to neglect, or simply ignore, the complexities of the acoustic-phonetic input to the processes they are studying.

We can take for granted that psycholinguists should pay more attention to acoustic-phonetic issues. What is less straightforward is the claim that phoneticians should pay more attention to psycholinguistic issues. Nonetheless, this is what I will try to establish here. I will do so with particular reference to the relationship between the acoustic-phonetic analysis of the speech signal and the perception and identification of spoken words.

Two questions need to be examined here. First, how far does the study of spoken word-recognition also raise important acoustic-phonetic questions? Second, how far has research in acoustic-phonetics in fact provided an adequate basis for an approach to these questions?

2.

The first point to be made concerns the extent to which further progress in understanding spoken word-recognition depends on developments in acoustic-phonetics. In the past, research on spoken word-recognition has been so general in the kinds of claims it made about the recognition process that it was not necessary to pay close attention to the acoustic-phonetic substrate for this process. It did not really matter what the imput to the word-recognition process was since the issue never really arose of how individual spoken words were discriminated from each other (although this question certainly did arise very early on in research on machine recognition of fluent speech). Recent research, however, has led to the development of psycholinguistic theories of spoken word-recognition that do require a much more precise specification of the properties of speech analysis.

These developments arise from some observations of the rapidity and the immediacy with which the speech signal is mapped onto the mental lexicon (c.f. Cole and Jakimik, 1980; Grosjean, 1980; Marslen-Wilson 1975; 1980; 1983; Marslen-Wilson and Tyler, 1975; 1980). A wide variety of different experiments converge on a highly consistent estimate of the average "recognition-time" for words heard in a normal utterance and discourse context – where the term "recognition-time" refers to the amount of sensory input, measuring from word-onset, that needs to be heard before a listener can start behaving as if he or she has correctly identified the word in question. The estimate of this average recognition-time for words in context is of the order of 200 msec.

Not only is this remarkably fast, but also it is remarkably *early*, relative to the total duration of the words being identified. For the kinds of experiments involved, the words averaged 375-420 msec in length. This means that words in context can reliably be identified when little more than half of the acoustic input corresponding to that word in the signal could have been heard. This in turn implies that listeners are highly efficient in their use of the acoustic-phonetic information carried by the speech signal. More recent results (Marslen-Wilson, 1983) show that listeners are in fact *optimally* efficient in their use of this information.

The notion of optimal efficiency can, in principle, be defined as the extraction of the maximum information-value from the signal, in real-time as it is heard. The term "information-value" can itself be related to the definition of information in terms of the number of alternatives between which a given signal can allow a receiver to discriminate (Shannon and Weaver 1949). If we assume some set of possible messages that a given signal can transmit, to a given listener in a given context, then the speech signal can be viewed as providing a continuous flow of potential discriminative information with respect to this set of possibilities.

If the set of possibilities involved is the complete set of words in the language, known to a given listener, then the information-value of the signal is defined with respect to the information that the signal provides, over time for the discrimination of the correct word from among the initial total set of alternatives. Experiments using an auditory lexical decision task show that listeners are indeed able to identify the word being uttered at precisely that point in the word at which the theoretically sufficient acoustic-phonetic information becomes available (Marslen-Wilson, 1983; see also Tyler and Wessels, 1983).

These results, and other considerations, lead to a model of spoken wordrecognition in which there is a multiple accessing of possible word-candidates early in the word. The subsequent selection of the correct candidate depends on the manner in which the accumulating sensory input not only matches the specifications (in the mental lexicon) of the correct word, but also fails to match the specifications of the incorrect words. The recognition of the correct word becomes possible, as experimentally demonstrated, as soon as the signal diverges sufficiently from the specifications of all other possible words.

An approach of this kind therefore stresses the implications for the identification of individual words of the discriminative information accumulating as the signal is heard. It is clear that the evaluation and development of such an approach depends on a satisfactory analysis of the nature of the input to these word-discrimination processes. Under what description are the products of acoustic-phonetic analysis delivered to the word-recognition system? What aspects of the original signal are preserved or discarded in the process of analysis? With respect to which set of discriminative categories should the information-value of the signal be evaluated?

3.

If, for an answer to these questions, we now turn to the main body of acoustic-phonetic research, we do not receive a coherent answer. One is faced with a remarkable diversity of different and incompletely specified proposals, where the products of speech analysis range from strings of phonemic labels, to bundles of probabilitically weighted features, to direct perceptions of speech events.

At least one distinguished acoustic-phonetician, confronted with these difficulties, has concluded that the best approach to the question of how the signal is mapped onto lexical representations is, in effect, to renounce the whole framework of classical phonetics (Klatt, 1979; 1980). Instead one should opt for the kind of "brute force" computational solution, based on direct matching to spectral templates without any intervening phonetic analysis, that is exemplified in the **harpy** speech recognition system (Lowerre and Reddy, 1978). It is likely that this conclusion is too pessimistic. Nonetheless, it is clear that acoustic-phonetic research, for all its advances over the past thirty years, has failed to satisfactorily answer those questions that are most critical for researchers working on other aspects of language processing. In part this is no doubt due to the fact that acoustic-phonetics, just like any other branch of the study of human language, is extremely difficult; that it can't be expected to have found all the answers yet. But in part it may also be the consequence of the set of assumptions that permit, and even encourage, the current *de facto* separation between research on speech analysis and research on spoken word-recognition.

The most important of these assumptions seem to be the following. First, one must assume that there are two distinct levels of perceptual representation computed during speech analysis. These correspond, respectively, to an acoustic-phonetic level of analysis and to a lexical level. Secondly, and crucially, one must assume that the properties of the acoustic-phonetic level, and of the processes that map from the speech signal onto this level, can be determined solely with reference to phenomena internal to this level, and without reference to the functional goal of these processes. Without reference, that is, to the role of these processes in providing the basis for a further mapping onto the mental lexicon (which in turn provides the basis for the extraction of communicative meaning).

Thirdly, one has to accept the direct translatability of results obtained in the phonetics laboratory, typically using either citation forms of synthetic speech, to the perceptual situation of the listener hearing fluent conversational speech. That is, one must assume that the kinds of relationships observed in the laboratory between a given speech signal and a given phonetic contrast, will also hold in the often different conditions of normal speech production and comprehension.

4.

It is not possible to state categorically that these assumptions are either false or misleading. But they are at least open to serious question. Consider, in particular, the second assumption, that speech analysis is most appropriately studied in functional isolation. In the case of spoken word-recognition, for example, one finds that it is by studying word-recognition in its functional context – as it contributes to the processes of language comprehension in utterances and discourses – that one can place the strongest constraints on possible models of lexical access (Marslen-Wilson, 1983; Marslen-Wilson and Welsh, 1978). In the same way, it may be that by examining the processes of speech analysis in their proper functional context – as part of the process of speech understanding – that one can place constraints on theories of speech analysis that could not be derived just by attempting to study these processes in isolation.

If, for example, as current analyses of spoken word-recognition suggest, one can predict precisely when a given word should become discriminable, then it should also be possible to determine just which aspects of the sensory signal are employed in making these discriminations. This, in turn, would surely have implications for one's assumptions about the speech analysis process that produces the basis for these effects.

Whether or not this particular strategy turns out to be fruitful remains to be seen (but see Streeter and Nigrom 1979). But the general point remains. Many psycholinguistic questions about the processes of spoken word-recognition are inescapably acoustic-phonetic questions as well. And it seems most unlikely that these questions can be resolved without a proper contact between the two disciplines – both in theoretical analysis and in experimental practice.

References

Cole, R.A., and Jakimik, J. (1980). A model of speech perception. In: R.A. Cole (Ed.), *Perception and Production of Fluent Speech*. Hillsdale, NJ: LEA.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics* **28**, 267-283.

Klatt, D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, **2**, 279-312.

Lowerre, B.T., and Reddy, D.R. (1978). The Harpy speech understanding system. In: W.E. Lea (Ed.) *Trends in Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall.

Marslen-Wilson, W.D. (1975). Sentence perception as an interactive parallel process. *Science*, **189**, 226-228.

Marslen-Wilson, W.D. (1980). Speech understanding as a psychological process. In: J.C. Simon (Ed.), *Spoken Language Generation and Recognition*. Dordrecht: Reidel.

Marslen-Wilson, W.D. (1983). Function and process in spoken word-recognition. In: H. Bouma and D.G. Bouwhuis (Eds.), *Attention and Performance X*. Hillsdale, NJ: LEA.

Marslen-Wilson, W.D., and Tyler, L.K. (1975). Processing structure of sentence perception. *Nature*, 1975, **257**, 784-786.

Marslen-Wilson, W.D., and Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8** 1-71.

Marslen-Wilson, W.D., and Welsh, A. (1978). Processing interactions and lexical access during word-recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Nooteboom, S.G. (1979). More attention for words in speech communication research? In: B. Lindblom and S. Ohman (Eds.), *Frontiers of Speech Communication Research*. London: Academic Press.

Shannon, C.E., and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Streeter, L.A., and Nigro, G.N. (1979). The role of medial consonant transitions in word perception. *Journal of the Acoustical Society of America*, **65**, 1533-1541.

Tyler, L.K., and Wessels, J. (1983). Quantifying contextual contributions to word recognition processes. Manuscript, MPI for Psycholinguistics, Nijmegen.

Spontaneous Self-Repairs in Speech: Processes and Representations

Willem J.M. Levelt
Nijmegen, the Netherlands

1. Phonetics, Psycholinguistics, and Self-Repairs

Psycholinguistics study the acquisition, comprehension and production of language, but the study of language production has not advanced at the same pace as the other two main areas of inquiry. While there are notable exceptions, the impetus to study the process of speaking does not usually come from within the discipline itself, but from neighboring fields, such as ethnomethodology, aphasiology, and last but not least, phonetics. No approach has informed the psychological study of language production to the same extent as the systematic analysis of spontaneous speech errors, a technique which was reintroduced by Cohen and Nooteboom during the sixties. Traffic in the reverse direction has been light: the psychology of speaking has not influenced phonetics to the same degree.

It is only honest, therefore, to speak about the *potential* contributions to phonetics of psycholinguistic production research. The study of spontaneous self-repairs may well develop into such a contribution. Self-repairs are, on first view, rather complex phenomena. And surely, they involve quite disparate phonetic processes, such as self-monitoring, the production and detection of phonetic, lexical and other types of speech errors, self-interruption, prosodic marking of the correction, etc. This complexity on the surface, however, does not preclude systematicity at a deeper level, a systematicity which may reveal principles of organization of the speech production process that would be hard to discover on the basis of laboratory data alone.

The more specific psycholinguistic contribution here is to clarify the character of this underlying systematicity.The psycholinguist will, more in particular, try to analyse the levels of representation involved in the generation of a speech repair. What are the relevant entities for the analysis of a speaker's self-monitoring, self-interruption, re-starting, etc? Are they phonetic features, phonemes, words, clauses, concepts, intentions, or several of these at the same time? And closely related to this is the issue of how these entities are stored and addressed during the process of repairing. What sort of memory structures are involved, for instance, in repeating part of the same utterance, or in aligning the prosody of the correction to that of the interrupted utterance? The psycholinguist will try to explain self-repairs in terms of the same representations and processes which underly normal fluent speech.

Phonetic complexity may turn out to be psycholinguistic systematicity. Let us therefore turn to normal fluent speech first.

A speaker usually constructs and uses an utterance in order to realize an intention, such as to inform or convince an interlocutor, to request action or information from him, etc. The psycholinguist tries to follow the flow of information from the initial conception of an intention to the articulation of speech. The processing is done in stages involving different levels of representation. An early stage concerns the retrieval and selection of information whose expression may fulfill the intention. The speaker takes a certain *perspective* with respect to this information; the information will have a focus, it may contrast with earlier expressed information or rather be a further elaboration of what was previously said, there will be a spatio-temporal relation between the speaker and the information selected, and the speaker will have an attitude of belief, want, distrust etc. with respect to that information. The information selected for expression is usually called the *message*. The subsequent levels of representation result from mapping the message onto linguistic form. Major processes here are the retrieval of lexical items, the creation of syntactic configurations such as phrases and clauses, the realization of morphonological structure. The final output of these processes can informally be called *inner speech*. Phenomenologically speaking, inner speech *is* a level of representation; whether it can be used as a theoretical construct is still to be seen.

The last stage in the flow of information is the planning and execution of articulatory procedure. Its output is *overt speech*.

How do the observable phenomena in self-repair proceed from different levels or types of representation involved in normal speech? In the following, I will discuss some of these issues in the light of a corpus of 959 tape-recorded repairs. The corpus was obtained in an experiment designed for a different purpose. In this experiment 53 subjects were asked to describe spatial patterns consisting of colored dots, connected by horizontal or vertical arcs, such as the ones in Figure 1. There was, on the average, one repair in every third pattern description.

The main analysis of this corpus of repairs can be found in Levelt (1983), a subsequent study of prosodic aspects of these repairs appears in Levelt and Cutler (1983). Others, especially Ewald Lang (unpublished) have also contributed to the analyses of this corpus.

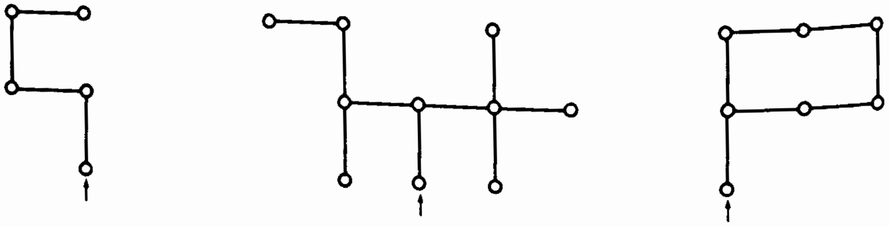


Figure 1. Examples of patterns described by subjects. Dots in the patterns were differently colored.

2. The Structure of Repair

A repair consists, typically, of three phases. In the first, trouble occurs and is detected by the speaker, who decides to interrupt the flow of speech. The segment of speech affected will be called the *original utterance*. In example (1) the original utterance is 'right of pink is a black'.

1. right of pink is a black, er a blue point

The original utterance contains a lexical error ('black' for 'blue') which is apparently detected by the speaker, who interrupts immediately after the error.

The second phase is one of filled or unfilled pausing. In (1) the pausing is filled by 'er', but other *editing terms* are also frequently used. They are systematically related to the source of trouble, and how recently it occurred.

The third and last phase consists of the *correction* itself. In (1) it is 'a blue point'. Important events take place in this phase. The speaker tells the listener how to relate the repair to the original utterance, by restricted syntactic, lexical, and prosodic means. In this way the speaker establishes on-line interpretability of the repair for the listener.

Let us now turn to these three phases in some more detail.

2.1. Interrupting the Utterance

There are many possible reasons why a speaker might want to interrupt the flow of speech, but two major sources of trouble were found in the corpus of self-repairs on which the present study is based. The first is the appearance of error, be it a phonetic error (seldom, less than 1% of the data), a lexical error as in (1) (frequent, 38% of the corpus), or other (2%).

The second most frequent situation is one in which, although what was said was correct, it was not fully appropriate. An example is given in (2):

2. a line to the yellow disc, to a yellow disc

Here the yellow disc had not been introduced before by the speaker, and it is thus more appropriate to use the indefinite article.

Another appropriateness repair is given in (3):

3. right thereof, of the orange one a blue dot

The demonstrative 'thereof' refers to a dot mentioned in a previous utterance, but the speaker realizes that there may be an ambiguity of reference here, and decides to name the referent explicitly. In neither (2) nor (3) anything was said that was *false*. The new utterance is only more appropriate in the discourse context (the precise criteria for this category of repairs are outlined in Levelt, 1983). Thirty percent of the corpus consists of appropriateness repairs. These two main categories of trouble, error and inappropriateness, are very different in their consequences for the way in which the repair is made. It should be added, for completeness' sake, that there is a third major category of repairs in the corpus whose etiology cannot be decided unambiguously; we called them *covert* repairs or hesitations. An example is given in (4):

4. up is, er blue

It is unclear whether the speaker had a perceptual problem, had difficulty finding the color name, intercepted an erroneous lexical item that he was about to pronounce, or otherwise. These covert repairs account for 25% of the corpus. (the remaining 4% of the corpus consists of minor categories which are of no interest for the present purposes).

How does the speaker detect trouble? Or in the above terminology: which levels of representation are accessible to the speaker? My conjecture is that the speaker can attend to messages, to inner speech and to overt speech, but to nothing else. He has no direct way of monitoring his own lexical access procedures, the construction of phrasal configurations, the assignment of syntactic agreement, the construction of tone groups, the generation of coarticulation, etc. The speech production apparatus is *cognitively impenetrable*, to use Pylyshyn's (1980) terms. If this is correct, there are exactly three levels of representation which allow for monitoring. The first one is the message level. A speaker may want to replace or change a message before it enters the next stage of processing. This may lead to delays or hesitation, such as in (4).

The second one is at the level of 'inner speech', and the third one is at the level of overt speech. For the latter two levels I suppose that the speaker uses his normal speech perception apparatus; he or she will parse the speech produced, and derive the message, as if listening to someone else. Self-monitoring will in the first place consist of comparing this derived message to the intended message. If there is a major or communicatively important difference in truth, reference, clarity, etc. between derived and intended message, the speaker may want to interrupt speech and make a correction. The speaker can also monitor for certain well-formedness aspects of the self-produced speech, such as phonetic errors. Although the processes involved are quite mysterious, they need not be different from those involved in detecting ill-formedness in the speech of others.

The moment of interruption can vary widely with respect to the trouble spot. In (1) it follows the trouble item immediately, and there are many instances in the corpus where interruption is even faster, namely *within* the trouble item itself, as in example (5) below. One major theoretical issue is why speakers do not always interrupt immediately. In (2), the trouble item 'the' is followed by two more words before interruption. Is this due to inertia in the production apparatus, i.e. a tendency to complete some linguistic unit (a clause, a phrase, a tone group)? Or is it rather the case that the speaker did not *detect* the trouble until two words later?

The analyses point largely to the latter explanation. We obtained statistical evidence for a slight tendency on the part of the speakers to interrupt more often (in 71% of the cases) at the end of a surface phrase, such as an NP, a VP, a PrepP, than was to be expected statistically (we found that a random point of interruption in these pattern descriptions completed a phrase in 58% of the cases). The inertia theory predicts that a speaker tends to complete a phrase after detection of trouble. Delayed interruptions will therefore respect phrase boundaries more often than immediate interruptions. This is, however, not what was found. Phrase structure was respected in 66% of the delayed interruptions, but in as much as 74% of the immediate interruptions. This argues against the inertia theory. The tendency to respect constituent structure should rather be interpreted as resulting from a detection mechanism. There is an increased chance of detecting trouble towards the end of a surface phrase, and indeed we found that the rate of detected versus non-detected errors increases sharply towards the ends of phrases (cf. Levelt, 1983). The speaker's attention apparently fluctuates between constructing the message and monitoring the inner or overt speech. Ends of phrases are natural points for checking their contents, or in other words, the phrasal structure of inner and/or overt speech dictates the rhythm of attention shifts. This phrasal constraint is not due to formulating inertia, but to trouble *detection*, i.e. to perceptual parsing. As far as this goes, we can maintain that *speakers interrupt their speech immediately upon detection of trouble.* This rule is in full correspondence with Nooteboom's (1980) analysis of the Meringer data. The rule predicts that speech can be stopped at any point after detection of trouble. In one analysis we checked whether a speaker respects phonotactic boundaries while interrupting his speech. There were 172 within-word interruptions in the corpus. An example is given in (5):

5. rechtsaf naar /z/-, wit (right to /b/-, white)

Here the speaker started saying 'zwart' (black), but then interrupted the incorrect word to replace it by 'wit' (white). The interruption in (5) violates phonological well-formedness in Dutch, /z/ is, phonotactically speaking, not a possible word. We used the 'possible word' criterion to listen to all cases of word-interruption in the corpus. Although such judgments were not equally straightforward, there were 67 cases where we felt certain that the interrupted fragment was phonotactically not a possible word of Dutch, as in (5). It is hard to evaluate this finding statistically, but one thing can be said with confidence: nothing prevents the speaker from interrupting speech at phonologically odd places. Again, inertia of the production apparatus, in this case at a phonotactic level of representation, does not seem to be an important determinant of moments of interruption.

What we did find, however, is that speakers tend to complete words before interruption. Only 20% of appropriateness and error repairs involve halting within a word. But there is an interesting rule here; these 20% are not evenly distributed over repairs. Of the appropriateness repairs only 10% involve word interruption, whereas 28% of error repairs do. Is it the case that speakers are willing to interrupt an *erroneous* word, but not a correct (though potentially inappropriate) one? This is easily checked. It predicts that the percentage of word interruption should be high for immediate halts in error repair, i.e. halts during or right after the trouble item such as in (1) and (5) above; in these cases a within-word interruption is indeed interruption of an erroneous word. The percentage should, however be low for error repairs with *delayed* interruption, such as (6):

6. and left of the black disc, no right of the black disc...

Here *left* is erroneous, but the subsequent words till interruption are all correct. If the rule 'Do not interrupt correct words' holds, there should be a low incidence of within-word interruptions in these delayed cases. What we found was 47% word-interruptions when halting was immediate, against 17% when halting was delayed in error repairs. This is in agreement with the rule. It should be noted that this qualifies the interruption rule we gave a moment ago: there are cases where a speaker does not immediately interrupt upon detection of trouble. Speakers tend to complete words in all cases, but they are willing to interrupt ones that are erroneous. This finding is in agreement with Nooteboom's (1980) analysis of the Meringer corpus, and one may conclude that the interruption process is sensitive to representations of meaningful words.

2.2. Editing Terms

Editing terms vary in the degree of contrast they establish. If the term is 'no', 'rather' ('of' in the Dutch corpus), or 'sorry', it involves an explicit rejection of what was said. But if it is 'therefore' ('dus' in the Dutch corpus) it rather confirms the previous expression. The degree of contrast set up depends on the occasion for repair: error releases much more contrast than does inappropriateness. Indeed we found a much higher incidence of terms such as 'no' 'or', 'sorry' in error repairs than in appropriateness repairs. Inversely, Dutch 'dus' occurred exclusively in appropriateness repairs. More generally, error repairs released more than twice as many editing expressions than appropriateness repairs (62% versus 28%).

The degree of semantic contrast in a repair is not only expressed by the editing term, but also by prosodic features. In a recent paper Cutler (1983) proposed to make a distinction between repairs that are prosodically marked versus those that are unmarked. Pitch, amplitude and relative duration of an unmarked repair closely mimic those of the trouble segment. A repair is marked when repair and trouble item differ prosodically. Levelt and Cutler (1983) applied this distinction to a subset of 299 lexical repairs in the present corpus, i.e. repairs where a single trouble word was replaced in the repair. We found that error repairs are far more often prosodically marked (in 53% of the cases) than appropriateness repairs (only 19%), and we could show that this is indeed due to the degree of semantic contrast established in the error repairs. Ewald Lang (unpublished) drew our attention to a correlation between editing term used and prosodic markedness. If we compare the 'contrast establishing' editing terms 'nee' (no), 'of' (rather), 'sorry' (sorry) to the 'neutral' editing term 'eh' (er) and the non-contrasting 'dus' (therefore), we find 55% prosodically marked cases among the former repairs, but only 32% among the latter. Editing terms and prosodic marking thus seem to arise from a semantic level of representation, but this does not exclude the existence of other determinants. It was argued in Levelt (1983) that the interjection 'er' entertains a rather mechanical relation to the interruption process. The faster the interruption after trouble, the higher the incidence of 'er'.

2.3. The Correction and its Relation to the original Utterance

There are at least three determinants of the way in which the speaker constructs the correction. I will call them *intentional, interactional* and *structural.*

2.3.1. Intentional Determinants

The intentional determinants have to do with what a correction is made *for*, especially whether the speaker intends to patch up an error or rather to find a more appropriate way of expressing the same state of affairs. We found major differences between corrections for error and corrections for appropriateness. I already mentioned the difference in prosodic marking between these two types of repair. A major finding is furthermore that corrections for error are highly conservative, closely copying the wording of the original utterance. This is far less so for appropriateness repairs. A detailed account can be found in Levelt (1983).

2.3.2. Interactional Determinants

I will also be short on the interactional determinants. The main point is this: The speaker's sudden interruption of the flow of speech leaves the listener with a so-called 'continuation problem': how is the new utterance to be related to the interrupted utterance? The listener must decide how to 'splice' the two parts together, so to say. We were surprised to find that speakers construct their repairs in such a way that the listener can solve this 'continuation problem' on-line, i.e. no later than upon hearing the very first word of the correction. The main rules the speaker adheres to in order to achieve this for the listener are given on page 48 of the Abstracts of this Congress, and further details are to be found in Levelt (1983)

2.3.3. Structural Determinants

There are restrictions on the ways in which the speaker can make a repair which are purely structural in character. Example (7) is a well-formed pair:

7. Is he seeing, er interviewing patients?

A grammatically ill-formed way of correcting for the same trouble would be (8):

8. *Is he seeing, er he interviewing patients?

There is nothing semantic or pragmatic which forbids (8), it is fully transparent but still ill-formed. Is this ill-formedness a consequence of the structure of the correction itself, i.e. 'he interviewing patients'? One might argue that it doesn't have the right constituent structure for being used as a correction. But this is not so; the same correction is all right in the following example:

9. Is she, er he interviewing patients?

Apparently, the grammatical well-formedness of a repair is a function of the structural relation between the original utterance and the correction. In Levelt (1983) I have argued that this structural relation is essentially the same as the relation between conjuncts in a coordinate structure, and a precise rule was given how to derive the well-formedness of a repair from the well-formedness of a corresponding coordination. This will not be repeated here. On this occasion I would rather address the issue of how this structural relation is realized in the process of speaking.

The speaker has certain 'grammatical commitments' at the moment of interruption. In (7), for instance, at the moment after 'seeing' there is the commitment to complete the verb phrase by either a noun phrase or a complement clause. Another way of putting this is that the production process is 'under the control of VP' at the moment of interruption. Hoenkamp (1982) suggests that this production process is held 'in suspension' during the editing phase, and subsequently reactivated in order to produce the correction. In other words, there would be a way to store the control structure of the interrupted utterance. This control structure is at a different level of representation than the message. The speaker often changes (adapts, specifies) the message in making a repair, but that change is executed under the same control structure. And if only the message, but not the original utterance itself, were kept in store, one could not prevent a repair such as (8) which is, as was argued, semantically and pragmatically fully transparent.

The suspension theory puts the storage of the relevant features of the original utterance on the output side: the formulation program is in a state of abeyance. But one could also put storage more in the input side. A possibility suggested by Levelt (1983) is that the speaker, who is his own listener, keeps a trace of his own recent inner or overt speech. This trace can then be referred to in the construction of the repair proper. This would involve a very different type of representation. One would either expect the trace to be in echoic memory, which is a rather short term acoustic storage, or in working memory, i.e. it is at least phonemically coded or else semantically.

It is not an easy task to find out which representational system mediates between the structure of the original utterance and the structure of the repair, and I have no definite answer to offer. Still, I feel that a phonetic analysis may clarify these issues. Consider first working memory as the locus of storage. It is well known that the fine acoustic shape of speech is not represented in working memory; the level of coding is phonemic or semantic rather than phonetic. One would therefore not expect the repair to be an *acoustic* continuation of the interrupted original utterance, though grammatical contiguity as in (7) or (9) would be possible. Echoic memory as the locus of storage leads to a different prediction. In this case there exists a faithful auditory trace, but it is short-lived. The most accurate measurements of 'brief auditory storage' are those by M. Treisman and Rostron (1971); they found that the auditory trace was lost in about 1 s., confirming earlier data for storage of running speech obtained by A. Treisman (1964) who found a value of 1.3 s. One would therefore predict good acoustic contiguity for short delays between trouble item and replacement, but diminishing contiguity for longer delays, with an asymptote at 1 to 1.3 s.

The suspension theory, finally, predicts that acoustically the new utterance should fit seamlessly into the original utterance, even for longer delays (or at least there is no known limit on the persistence of an interrupted speech program).

The obvious phonetic analysis to undertake for distinguishing these three loci of memory is to splice the new utterance, i.e. the correction itself, into the original interrupted utterance at the appropriate place, that is deleting all the repeated material and the whole editing phase, and then to listen whether the resulting utterance is phonetically natural (I am grateful to Anthony Cohen who suggested this way of splicing to me).

If the locus of storing the original utterance is working memory, naturalness of the spliced utterance will be a matter of accident, since no acoustic or motor information is preserved in working memory. In other words, naturalness will not generally result.

The echoic memory theory predicts naturalness for cases where the repair had a short interval between trouble and replacement. Naturalness will however, break down for cases where that interval exceeded 1 to 1.3 s,

If naturalness is preserved for cases where the interval substantially exceeded the 1 to 1.3 s. limit, the suspension theory is the remaining alternative. It should not be expected, of course, that the interrupted speech program will be preserved indefinitely, and it is an interesting empirical issue to find out what size of interval can be bridged by the program in abeyance. I would, finally, like to express my awareness that the splicing test is not a definitive one in any sense, but it is probably as far as one can get on the basis of natural data.

We took special care in selecting utterances for this splicing test. It was, firstly, important to take repairs with a fair range of delays between trouble item and replacement. Secondly, the sample should contain both error and appropriateness repairs; they are *intentionally* quite different, but this should be immaterial for any of the three forms of storage. Thirdly, it was decided to splice both marked and unmarked repairs. Since marking involves a change of the prosodic structure, marked repairs should be very vulnerable to phonetic discontinuity. Fourth, we tried to select a wide range of splicing points. In repairs such as (5), (6), (7), (9), the first word of the repair proper (i.e. after the editing term) is the replacement for the trouble item. They are called *instant repairs*. In other repairs, such as (1) and (2), the speaker retraces to an earlier word so as to ‘lead in’ the replacing item. These *retracings* can be substantial (in number of words). The amount of retracing determines the distance between the splicing point and the point of trouble/replacement. Finally, care was taken to select repairs from a wide range of different speakers.

Guided by these five requirements we selected a set of repairs from the transcripts, i.e. without listening to the tapes. Subsequently we checked which of these were of sufficient acoustic quality to splice them by means of the Max-Planck speech editing system SPED. The surviving twenty-one candidates are listed in Table I.

The repairs are ordered in terms of increasing delay, and for each repair the places of splicing are marked by ‘/’. In other words, the stretch of speech between the first and the second occurrence of ‘/’ was deleted in the splicing procedure. Also, the table lists the duration of this deleted stretch of speech, which is a measure for the delay between trouble item and replacement; the range is from 0.55 to 6.34 s. Furthermore, the repairs’ status as error/appropriateness and as marked/unmarked are given in the table. All repairs are from different speakers, except for the pairs 1 and 12, 5 and 6, 9 and 15, and for the triple 3, 7, 19.

We (the author and Ger Desserjer – see acknowledgements) listened to the resulting utterances, and found out that up till item 18 they were all perfectly natural phonetically (items 3, 6, 9, 12, 15, and 18 were demonstrated during the presentation of this paper at the Congress of Phonetic Sciences, so that a wider audience could convince itself). It is especially important to notice that no loss of naturalness occurs for cases where the spliced interval exceeded the 1.3 s. asymptotic value of the ‘brief auditory trace’ (items 9 through 18). This finding makes both echoic memory and working memory less likely loci for the preservation of the original utterance’s relevant features. By default the results support the suspension theory: the speaker can interrupt the flow of speech at any moment, but hold the control structure in abeyance over substantial stretches of time which can be filled with other speech activities.

What is the limit for this persistence? Cases 19 to 21 suggest that the critical interval could be about 3 or 4 s. Though case 21 sounds natural when spliced, both cases 19 (demonstrated at the Congress) and 20 are clearly unnatural.

Table I. Repairs used for splicing. The parts between ‘/’ and ‘/’ are deleted. (*d* = delay in seconds, E = error repair, A = appropriateness repair, M = prosodically marked, U = prosodically unmarked)

1. (*d* = 0.55, A, U) DAAR LINKS VAN / een blank of / EEN WIT KRUISPUNT
(left thereof / a blank or / a white crossing point)
2. (*d* = 0.68, E, M) VANUIT HET GROENE GA JE NAAR / links / RECHTS EN DAAR LIGT EEN EH GEEL KRUISPUNT
(from the green you go to the / left / right and there is a er yellow crossing point)
3. (*d* = 0.69, A, U) DAAR KUN JE ALLEEN MAAR RECHTSAF, DAN / gaan we n- /DAN KOMEN WE BIJ HOEK ORANJE
(you can only go right, then / we go t- / then we come to corner orange)
4. (*d* = 0.80, E, U) EN RECHTS VAN BLAUW IS EEN WEG / naar een grijze / NAAR EEN ROZE PUNT
(right of blue is a way / to a gray / to a pink point)
5. (*d* = 0.83, A, U) NAAR RECHTS / gaan we / LOPEN WE DOOD OP EEN ZWART KNOOPPUNT
(to the right / we go / we get stuck at a black node)
6. (*d* = 0.90, E, M) ANDERE MOGELIJKHEID VANUIT HET / groen-, eh / BRUINE PUNT LINKSAF NAAR EEN T-KRUISING GEEL
(other possibility from the / green, er / brown point left to a yellow T-crossing)
7. (*d* = 1.05, E, U) DAN GAAN WE EERST MAAR EVEN RECHTSAF NAAR PUNT / rood, eh sorry / ORANJE. DAT IS EEN EINDPUNT
(then we go first for a while to the right to point / red, er sorry / orange. That is an end point)
8. (*d* = 1.25, A, U) DE WEG OMHOOG / dat is een / DIE KOMT UIT BIJ EEN ZWARTE KRUISING
(the way up / that is a / that ends at a black crossing)
9. (*d* = 1.51, E, M) DAAROP VOLGT / een horizon- nee / EEN VERTIKALE LIJN WAARBOVEN EEN WIT BOLLETJE ZIT
(therafter follows / a horizon- no / a vertical line above which is a white ball)
10. (*d* = 1.52, A, M) VANAF HET GELE KNOOPPUNT / gaan we n- / TREKKEN WE EEN VERBINDINGSSTREEPJE NAAR BENEDEN NAAR HET BLAUWE KNOOPPUNT
(from the yellow node / we go t- / we draw a connecting line downward to the blue node)

There ıs a lack of long-interval data in our corpus, and new data will be necessary to substantiate these values. Also, the persistence theory is clearly in need of further theoretical specification. One would like to know more about the precise nature of the stored code, about its sensitivity to interference etc.

Considering, finally, the close correspondence between the structure of repairs and the structure of coordination, it is of great interest to apply the same splicing test to coordinations, such as in (10):

10. JOHN COOKED / and Mary ate / THE DINNER

Will one find comparable values for the persistence of the control structure in these cases?

Table I (cont.)

11. (*d* = 1.70, E, U) HIER RECHTSAF NAAR / blauw of / PAARS
(here right to / blue or / purple)
12. (*d* = 1.71, A, U) VANUIT DAAR / naar onderen / NAAR OMLAAG EEN GRIJS KRUISPUNT
(and from there / downward / descending a grey crossing point)
13. (*d* = 2.00, E, M) TWEE MOGELIJKHEDEN. LINKSAF / naar or- naar paars eh / NAAR ROSE
(two possibilities. left / to or- to purple er / to pink)
14. (*d* = 2.05, E, M) EEN VERBINDING / tussen blauw en ro-, nee wacht / TUSSEN GROEN EN ORANJE
(and a connection / between blue and re-, no wait / between green and orange)
15. (*d* = 2.12, A, M) DAARONDER EEN LIJN / naar het gele rondje / NAAR EEN GEEL RONDJE
(thereunder a line / to the yellow disc / to a yellow disc)
16. (*d* = 2.21, E, U) NAAR RECHTS / naar het gele rondje of naar het gele / NAAR HET GROENE RONDJE
(to the right / to the yellow disc or to the yellow / to the green disc)
17. (*d* = 2.62, E, U) DAN WEER EEN ZWARTE RECHTE LIJN NAAR BOVEN TOE MET / een paars bolletje of nee sorry / EEN GRIJS BOLLETJE
(then again a black straight line upwards with / a purple ball or no sorry / a grey ball)
18. (*d* = 2.68, first A, then E, U) IK MOET STEEDS / rechte strepen maken, rechte wegen megen eh eh / RECHTE WEGEN MAKEN
(I must always / make straight lines, moak straight roads er er / make straight roads)
19. (*d* = 3.73, E, U) DAN RECHTDOOR NAAR / paars, eh eh sorry hoe heet 't eh / ROSE
(than straight on to / purple, er er sorry what's it er / pink)
20. (*d* = 4.45, A, M) UIT HET VEL KOMEND / ligt halverwege en ook gezien van beide zijkanten / LIGT HALVERWEGE PUNT BLAUW
(rising from the sheet / is halfway and also viewed from both edges / is halfway point blue)
21. (*d* = 6.34, A, M) DAAR / heb je / KUN JE NAAR RECHTS AFDRAAIEND EEN VIERKANTJE RIJDEN
(there / you have / you can turn to the right and drive a square)

3. Conclusion

Phonetics has contributed more to a psycholinguistic theory of speaking than inversely. But one may hope that this situation will change in the near future. The general direction of psycholinguistic contributions to phonetics will probably be one of delineating underlying representational systems which are responsible for the generation of surface phonetic phenomena. This approach was exemplified by an analysis of spontaneous self-repairs in speech. The phonetically rather chaotic surface behavior of speakers can in many respects be related to underlying levels and processes of representation.

Acknowledgements

Thanks are due to Ger Desserjer who assisted in analyzing the corpus of repairs and in splicing the utterances in Table I. Anne Cutler and Ewald Lang have opened my eyes for important aspects of self-repairs which were ignored in Levelt (1983). Discussions with Anthony Cohen and Sieb Nooteboom have been essential experiences in interacting with phoneticians. I am, finally, very grateful to Uli Frauenfelder, who read and corrected an earlier version of this paper.

References

Cutler, A. (1983). Speakers' conceptions of the functions of prosody. In: *Prosody. Models and Measurements.* (A. Cutler and D.R. Ladd, eds.). Heidelberg, Springer.

Hoenkamp, E. (1983). Een computermodel van de spreker. Psychologische en Linguistische aspecten. Dissertation Nijmegen University.

Levelt, W.J.M. (1983). Monitoring and self-repair in speech. *Cognition,* **14**, in press.

Levelt, W.J.M. and Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics,* **2**, in press.

Nooteboom, S.G. (1980). Speaking and unspeaking: detection and correction of phonological and lexical errors in spontaneous speech. In: *Errors in Linguistic Performance.* (V.A. Fromkin, ed.). 87-96.

Pylyshyn, Z. (1980). Computation and cognition: Issues in the foundations of cognitive science. *The Behavioral and Brain Sciences,* **3**, 111-132.

Treisman, A.M. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior,* **3**, 449-459.

Treisman, M. and Rostron, A.B. (1972). Brief auditory storage: a modification of Sperling's paradigm applied to audition. *Acta Psychologica,* **36**, 161-170.

5. Speech Technology in the Next Decades

Speech Technology in the Coming Decades

J.L. Flanagan
Murray Hill, N.J., USA

Especially in science, prognostication is at best risky – at worst, futile. Nevertheless, my assignment is to comment on speech *technology* in the coming decade*s*, and I take up this gauntlet.

To achieve perspective, it seems prudent to look over the recent past and take note of advances that have been key in speech research, and that have significantly impacted speech technology. These may suggest the nature of accomplishments to look toward in the future. To make such assessment, some criterion of focus is naturally implied. The choice invariably is conditioned by personal experience with the field.

First, why do we do speech research? Many would say to provide greater capabilities for human communication. And, I believe this to be a moderately universal and valid motivation. How do we improve capabilities for human communication – both between humans, and between humans and machines? The possibilities branch in many directions. My choice lies with techniques for telecommunications and voice processing. Others devote effort to communication aids for the handicapped, to speech teaching and therapy, to studies of language and language acquisition, to diagnostic methods for voice disorders, and to the many areas typified in the literature of the phonetics journals. Let us agree, though, that our common motivation is betterment of human communication, and against this backdrop presume to assess – and extrapolate – contributions in speech technology. In so doing, we can try to correlate the technological needs, the advances to meet the needs, and the acquisition of fundamental understanding to support the advances.

The era of the 1940's rode the swell of the evolving electronics age, and it seems not unreasonable to commence comment here. Undoubtedly the prominent technology of this time must include the vocoder – the first practicable analysis/synthesis system for bandwidth conservation in telephony. This achievement was spurred by the desire to transmit voice over the early transatlantic cable. But, this cable (before the time of integral, submerged amplifiers) could only support a bandwidth of a couple hundred Hertz, only enough for telegraphy. The desire to have the speed and convenience of voice communications therefore gave rise to the vocoder technique for a 10-fold reduction in the bandwidth of a speech signal.

Stemming also from this motivation were the fundamental concepts of the 'carrier nature of speech', the 'source-system' model of the signal, and the 'information-bearing' properties of the short-time amplitude spectrum. And while the vocoder was never put to work on telegraph cables (because bandwidth improvements progressed more economically), it later found extensive use for voice encryption purposes. Also, its synthesizer component evolved into a human-controlled electronic speaking machine – the voder. Its analyzer component influenced the design of the sound spectrograph, a fundamental instrument commonly found in most phonetics laboratories, and the design of the visible-speech translator, a useful tool for articulatory training of hearing-impaired individuals. On the perceptual side, the need to characterize and analyze the performance of speech processing systems gave birth to the concepts of articulation testing and articulation index.

The era of 1950 continued the interest in efficient voice communications, but recognized the need for better fundamental underpinnings. The wave nature of sound propagation in the vocal tract was put on a firm basis, as was similar understanding for auditory function; i.e., for the basilar membrane. Transmission-line models – bilinear, passive circuits – were introduced to good effect as analog computers. The non-independence of speech-spectrum amplitudes, and the information properties of formants were firmly established. Engineers moved to exploit this knowledge in automatic formant trackers and in formant analyzers and synthesizers. Concomitantly, the field of electronics experienced major progress with the introduction of the transistor and solid-state circuitry – a harbinger of greater vehicles for speech technology (the digital computer and integrated electronics).

The 1960's witnessed the impact of digital computers in partnership with sampled data theory as formidable tools for speech research. Previously, great limitations were imposed on the complexity of algorithms that could be implemented in electronic circuitry, and on the speed with which new ideas could be realized for test in traditional analog electronics. Digital simulation significantly relaxed these restrictions, and allowed much greater sophistication in processing. Speaking machine programs were of immediate interest, and formant synthesizers with discrete phonetic control of segmental and supra-segmental features attracted early interest. Eventually, complete formant vocoders were implemented in the laboratory, with real-time formant tracking accomplished by dedicated computer. The traditional vocoder concepts, and the extensions to pattern-matching vocoders (now given the more prestigious term vector-quantized spectra), were also cast into digital forms. On the practical side, transistor circuitry supported the electronic artificial larynx, which was built upon fundamental understanding of vocal-cord function. And, the vocoder concept of the source-system signal model was extended to its most sophisticated level in the form of linear-predictive-coding (LPC).

By the early 1970's, with broadband transmission technologies such as coaxial cable a reality, light guide showing great promise, and digital machines increasing in capability, the needs in speech technology largely shifted away from band-conservation and toward human/machine communications. Giving machines the ability to speak stored information to a human, and to respond to human-spoken commands (even to confirm the identity of the talker) became central foci of research. Initially, the accumulated understanding from the vocoder art, and its direct derivatives, supported these efforts. But the sophistication of the machines permitted much more. Complete systems for speech synthesis from printed text were demonstrated and tested, for information retrieval purposes and as reading machines for the visually handicapped. Waveform coding methods such as adaptive differential-PCM (ADPCM) were devised for transmission economies, but used initially for multi-line computer voice answerback systems. Isolated word recognition systems of high performance, and talker verification systems of high accuracy filled in the developing picture. Fundamental studies to support more ambitious undertakings did not languish either. Detailed computer models of vocal-cord and vocal-tract function were established for speech synthesis. Sub-languages, having usefully-large vocabulary size and quantitatively-delineated grammar, were designed and programmed for automatic syntax analysis in speech recognition systems. But in all of this work, the central tool, the laboratory digital computer and its elaborate peripherals, remained large, expensive and oftimes not fast enough for real-time simulations.

Around 1980 this picture changed dramatically, with explosive advances in microelectronics. Already in the early 80's we have single-chip computers that are more powerful than the dedicated laboratory computers of the 70's. Integrated speech synthesizers are pervasive, and even provide convenient test beds for phonetics laboratories. Chip-set speech recognizers are appearing, and most of the designs can be made compatible with the communication protocol of existing microcomputers.

As we approach the mid 80's, activities in speech technology are still dominated largely by the needs of human/machine communication. The advances necessary to meet these needs are in the areas of higher-quality synthetic voice, automatic recognition of connected speech, and simultaneous speech and talker recognition. By the end of the decade, 1990, it seems reasonable to expect significant advances in each sector. We will have text-to-voice converters that will deal reliably with virtually unlimited vocabulary and will produce intelligible, natural-sounding output. We will not be able to specify and duplicate the subtleties of dialect and accent, but we will be substantially past the stage of the inept automaton. Similarly, recognizer-/synthesizer systems will be able to carry on intelligent, interactive conversations with humans. Not fluently, nor with all talkers on all subjects, but constrained to vocabularies, grammars and topic areas that are nevertheless comfortably large. The applications outside telecommunications, such as aids for handicapped and speech teaching, are apparent.

The fundamental studies to support these advances are numerous and do not differ much from the objectives of the recent past. Accurate, agile, models of articulation are needed. Letter-to-sound conversion – utilizing the proper marriage of stored pronouncing dictionaries and grapheme/phoneme rules – will be refined. Connected speech recognition will utilize stored word templates initially. But, as aspirations expand to vocabularies on the order of 1,000 words, or more, feature labelling and statistical modelling may prove more attractive.

While the need for bandwidth conservation, of late, has been in the background, it will not remain so. Telecommunication transmission systems are rapidly evolving to digital techniques. And while light guide will provide enormous bandwidths, specific considerations of access and switching make band conservation attractive, and in some cases necessary. Already 32K bits/sec ADPCM is on the threshold of use to achieve 2:1 savings in transmission capacity. Other techniques, such as sub-band coding (SBC), perceptually-weighted multi-pulse LPC, and adaptive transform coding (ATC), are in advanced stages of research and address the transmission ranges below 32K bits/sec. A strong, emerging need is to encode and transmit high-quality speech at data speeds, 9.6K bits/sec and lower. This low bit-rate coding makes speech signals adaptable to networks that also handle data and low-rate non-speech signals. The end of this decade will see high-quality speech transmission at rates in the range of 9.6K. High-quality at much lower rates, for example 2.4K, is possible, but probably will be longer in coming. Fundamental understanding, that significantly surpasses the traditional source-system signal model, must first be acquired.

If we look toward the 2000 era, predictions indeed become risky. But, I believe this time scale can see the beginning of a unified 'ultimate' solution to the synthesis, recognition and coding problems. At this moment, I believe this solution must be based upon a speech signal model that much surpasses the traditional vocoder source-system model – one that allows exquisite, dynamic representation of the details of laryngeal and tract functions and which is controlled adaptively to 'mimic' an unknown input (either to duplicate it for synthesis, to categorize it for recognition, or to parameterize it for coding and transmission). The adaptation algorithms obviously must contain voluminous built-in information about speech constraints and conventions, and about the mechanism of speech perception. Some of this insight is being accumulated. But the complete solution is some years away. In part, it will depend upon quantum advances in computer capability. For practical application, inexpensive processors with arithmetic capability in excess of 100 Mips will be needed to support this complexity.

Speech Technology in the Next Decades

J.N. Holmes
Cheltenham, United Kingdom

1. Introduction

This talk is about likely technological applications related directly to future speech research. Because the subject is so vast I will be forced to restrict my discussion of the applications to three main areas – automatic speech synthesis, authomatic speech recognition and digital speech coding for transmission or storage. I will divide my projections into two parts: short term (less than 10 years) and long term (significantly more than 10 years). While the research in these subjects is progressing there will be many diversions to apply intermediate results to immediate practical problems as they become soluble.

It is unavoidable in making projections of this type that I will be adopting my own personal viewpoint, based on my particular research experience and the research that is currently going on in my own group. I would expect people with a different background to see things differently.

I predict that the really advanced use of the results of speech research in technological products for all of the above three application areas will be very dependent on models of speech production and perception, and it is these aspects that are likely to be of most interest to a phonetics conference. I will discuss such models in the relevant sections of this paper. Because speech production models will be important not only in speech synthesis but also in automatic speech recognition and in digital coding of speech, these models will be discussed at some length in the first section.

2. Automatic Speech Synthesis - short term

Up to now there has been a dichotomy in approach to message synthesis for machine voice output, between methods that concatenate signals representing coded human speech, and those that generate the speech signals by rule. It is also possible to have a hybrid system, using rules for the higher levels of the process, but still using coded forms of particular human utterances for the lower levels. At one extreme complete messages or phrases of coded human speech are used, and such systems can reproduce all the speech quality features of the talker who provided the original speech material. In these cases the speech quality is limited only by the properties of the coding scheme, and can be very much better than the best that has so far been obtained by any rule-based system. This method is adequate for a set of fixed messages, but for more general applications in which flexibility of message structure is required, means must be found for converting by rule from linguistic units to phonetic specifications, whether rule-produced or coded speech is used for the final stages of the signal generation process.

The success of an all-rule system depends on having the right sort of models for all the various stages of the speech generation process, but particularly for converting from a detailed phonetic specification to the speech waveform. It is then necessary to choose the parameters that govern the operation of the rules, not only to achieve the desired phonetic properties of the speech segments, but also to achieve the desired voice quality. Once a really good model has been found one has the basis for a completely flexible speech generation method that should be able to provide any type of speech quality required, without ever having to choose particular human talkers to provide the raw material for a new voice.

Choosing a computational model for generating the acoustic waveform from a specification of the operation of the vocal tract is not enough. The range of model parameters that govern speech generation from linguistically specified units depends on language and dialect, on the type of voice quality, on carefulness of articulation and subtleties of prosody. To be generally useful, therefore, such a model has to have a method of determining the parameters for any new requirement. Up to now the traditional method of choosing the parameters defining the operation of a speech synthesis-by-rule system has been by laborious human trial-and-error methods, guided by phonetic knowledge, spectrographic analysis of human speech, perceptual experiments, etc. The results more than 20 years after such methods were first started still leave much to be desired; and this is one of the chief reasons why coded human speech methods remain in such widespread use for machine voice output. What is obviously needed is a model for which it is practicable to optimize the parameters automatically by iterative adjustment. The aim would be to make the model reproduce human performance as best it could, when judged against large amounts of good quality natural speech.

For a good representation of the speech production process it might at first seem essential to use a model that is a close analogue of the human articulatory system. With such a model the co-articulation effects should arise naturally, and in principle it should be possible to deal correctly with glottal source properties, interaction between vocal tract and vocal folds, and the contribution of the sub-glottal system, nasal cavities etc. Some very good research has been done making a useful start in this direction.

However, there are some disadvantages with an articulatory mode. In human language acquisition the criterion of speech production success is inherently auditory, and the precise articulatory strategy that a human being will use will depend partly on the particular anatomy of the speaker's vocal tract, and partly on chance initial choices of trial strategies aimed at producing particular speech sounds. The great complexity of motion of the vocal folds, the interaction between this motion and the supra- and sub-glottal systems, and the mere complexity of shape of the vocal tract itself mean that it will be very difficult to make a really good articulatory model, particularly when a wide range of different voice qualities is required. An even stronger reason against using an articulatory model for machine voice output is that the relationship between articulatory gestures and the acoustic signal is very complex. This complexity would make it very difficult to generate automatically the details of articulatory control needed to produce a synthetic copy of a given sample of human speech. Articulatory models will continue to be of great importance for research purposes, to provide insights into how the various acoustic features of human speech arise, but I believe acoustic-domain models are much more likely to be successful in machine voice output applications.

The output from a simple all-pole terminal analogue model of speech using a cascade connection of formant resonators is theoretically equivalent to that of an ideal unbranched acoustic tube of appropriate dimensions, and yet its parameters are more directly related to measurable speech properties than are the parameters of articulatory models. Such an all-pole model cannot, however, be justified for nasal and obstruent sounds. Even for non-nasal vowels there can be very significant differences between the properties of real speech and the idealized assumptions that are used to justify this model of speech production. For this reason, using a cascade formant model does not give any advantage over a true articulatory model except in terms of implementation.

By contrast, acoustic-domain models using a small number of parallel formant resonators have the formant control signals for both frequency and amplitude very directly related to easily measured properties of human speech. Such a model clearly has no simple relationship to an articulatory specification of the vocal tract, nor has it a good theoretical justification as a representation of the human articulatory system. However, if implemented with sufficient attention to detail, it has already been demonstrated to be capable of producing output that is subjectively extremely close to human speech, when provided with control signals that copy the measured properties of human utterances (Holmes, 1973). The general configuration of formant resonators in the latest form of the Holmes synthesizer (Holmes, 1983) is shown in Fig. 1. Although such a model cannot properly represent the effects of varying glottal impedance and sub-glottal coupling, the subtleties of vocal fold motion, etc., it is possible to make a functional approximation to these effects by storing a typical glottal pulse shape and by letting the derived glottal flow waveform modify the formant parameters in addition to exciting the formant filter system.

The value of such a synthesizer as a possible future voice output device for speech synthesis by rule depends upon whether it is practicable to devise a successful control strategy relating a phonetic description to the formant

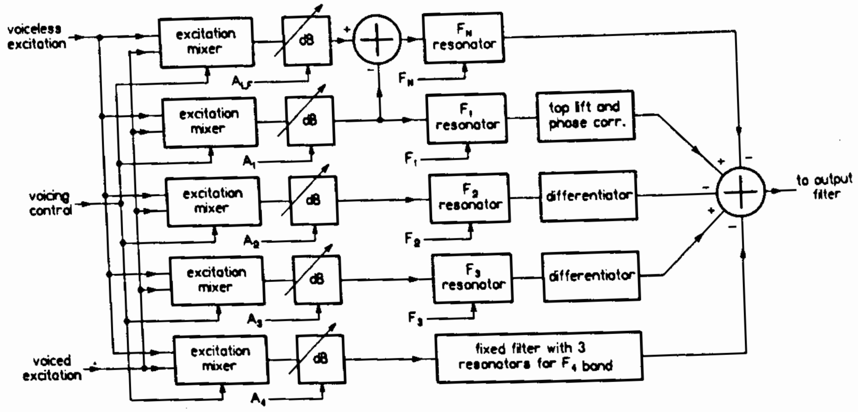


Figure 1. Arrangement of resonators in parallel-formant synthesizer. After Holmes, 1983.

description. Early attempts to produce rules relating a phonetic specification to formant parameters were described many years ago (Liberman et al., 1959; Holmes et al., 1964), and in general were adequate to enable subjects to make the right phonemic interpretations in the target language, but were otherwise very inadequate, both in terms of speech quality and in phonetic detail. The question is whether these poor results were an inherent property of a rule system lacking intrinsic articulatory constraints, or whether it was merely that the details of the rules were inadequate.

The rule system of Holmes et al. (1964) used sets of target values for the various formant parameters for each allophone in its inventory, and simple rules for calculating parameter values at nominal boundaries between the time segments associated with each allophone. Transition durations were specified between every set of boundary values and their adjacent targets. Linear interpolation between these values was used in all parameters. Thus the speech was completely specified by about 20 – 30 sets of parameter values every second and their times of occurrence.

Using a parallel-formant synthesizer and hand-derived formant control signals, it was shown by McLarnon et al. (1975) that very high quality speech was obtainable merely by specifying the formant parameters only at selected time instants, and using linear interpolation between them. Using an automatic algorithm for selecting the instants when the formant values were specified, they found that an average of 25 sets of formant values per second was adequate for normal talking rates.

The above facts are a strong indication that the very crude rule system of Holmes et al. (1964) might be elaborated to provide natural-sounding speech if a sufficient number of allophones were specified, and suitable numbers were used in the tables of formant target values and transition parameters.

This situation has led Bridle and Ralls (1983) to investigate an automatic method of adjusting the tables of the Holmes et al. rule system to make it copy particular utterances of human speech, as specified by automatically derived formant data. For reasons concerned with the mathematical process involved, a parabolic rather than linear interpolation was used, which undoubtedly makes formant tracks look more realistic, but has been found to be subjectively insignificant (J.A. Edward, private communication). A typical result is shown in Fig. 2. Fig. 2a shows a pseudo-spectrogram representation of the formant parameters of the words 'an apple a day', derived by automatic formant analysis. Resynthesis from these parameters produces very natural-sounding speech, immediately recognizable as the original talker. Fig. 2b shows the same passage generated by rule from a phonetic specification, after the rule tables had been automatically optimized for this

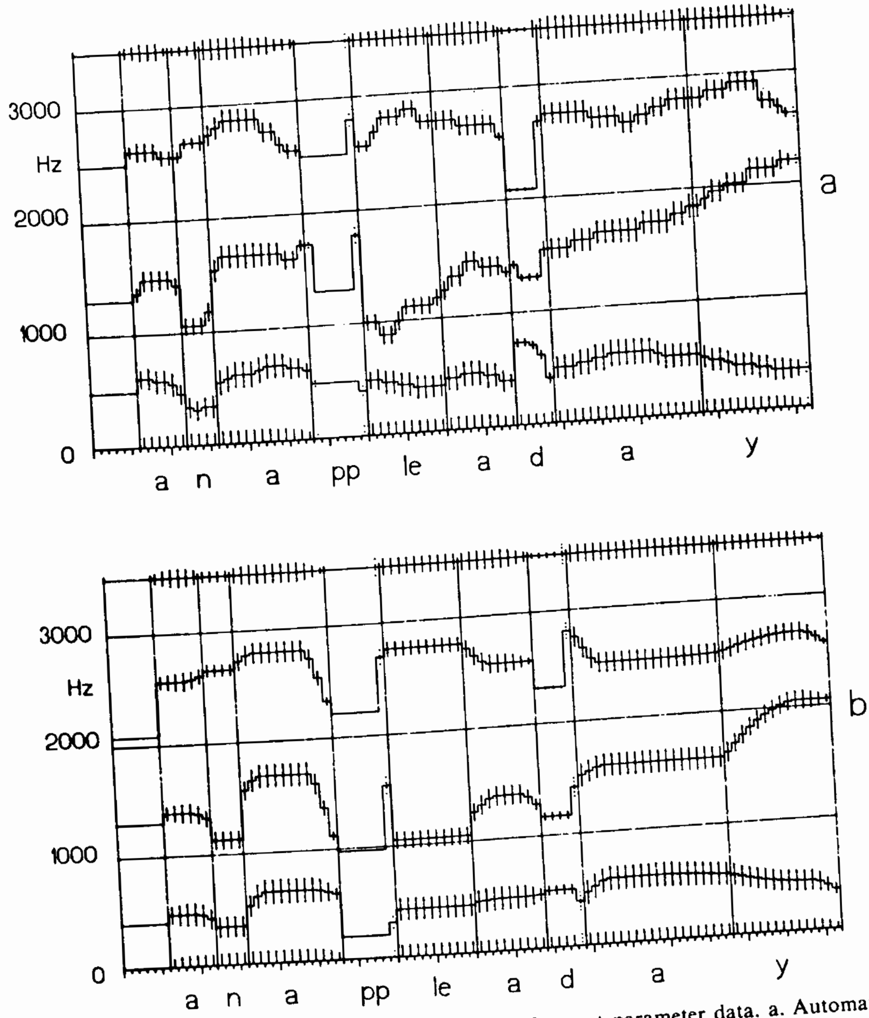


Figure 2. Pseudo-spectrographic reresentation of formant parameter data. a. Automatically derived from natural speech. b. Generated by rule using tables adjusted to suit utterance. After Bridle and Ralls, 1983.

utterance. The speech from these rule-generated parameters is similar in quality to that associated with Fig. 2b, and again the speaker characteristics are preserved.

These results appear to open the way to a formant-based speech production model that, from a low data rate phonetic description, will be capable of generating speech that copies the subjective quality of any normal speaker. Because this model would produce spectral peaks in the signal at the correct frequencies and amplitudes, it should preserve the perceptually important features, even though it will give no insight into how the speech sounds are produced. An aspect of voice output systems that could be important for many applications is the type of voice quality required. The type of automatic adjustment scheme described above makes it possible to match the synthesis rules to any particular talker. Alternatively, by using a large number of speakers of the same dialect for the table adjustment, it would be possible to specify an 'average' talker of that dialect. By analysing the relationship between the individual rule synthesis tables for a large number of talkers, however, it should also be possible to determine any systematic trends in how the entries for different phonemes are related for different talkers. Knowledge of these relationships will imply some ability to separate those features of the rules that relate to the phonemes and those that relate to the speaker. If this separation can be performed, it could also provide a powerful basis for automatic talker recognition. It should also be possible to make global modifications of the formant parameters to produce different voice qualities (man, woman or child, etc.), without having to modify the phonetic rules.

Although the optimization process described above has not yet given a completely satisfactory performance, first experiments have shown sufficient promise to make me believe that this approach will adequately solve the problem of generating the lower level phonetic features within a very few years. It should also be easy to modify the system for any new language for which the phonology is adequately specified.

The performance of existing prosodic rule systems and systems for converting from an orthographic to a phonemic representation are not yet as good as one would like, but currently this is not the main limitation to the overall performance of speech synthesis-by-rule systems. I see no reason why a similar automatic optimization technique should not be developed for these stages also, based on adjusting the properties of the model to match human performance for a large body of data. Lucassen (1983) has already shown promising results using such a technique for spelling-to-sound rules in American English.

Although the tables controlling this type of rule synthesis are very large, and the programs are fairly complicated, the implementation of the rules in real time is well within the capabilities of current single-chip microprocessors. With a programmable signal processing chip to implement a formant synthesizer, this means that a fairly low cost implementation of the most powerful rule systems that we know how to specify is even now not restricted by the technology. I therefore predict that as the rule development progresses, systems of this type will displace stored human speech methods of voice output for almost all applications within a few years.

3. Automatic Speech Synthesis - long term

The main problem in speech synthesis that will only be solved in the long term is to deal correctly with all conventionally spelled text input. Solutions for this problem will require machines to have linguistic knowledge comparable to that of the skilled human reader, particularly for choosing the correct prosody to suit the semantics of the message, and for choosing between alternative pronunciations of words. This capability has to wait for artificial intelligence research to progress much further than it has so far, and probably is more than 10 years away.

The problems of speech synthesis for general forms of man-machine dialogue are very different, because in this case one has to generate messages from abstract concepts in the machine. Again artificial intelligence will be involved, to formulate the messages in linguistic form, but the synthesis problems should be easier because the process of choosing the words would be intrinsically accompanied by knowledge of their pronunciation and the required prosody. I therefore expect that the problems of formulating the utterances will prevent completely natural language from being used for general dialogue until after the next decade, rather than the problems of speech synthesis itself. Some early ideas about speech synthesis from concept have already been published (Young and Fallside, 1979).

4. Automatic Speech Recognition - short term

Current automatic speech recognition systems take very little account of acoustic-phonetic knowledge and early attempts to make 'phoneme recognizers' were, of course, doomed to failure, because the identities of the phonemes of speech are not contained unambiguously in the local properties of the acoustic signal. It now seems to be fairly generally accepted that humans recognize larger units (words or syllables) before they can decide on the identities of the phonemes. It therefore follows that effective automatic speech recognition should recognize these larger units, and should make extensive use of linguistic knowledge. Even in present-day isolated word recognizers linguistic knowledge is used to some extent - knowledge of the permitted vocabulary and any word sequence constraints.

The inconvenience of speaking isolated words and the existence of computationally efficient algorithms for dealing with connected pattern sequences (Bridle et al., 1983) should make isolated-word recognizers obsolete within a very few years. Progress in improving current connected-word recognizers will occur in several areas:

(i) The acoustic analysis (already often simulating the frequency resolution of the ear) will be extended to highlight those types of acoustic features known to be phonetically significant (such as sudden increases of level, or formant transitions). The acoustic analysis will not, in itself, try to make any phonetic decisions, but will ensure that phonetically important features are given sufficient weight in any subsequent pattern-matching process. An example of this sort of process is shown in the work of Darwin (private communication, 1983), who has convolved Bark-scale spectrograms (Fig. 3b) with a series of masks, each designed to detect spectral peaks with a particular rate of formant transition. The corresponding crosses on Fig. 4a show an indication of the positions, intensities and rates of movement of spectral peaks, and highlight phonetically important features that are not so immediately apparent in the simple representation of a spectrogram. The further process displayed in Fig. 4b shows the effect of plotting the time-derivative of the amplitude indicated in Fig. 4a. This process gives prominence to features such as stop consonant bursts.

(ii) The distance calculation in the pattern-matching process will be more closely related to perceptual criteria. Already various workers have developed distance metrics related to perceived psycho-physical distance (Bladon and Lindblom, 1981), but for speech recognition it is phonetic distance that is important (Klatt, 1979). Improved distance metrics will receive input from the sort of processes described in (i) above, but will also include methods of reducing the importance of the

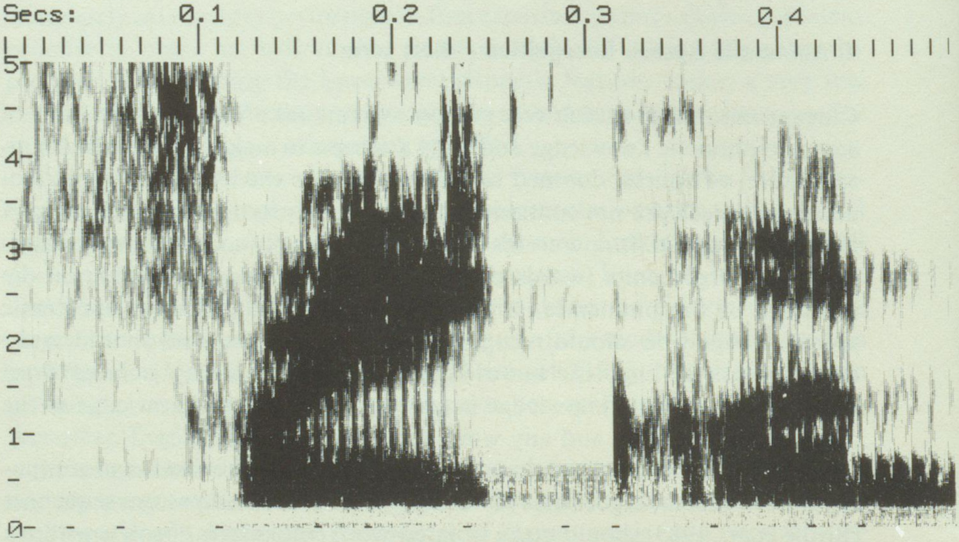


Figure 3a. Conventional display of the first part of the word "frequency" spoken by a male talker.

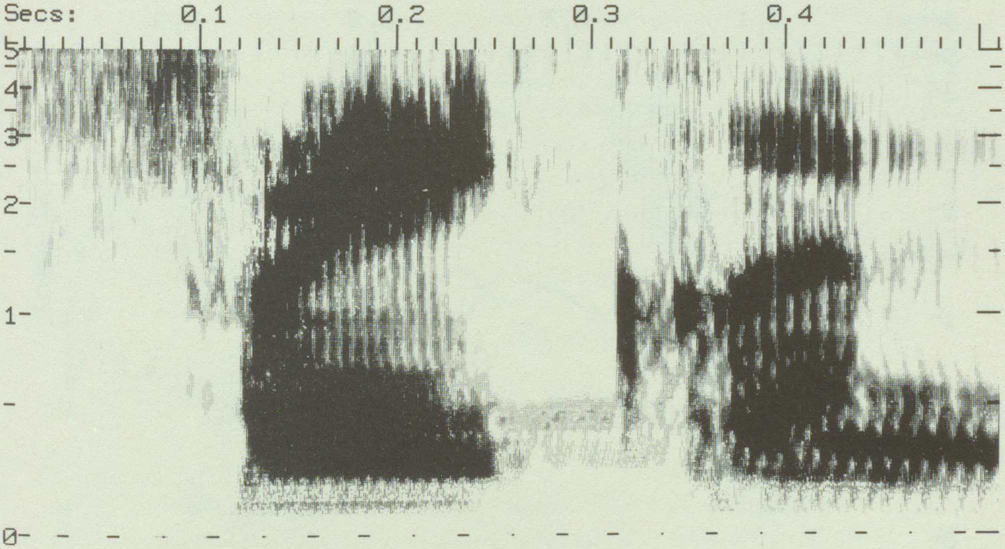


Figure 3b. Spectrogram of Fig. 3a modified to a Bark scale. After Darwin.

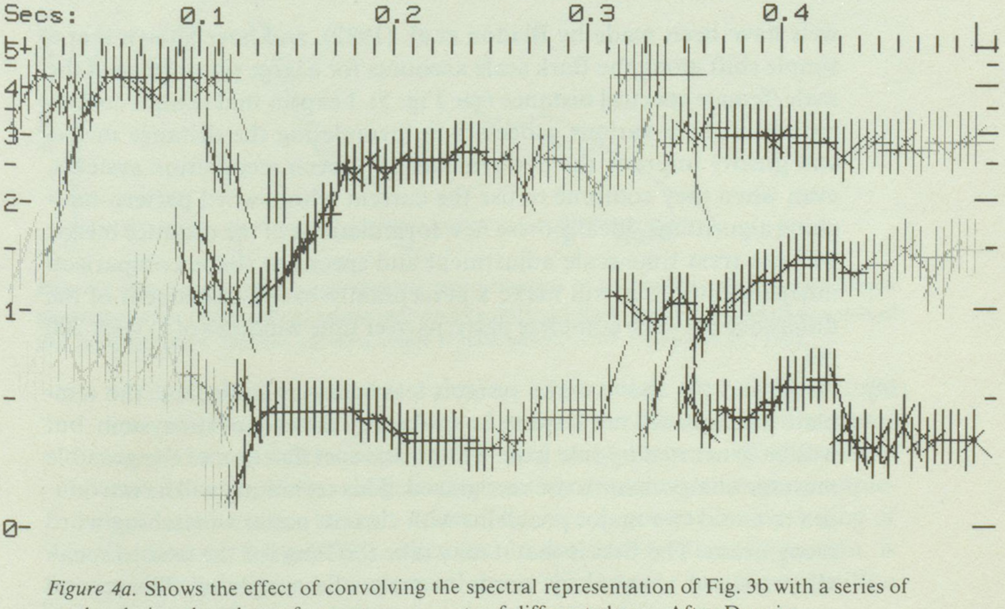


Figure 4a. Shows the effect of convolving the spectral representation of Fig. 3b with a series of masks, designed to detect formant movements of different slopes. After Darwin.

sort of formant intensity variations that arise from changes in glottal source spectrum, acoustic environment etc. In addition, they will need to include some normalization to accommodate the effects of anatomical differences between talkers, both within one sex and between sexes. Some studies of male/female spectrum differences for equivalent vo-

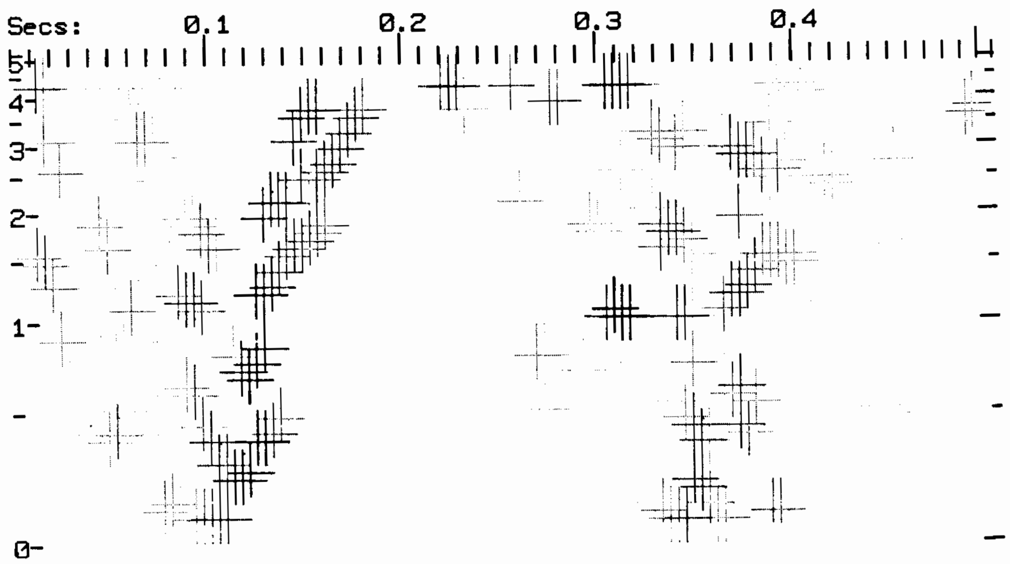


Figure 4b. Shows the effect of plotting the first difference of the signal derived using a Gaussian mask, to highlight the onsets of energy. After Darwin.

wels have been made by Bladon et al. (1982), and have shown that a simple shift along the Bark scale accounts for a large proportion of the male/female spectral distance (see Fig. 5). I expect that using a combination of these various processes in formulating the distance metric will greatly improve the performance of speech recognition systems, even when they continue to use the current whole-word pattern-matching algorithms. Ideally these new formulations of the distance metric will not treat time-scale adjustment and spectrum shape comparison independently, but will make a perceptually-based assessment of the difference between complete patterns over time windows of at least 200 ms.

(iii) As (i) and (ii) above make systems less speaker-dependent, the template patterns will not have to be spoken by the user of the system, but will be generated by rule from a linguistic specification of the possible message components to be recognized. This technique will be introduced to avoid two major problems with current pattern-matching word recognition. The first is that it may take too long for the user to speak all permitted vocabulary words to make the templates. The second problem is that recognition errors may arise because chance variations in production of phonemically identical sequences occurring in different words may be greater than the differences caused by the intended phonetic distinctions. Later, it will become more convenient to generate the desired templates dynamically by rule as they need to be used, rather than having them stored as acoustic patterns. When this is done the correct form of co-articulation at word junctures will arise automatically from the synthesis rules, so overcoming one of the present limitations of concatenated-word recognizers. Further performance improvement will be achieved by using early words that are known to have been correctly recognized to modify the rule synthesizer parameters to match the talker's voice. The speech generation modelling of Bridle and Ralls (1983), discussed in the speech synthesis section, is primarily intended for eventual use in this way for speech recognition applications.

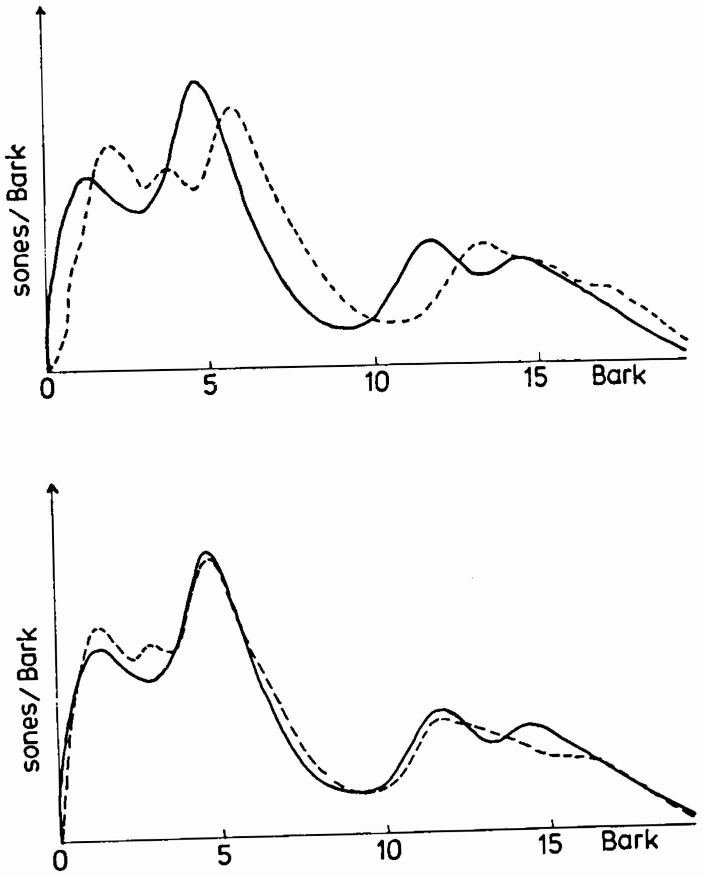


Figure 5. Psychophysically-based spectral representation of an English vowel, before (upper panel) and after (lower panel) normalization by a shift of 1 Bark. Solid curve: average vowel of five male speakers. Dashed curve: a single female vowel. After Bladon et al., 1982.

(iv) Powerful statistical techniques, based on principles described many years ago by Baum and Petrie (1966), have already been used in automatic speech recognition studies by a few research groups (Baker, 1975; Bahl et al., 1983; Levinson et al., 1983). Such techniques will become much more widely used, and will make a major contribution to identifying the underlying linguistic units from the surface structure of the speech signal, with manageable amounts of computation.

5. Automatic Speech Recognition - long term

Most of the improvements in automatic speech recognition outlined above should have been achieved, at least in the laboratory, within 10 years. The longer-term improvements will be in the ability to incorporate artificial intelligence and more advanced natural language models into systems. If current trends in reduction of computation costs continue there will be no great difficulty in providing sufficient computational power for these purposes, but it will not be a trivial task to devise suitable machine architectures to use this power effectively. Assuming these problems are overcome, these techniques will probably become cost-effective for many applications.

6. Digital Coding - short term

Digital coding for transmission and storage of speech signals divides into two classes, dependent on the application. In the first class the input has to be accepted from any member of the general public, perhaps in a noisy environment, and it is possibly transmitted to the coder via a poor quality local telephone line. In the second class the providers of the coder have some control over the users and their local equipment, such as in military systems, office systems or systems providing telecommunications between separated sites of a single organization. In the former case I do not see more than modest coding advances in the short term. These will include gradually changing the present 64 kbit/s PCM coders of commercial telephony to coders using about 32 kbit/s, by exploiting some of the more obvious signal redundancy. However, significantly lower digit rates than this will have to wait for major changes in telephone network organization: until then systems will have to cope with poor quality input, and may have to return the signal to analogue form and recode it several times along its route. These system difficulties will mean that the actual signal presented to the coder will not conform well to the sort of speech production model discussed earlier in this paper, and so the more powerful coding algorithms will not be generally usable.

The situation for restricted users is very different. It will often be possible to ensure that good quality speech signals are provided as input. User training or selection can prevent problems with difficult speakers. Under these circumstances analysis/synthesis methods, using a good acoustically-based model of speech production, vector quantization and variable-frame-rate transmission, will be able to yield very good speech quality at 600 - 800 bit/s. In contrast to the speech production models using phonetic rules, the algorithms for general speech coding must work for a wide variety of speakers and languages. However, such models will still be able to use the parallel formant acoustic model of the speech process, and simple linear interpolation between spectral patterns specified in the formant domain at irregular time intervals. To work really well, such systems will use analysis-by-synthesis to choose that sequence of patterns that minimizes an error score specifying the distance in perceptual terms between the input speech and the synthetic speech reproduced from the pattern sequence (B.C. Dupree, private communication). Work is currently in progress in my laboratory to produce a computer simulation of such a scheme. The process is illustrated in Fig. 6. This process will be very expensive in computation and especially in memory, but will be technologically practicable within 10 years. However, because of the cost, I expect it to be deployed only where are very great advantages in lowering the digit rate. A much cheaper alternative, which would be significantly more robust for poorer quality input, will be a medium-bit-rate system such as adaptive predictive coding, adaptive transform coding or sub-band coding, which should give good results at transmission rates in the 8 - 16 kbit/s range.

Although there is current research aimed at providing vocoder speech at 200-300 bit/s (Roucos et al., 1982), I see no short-term prospect of performance being adequate for speech at normal conversational speed and for a wide variety of speakers. To achieve such low rates would require reliable identification of phoneme-sized units, and the information needed is often not available in the signal without first determining the words spoken. When the input vocabulary can be sufficiently constrained it is already possible to achieve a sort of speech transmission at very low digit rates by connecting a speech recognizer to a speech synthesizer. The expected improvements in recognition and synthesis will soon make this possibility much more practical for special applications.

7. Digital Coding - long term

In the much longer term it will become possible to incorporate the linguistic

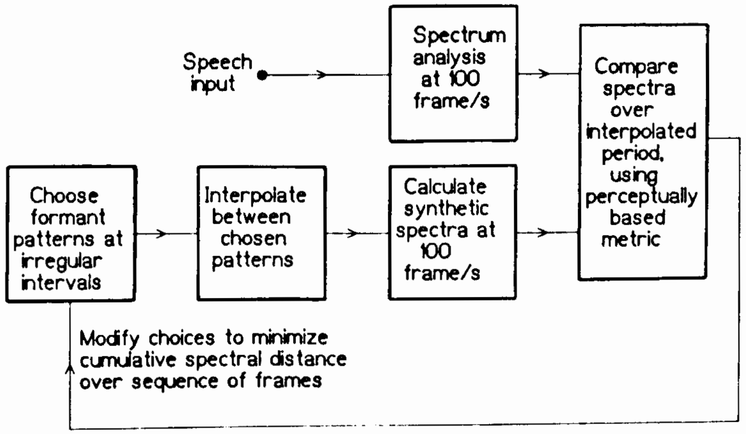

Speech input
Spectrum analysis at 100 frame/s
Compare spectra over interpolated period, using perceptually based metric
Choose formant patterns at irregular intervals
Interpolate between chosen patterns
Calculate synthetic spectra at 100 frame/s
Modify choices to minimize cumulative spectral distance over sequence of frames

Figure 6. Analysis-by-synthesis system for formant coding. After Dupree.

knowledge of an intelligent human being into a coding equipment. When that time comes such a coder will be able to understand the messages, and therefore to code them as words or phonemes for subsequent synthesis. It will also be possible to deduce the subtleties of prosody and the characteristics of the speaker. All this information should not require more than about 200 bit/s, to achieve completely natural-sounding speech. This performance should be achievable for any input of adequate quality for a human to understand reliably. It will then be possible in principle to produce good quality at the receiver from poor quality input. The problem with this approach (which applies equally if a human being instead of a machine is asked to perform the relay function) is that a poor quality input may actually be misunderstood, and so be transmitted incorrectly. The listener, hearing excellent quality output, will be unaware of the errors.

Such a coder is not likely to be cost-effective or even desirable for civil telephony transmission. Telecommunications circuits will, in any case, be available with a digital capacity of many tens of kilobits per second at any location where there is a normal telephone. End-to-end digital transmission of the waveform would guarantee correct reproduction of the microphone signal at the receiving telephone. It is thus likely to be more acceptable to users to transmit the speech signal as produced, with whatever background noise is present. I do not expect it will ever be possible to lower the data rate for perceptually-transparent transmission of such signals to below about 8 kbit/s, but for most civil telephony applications this would be acceptable. The big advantage of much more complicated very-low-bit-rate coders in civil telephony will be for store-and-forward applications, and possibly also for very long distance transmission.

8. Conclusion

The future developments predicted in this paper should make the performance of man-machine communication by speech approach that of communication between people. It has been common in the past for considerable importance to be attached to the 'naturalness' of speech as a method of communicating, and so for people to assume that speech is necessarily better than other forms of communication. There are many cases where this assumption is undoubtedly justified, and many other cases where speech is the only medium available, such as when an ordinary telephone is involved, or for people with visual or motor disabilities. However, even for human-to-human communication, it is often better to employ other means, such as when using a map to show land features, or using a graph to illustrate the form of a mathematical function. When a machine is involved there are even more cases where speech is unsuitable (e.g. for controlling the steering of a car).

I therefore think it is very important that, in parallel with the research on speech technology of the next decades, there should also be careful study of

the human factors aspects of using the new speech devices. These devices would then be able to be used as soon as possible for those tasks for which they are suited, and they would not acquire a bad reputation merely as a result of people unjustly expecting them to solve all their communication problems.

References

Bahl, L.R., Jelinek, F. and Mercer, R.L. (1983). A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Anal. and Mach. Intelligence,* PAMI-**5**, 179-190.

Baker, J.K. (1975). The DRAGON system - an overview. *IEEE Trans. Acoust. Speech and Signal Process.* ASSP-**23**, 24-29.

Baum, L.E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov processes. *Ann. Math. Stat.* **37**, 1559-1563.

Bladon, R.A.W., Henton, C.G. and Pickering, J.B. (1982). Towards an auditory basis for speaker normalization. Institute of Acoustics Speech Group Meeting, Keele.

Bladon, R.A.W. and Lindblom, B. (1981). Modelling the judgment of vowel quality differences. *J. Acoust. Soc. Am.* **69**, 1414-1422.

Bridle, J.S., Brown, M.D. and Chamberlain, R.M. (1983). Continuous connected word recognition using whole word templates. *Radio and Electron. Eng.* **53**, 167-175.

Bridle, J.S. and Ralls, M.P. (1983). An approach to speech recognition using synthesis by rule. In: F. Fallside and W.A. Woods (eds.) *Computer Speech Processing*, to be published.

Holmes, J.N. (1973). The influence of glottal waveform on the naturalness of speech from a parallel-formant synthesizer. *IEEE Trans.* AU-**21**, 298-305.

Holmes, J.N. (1983). A parallel-formant synthesizer for machine voice output. In: F. Fallside and W.A. Woods (eds.) *Computer Speech Processing*, to be published.

Holmes, J.N., Mattingly, I.G. and Shearme, J.N. (1964). Speech synthesis by rule. *Lang. and Speech* **7**, 27-143.

Klatt, D.H. (1979). Perceptual comparisons among a set of vowels similar to /ae/; some differences between psychophysical distance and phonetic distance. *J. Acoust. Soc. Am.* **66**, S86.

Levinson, S.E., Rabiner, L.R. and Sondhi, M.M. (1983). An introduction to the application of theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell Syst. Tech. J.* **62**, 1035-1074.

Lucassen, J.M. (1983). Discovering phoneme base forms automatically: an information theoretic approach. M.S. Dissertation, M.I.T., Cambridge, Mass.

McLarnon, E., Holmes, J.N. and Judd, M.W. (1975). Experiments with a variable-frame-rate coding scheme applied to formant synthesizer control signals. In: G. Fant (ed.) *Speech Communication.* Stockholm, Almqvist and Wiksell, 71-79.

Roucos, S., Makhoul, J. and Schwartz R. (1982). A variable-order Markov chain for coding of speech spectra. *IEEE Int. Conf. Acoust. Speech and Signal Process.*, Paris, 582-85.

Young, S.J. and Fallside, F. (1979). Speech synthesis from concept, a method for speech output from information systems. *J. Acoust. Soc. Am.* **66**, 685-695.

Semi-Plenary Sessions:

Symposia

SYMPOSIUM 1: Semantics, Syntax and Prosody

Chairman: *Ilse Lehiste, Columbus, U.S.A.*

Panel members: *E. Gårding, P. Martin, A. Cutler, V. Fromkin*

Discussant: *H. Fujisaki*

The symposium started with a brief introduction by the chairman, who observed that the relationship between prosody and syntax has been explored quite intensively in the past few years, but that the relationship between prosody and semantics has not been in the focus of research to a comparable degree. Contributors to the symposium have tried to address themselves to both questions. Since the topic - the relationship between semantics, syntax, and prosody - is very broad, some limitations have to be imposed on its treatment. We will not attempt to define precisely the relationship between semantics and syntax. Neither will we try to draw internal distinctions within the broad field of semantics. Prosody will be discussed at sentence level. Obviously, prosodic features are used in language also to convey both lexical and grammatical meaning at the word level: a tonal pattern, a quantity pattern, or difference in stress can distinguish between two lexical items or indicate, for example, a difference between two case forms of the same word. These functions of prosodic features will be taken for granted. Context more extensive than a single sentence will enter into the discussion by implication in connection with pragmatic considerations, i.e. when prosody is used to relate the sentence to a particular situation.

Sentence-level prosody comprises aspects of timing, accent, and intonation. Two of the four contributions (those by Gårding and Martin) address themselves primarily to intonation, and two (those by Cutler and Fromkin) focus on accent. Since timing was not discussed to any great extent by any member of the panel, the chairman started the symposium by saying a few words about the sentence-level function of the time dimension of spoken language.

In a series of papers summarized in Lehiste's contribution to the 13th International Congress of Linguists (held in Tokyo in the autumn of 1982) she has explored the relationship between timing and syntax, and has come to the conclusion that timing is the primary cue to the syntactic structure of a sentence. Boundaries of syntactic constituents can be signalled by pauses; they are also signalled by preboundary lengthening. At least in English pauses are not obligatory. It has been Lehiste's claim that lengthening does not achieve this effect (of signalling the presence of a boundary) by itself, but indirectly through its effect on rhythm: both preboundary lengthening and insertion of a pause disrupt the expected rhythmic structure of a sentence. In a more recent paper (about to be published in *Folia Linguistica*) Lehiste has shown that the presence of fundamental frequency is not necessary for signalling syntactic boundaries - speakers and listeners are equally successful in disambiguating whispered sentences as they are in producing and perceiving normally phonated speech.

The existence of a link between syntax and prosody seems to be generally accepted by now. Two issues are currently being debated. The first is the question whether prosodic structure is entirely predictable on the basis of syntactic structure, or whether there exists an independent prosodic structure that may or may not correspond to the syntactic structure of a sentence. The second issue is the question whether the syntactic structure of a sentence is reflected primarily in timing or whether it is primarily reflected in some other prosodic feature, such as accent or intonation. Lehiste has come out both in favor of an independent prosodic structure, namely sentence-level rhythm described in terms of metric feet, and in favor of timing as the primary means for signalling syntactic structure - by controlled disruption of the rhythmic structure of the spoken utterance.

What does not seem to have been given any serious consideration is the relationship between speech timing and semantics. It appears that duration has only an indirect role here: if it can be shown that *accent* is directly connected to semantics, then duration enters in as one aspect of the phonetic manifestation of accent. If it can be shown that *intonation* is directly connected with semantics, duration could conceivably play a role in helping establish the domains over which particular intonational patterns are manifested: intonation units must have a duration. At the present time, there seems to be no immediate connection between semantics and sentence timing.

Gårding in her contribution developed a theory of intonation. Using examples from Swedish, French and Chinese, she showed how intonation can serve to express modality and syntactic and information structure.

An intonation curve has local maxima and minima, turning points. These turning points are parts of a larger pattern, the grid, which is obtained by connecting the main maxima by a topline and the main minima by a baseline. A grid may be rising, falling, or level. Such grids may in turn be part of a global pattern which may be rising, falling or level. A grid may have a normal, expanded or compressed width even to the extent of being best represented by one line only. The part of the grid where the direction or width are changed, or where the grid takes a jump, is called a pivot. An intonation unit is a piece of an utterance with an unbroken grid.

These concepts - the local turning points, pivots, direction and width of the grid - are associated with the syntax and semantics of the utterance in the following way. The local turning points signal words and morphemes, that is semantic and syntactic entities. The pivots signal semantic constituents, like the theme and the rheme, or syntactic constituents like the subject and the predicate. The general direction of the grid over the utterance, often in combination with the direction of the last intonation unit, determine the speech act type, which may be declarative, interrogative and imperative, to use classical terms. In all cases, this is also a syntactic sentence type but not necessarily of the corresponding class. An interrogative speech act type may very well correspond to a declarative sentence type.

The width and position of the grid of an intonation unit signal its information weight relative to other intonation units. In other words, these intonation units may be in focus and out of focus. In this way the width and position may express coordination and subordination of semantic constituents and of syntactic ones.

Again, the classes are not isomorphic. We may very well focus a subordinate clause by prosodic means, and we may focus any clause or part of a clause by syntactic means, e.g. word order, without assistance from prosody.

Gårding summed up her presentation by asserting her belief that the grid, pivot and intonation unit are useful tools in intonation analysis. They are strongly connected with syntax and semantics in a similar way in different languages. Whenever there is a conflict between syntactic and semantic categories, it is the semantic function that has the greater impact on intonation.

Martin defined intonation as the sequence of prosodic contours located on the stressed vowels of an utterance. Intonation is perceived to be linked in certain ways to both the syntactic structure of a sentence and its semantic content. Using examples from French, Martin demonstrated that intonation can be constrained by syntax and semantics; nevertheless all three are considered to be independent systems, functioning according to their own rules. Martin developed a theory of intonation, according to which utterances have both a syntactic structure and a prosodic structure. Both structures are hierarchical, and they are not necessarily isomorphic, although some connection must exist between the two. Sentences are divided into *rheme* and one or more *themes*; prosodic division into rheme and theme is indicated by a specific prosodic contour located on the last stressed syllable of the rheme. The prosodic division into rheme and theme can be either prevented or forced by a semantic or syntactic constraint. According to Martin, pauses are direct manifestation of syntax and should not be considered as part of an independent prosodic system.

The papers by Cutler and Fromkin were primarily concerned with the relationship between sentence accent and syntax and semantics: whether the position of accent is determined by syntax, or whether the speaker can assign it more or less freely to express a particular communicative intent. Both papers used slips of the tongue as relevant research data; it was interesting to note that the same kinds of data served to support rather different conclusions.

Cutler finds that in producing accent patterns, language users behave as if sentence accent placement were concerned with the semantic and pragmatic structure of utterances rather than with their syntax. Speakers place accents to reflect the information structure of the message they wish to convey;

listeners use accentual structure to locate points of information focus. Cutler adduced evidence from children's acquisition of the production and comprehension of accent to strengthen her claim that accentual structure is associated with the meaning of a message rather than its form.

Fromkin argued that the semantic function of accents does not exclude a dependence on syntax, and that no new evidence exists to counter the claim that phrasal stress (which can coincide with accent) is determined by syntactic structure. According to Fromkin, normal, non-contrastive intonation, too, is determined syntactically. Primary stress or accent must be assigned after the syntax is determined; accent placement is independent of the particular lexical items on which it falls even if the semantic focus is thereby confused or changed. Nevertheless, Fromkin concluded that assignment of prosodic structure depends both on the syntactic structure and information structure of utterances.

Commenting on Gårding's paper, *Fujisaki* agreed with her concerning the need for a generative model in order to be able to interpret correctly a given F_0 contour. This is so because an F_0 contour is generated as a consequence of control of the vocal fold vibration by a set of commands that are directly related to the linguistic structure of an utterance. As Fujisaki has shown elsewhere, the underlying linguistic structure becomes more explicit if we *deconvolve* the mechanical/physiological characteristics of the laryngeal control mechanism from an observed F_0 contour. In this sense, the use of a piecewise-linear tonal grid is a rather crude and unrealistic approximation of the observed characteristics of F_0 contours. In Fujisaki's opinion, an objective and quantitative way to extract the underlying linguistic information is possible not by stylization but by deconvolution, using analytically expressible transfer functions (describing the quantitative relationships between the linguistic information as input and an F_0 contour as output) with physically and physiologically meaningful formulations.

Fujisaki then presented his own model for the mechanism of generating an F_0 contour from a set of linguistically meaningful commands. According to this model (based originally on an idea by Öhman, but elaborated by Fujisaki), an F_0 contour – expressed in terms of log F_0 as a function of time – can be decomposed into two types of components: 1) phrase components corresponding to prosodic phrases, and 2) accent components corresponding to prosodic words. Each of these components can be considered as a consequence of control of the laryngeal mechanism by a separate linguistic command: 1) the phrase command, which is a sharp pulse (an impulsive force) applied to the laryngeal structure and which generates a phrase component, and 2) the accent command, which is a stepwise force applied to another part of the laryngeal structure, which generates an accent component. Responses of the laryngeal mechanism to these two kinds of commands are different, but each of them can be approximated very closely by the transfer function of a second-order linear system. By expressing these transfer functions in a quantitative way, one can decompose a given F_0 contour into its constituent components (i.e. phrase components and accent components) and deconvolve the inputs and the respective transfer characteristics.

Phrase components are the causes of declination and constitute what some people call the baseline - roughly corresponding to the lowest line of Gårding's tonal grid. Both a complete resetting (at a respiratory pause) and a mere superposition of a new phrase upon the old one are possible, and this corresponds roughly to Gårding's pivot. Fujisaki's research has shown that his model can approximate quite closely F_0 contours of spoken words and sentences of Japanese, English, and Estonian. The analysis of Japanese sentences has shown that speakers of Japanese generally use three levels of phrase components and two levels of accent components; these numbers, however, may differ from language to language.

Commenting on Martin's paper, Fujisaki expressed his readiness to agree with Martin that the prosodic structure and the syntactic structure are two different systems. He also liked the concept of 'prosodic word', which is appropriate for describing the phenomenon of 'accent sandhi' in Japanese, where two or more lexical items, each of them with its individual word accent types, are often *concatenated* (not *compounded*) and behave as if they were one long word with its own word accent type.

In the case of Japanese, however, accent sandhi is not necessarily all or none. Fujisaki has encountered cases where the coupling between two words is somewhat intermediate and where it is difficult to decide whether they should be regarded as one prosodic word or two prosodic words. He suspects that this may also be the case in spoken French. He asked Martin to define a prosodic word, to describe its acoustic-phonetic manifestation, and to say to what an extent prosodic words are stable or variable.

Commenting on the papers of Cutler and Fromkin, Fujisaki expressed his concern about the discrepancy of conclusions drawn by Cutler and Fromkin from the same kind of speech error data. Although he has no doubts about the validity of the research technique - of collecting and interpreting speech error data, - it nevertheless seems that the patterns of speech errors are of such a great variety that it is possible to draw two entirely different conclusions. He would interpret the apparent discrepancy between Cutler's conclusions and Fromkin's conclusion in the following way: the pattern of errors in stress placement could sometimes depend mainly on the syntactic structure, but sometimes mainly on the semantic and pragmatic structure of the utterance.

Generalizing from all four papers, Fujisaki observed that there exist at least two types of languages which differ in the ways of expressing information concerning the focus. Judging from Gårding's paper, focal information in Swedish seems to be signalled mainly by pivots (or by the presence of a new phrase command in Fujisaki's model), while in English it seems to be signalled mainly by accent placement (or by an increase of the accent command on a particular word) according to Cutler and Fromkin. Fujisaki added that in the case of Japanese, phrase commands are mainly determined by syntax and there is very little chance for focal information to be signalled by intonation.

While agreeing with the idea that the prosody of a sentence has its own structure, Fujisaki pointed out the differences in the constraints imposed on these structures. Syntactic and semantic structures of a sentence are constrained by the limited capacity of the short-term memory and the limited depth of, e.g., syntactic embedding; the prosodic structure; however, is more severely constrained by the limited characteristics of peripheral mechanisms, both in production and in perception. For instance, one cannot utter a very long sentence without pause, even though the syntax may allow it. One cannot have more than three or four different sizes of phrase commands to signal the presence of syntactic or semantic units of different size or different complexity, and so on. In other words, the prosodic structure is more constrained as compared with the syntactic structure or the semantic structure. Therefore one often finds a degenerate (incomplete) representation of the syntactic or the semantic structure. Whether one tries to preserve (a part of) the syntactic information or to preserve (a part of) the semantic information in the prosodic structure will depend on the individual, on the situation, and on the language.

Responding to Fujisaki's comments, Gårding explained the notion of grid in a more detailed way. She pointed out that her model and that of Fujisaki are similar; both are based on Öhman's model. Grid is a concept that is useful for describing focus over part of the utterance. Responding to a question by Klaus Kohler about the degree of abstractness of her model, Gårding stated that the model is quite concrete: it is used both for analysis and for the generation of a concrete pitch contour.

Martin answered Fujisaki's question about prosodic words by reference to the relative amplitude of melodic variation. There is one lexical stress per prosodic word; in French, the prosodic word comprises the accented syllable and preceding unaccented syllables.

Cutler and Fromkin responded to Fujisaki by emphasizing the validity of the technique involving collection and interpretation of speech errors.

The discussion now became more general. The chairman had requested that those contributors to the discussion who wanted their comments included in the report provide a written version before the end of the congress; not all speakers complied with this request, and their comments can therefore be included only in a general way.

Tore Janson commented about a statement made by Fujisaki, stressing that even if prosody is implemented at a low level in articulation, it reflects complicated syntactic and semantic/pragmatic facts. The basic problem in speech communication is that the speaker has to convey information to the listeners about complex hierarchical structures over a channel that permits only linear order. The role of prosody is largely to hint the nature of the hierarchical structure. This can be done only partially and imperfectly, due to the relatively poor expressive capacity of the prosodic signals. Therefore, the listeners will have to do much guesswork when evaluating this information. For that reason, it seems certain that prosodic processing is going on even at the highest levels at the listeners' end.

C.W. Temu criticized the members of the panel for the narrow choice of examples from European languages. He asked also about differences in the way prosody works in languages in which there are changes possible in word order, versus those languages in which there is no change in word order. Fromkin responded to this comment with examples from Twi; Gårding's Chinese examples were likewise relevant in that context.

Mario Rossi, in his extensive comment, concentrated on the functions of intonation: expressive, demarcative and hierarchical. The units of expressive function are specific contours; the units of demarcative function are continuative and terminal intonemes; the units of hierarchical function may be units of demarcative function and/or semantic accent. The demarcative function is constrained by syntax. The hierarchical function is not always congruent with syntax, because it is mainly constrained by semantic organization of focus/presupposition and rheme/theme. To the extent that the theme/rheme structure corresponds to subject/predicate structure, intonation is congruent with syntax; when rheme indicates information weight (i.e. focus), intonation is used to scale the semantic content (as demonstrated many years ago by Mathesius and Karčevsky), and intonation is not necessarily congruent with syntax.

Several speakers suggested that if there is a conflict between semantics and prosody, semantics has the upper hand. Lehiste responded to that by arguing that prosody has to be independent of semantics at least to a certain degree, since it can be used to turn the meaning of a sentence into its opposite in the case of conscious expression of irony. It was suggested that prosody, syntax and semantics should be considered three 'channels of communication', all three of which are independent in a certain way.

It remains to assess the state of the art, as it is reflected in the four contributions to the symposium, the comments of the discussant, and the contributions made from the floor. It appears to be generally accepted that a relationship exists between prosody, syntax and semantics. The disagreements pertain to the relative independence of these three aspects of language. In their written contributions, Gårding had viewed intonation as basically independent of syntax, and Martin as independent of both syntax and semantics. Gårding, however, associated pivots (i.e. changes in the direction of fundamental frequency movements) with syntactic boundaries as well as with the information structure of the utterance. Cutler saw accents as determined primarily by the semantic import of an utterance, while Fromkin argued for a more important role for syntax in determining the location of the accent. Lehiste claimed that the rhythmic structure of a sentence is independent of its syntactic structure, but can be modified to reflect differences in the placement of syntactic boundaries. Fujisaki viewed prosody as a rather severely constrained way to represent syntactic and semantic structure. As a result of the discussion, panel members seemed willing to concede that alternative interpretations are possible; it appears rather remarkable that no strong claims for universality were made, and the opinion seemed to prevail that languages may differ in the ways in which they relate prosody, syntax and semantics.

References

Lehiste, Ilse (1982). The role of prosody in the internal structuring of a sentence. *Preprints of the Plenary Session Papers*, The XIIIth International Congress of Linguists, Tokyo, 1982. Pp. 189-198.

Lehiste, Ilse (1983). Signalling of syntactic structure in whispered speech. To appear in: *Folia Linguistica*, Vol. XVII.

SYMPOSIUM 2: Units in Speech Synthesis

Chairman: *J. Allen, Cambridge, Mass., U.S.A.*

Panel members: *O. Fujimura, J. Holmes, R. Carlson, B. Granström, J. Pierrehumbert*

The goal of synthetic speech algorithms is to provide a means to produce a large (infinite) set of speech waveforms. The major questions posed to a designer of such a system include the question as to what input is to be used, by what set of transformations is the speech to be produced, and by what quality measures can success be judged. At a general level, the issues are the choice of *representations* and the provision of *transformations* between these representations. The complexity of the relation between input (e.g., text) and output (the speech waveform) forces the introduction of intermediate levels of representation. These levels of representation are arrayed in two hierarchies. First, there is the structural hierarchy, wherein bigger constructs are made from smaller ones. There is a feeling that over the last 20 years during which speech synthesis by rule has been studied, the level of rule complexity at any particular level is staying about the same. This means that in order to introduce additional constraints, greater modularity must be introduced into the overall speech synthesis system together with narrowly constrained interaction between these domains. By such a means the structural hierarchy becomes richer, but the level of complexity at any level of the hierarchy does not grow to an unreasonable level. The second kind of hierarchy is the qualitative structure that exhibits the differing nature of constraint domains that are needed to specify the output speech waveform. These constraint domains include phonetics, phonology, syntax, semantics, acoustics, anatomy, physiology, and computation.

In the following paragraphs, the desiderata for the choice of units for speech synthesis are discussed.

1. The large number of utterances that must be created by a speech synthesis system forces *composition* from a smaller number of basic units.
2. The larger the unit, the more items of each unit there are. Thus at the small end, there are relatively few phonemes, but as the units become larger through diphone, demisyllable, syllable, word, phrase; . . . the number of items of the unit grows without limit. Furthermore, the smaller the unit, the more abstract it is likely to be. There is a general tendency towards increased abstraction as additional knowledge about speech becomes codified.
3. The choice in representation of units must allow sufficient degrees of freedom to control all significant aspects of the waveform. Questions as to

what is significant must be answered through perceptual tests and there is a great need for increased investigations in this area.

4. It must be possible to *analyze* speech in terms of the selected units and their parameters. Clearly, as the units become more abstract, there is increasing difficulty in relating surface measurements to the abstract structural relationships. Nevertheless, there is an increasing trend to introduce these structures in order to be able to generate the rich variety of surface phonetic detail. This phenomenon can be readily observed in the area of suprasegmental control, where the complexity of specifying a wide variety of fundamental frequency contours requires both a rich structural framework and a substantial set of complementary procedures. Another reason for the introduction of abstract structural models is the need for automatic training. Large databases of speech waveforms must be analyzed in terms of these models in order to accurately characterize the needed control parameters. While substantial abstraction can be introduced in order to characterize generalizations in the underlying framework, these abstractions cannot become too distant from the surface phonetic reality if automatic training is to be possible. Furthermore, substantial improvements in speech synthesis quality are probably impossible without the introduction of semi-automatic techniques for improving the quality of the control parameters.
5. It must be possible to *compose* an utterance by interpretive processes on the chosen units. Here a tension between internal vs. external structure and cohesion can be observed. Thus, the phoneme has relatively little internal structure, but considerable external structure is needed to specify the transitions between phonemes at their boundaries. Diphones and syllables, on the other hand, have a richer internal structure, and require less specification at their boundaries. Advocates of larger units, such as the diphone and the demisyllable, maintain that 'hard' (or physiologically determined) coarticulation is captured internally within the unit, thus leading to simpler composition functions. On the other hand, there is a growing feeling that the units should be sufficiently rich internally to allow for the automatic centralization of vowels and durational adjustments without the necessity for these changes to be imposed entirely externally. The search for the 'natural joints' of language will certainly continue in an effort to determine the optimal units, but high quality synthetic speech probably requires the same amount of detailed knowledge of speech and language whether the knowledge is represented in terms of either internal unit complexity, or complexity in the external composition function.
6. Consideration of the *composition function* indicates that the chosen units can be either used in *compiled*, static form or *interpretively*. Thus the choice of units is related to the question as to whether perceptually significant knowledge of speech should be represented structurally in static form or procedurally in terms of interpretive rules. This contrast does not have to be viewed as a necessary tradeoff, since it is certainly possible to have increased complexity in terms of the structure of the units together with increasingly rich procedural interpretive processes. In fact, the choice of increasingly complex abstractions demands a correspondingly complex interpretive process to realize the complete variety of surface phonetic waveforms.

7. The *binding time* of decisions must be specified. Some units imply early binding time. Thus compiled structural lexical forms, such as diphones, imply less complicated procedures at composition time. Such a choice may imply substantial effort to obtain the original compiled structural forms, but less computation when the speech waveform is created. Alternatively, late binding implies greater flexibility to modify aspects of the waveform. The units of speech synthesis tend to be more abstract in this case, and there is more emphasis on procedures to compose the units. In terms of general tendencies among computational procedures in the large, there is a tendency towards late binding. Modern computational resources support this flexibility, but it demands comprehensive modeling and training procedures.
8. The units should have a *natural relation* to the waveform production model. There is still no agreement as to the choice of model for constraining the output speech waveform. Most schemes employ a source/filter model, but there are differences as to whether cascade, parallel, or other combinations of resonators are used. Furthermore, the source model varies substantially, and it is clear that a great deal of work is needed to improve this model. Articulatory models are used in some instances, but the relative lack of articulatory data has slowed progress in this area. Whatever model is selected, it is important that it be possible to readily compute the values of the control parameters from this model through relatively straightforward computations based on the speech waveform.
9. The units should provide *insightful* relevance to the research literature. For this reason many synthesis schemes rely on the specification of formant frequencies and bandwidths. When linear predictive coding is used, it is often necessary to transform back to the formant space in order to edit parametric representations for improved speech quality. It is possible that new more complex units and representational frameworks may provide good synthesis, new insight, and also stimulate new research. The tendency over time is likely to involve the development of increasingly complex unit frameworks which in turn imply new classes of perceptual experiments and training procedures which should both improve speech quality and provide greater understanding of underlying speech processes.
10. There are several important *levels* of units that are of varying *scope*. In general, all aspects of linguistic structure are reflected in the acoustic waveform. Thus there is no one unit which is more important than all others. The levels of structure currently recognized are: discourse, sen-

tence, clause, phrase, word, morpheme, metrical foot, syllable, phoneme, and feature. These have all been found useful in linguistic analysis for reasons of distribution and contrast, but they each exhibit some intrinsic cohesion and place focus on the units themselves. Thus each of these units implies an internal theory, together with a relation to other units and their corresponding theory. There is a tendency to think of these different levels of units with their attendant scopes as being arrayed over a temporal axis. While such a view provides a convenient geometric intuition, contemporary notions of abstraction may not rely on notions of temporal scope at all levels. Thus at the individual segment feature level the initial specification may include an unordered set of features which is then transformed at a higher level to derive temporal extent. Research in the asynchronous control of these features reveals the need for this varying scope at the feature level, as in the control of nasality.

11. The hierarchy of units should provide for a *gradual* progression through the various representations. That is, a sufficient number of units should be available so that the transformation from one level to another through the hierarchy is relatively small and direct. Thus for example, phonemes are transformed to allophones and then to targets, parameters and finally to the speech waveform. In this way, understanding of the speech process is modularized at each level, thus simplifying the rule structure at each level, even though the total number of rules over the entire system may be large.

12. The notion of *target* serves as a bridge between abstraction and physical properties. There is a tendency for the targets themselves to become complex objects exhibiting composition, scope, and internal cohesion. For example, fundamental frequency contours must be characterized at both the phrase level and in terms of 'segmental' effects. It is seen that the term 'target', originally motivated by the characterization of formant trajectories, is increasingly synonymous with 'abstract representational structure'. The role of *variability* in the choice of units must be considered in the design of any speech synthesis system. It is often felt that many attributes of these units are necessary, and hence admit of no variability. Recent understanding, however, indicates that many cues are redundant, and that some contextually determined integrative process that may be deterministic in nature determines the cues that must be instantiated in order to derive the intended percept. Thus there may be a level of variation on the surface that appears to be free or random, but which in fact is a highly complex manifestation of underlying regularity. There are no contemporary speech synthesis systems that begin to approach the level of surface phonetic variability observed in natural speech. Models are needed that characterize this variability in terms of redundant cues, idiosyncratic gestures associated with a given speaker, and true random variability. The level of understanding needed to provide this control is immense, and will doubtlessly elude practical systems for many years to come.

13. The units of speech synthesis should be able to support a wide class of *voices*. Most contemporary systems focus on the adult male voice, but child, and female voices must be obtainable by any general system, and there is a need for the ability to synthesize several languages and diverse dialects within the context of one overall synthesis framework. A few contemporary systems are capable of synthesizing utterances in several languages within the same framework, but these systems generally sacrifice quality of synthesis in any particular language for breadth of coverage over the diverse linguistic forms.

14. The choice of units is intimately related to questions of *intelligibility*, *naturalness*, and *reduction* phenomena. To date, most research in speech synthesis has aimed at producing intelligible speech, without as much attention being focused on issues of naturalness and vowel reduction. There is still much to be learned about vowel reduction, and some feeling that such reduction may come from redundancy when viewed in the context of the sum total of all constraints in force at any point in the utterance. There is a tension between intelligibility and naturalness in that if a system aims for substantial reduction in the name of naturalness, intelligibility may suffer. It is also very difficult to evaluate naturalness, although some success has been achieved using cognitive loading techniques. Nevertheless, this class of psycholinguistic experimentation is still in its infancy and a great deal of work needs to be performed. At the discourse level, there is increased need for naturalness since attributes of focus and old/new information cause substantial stressing and destressing over that which would be predicted for a sentence spoken in isolation. It should also be remarked that individual segment intelligibility still leaves much to be desired, and the codification within an insightful rule structure of the vast amount of segmental phonetic detail must be substantially improved if there is to be any appreciable gain in synthetic speech quality.

15. The units may exploit available technology, but they should not be overly determined by it. Computational technology typically involves space/time tradeoffs. Thus increased availability of memory tends to favor compiled strategies, whereas increased processing capability favors interpretive strategies. As has been suggested, both will be needed for high performance systems of the future. It is well to remember that technology is progressing much faster than knowledge of speech. For this reason, a *correct theory* is the central need of research. The technology will easily rise to support such a theory. Through comprehensive study of extensive databases, new and insightful complex abstract unit hierarchies will be developed which when interpreted procedurally within domains of varying scope can be expected to give rise to synthetic speech of very high quality.

tence, clause, phrase, word, morpheme, metrical foot, syllable, phoneme, and feature. These have all been found useful in linguistic analysis for reasons of distribution and contrast, but they each exhibit some intrinsic cohesion and place focus on the units themselves. Thus each of these units implies an internal theory, together with a relation to other units and their corresponding theory. There is a tendency to think of these different levels of units with their attendant scopes as being arrayed over a temporal axis. While such a view provides a convenient geometric intuition, contemporary notions of abstraction may not rely on notions of temporal scope at all levels. Thus at the individual segment feature level the initial specification may include an unordered set of features which is then transformed at a higher level to derive temporal extent. Research in the asynchronous control of these features reveals the need for this varying scope at the feature level, as in the control of nasality.

11. The hierarchy of units should provide for a *gradual* progression through the various representations. That is, a sufficient number of units should be available so that the transformation from one level to another through the hierarchy is relatively small and direct. Thus for example, phonemes are transformed to allophones and then to targets, parameters and finally to the speech waveform. In this way, understanding of the speech process is modularized at each level, thus simplifying the rule structure at each level, even though the total number of rules over the entire system may be large.
12. The notion of *target* serves as a bridge between abstraction and physical properties. There is a tendency for the targets themselves to become complex objects exhibiting composition, scope, and internal cohesion. For example, fundamental frequency contours must be characterized at both the phrase level and in terms of 'segmental' effects. It is seen that the term 'target', originally motivated by the characterization of formant trajectories, is increasingly synonymous with 'abstract representational structure'. The role of *variability* in the choice of units must be considered in the design of any speech synthesis system. It is often felt that many attributes of these units are necessary, and hence admit of no variability. Recent understanding, however, indicates that many cues are redundant, and that some contextually determined integrative process that may be deterministic in nature determines the cues that must be instantiated in order to derive the intended percept. Thus there may be a level of variation on the surface that appears to be free or random, but which in fact is a highly complex manifestation of underlying regularity. There are no contemporary speech synthesis systems that begin to approach the level of surface phonetic variability observed in natural speech. Models are needed that characterize this variability in terms of redundant cues, idiosyncratic gestures associated with a given speaker, and true random variability. The level of understanding needed to provide this control is immense, and will doubtlessly elude practical systems for many years to come.

13. The units of speech synthesis should be able to support a wide class of *voices*. Most contemporary systems focus on the adult male voice, but child, and female voices must be obtainable by any general system, and there is a need for the ability to synthesize several languages and diverse dialects within the context of one overall synthesis framework. A few contemporary systems are capable of synthesizing utterances in several languages within the same framework, but these systems generally sacrifice quality of synthesis in any particular language for breadth of coverage over the diverse linguistic forms.
14. The choice of units is intimately related to questions of *intelligibility*, *naturalness*, and *reduction* phenomena. To date, most research in speech synthesis has aimed at producing intelligible speech, without as much attention being focused on issues of naturalness and vowel reduction. There is still much to be learned about vowel reduction, and some feeling that such reduction may come from redundancy when viewed in the context of the sum total of all constraints in force at any point in the utterance. There is a tension between intelligibility and naturalness in that if a system aims for substantial reduction in the name of naturalness, intelligibility may suffer. It is also very difficult to evaluate naturalness, although some success has been achieved using cognitive loading techniques. Nevertheless, this class of psycholinguistic experimentation is still in its infancy and a great deal of work needs to be performed. At the discourse level, there is increased need for naturalness since attributes of focus and old/new information cause substantial stressing and destressing over that which would be predicted for a sentence spoken in isolation. It should also be remarked that individual segment intelligibility still leaves much to be desired, and the codification within an insightful rule structure of the vast amount of segmental phonetic detail must be substantially improved if there is to be any appreciable gain in synthetic speech quality.
15. The units may exploit available technology, but they should not be overly determined by it. Computational technology typically involves space/time tradeoffs. Thus increased availability of memory tends to favor compiled strategies, whereas increased processing capability favors interpretive strategies. As has been suggested, both will be needed for high performance systems of the future. It is well to remember that technology is progressing much faster than knowledge of speech. For this reason, a *correct theory* is the central need of research. The technology will easily rise to support such a theory. Through comprehensive study of extensive databases, new and insightful complex abstract unit hierarchies will be developed which when interpreted procedurally within domains of varying scope can be expected to give rise to synthetic speech of very high quality.

SYMPOSIUM 3: Models of the Larynx

Chairman: *C. Scully, Leeds, United Kingdom*

Panel members: *G. Fant, I.R. Titze, M. Hirano, F. MacCurtain*

Discussant: *J. Sundberg, K. Stevens*

1. Introduction

Modelling of the larynx can be directed at different goals. One is the provision of better synthetic voices with increased naturalness. The emphasis here has been on the representation of one speaker of one language. A more distant - because it is more ambitious - goal, is the deeper understanding of the regularities and constancies of real speech for many different languages and for different speaker types. Improved quantitative phonetic descriptions should result; and a better understanding of the meshing together of common phonetic properties and individual characteristics. A third major aim is the provision of methods for the diagnosis of pathological states of the vocal folds.

A good model is a simplified representation of a process or system. In the case of speech production these are undoubtedly extremely complicated (see Titze, 1983). The main factors need to be identified, but until we have a better understanding of the fundamentals, our models are likely to contain large numbers of parameters. The modellers need as much quantitative information on natural human speech as their colleagues are able to provide. In return, the modellers may be able to make predictions beyond what can be measured in real speech. It is helpful to identify the larynx as a quasi-independent component of the speech producing system. Its complexity and importance in speech are indicated by the number of scholars investigating larynx activity as seen, for example, in the papers given at the Tenth International Congress of Phonetic Sciences and by the wide variety of techniques employed. Since the symposium on the larynx at the Eighth International Congress of Phonetic Sciences (Fant and Scully, 1977), several conferences and books have been partly or entirely devoted to the larynx in speech and singing (Fink, 1975; Carré, Descout and Wajskop, 1977; Fink and Demarest, 1978; Boë, Descout and Guérin, 1979; Lass, 1979, 1981; Lawrence and Weinberg, 1980; Stevens and Hirano, 1981; Abbs et al., 1983). Modelling is one path towards a greater understanding of the larynx, but it needs to be considered as several systems: (1) neural control mechanisms; (2) anatomical structures, tissue properties and muscle mechanics; (3) articulation; (4) aerodynamics; (5) acoustic sources; (6) acoustic filters. The actions are directed towards the achievement of auditory goals. Not surprisingly, therefore, in the symposium it was necessary to discuss articulation (whether subglottal, laryngeal or supraglottal) with reference to acoustic attributes such as fundamental frequency, voice source characteristics or frequency spectra for different vowel qualities. The larynx does not operate in isolation. To different degrees, in each of the systems listed above, links between the larynx and both subglottal and supraglottal regions need to be considered. Much of the discussion in the symposium concerned the effects of supraglottal changes on the voice source; and the changes in (supraglottal) vocal tract shape associated with different voice qualities the primary control of which is ascribed to the larynx.

In this report, the reviews of the four Panelists are summarised first (see Cohen and Van den Broecke, 1983). A report on the symposium follows. Statements are credited to individual scholars, but any misrepresentation is the responsibility of the Chairman. An attempt is made to relate discussion in the symposium to some current research issues.

2. Panelists' reviews

Hirano reviewed recent advances in research on the structure, mechanical properties and adjustments of the vocal folds.

From a mechanical point of view, the five layers of the vocal folds can be reclassified as: the cover (epithelium and superficial layer of the lamina propria); the transition (intermediate and deep layers of the lamina propria); and the body (vocalis muscle). The layer structure varies along the length of the vocal fold, with gradually increasing stiffness towards the ends. The membranous vocal fold is most pliant at the midportion because of its structure as well as its location. Vocal fold structures change with age: newborns have no vocal ligament; maturation of the layer structure appears to be completed around the end of adolescence; aging changes the layer structures, somewhat differently in women and men.

Vocal fold tissues exhibit non-linear stress-strain curves. Stiffness increases with elongation and is greater in the longitudinal direction (along the length of the vocal fold) than in the transverse direction; saturation effects are found in the longitudinal direction. Other mechanical properties have been measured also. Electrical stimulation of intrinsic muscles demonstrates their effects on the position, shape and structure of the vocal fold (Hirano, 1975).

Vocal fold adjustments in speech have been extensively investigated by Hirose and his co-workers (see Hirose and Sawashima, pp. 137-154 in Stevens and Hirano, 1981). A summary of their findings follows: (1) CT primarily controls F_0. Occasionally, CT contraction in segmental glottal adjustment for enhancing voicelessness is suggested. (2) PCA is active for voiceless consonants. The PCA activity is positively related to the degree of glottal opening. For voiced phonemes, PCA is inactive. (3) IA presents a pattern reciprocal to PCA. (4) The activity level of VOC and LCA is influenced by context. Generally speaking, their activity is decreased for consonants. The suppression of VOC and LCA is occasionally more significant for voiceless consonants than for voiced consonants. The pattern of the activity is different between the two muscles in some specific cases.

Titze reviewed the difficulties in computer stimulation of biomechanical processes of speech production, the progress being made to overcome them, predictions made by modelling, and future expectations.

Although studies on speech production can borrow from research into other biomechanical processes, phonation presents unique problems associated with its rapid tissue and air movements, well into the kHz range, where low frequency measurements of viscoelastic properties cannot be trusted. The structures must be considered as three-dimensional, because their dimensions are comparable to wavelengths. This necessitates a distributed-parameter (continuum) approach and tbe inclusion of displacements transverse to as well as along the tissue fibre direction. Computational problems arise as glottal area approaches zero. Both finite-difference and the current finite-element approaches to representing the vocal folds have computational difficulties, but these should be eased by developments in computing. The major obstacle remains the refinement of theories of pressure and flow development in the glottis, tissue deformation under these pressures, and the effects of laryngeal muscle contractions on the viscoelastic properties.

Successes include a clear demonstration of the flow-induced self-oscillatory nature of vocal fold vibration. Asymmetry in the net aerodynamic driving pressure acting outwards on the medial surface of the vocal folds with respect to opening and closing keeps the vocal folds oscillating. This asymmetry derives from the resistance and inertance of the air in the glottis and is enhanced by an inertive vocal tract load, while it is reduced by a capacitive vocal tract load (Ishizaka and Flanagan, 1972). The most efficient way of maintaining asymmetry is through vertical phase differences, as in two-mass or multiple-mass models.

Models have quantified the effects of vocal fold dimensions and configurations upon glottal area shape, vocal fold contact area, aerodynamic variables and the properties of the acoustic output. They have begun to predict the biomechanical significance of the layered structure of the vocal folds and of tissue anisotropy (Titze and Talkin, 1979). As more data become available modelling will help to explain the effects of variations in size, shape, viscoelastic properties, muscular contraction, lesions, stress relaxation and unsteady flow conditions.

Fant considered recent significant advances in the source-filter concept of voice production.

In current models source and filter are assumed to be short-time invariant. A firmer theoretical basis is being constructed in which acoustic and mechanical interactions between filter and source are included. The major complication for modelling is that, in the glottal open state (a phase of acoustical energy charging), the sub- and supraglottal parts of the system are acoustically coupled, whereas, with the glottis closed, (an energy discharge phase), they each execute approximately free and separate oscillations.

In the most complete speech production model (Ishizaka and Flanagan, 1972), lung air pressure provides the power; the system is self-oscillating and self-adjusting, without a source in the linear network sense. The shape of the glottal area function Ag(t) may be assumed instead, and is combined with lung air pressure to derive glottal volume velocity Ug(t), defined as the voice source (Ananthapadmanabha and Fant, 1982). Alternatively, a 'short circuit' or 'no load' source Us(t) can be derived. These two voice sources Ug(t) and Us(t) differ in two main respects: Ug(t) is skewed to the right, with a slower rising branch and a faster falling branch than Us(t); oscillatory ripples, usually close to F_1 in frequency, are superimposed on Ug(t). Us(t) provides a simple source but a complicated, time-variable filter function while, with Ug(t) as source, the filter function is simple, being linear and only slowly time-variable, but Ug(t) depends on the entire sub-and supraglottal configuration and has to be derived by complex calculations. In a third, approximate, model (Fant, 1979) a smoothed version of Ug(t) applies a constant current source to the vocal tract in parallel with the time-varying glottal plus subglottal impedance. Two stage synthesis has been proposed (Mryati et al., 1976). Here an Ag(t) voltage, with an F_1 and F_2 impedance load, gives a current function which is used as the source in a formant synthesizer. The perceptual importance of the various approximations has not yet been fully investigated.

Parameterisation of the glottal flow or glottal area functions was discussed, and extensions to the Fant (1979) model for better matches with real speech data. Important characteristics of the Ug(t) function include: the area under the curve, related to low frequency acoustic amplitude; a low frequency spectral maximum Fg which is prominent at low voice effort; and the amplitude of the negative spike of the flow derivative at closure, which controls formant amplitudes.

Problems in the use of different kinds of inverse filtering to obtain reliable estimates of glottal flow were reviewed. More research is needed to improve the modelling of women and children's voices, speaker specific factors and speech source dynamics, which are probably carried more by vocal fold actions than by lung pressure variations. Rules for voice sources need to be based on an understanding of production events and underlying physiological parameters.

MacCurtain presented evidence of the pharyngeal gestures and associated larynx adjustments which occur in the formation of different voice qualities.

A 5-parameter descriptive framework is proposed, based on data from seven languages. Phonetic theory as currently formulated fails to account for some voice quality contrasts. The techniques used, xeroradiography (combined with acoustic and electro-laryngographic recordings) show soft tissue in extremely fine detail, even indicating individual muscle groups.

English informants produced different voice qualities for an [i] vowel. Out of a descriptive framework of 12 anatomical points for the larynx and pharynx, 5 points altered significantly between rest position (expiration) and phonation. Large inter-speaker variations in muscle patterning were found for a common auditory target of, for example, 'harsh' voice; it may be misleading to attribute one set of componential elements (Laver, 1980) to one voice quality.

The 5-parameter framework was applied also to minimal pairs of phonological oppositions in which contrastive voice qualities are required. As an example, results for a speaker of Dinka were cited: in moving from 'brassy' voice quality, associated with high tone, to 'hollow' voice quality, associated with low tone, the speaker greatly widened both the laryngopharynx and the lower oropharynx; the thyroid cartilage was lowered to 'resting' height, but the dimensions of the laryngeal airway (from the anterior end of the vocal folds to the supero-posterior cricoid tip) remained unchanged, in spite of the required frequency contrast.

The parameters are being applied to the evaluation of voice disorders (Berry et al., 1981) and to a study of singing.

3. Neurological control

The modelling of the larynx as part of a neural system has been discussed by Muller, Abbs and Kennedy (see pp. 209-227 in Stevens and Hirano, 1981).

Hirano: The role to sensory feedback and reflexes of the PNS has been over-emphasised; more attention should be given to central mechanisms, including the extra-pyramidal system.

Hirose: Part of the cerebellum and the basal ganglia participate in the preprogramming of movement before the motor command is realised at the level of the motor cortex. More significantly, another part of the cerebellum is checking the outflow of the motor command from the cortex to the periphery before the commands reach the peripheral organ itself. Peripheral feedback mechanisms are certainly important, particularly in the course of learning skilled movements, including speech. But after a certain period of learning, the CNS knows precisely what should be done according to the template stored in the brain and updating mechanisms provided by cerebello-cerebral interactions.

4. Biomechanical interactions and aerodynamic forces in the control of F_0 and the voice source

Stevens: It is important to make a distinction between effects seen across languages and those seen across individuals. There are relatively few possibilities for creating contrasts which are phonologically relevant; these should conform to quantal principles. On the other hand, inter-speaker differences are likely to be characterised by fine, complex gradations. Three or four basic parameters are needed to characterise the controlled variables of F_0 and the voice source waveform with its associated spectral emphasis, ranging from 'breathy' to 'pressed' phonation. Because the muscles which alter F_0 and those which control the waveform interact, these two acoustic aspects of speech interact also. A speaker may change the vocal tract shape in order to enhance contrasts created primarily in the larynx; larynx actions, on the other hand may be able to create auditory effects which are assumed to be associated with the vocal tract: an example might be the auditory property of nasalisation created by abduction of the vocal folds.

Fujisaki: His functional model of F_0 control by the cricothyroid and vocalis muscles represents F_0 changes as the response of a second-order linear system to step functions. With suitable values for the damping factor (different for speech and for different kinds of pitch transition in singing) good matches to F_0 contours are generated. With the assumption that the vocal folds vibrate like simple elastic membranes, so that frequency is proportional to the square root of tension, and using the known non-linear relationship between tension and elongation for skeletal muscles, it has been shown that, to a first order of approximation, the logarithm of F_0 is proportional to the elongation of the vocalis muscle. Data from real speech (Honda et al., 1980) support the analysis.

Another implication of the non-linear stress-strain curve of muscle tissue was discussed.

Titze: It should not be assumed that an increase in any aerodynamic variable operating at the larynx will necessarily raise pitch. Indeed, an increase in volume flow rate accompanies decreased longitudinal stiffness of the vocal folds and under these circumstances increased airflow accompanies a decrease in F_0. The empirically observed correlation between increased transglottal pressure drop and higher F_0 is not simple to explain. The effect is indirect. The higher aerodynamic forces on the vocal folds due to increased pressure drop result in increased amplitude of vocal fold vibration. This would not increase F_0 if the vocal folds constituted a linear oscillator. A pendulum, or a child on a swing for example, has the same frequency of oscillation regardless of amplitude (over a limited range at least). But, in the case of vocal fold tissue, increased deformation is associated with greater mechanical stiffness. As the amplitude of vibration of the vocal folds increases the effective stiffness of the vocal folds increases also; for this reason F_0 rises as transglottal pressure drop increases. Control of F_0 is made more complicated by relaxation effects. After vocal fold stiffness has been suddenly increased, it declines gradually.

The declination of F_0 contours in speech can probably be accounted for by an observed gradual decrease in subglottal pressure. It is not necessarily possible to ascribe a fixed change in F_0 to a fixed change in pressure drop. $\delta F_0/\delta \Delta P$ may well vary depending on the biomechanical state of the vocal folds. This is an example of the physiological variables interacting in rather complex wavs to determine acoustic results (Collier, see p. 440 in Cohen and Van den Broecke, 1983, with comments by Titze).

Ohala: Non species-specific biological explanations may be offered concerning pitch control. Some survival value may attach to the possession of rapidly operating mechanisms for both F_0 raising and also F_0 lowering, to exaggerate the animal's body size, for instance.

Löfavist: Caution is needed in assigning functions to particular muscles. For example three kinds of articulatory actions may be used to make voicing cease. The vocal folds can be widely abducted, or strongly adducted, or abruptly stiffened. Interactions between vocalis and posterior cricoarythenoid muscles may be just as important as the reciprocal activity observed for interarytenoid and posterior cricoarytenoid muscles in the control of the acoustic feature ± voice.

Published research for natural speech (see, for example, Rothenberg, 1973; Gauffin and Sundberg, 1980) suggests that, within a 'plateau' region for phonation, the form of the total volume flow rate through the glottis varies. At the edge of the plateau, voicing dies away, for a variety of inimical articulatory and aerodynamic reasons. Covariation in the a.c. (acoustic) and d.c. components of air flow through the glottis implies that voice and aspiration sources covary. Published data for natural speech have been incorporated in a functional model for the voice source, with three controlling variables: pressure drop, average glottal area and a vocal fold stiffness factor (Scully and Allwood, 1983).

5. Larynx - vocal tract physiological interactions

With models, individual factors may be identified and manipulated as they cannot be in real speech. For example, acoustic coupling and anatomical links between larynx and vocal tract can be dissociated. Modelling of acoustic coupling alone gives variations of fundamental frequency with vowel type in the opposite sense to that of real speech so physiological links presumably give overriding effects (Guérin, Degyrse and Boë, pp. 263-277 in Carré et al., 1977).

The hyoid bone has been observed in continuous speech (see, for example, Botherel, 1975) and for isolated vowels (see, for example Rossi and Autesserre, 1981). The forces acting on the hyoid bone were considered at the symposium. Eight pairs of muscles are attached to the hyoid bone, and five pairs of ligaments or membranes: an astonishing number for this small bone. Through the geniohyoid, mylohyoid and digastricus muscles, the larynx is likely to be affected by jaw movements; and the larynx is connected to the tongue through the hyoglossus muscle pair. Even the soft palate is connected to the larynx, via the palatopharyngeus muscles, which attach to the laminae of the thyroid cartilage.

Whereas it is generally accepted that longitudinal stretching of the vocal folds increases the stiffness and raises F_0, the effects of vertical stretching are not so clear. Interestingly, the association between high F_0 and advanced tongue root with wide pharynx discussed by Honda, below, is in contrast to the low F_0-wide pharynx association found by MacCurtain for Dinka (MacCurtain, 1983).

Honda: Observed geniohyoid muscle activity associated with pitch raising and correlations between high F_0 and horizontal position of the hyoid bone suggest additional mechanisms for pitch raising besides intrinsic larynx muscle activity. The intrinsic pitch of vowels can be explained by the same mechanism of the laryngeal external frame function. The hyoid bone is pulled forward and downward by the geniohyoid and genioglossus (posterior fibres) muscle forces, the latter being the more important in reducing the anterior-posterior length of the tongue in high vowels. As a result of the hyoid bone shift, the thyroid cartilage is rotated forward about the cricothyroid joint, thus stretching the vocal folds and raising F_0. Tongue movements can influence vocal fold state indirectly (via the epiglottis) also. When the epiglottis is pulled upright as the tongue root moves forward, the aryepiglottic fold is stretched. Increased tension here may raise F_0 through increased stabilisation of the arytenoid cartilages. But the (vertically) wide laryngeal ventricle seen for high vowels may perhaps not stiffen the vocal folds for two reasons: first, the vocal folds are not tightly coupled to the false vocal folds; secondly, the vocal folds are more compliant in the vertical direction than in the horizontal direction.

MacCurtain: Because of the large number of muscle attachments, there is much redundancy in the positional control of the hyoid. For some speakers the cornu shifts horizontally. Different angles of tilt of the hyoid are seen for different voice qualities.

Liberman: Emg activities for geniohyoid and other extrinsic laryngeal muscles show a positive correlation with F_0, especially in the upper quartile of the range.

Fischer-Jørgensen: Data published by Alfonso show that F_0 variations for different vowels correlate better with the position of the jaw than with that of the tongue.

6. Aerodynamics

If the slowly changing (d.c.) components of air pressure and volume flow rate of air are considered separately from the acoustic (d.c.) components, these former are variables in an irreducible aerodynamic system comprising the whole respiratory tract. Since aerodynamics links actions to sounds, a single articulatory change results in multiple acoustic changes. With this approach, the average flow rate through the glottis reflects the articulation at the larynx and elsewhere. A less simplified approach would need to incorporate the effects of aerodynamic forces on glottal area directly. Average glottal area is not necessarily a good representation of vocal fold abduction-adduction. Modelling shows that where longitudinal stiffness of the vocal folds is increased at constant vocal fold setting and subglottal pressure, average airflow decrease; so that average glottal area derived from the aerodynamic

orifice equation would decrease (Titze and Talkin, 1979). This kind of effect needs to be taken into account if a model maps from average airflow onto the airflow waveform for the voice source. A sophisticated model of respiratory control needs to be developed. Should average subglottal pressure and glottal area be considered as two independent controlling factors or should average subglottal pressure, on the contrary, be treated as a dependent variable, to be derived from average glottal area combined with a lung volume control model (in which lung volume decrement or pressure in the alveoli of the lungs or nett expiratory force could be the independent variable)? Models need to be able to explain why subglottal pressure falls during an utterance.

7. The voice source and acoustic coupling

The quasi-periodic waveform of volume flow rate at the glottis can be recovered for vowels in natural speech by inverse filtering. The complex shapes and structures of the vocal folds have been represented in a number of different ways in models.

Flanagan described the two-mass model of the vocal folds. When each vocal fold was represented as a single mass the system was self-oscillating but was unduly sensitive to vocal tract loading. Synthetic speech, obtained by coupling the two-mass model to a transmission line model of the vocal tract, was demonstrated. This modelling was for speech synthesis, without the intention of duplicating fine physiological detail. With the vocal folds set to a suitable operating glottal area and with subglottal pressure applied, dominant acoustic effects of natural speech were produced (see Ishizaka and Flanagan, 1972).

As articulatory synthesis of speech this research has remained unsurpassed. Other approaches to modelling have sought insight into the wide range of normal and abnormal vibration modes found in real speech. The vocal folds have been represented as many masses and springs, and as continuous viscoelastic media (see Titze, 1983). The vocal folds have been represented as a pair of beams (Perrier, 1982). It is hoped that this research may resolve some of the problems of commands for two-mass models, while providing a closer match to the complexities of the natural physiology (Guérin, pers. comm., 1982). The forces acting upon the parallel structures of the body and the cover of the vocal folds have been analysed by Fujimura (see pp. 271-288 in Stevens and Hirano, 1981).

Fant considered ripple effects seen in voice waveforms and skewing effects. The flow rate waveform is skewed relative to the glottal area waveform, with sharpening of the flow cut-off. Cut-off steepness is modified by the inclusion of frictional losses in the analysis. Since skewing of the airflow waveform is associated with inertance of air above as well as in the glottis, the skewing effect is different for different vowels. The effects have been modelled by Rothenberg (see pp. 305-328 in Stevens and Hirano, 1981).

Fujimura: It is interesting to consider whether the different pharynx shapes found for contrasting voice qualities (see MacCurtain, 1983) may influence the acoustic output in several ways which could be separated by modelling. Besides the different filter shapes, the voice source could perhaps be affected in two ways: through the physiological links of the external frame function of the larynx and in a direct, acoustic, way.

Titze: Yes, acoustic coupling effects would be expected: a large cross-section area in the lower pharynx, just above the vocal folds, reduces the skewing effect, while a constricted pharynx enhances it.

Fant: In addition. differences of shape in the vestibule and lower pharynx may influence acoustic sources because of their effects upon the total pattern of airflow just above the glottis.

8. Singing versus speech

Sundberg: Trained singers are valuable subjects for research into larynx behaviour. The normal and natural way for speech and untrained singing is with a physiological emphasis. Covariations are allowed. For example, louaness and pitch generally rising and falling together gives a 'natural' effect. Singing, on the other hand, is oriented towards the side of perception, with more stringently controlled auditory goals. Singing demonstrates phenomena not found in speech, with more demands made on physiological control. For example, gastric and oesophageal pressure measurements show that the diaphragm muscle, found to be inactive in previous studies on speech, is active during professional singing. Muscle forces controlling movement act faster in singing, giving more rapid changes in subglottal pressure than those found in speech (von Euler et al., 1983). The automatic results of physiological and acoustic coupling between the larynx and other structures may be unacceptable in professional singing. Singers often experience vowel-dependent difficulties in singing on a given pitch; these and other voice source effects which are revealed in singing need to be investigated further. In the course of their training singers probably learn to break up the habitual relationships between the voice source on the one hand and pitch and loudness on the other. Control of the voice source spectrum independently of loudness and pitch may not be required in speech but probably is important for opera singers.

9. Inter-speaker variations

Models ought to conform to physical constraints yet be capable of simulating different speaker types, whether the inter-speaker differences arise from anatomical and physiological differences or from the options selected. Apart from contrasts between speech and professional singing, two other aspects were discussed.

10. Voice source waveforms for women speakers

Karlsson (Stockholm) showed voice waveforms obtained by inverse filtering, for four Swedish women speakers using a normal F_0 range with normal and loud voice. There is a wide range of waveshape parameter values across speakers; greater than any inter-sex differences.

Price: By contrast with Karlsson's data, American women speakers gave voice waveshapes with a 40% shorter 'closed' fraction than those for men. The 'closing' time was three times as long for women, with a much more rounded 'corner'.

These differing results could possibly be ascribed to the techniques of inverse filtering used (from mouth flow rate with a Rothenberg mask for the Swedish data; from acoustic pressure for the American English). However, it may be that the differences are genuine cross-language ones, possibly reflecting different social attitudes and conditioning; or even simply inter-speaker differences. Much more emphasis on the analysis of women's speech is called for.

11. Pathological conditions

Hirano gave an account of pathological conditions of the vocal folds, including paralysis, polyps and carcinoma. Each type of pathology creates its own kind of abnormal mechanical state for the pair of vocal folds with implications for their modes of vibration.

As acoustic and other accessible measures are related to specific pathologies, empirically based non-invasive assessment techniques will be developed. Laryngograph waveforms differ for a small single area of vocal fold contact and for a small total area associated with numerous very small contact areas for rough surfaces (Fourcin, pers. comm., 1983). Acoustic analyses of speech outputs are being matched to histological examinations, so that changes in vocal fold cell structure may be inferred from acoustic data (MacKenzie et al., 1983). Within the descriptive framework of MacCurtain, some of the vocal tract parameters are likely to prove particularly relevant for the diagnosis of disorders of the larynx.

Hirano raised the question of the effect on source-filter acoustic interaction when there is no closed phase, as in some laryngeal pathologies.

Fant: The increase in acoustic interaction between the supraglottal and subglottal systems affects resonance frequencies and bandwidths. The problems of obtaining true glottal flow are increased, since frequencies and bandwidths must represent closed glottis conditions if inverse filtering is to be an exact inverse of the transfer from glottal flow to the speech wave.

12. Quantitative phonetic descriptions

In the time available, questions of appropriate phonetic features for the description of larynx behaviour in models and natural speech could not be considered in detail.

Ladefoged reviewed the variety of approaches needed for larynx descriptions, from both physiological and acoustic bases. The vocal folds can be investigated as vibrating structures or as the voice waveshape, related to phonetic properties of the output speech. Material from a wide range of languages needs to be analysed. In some languages the phonetic properties of breathy voice and high F_0, creaky voice and low F_0 hang together; in other languages they 'hang apart' and are not found to be correlated.

13. Conclusions

The symposium offered a rapid scan and overview of a number of areas of research relevant for improved larynx modelling. Some of these areas merit, and are already the subjects of, more specialised discussion meetings.

Acoustic and physiological data can be correlated with auditory phonetic descriptions, preferably for many different languages. The growing interest in the analysis and synthesis of women's voices should be applauded and encouraged. Physical modelling of abnormal conditions of the vocal folds should assist acoustic differentiation of pathological states of the larynx. The valuable new tool of inverse filtering (whether from volume flow rate or from acoustic pressure) will provide voice source data. These real speech waveforms can provide a means of assessment of outputs from vocal fold models and more natural source inputs for acoustic models of speech, whether for terminal-analog or for line-analog synthesis. Acoustically significant features of the glottal flow waveshape have been identified. These should be described quantitatively in future analyses, whatever additional parameters may be of interest to the experimenter. Future modelling can contribute to an increased understanding of turbulence noise source properties and of their interactions with the voice source. Acoustic modelling of interactions between the voice source and the vocal tract filter is well advanced and the main factors seem to have been identified; the contribution of the subglottal acoustic filter is becoming better understood. There is a clear need for models of anatomical and physiological interactions between the larynx and supraglottal structures, supported by physiological experiments on real speech in which acoustic effects such as F_0 changes are recorded. Fink and Demarest (1978) have provided a possible appraoch to this kind of modelling.

Unification of terminology must probably await increased understanding. Registers in speech, for example, may not correspond to those of singing (see Hollien, p. 487 in Cohen and Van den Broecke, 1983), but different voice qualities, phonation types and registers in speech need to be labelled auditorily and related to vibratory patterns for the vocal folds (see Stevens and Hirano, 1981, p. 168).

References

Abbs, J. et al. (1983). *Vocal Fold Physiology Conference*, Madison, Wisc. (in press).

Ananthapadmanabha, T.V. and Fant, G. (1982). Calculation of true glottal flow and its components. *Royal Inst. of Technol. Dept. of Speech Communication and Music Acoustics, Stockholm, QPSR 1/1982*, 1-30.

Berry, R.J., Epstein, R.L., Fourcin, A.J., Freeman, M., MacCurtain, F. and Noscoe, N. (1982). An objective analysis of voice disorder, parts 1 and 2, *Brit. J. of Disorders of Comm.* **17**, 67-85.

Boë, L.J., Descout, R. and Guérin, B., eds. (1979). *Larynx et Parole*. Proceedings of a GALF Seminar, Grenoble, February 8-9, 1979. Grenoble: Institut de Phonétique de Grenoble.

Bothorel, A. (1975). Positions et mouvements de l'os hyoïde dans la chaîne parlée, *Travaux de l'Institut de Phonétique de Strasbourg* **7**, 80-132.

Carré R., Descout, R. and Wajskop, M., eds. (1977). *Articulatory Modelling and Phonetics*. Proceedings of a Symposium, Grenoble, July 10-12, 1977. Brussels: GALF Groupe de la Communication Parlée.

Cohen, A. Van den Broecke, M., eds. (1983). *Abstracts of the Tenth Intl. Congr. of Phonetic Sciences*, Foris Publications, Dordrecht, Holland.

Euler, C. von, Lagercranz, H., Leanderson, R. and Sundberg, J. (1983). Diaphragmatic activity during singing. *Proc. of the Stockholm Music Acoustics Conf. July-Aug. 1983*, to appear.

Fant, G. (1979). Glottal source and excitation analysis. *Royal Inst. of Technol. Dept. of Speech Communication and Music Acoustics, Stockholm QPSR 1/1979*, 85-107.

Fant, G. (1983). The voice source - acoustic modelling. In: Cohen and Van den Broecke, 1983, 151-177.

Fant, G. and Scully C., eds. (1977). *The Larynx and Language. Proceedings of a Discussion Seminar àt the 8th International Congress of Phonetic Sciences Leeds August 17-23, 1975 Phonetica* **34**, 4.

Fink, B.R. (1975). *The Human Larynx, A Functional Study*. New York: Raven Press.

Fink, B.R. and Demarest, R.J. (1978). *Laryngeal Biomechanics*. Cambridge, MA: Harvard University Press.

Gauffin, J. and Sundberg, J. (1980). Data on the glottal voice source behaviour in vowel production. *Royal Inst. of Technol. Dept. of Speech Communication and Music Acoustics, Stockholm, QPSR 2-3/1980*, 61-70.

Hirano, M. (1975). Phonosurgery - Basic and clinical investigations. *Otologia Fukoka* **21**, 239-440.

Hirano, M. (1983). Structure, mechanical properties and adjustments of the vocal fold. In: Cohen en Van den Broecke, 1983, 187-193.

Honda, K., Hibi S.R., Kiritani, S., Niimi, S. and Hirose, H. 980). Measurement of the laryngeal structures during phonation by use of a stereoendoscope. *Ann. Bull. RILP, Univ. of Tokyo* **14**, 73-78.

Ishizaka K. and Flanagan, J. (1972). Synthesis of voice sounds from a two-mass model of the vocal cords. *Bell Syst. Tech. J.* **51**, 1233-1268.

Lass, N.J., ed. (1979, 1981). *Vols. 2 & 5, Speech and Language: Advances in Basic Research and Practice*. New York: Academic Press.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: C.U.P.

Lawrence, V.L. and Weinberg, B., eds. (1980). *Transcripts of the Eighth Symposium Care of the Professional Voice, Parts I, II, and III*, June 11-15, 1979, The Juilliard School, NYC. New York: The Voice Foundation.

MacCurtain, F. (1983). Pharyngeal gestures in contrastive voice qualities. In: Cohen and Van den Broecke, 1983, 199-213.

Mackenzie, J., Laver, J. and Hiller, S.M. (1983). Structural pathologies of the vocal folds and phonation. *Edinburgh Univ. Dept. of Linguistic Work in Progress* **16**, 80-116.

Mryati, M. Guérin, B. and Boë, L.J. (1976). Etude de l'impédance d'entrée du conduit vocal. Couplage source-conduit vocal. *Acustica* **35**, 330-340.

Perrier, P. (1982). Etude d'un modèle continu des cordes vocales sous forme de deux poutres bi-articulées: premières simulations, unpublished Diss. Doct. Ingén., ENSERG l'Inst. Natl. Polytechn. de Grenoble.

Rossi, M. and Autesserre, D. (1981). Movements of the hyoid and the larynx and the intrinsic frequency of vowels. *J. of Phonetics* **9**, 233-249.

Rothenberg, M. (1973). A new inverse-filtering techniaue or deriving the glottal air flow waveform during voicing. *J. Acoust. Soc. Amer.* **53**, 1632-1645.

Scully, C. and Allwood, E. (1983). Simulation of singing with a composite model of speech production. *Proc. Stockholm Music Acoustics Congress 1983*, to appear.

Stevens, K.N. and Hirano, M., eds. (1981). *Vocal Fold Physiology.* Proceedings of a Conference, Kurume. January 15-19, 1980. Tokyo: University of Tokyo Press.

Titze, R. (1983). Approaches to computational modelling of laryngeal function: successes and prevailing difficulties. In: Cohen and van den Broecke, 1983, 179-186.

Titze, I.R. and Talkin, D.T. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *J. Acoust. Soc. Amer.* **66**, 60-74.

SYMPOSIUM 4: Auditory Analysis and Speech Perception

Chairman: *A. Fourcin, London, United Kingdom*

Panel members: *A. Risberg, J.M. Pickett and S.G. Revoile, R.J. Ritsma and J.W. Horst, J.L. Goldstein*

This symposium was divided into two clear parts. In the first, the five discussants, who had contributed invited abstracts to the main programme, each presented a summary of their individual work and gave a further overview of its more general implications.

In the second part, the general discussion touched on five main areas in which work on hearing has practical and theoretical bearing on human speech processing. Several themes were common to both aspects of the symposium. First, there was a bias towards the use of research with the hearing impaired as a means of better understanding the mechanism of normal hearing. Second, there was a general interest in models of hearing with special reference to frequency and temporal analysis and peripheral rather than cognitive processing. Third, the use of speech contrastive stimuli as an analytic - experimental - tool was discussed with particular reference to the use of recoded natural syntheses rather than complete natural speech.

It is best to follow the logical rather than the temporal organisation of these various elements.

Julius Goldstein gave a discussion of the possible mechanisms underlying our perception of voice pitch, and evaluated the broad ways in which a more complete speaker-listener model would be influenced by these basic peripheral processing issues. The essential proposal is that voice pitch is mediated by a two- fold process. First, auditory filtering corresponding to what is experimentally found neuro-physiologically in animals (Kiang, 1965) and psychophysiologically determined in man (Patterson et al., 1982), processes the complex acoustic input. Second, auditory fibres close in characteristic frequency to this region of maximum place response then give a temporal pattern of activity which is used as an input to an 'optimum' central processor of fundamental frequency (Goldstein, 1983; Goldstein and Srulovicz, 1983). Although temporal processing is, here, prominent in the last part of the peripheral sequence of operations this hypothesis does not rely on time patterning as a primary basis for sound analysis. Indeed, in the overall discussion Hoekstra's work was often referred to as demonstrating the relative unimportance of any pitch mediating algorithm which did not rely on place analysis. The relevance of situations in which no place analysis is possible and yet temporal patterning elicits useful prosodic contrastive ability (Fourcin et al., 1983) was pointed out. In normal listeners, however, this temporal feature of auditory analysis may be of lesser importance in the developed adult. It was proposed, however, that it might be crucial both for the normally developing infant and for the congenitally hearing impaired (Fourcin, 1977).

The general discussion went from these considerations relating to pitch to deal with timbre analysis and the relevance of different auditory models. In particular the 'Stockholm' zero crossing model was compared with Goldstein's (and, effectively, Sachs and Young's 1982) model with particular reference to the possibility not only of analysing spectral prominences but also assessing gross spectral weighting.

Current work in Stockholm has led to the 'DOMIN' model (Carlson & Granström, 1982) and whilst this has the desirable feature of delineating spectral prominences the point was made in the discussion that timbre effects coming from level variations within a spectrum were possibly not sufficiently well represented to cater for all phonetic contrasts. The trend, however, from a largely physical model to one which was substantially determined by observational parameters was considered essential to future progress - in work on both man and machine.

The contributed papers by Ritsma and Horst (1983) and Pickett and Revoile (1983) related to speech processing by the normal and by the impaired auditory system. An essential implication of the auditory models discussed immediately above is that a reduction in peripheral frequency selectivity will inevitably entail a corresponding impairment of speech receptive ability. Ritsma's presentation gave experimental evidence to show, however, that essentially normal responses could be obtained even for the hearing impaired with a demonstrably degraded auditory filter bandwidth in quiet listening to a single formant changed in centre frequency with bandwidth as a control parameter. When the signal/noise ratio of the stimuli was degraded, the difference between normal and hearing impaired listeners became apparent. The immediate conclusion to be drawn from these experiments is that 'frequency selectivity primarily influences (the) speech-to-noise ratio and not (the) speech perception as such.' In the broader context of this Symposium, however, it emerges that it may be more accurate to postulate that frequency selectivity influences the complexity of speech processing available to the listener (see particularly the contributions by Risberg, 1983; Pickett et al., 1983 and by Fourcin, 1983.

Arne Risberg's contribution complemented the work presented by Ritsma and Horst in that he gave results relating another measure of frequency selectivity to speech receptive ability for a range of hearing impaired subjects. Here, scores tended to be lower as selectivity decreased even though the tests were performed without added noise. The subjects with poor selectivity all complained, however, that especially in a noisy situation they had speech perception difficulties.

Risberg also discussed the relation between prosodic receptive ability and his (frequency modulated tone) assessment of frequency. Here, a fairly clear correlation between low frequency discrimination ability and response accuracy in stressed word identification test emerged. This finding was especially relevant to the topic of frequency versus temporal processing which was a central discussion theme and it seemed possible that, contrary to the case for normals as shown by Hoekstra (whose results were given special attention by several speakers), temporal processing might be basic to these findings (and this correlates with Fourcin's report on the EPI group's work on electro-cochlear stimulation with the totally deaf - where only temporal processing is possible - (cf. Fourcin et al., 1983)).

Measurements of this type with congenitally hearing-impaired subjects are difficult to interpret, however, since response ability is conditioned by speech knowledge as well as sensory processing.

Fourcin's contribution (1983) was especially concerned with the selective use of simple minimal pair contrasts synthesized so that only a small number of acoustic elements were involved. Using auditory tests based on the interactive use of these stimuli it was possible to follow the course of speech acquisition in both normal and hearing- impaired children and to show that for different ages, and hearing abilities, different pattern aspects of the speech signal were of importance. The tendency being for temporal and low frequency components to dominate initially and to be reinforced by higher frequency constrastive information at later stages of development. This analytic use of speech material was also of value in speech aids and examples were discussed of the use of only fundamental frequency information in both acoustic and electrical stimulation prostheses. In each case quite substantial improvement was possible not only in lip-read receptive ability but also in speaking ability.

These techniques seemed likely, in the future, to be of much wider applications since they provide ways of training and assessment which are not otherwise available and link psycho-acoustic and phonetic levels of processing.

Pickett and Revoile (1983) presented the results of experiments in which natural speech was used as the basis of tests of speech perceptual ability in both normal and hearing impaired subjects. Initial and final voiced voiceless cognate consonantal contrasts were controlled, for example, in regard to frication; appropriate onset and offset transitions; vowel duration. Fairly consistent differences in overall response strategies were found between the two groups of listeners in regard to the use of durational and low frequency speech pattern elements although some individuals showed marked differences in performance. This aspect of speech perception was emphasized also by Risberg and by Fourcin and for the latter arose from the intrinsic auditory nature of speech contrasts. When a number of potentially useful contrastive features exist - and this is always the case for normal speech - the listener must make a choice. The act of choice creates a feature sub-set which, whether synthesis modification of natural speech is used, may map quite differently from one subject to another, onto the information available in the stimulus presentation.

A final, but related, discussion topic concerned the dimensionality of the auditory representation of speech. One argument proposed the use of factor analysis techniques (cf. Pols and Stoop, 1983) to arrive at a pertinent set of contrastive features which would be relevant to the set of listeners investigated. Another proposition was that the study of normal development could provide a basis for a specification and hierarchical ordering of important features. Both of these approaches were recognised, however, as requiring contributions towards the better definition of the dimensions from both basic auditory processing and phonological constraints.

References

Carlson, R. and Granström, B. (1982) Towards an auditory spectrograph, in R. Carlson and B. Granström (Eds) *The Representation of Speech in the Peripheral Auditory System* Amsterdam; Elsevier Biomedical Press, 109-114.

Fourcin, A.J. (1977) Acoustic patterns and speech acquisition, in N. Waterson and C. Snow (Eds) *The Development of Communication.* London: J. Wiley, 144-171.

Fourcin, A.J. (1983) Perception of speech pattern elements, in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences,* Dordrecht: Foris 217-226.

Fourcin, A.J., Douek, E.E., Moore, B.C.J., Rosen S., Walliker, J.R. Howard, D.M., Abberton, E. and Frampton, S. (1983 Speech perception with promontory stimulation. *Annals of the New York Academy of Sciences, Vol. 405.* New York: New York Academy of Sciences 280-294.

Goldstein, J.L. (1983) An outline of recent research progress on fundamental pitch perception, in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences* Dordrecht: Foris, 245-247.

Goldstein, J.L. and Srulovicz, P. (1983) A central spectrum model: a synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. *Journal of the Acoustical Society of America* , 73, 1266-1276.

Kiang, N. Y-S (1965) Discharge patterns of single fibers in the cat's auditory nerve. *MIT Research Monograph no. 35.* Cambridge, Mass: MIT Press.

Patterson, R.D. Nimmo-Smith, I., Weber, D.L. and Miroy, R. (1982) The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journal of the Acoustical Society of America,* 72, 1788-1803.

Pickett, J.M. and Revoile, S.G. (1983) Speech cue use by the hearing-impaired, in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences* Dordrecht: Foris, 233-237.

Pols, L.C.W. and Stoop, K. (1983) Confusions between Dutch consonants under various conditions of noise and reverberation. in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences.* Dordrecht: Foris, 515.

Risberg, A. (1983) Frequency discrimination and speech perception in the hearing impaired, in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences.* Dordrecht: Foris, 227-232.

Ritsma, R.J. and Horst, J.W. (1983) Speech perception and frequency selectivity, in A. Cohen and M.P.R. van den Broecke (Eds) *Abstracts of the Tenth International Congress of Phonetic Sciences.* Dordrecht: Foris, 239-244.

Sachs, M.B. , Young, E.D. and Millerm M.I. (1982) Encoding of speech features in the auditory nerve, in R. Carlson and B. Granström (Eds) *The Representation of Speech in the Peripheral Auditory System,* Amsterdam: Elsevier Biomedical Press, 115-130.

SYMPOSIUM 5: Phonetic Explanations in Phonology

Chairman: *J.J. Ohala, U.S.A.*

Panel members: *E. Fischer-Jørgensen, K. Kohler, L. Goldstein*

1. Chairman's opening remarks

The purpose of this symposium is to demonstrate that aspects of the structure and behavior of speech, i.e., its phonology, can be explained by reference to phonetic facts. Although this is a position with a respectable history (Passy 1890; Rousselot 1891; Grammont 1933) it has recently become controversial due to some well-articulated challenges which argue that there are important limitations on our ability to explain phonological facts (Lass 1980; Anderson 1981; Ladefoged 1983). A few preliminary remarks may help to eliminate some misunderstandings and thus to channel the arguments on this issue into productive directions.

First, it must be said that the commitment to search for explanations in any area of science is an act of faith. There is no guarantee that the search will be successful. In the history of science, failures to discover explanations are much more abundant than successes. Nevertheless, the appetite for understanding the workings of the universe, i.e., to reduce the mysterious to the familiar, has been stimulated in us by the classical Greek philosophers. They believed that by the application of rational means of inquiry it was possible to penetrate the seeming chaos of the universe and to discover a system - some small set of primitive entities and principles - from which the behavior of the universe could be derived. Other philosophical stances are possible, e.g., a belief in the inscrutability of the universe, that humans should try to achieve a harmonious union with the world rather than try to dissect it or to force it to reveal its secrets. It seems to me there is no way to show that one of these philosophical positions is better than the other. Thus it seems fruitless to attempt by persuasion and argument to make an advocate of one of these views renounce their faith.

Second, the explanations offered are *partial* explanations. As Hermann Paul (1880) noted, speech and language exist in three domains: the physical, the psychological, and the social. It follows that linguistic phenomena, like virtually every interesting topic of scientific study, exhibit behavior that has multiple determinants. Our situation in phonology, then, is similar to that of the medical researcher who ventures to show a causal relationship between heavy smoking and lung cancer. Lots of factors undoubtedly contribute to the development of lung cancer: diet, life style, environmental factors, heredity, etc. The causal connection between smoking and cancer can only be established on a statistical basis where all these other factors have been controlled, conceivably by being neutralized by varying randomly over the sample studied. This is, in fact, one of the ways such epidemiological studies are done. Thus, faced with an individual who was a heavy smoker and had lung cancer, can the epidemiologist say that the person contracted lung cancer because of the smoking? Not with as much certainty as one can establish for a large sample but nevertheless with a probability that is high enough that people will act on his advice (by giving up smoking, for example).

So, in our phonetically-based explanation for sound patterns we are on the safest ground when we find the same pattern in many diverse languages where the psychological and social forces vary in random ways. We can still offer plausible explanations for isolated cases but run a greater risk - an acceptable risk, many would maintain - of being wrong.

The third point is to address Roger Lass's (1980) criticism that the fact that we can't predict language change - sound change, as it concerns us - , means we are not a true science like the natural sciences such as physics and chemistry which are capable, according to him, of formulating deductive nomological (law-like) accounts of the behavior in their domain of interest. This claim is based on two misconceptions. Anyone who has ever done an experiment in a physics class knows that the measurements of phenomena never come out exactly the way the 'laws' say they should. This is true no matter how careful the operations are done. Physical laws, then, as many modern philosophers of science admit - and as the history of science demonstrates (where the laws get revised or replaced from time to time) - are a fiction. This does not mean that the theories of physics or chemistry don't explain phenomena; they do, but, as mentioned above, they are partial probabilistic or statistical explanations. Furthermore the ability of the physicist, for example, to predict the future is on a par with linguists' ability to predict the future. If a physicist were challenged to predict the trajectory of a billiard ball it would be necessary to impose a host of restrictions on the event: to specify that the ball would be hit (a social fact, as it were), how it would be hit (force, angle), the conditions of the respective surfaces of the stick, the ball, the table, the side cushions, etc., that there would be no air currents impinging on the ball, that the table would remain stationary (and not upset by an earthquake), etc., and even then the prediction would not be exact because 'unpredictable' factors almost always crop up, e.g., a piece of the ball being chipped away when it is hit. If a linguist had the luxury of being able to set down comparable conditions, e.g., monolingual speech community, no orthography, a listener learns the pronunciation of words from a single, non-redundant utterance, etc., then it might very well be possible to achieve some success in predicting language change. If physics and chemistry seem to do a better job at their predictions than linguists do, it may be because (a) their public relations effort is more intense, (b) they have been practising their science for a longer time and hence have accumulated more useful knowledge, etc. I don't think there are any *inherent* differences between the disciplines in their potential for achieving deductively-based (but not nomological) explanations for phenomena.

2. Summary of symposium papers

As the complete texts of the symposium papers were printed in Cohen and Van den Broecke (1983), only brief summaries of these papers will be given here. These and the discussion which follows are paraphrased; use of first person pronouns does not indicate direct quotations.

Ohala: Part of the 'lore' of phonology that has accumulated over the past two centuries or so has been an intuitive expectation for the favored direction of sound change. To the extent we can make give these intuitions an empirical, phonetic base, we should be able to do a better job at reconstructing linguistic history.

Sound changes of the sort, back velars > labials, e.g., Proto-Indo-European *gwi-wo- 'living' (cf. English *quick*) > Greek *bios*, can be explained by the acoustic-auditory similarity of the two sounds. But this leaves unexplained why the substitution is usually asymmetric, i.e., why labials do not often change to back velars. The answer, I suggest, may lie in the same perceptual factors which cause asymmetries in the confusion of letters of the alphabet in a visual identification task, where, e.g., 'E' is misidentified as 'F' more often than the reverse. Viewers may miss the 'foot' of the E and therefore report it as the letter which is graphically equivalent to the E minus the foot, i.e., an 'F'. Although when viewing an 'F' they might miss some of its features, too, these would not lead to an 'E' percept. Moreover, they are unlikely to 'add' missing features. Back velars may thus have some 'extra' acoustic feature which is absent in labials; failing to perceive this would lead listeners to think they had heard a labial, but labials would not be misidentified as back velars by the same process.

Dissimilation, which operates in the reverse direction of the more common assimilation, is due, I propose, to the listener invoking - inappropriately - perceptual rules he has developed to discount the effects of assimilation in speech. Normally these perceptual rules would operate to factor out non-distinctive phonetic features, e.g., the anticipatory labialization of the vowel before a labial or labialized consonant. In the case of Latin /K^{w}iŋk^{w}e/ > Italian /tʃiŋk^{w}e/, listeners apparently took the labialization on the first syllable to be entirely a manifestation of anticipatory assimilation to the second /k/ and they therefore factored it out of their lexical representation of the word, which, of course, would be the basis for their pronunciation of it. Various bits of support for this hypothesis can be found in the circumstances under which dissimilation occurs as well as in some laboratory studies (Ohala, 1981). The subtle prejudice that exists against dissimilation - since it seems to contradict the more common and phonetically motivated process of assimilation - can be dismissed: most assimilatory processes are the product of the speaker; dissimilation is the product of the listener 'second guessing' the speaker.

Fischer-Jørgensen: The traditional descriptive system for vowels, high vs. low, front vs. back, rounded vs. unrounded, and tense vs. lax, has usually been defined primarily in articulatory terms. This system has been criticized by Wood (1982), among others, for being articulatorily inaccurate or vague and for failing to account for certain interactions beteen vowels and consonants. Ladefoged (1980) finds a need for the traditional descriptive system but bases it on acoustic-auditory dimensions, not articulatory. I believe both of these views are wrong insofar as they discard the traditional articulatory-based system. Admittedly, the notion introduced by Daniel Jones (for pedagogical purposes) that it is the 'high point of the tongue' that is being described by these terms, 'front, low' etc., is open to criticism. But if we treat these terms as characterizing the overall location of the upper tongue surface with respect to the palate, then they are physiologically more accurate and they succeed, where the other systems fail, in giving an insightful characterization of such sound patterns as umlaut and vowel harmony. In these latter two processes, moreover, tongue position behaves independently of lip rounding, which they would not do if these had only acoustic-auditory correlates. Certainly acoustic-auditory correlates for vowels are needed e.g., they play an important role in vowels' role in sound symbolism (Fischer-Jørgensen, 1978), but not to the exclusion of the quite useful traditional articulatory correlates.

Goldstein: It is a very old idea that the 'seeds' of sound change may be found in synchronic variation in speech. But since the sound changes that recur in unrelated languages proceed in certain favored directions, it follows that synchronic variation is also constrained in its directionality. A proper account of the latter can explain much about the former. I have explored this notion as it applies to (unconditioned) vowel shifts, about which the following two generalizations seem valid: front vowels shift only along the height dimension and back vowels along the height and the front-back dimension. I hoped to find the explanation for these patterns in the constraints on the conversion from articulation to sound. To do this I used a synthesizer which takes articulatory parameters as input, including one which specifies tongue body center. A variety of vowel types were synthesized and for each type, 100 tokens in which the tongue body center was perturbed along a radius 2 mm from the target position for that vowel type. These perturbations were intended to represent the kind of variability vowels are subject to in normal speech. When plotted on the F_1 vs. F_2 space, the spectra of these perturbed vowels showed that, as predicted, the front vowels' acoustic perturbation was primarily along the 'height' dimension and the back vowels along both height and front-back. Central vowels showed no such directionality.

Kohler: I wish to stress the need to introduce the time dimension into phonological analyses and to illustrate this by reference to the so-called 'voiced- voiceless' opposition in obstruents. Although usually thought of as an a-temporal feature at a static point in a segment chain I suggest that it is better characterized as a 'fortis-lenis' distinction which can be defined by a single physiological property: greater vs. lesser physiological effort and thus by faster vs. slower articulatory movements. This results in a host of measurable phonetic differences: fortis consonants will have longer closure duration, shorter VC transitions and longer CV transitions, etc.; lenis consonants, the reverse. All of these patterns can be illustrated in a detailed phonetic analysis of minimal pairs such as the German *leiden* vs. *leiten*. The shorter closure duration of lenis consonants may make actual voicing more likely. The longer closure of fortis consonants leads to a higher oral pressure and a passive cessation of voicing. Of course, voicing may also be actively controlled, and languages may differ in the ways they implement this opposition. Sound changes of the sort p/b > b/β, etc., suggest that the fortis-lenis opposition may be preserved in spite of being shifted on the fortis- lenis continuum.

3. Discussion

Fischer-Jørgensen: (to Ohala) Why – in your paper – do you reject explanations for sound change based on 'ease of articulation'?

Ohala: It is a matter of research strategy. We know so little about the 'effort' involved in articulating sounds that it is a notion that is *too easy* to invoke and, frankly, it is a notion that has often been abused. We should exhaust the explanatory principles that are known and testable before using this 'wild card'.

Fischer-Jørgensen: (to Ohala) I think your account of dissimilation is an interesting one but I do not think it can explain all types of dissimilation, e.g. those involving n/l and r/l. In those cases too many similar sounds in a word --often these are borrowed words-- confuses the listener.

Kohler: (to Ohala) I agree with that. Although I find it neat that you associate assimilation with articulation and dissimilation with perception, I do not think it can handle all cases. Examples of the sort French *marbre* to English *marble* seem to point to the articulatory difficulty of pronouncing (what must have been) two rolled *r* in a short interval.

Goldstein: (to Ohala) One thing I like about your approach is that, insofar as the directionality of sound change is explained, there is no suggestion made that the later state is any better than the original, that is, that it somehow improves the language.

Ohala: (to Goldstein) If you had specified a greater perturbation from the vowel targets, say 3 mm instead of 2 mm, would you have obtained more fronting of the back vowels?

Goldstein: A little more, but not an appreciable amount.

Fischer-Jørgensen: (to Kohler) The different rates of movement you found for fortis and lenis obstruents in VC transitions is quite well documented, but I'm not so sure about the CV transition. Why, in any case, should it be slower for fortis than lenis? The evidence for this when C is initial is at best equivocal and at worst does not support your claim.

Kohler: Part of the basis for that claim was the finding by Öhman in 1966 of greater EMG activity in the lips following /b/ than /p/, however, I should perhaps have made a more careful distinction between initial CV transitions and those which are intervocalic. It was the latter that formed the bulk of my studies.

Tore Janson: Although I am in general sympathy with the purpose of this symposium, I am surprised at the title 'Phonetic explanation in phonology'. This is a contradiction in terms: it is *phonology* that should provide generalizations or explanations for phonetic facts. If phonetics and phonology do not always dovetail, it may be that the phonologist has disregarded certain phonetic facts or that phonological theories are imperfect. But the basic facts, which need to be explained, are phonetic.

Kohler and Fischer-Jørgensen offer solutions to the perennial problem of the relation between phonetic substance and phonological features. But trying to find a set of features which are universal, phonetically realistic, able to capture the distinctions made in all languages, etc., might be conflicting aims. A basic universal set of features may be possible but these will have to be supplemented by 'extra' features as needed.

Ohala's and Goldstein's contributions are explanations for sound change, not strictly speaking, explanation in phonology. Phonologists often cite data from sound change but this data is at heart phonetic. Also, it should be kept in mind that what they address is sound change at the very initial stages. Subsequently sound change becomes embedded in the grammatical system of the language and its further progress depends more on morphology than phonology. A desirable (if long-range) goal for phoneticians working on sound change would be rank ordering of potential changes from more to less expected.

Lehiste: (to Kohler) I question your assumption that V + C is some kind of basic building block in phonological structure and acts as a whole in manifesting the fortis/lenis distinction. How would this accommodate fortis/lenis distinctions in syllable initial position?

Kohler: The pre-vocalic consonants behave in a different way from the postvocalic ones, so it will be necessary to give a different account for the former.

Kenneth Stevens: À propos of Janson's remarks, I think it can be said that phonetics has at least provided phonology with a set of distinctive features. What we are doing here today is trying to discover *relations* between those features. E.g., there is a sense in which /u/ is a *labial*, in which /i/ is a *coronal* or *palatal*, etc. We need to find out how one feature can help another and (in sound change) gradually take over for another.

G. Heike: In accord with Kohler's findings, 23 years ago I studied the German dialect of Cologne and found that in intervocalic position there was no difference between stops in voicing, intensity of burst, or aspiration but they did differ in duration. Changing this duration by tape splicing could change a lenis stop into fortis.

À propos of Goldstein's study, I would prefer to see more work with *dynamic* articulatory models to answer phonological questions. The German 'r' in post-vocalic position has many variants: from a uvular trill to a glide. All variants have in common a dorso-velar closing gesture. The dynamic gestures in speech would make a better focus for phonological theory than discrete segments.

Björn Lindblom: (To Goldstein) Some work that Johan Sundberg and I did some years back with an articulatory model also revealed an asymmetry in the propensity of vowels to shift along the front-back dimension. We observed from X-rays that the physiological rest position has the tongue in a more or less front position. Thus when we made peripheral vowels [i, e, a, o, u] and let them glide to the neutral position, the front vowels changed along the height dimension but the back vowels exhibited fronting.

René Gsell: On the subject of phonetic explanation for sound change, it should be noted that the essence of language is change. The code of the speaker may differ from the code of the hearer and so there must be a shifting between the realization of one and the identification of the other. Thus there arise allophones in synchrony which leads to phonological change diachronically. But the selection of one variant as the norm is a social fact and cannot be explained phonetically. Phonetics only explains the 'birth' of the variant, not its subsequent fate in the language.

(to Ohala) The shift of labialized velars to labials may be due to acoustic-auditory reasons but I do not agree that the shift of palatalized labials to apicals has the same explanation.

Vicky Fromkin: From the title of this symposium I had hoped that there would be a great deal of discussion on phonology. I think we've had very little discussion on phonology--only some *aspects* of phonology. There are some tremendously exciting theoretical approaches being proposed currently. It would be interesting to find out whether phonetic evidence can be found for some of these, e.g., metrical phonology and autosegmental phonology.

Ohala: The evidence discussed for the spreading of features throughout a word does have relevance to autosegmental notation. There is, in fact, quite a bit of phonetic literature already accumulated on this and other points which have a bearing on these newer approaches, e.g., Öhman (1966), but there seems to be little interest or attention given to it by the proponents. The phonetic literature is quite vast and the data in it obtained through much hard work and ingenuity; it deserves to be mined for its phonological relevance.

References

Anderson, S. (1981).Why phonology isn't 'natural'. *Linguistic Inquiry* **12**, 493-589.

Cohen, A. and Van den Broecke, M.P.R., Eds. (1983) *Abstracts of the Tenth International Congress of Phonetic Sciences.* Dordrecht: Foris.

Fischer-Jørgensen, E. (1978). On the universal character of phonetic symbolism with special reference to vowels. *Studia Linguistica,* **31**, 80-90.

Grammont, N. (1933). *Traité de phonétique.* Paris: Librairie Delagrave.

Ladefoged P. (1980). What are linguistic sounds made of? *Language,* **55**, 485-502.

Ladefoged, P. (1983). The limits of biological explanations in phonetics. In *Abstracts of the Tenth International Congress of Phonetic Sciences, (A. Cohen and M.P.R. Van den Broecke, Eds.) pp. 31-37 Dordrecht: Foris.*

Lass, R. (1980). On explaining language change Cambridge: Cambridge University Press.

Ohala, J.J. (1981) The listener as a source of sound change. In: *Papers from the Parasession in Language and Behavior,* (C.S. Masek, R.A. Hendrick, and M.F. Miller, Eds.), pp. 178-203. Chicago: Chicago Linguistic Society.

Öhman, S. (1966). Coarticulation in VCV utterances: spectrographic measurements. *J. Acous. Soc. Am.* **39**, 151-168.

Passy, P. (1980). *Étude sur les changements phonétiques.* Paris: Librairie Firmin-Didot.

Paul, H. (1880). *Prinzipien der Sprachgeschichte.* Halle: Niemeyer.

Rousselot, P. (1891). *Les modifications phonétiques du language.* Paris: H. Welter.

Wood, S. (1982). X-ray and model studies in vowel articulation. *Working Papers, Lund.* **23**.A, 1-191.

SYMPOSIUM 6: Human and Automatic Speech Recognition

Chairman: *D.H. Klatt, Cambridge, Mass. U.S.A.*

Panel members: *V.W. Zue, S.M. Marcus, M. Liberman, R. de Mori*

In general terms, the session sought to answer two broad questions: (1) Do current speech recognition strategies have anything to tell us about speech perception, and (2) can current theories and data concerning speech perception guide and improve the performance of automatic speech recognition systems? The panel consisted primarily of scientists involved in building speech recognition systems, although several panel members have also worked on problems in speech perception. The audience, on the other hand, was biased toward greater expertise in the areas of speech perception, as evidenced by the questions asked during the discussion period.

The chairman began the session by asking a number of fundamental questions. For example, is one of the first stages of the process leading to lexical hypothesization one in which a phonetic analysis is performed, or is the acoustic input matched directly with acoustic patterns for familiar words? There are problems with either view. Performing a phonetic analysis means making decisions and discarding information that may have been useful at later stages in the process. The inevitable transcription errors that result are much harder to correct during lexical search. Furthermore, it is difficult to specify the nature of a good phonetic representation - the best one is perhaps so detailed that it is simply a recoding of the input rather than an information reduction transformation. On the other hand, if word perception is direct, one has to account for speaker differences and dialect differences when matching input with acoustic patterns for words, and one must invoke separate analysis procedures for novel words.

Additional problems face builders of speech recognition systems and models of speech perception. Are some phonetic decisions easier? If so, should one use these robust cues to narrow the search? Victor Zue described a system employing this approach. However, during the question period, he was challenged on the performance of such a system (he said no data are available as yet) and on how to overcome an error in the initial partial transcription (he said lexical redundancy may permit detection and correction of some errors). He was also asked whether such an approach is practical in continuous speech where there is less certainty as to the locations of word beginnings and endings. Zue responded that he was quite optimistic as to the feasibility of applying the approach to continuous speech.

The potential advantage of search reduction via robust cues is that the remaining lexical candidates may be distinguished in a sort of hypothesis-and-test verification scheme where the acoustic expectations can be narrowly specified because one is assuming a particular phonetic context. However, in the discussion, it was pointed out that the advantages of verification can be achieved in a bottom-up fashion by precompiling this sort of knowledge into an acoustic decoding network for words and word sequences (e.g. LAFS). When a member of the audience asked exactly how this might be done, Mark Liberman pointed out the practical difficulties by supposing that one had a machine that stored the sentence response for every possible ten-second digitized waveform (i.e. the number of different responses would be two raised to the power of 10,000 samples/second times 8 bits/sample times ten seconds). While such a machine can be conceptualized, it will never be built within our universe.

Renato de Mori described an elaborate system for speech recognition involving many levels of representation and multiple cues leading to decisions at any level. A member of the audience challenged whether this system, or any other speech recognition device, used strategies as complex as we know to be necessary from the literature on multiple cues to phonetic contrasts (such as the voiced/voiceless distinction where Lisker has catalogued over a dozen distinct cues). The panel readily admitted that current systems do not approach the sophistication required to take advantage of this knowledge, in part because it is so difficult to program strategies that involve interacting decisions (a change at one place in the program has ramifications, often unexpected, at many other locations in the code) and in part because the various constants needed to optimize such a strategy are usually not given and require incredible effort to discover from data.

Mark Liberman stressed the importance of extracting an appropriate representation of speech in order to achieve better phonetic/lexical identification performance than has been obtained to date. He indicated that formant trajectories are good candidates, and that new strategies may result in improved formant-tracking performance, but that our knowledge in this fundamental area is still quite primitive.

Stephen Marcus described an approach to speech recognition where words consist of unordered sets of spectral changes. He emphasized how remarkably well such a system works, meaning that spectral change (and/or phonetic change) is a concise summary of the most important aspects of the acoustic pattern for a word. However, he was quick to admit that this is not the complete story, and order information is needed to distinguish many words.

A member of the audience, Adrian Fourcin, drew our attention to how the infant begins to understand language, and suggested that we might build machines that mimic this process. The panel was able to pick up on this point and stress how little we know about the role of learning (versus innateness) in speech perception, or how to implement strategies that tune themselves from experience as well as discover new rules from experience.

A member of the audience, Mac Pickett, asked whether speech recognition has progressed as a science to the point where devices could be built to be used as aids for the hearing handicapped. The panel was unanimous in concluding that the systems entering the marketplace have very limited capabilities, particularly with respect to dealing with many speakers or dealing with large vocabulary continuous speech. Only if a small vocabulary isolated word recognition capability was useful would it be worthwhile to mount an effort in this direction.

Finally, a member of the audience, Karl Eric Spens, asked whether we should worry about the potential misuse of speech recognition technology, particularly in the area of surveillance and invasion of privacy. Mark Liberman responded that the technology is too primitive, as yet, to be really concerned. However, it is clear that we, as scientists most closely tied to this technology, have a duty to inform the public of the dangers as they arise, or before!

Section Papers

Section 1
Acoustic Manifestations of Speech

Die Hochdeutschen und Schweizerhochdeutschen Betonten Monophthonge

A.K. Iivonen
Helsinki, Finland

1. Ziel der Untersuchung

Als Teil eines Gesamtthemas über die regionale Variation der hochdeutschen Vokalqualitäten werden hier 15 hauptbetonte Monophthonge in einigen schweizerhochdeutschen und hochdeutschen Idiolekten auf Grund der F_1/F_2-Representationen diskutiert.

2. Material, Sprecher, Formantenmessung

Dreierlei Materialanordnungen wurden angewandt:

1. Wörter vom Typ/bVtən/ (= Minimalpaare) wurden in dem Satzrahmen *Sag ... nochmal.* je 10mal wiederholt. Sprecher MR). Zehn Belege pro Vokaltyp wurden analysiert.
2. Isolierte zweisilbige Wörter vom Typ /K(K)VK(K)ən/ wurden in Gruppen von 12 Wörtern vorgelesen, wobei das erste und das letzte Wort aus dem Material ausgeschlossen wurden (Sprecher MM und AR). Fünf Belege pro Vokaltyp wurden analysiert.
3. Isolierte Wörter desselben Typ, wie in 2 wurden so vorgelesen, dass eine deutliche Pause zwischen den Wörtern vorlag (sonstige Sprecher). Fünf Belege pro Vokaltyp.

Der erste Versuchtyp wurde angeordnet, um zu zeigen, wie gross die Vokalvariation in den Wiederholungen desselben Wortes im Satzrahmen ist und wie die Konstellation der Vokale in Minimalpaaren auf der F1/F2-Karte aussieht (vgl. Iivonen, 1982). Die zweite Materialanordnung und die entsprechende Analyses zeigten, dass eine Reduktion der Vokalqualität deswegen möglich ist, weil die Versuchsperson ihre eigene Sprechgeschwindigkeit beim Aufzählen der Einzelwörter bestimmen kann (s. unten). Im dritten Versuch zeigte der Versuchsleiter dem Sprecher in einem Aufnahmestudio einzelne Zettel in gemischter Reihenfolge. Auf jedem Zettel stand ein Versuchswort. Zwischen den Wörtern wurden dabei Pausen gelegt, um den Sprecher zu veranlassen, während der Pause zu atmen. Wegen der Unterschiede in den Versuchs-anordnungen sind nur die Resultate der Sprecher GT, RT, HPS, US und SZ völlig vergleichbar.

Der Instruktion nach sollten die Gewährsleute formal, doch natürlich sprechen. Sie wurden gebeten, Hochdeutsch zu sprechen, aber einen Stil anzuwenden, den sie selbst in einer realen Situation benutzen würden. Alle Sprecher waren Akademiker (s. Tabelle der Personalien).

Die Zeitfunktion der Wörter wurde durch einen A/D-Konverter digitalisiert, in einem Mikrocomputer gespeichert und der FFT-Analyse unterzogen (s. Karjalainen, 1980). Der zeitliche Mittelpunkt der Vokale wurde bestimmt und ein 30 ms langes Segment wurde um den Mittelpunkt herausgewählt. Für dieses Segment wurde ein FFT-Kurzzeitspektrum bis 4000 Hz berechnet (dazu wurde ein Hamming-Fenster benutzt). Die Formanten F_1, F_2 und F_3 wurden mit dem Kursor auf dem Bildschirm lokalisiert und ausgemessen, wobei die eventuelle Asymmetrie des Formanten mitberücksichtigt wurde. Hinter der Motivierung dieses Verfahrens liegt eine Reihe von Vorversuchen (u.a. Iivonen, 1982, 1983). Die auf Grund von fünf Belegen pro Vokaltyp gewonnenen Mittelwerte der Formanten F_1 und F_2 wurden auf einem Koordinatensystem mit logarithmisierten Skalen aufgeführt (s. Abb. 1 und 2; Viereck = lange, Kreis = kurze Vokale).

Die Vokale des Sprechers MR sind jedoch abweichend behandelt worden. Sie wurden mit der LPC-Methode in Kiel (Institut für Phonetik der Universität Kiel; s. Iivonen, 1982) analysiert, und ihre Mittelwerte wurden von je 10 Belegen pro Vokaltyp berechnet.

Die phonetisch-phonologischen Zeichen der Abbildungen entsprechen der normativen Notation (Mangold und Grebe, 1971; Kohler, 1977:176). nicht den tatsächlichen phonetischen Werten der Sprecher.

3. Resultate

In der heutigen Forschungslage können wir kaum mit Sicherheit die idiolektalen Züge vom Mundarteinfluss auf Grund der Formantenkarte unterscheiden. Die Normalisierung des Vokaltrakteffekts ist auch noch nicht befriedigend gelöst. Wenn aber viele idiolektale Formantenkarten vorliegen, können wir durch Vergleich die relativen Positionen der Vokaltypen feststellen und auf Grund der Unterschiede und Ähnlichkeiten der relativen Positionen zwischen den Idiolekten Schlüsse ziehen, ob die Ähnlichkeiten innerhalb je einer regionalen Sprechergruppe als regionale Züge interpretiert werden können.

Das Schweizerhochdeutsche unterscheidet sich bewusst vom Hochdeutschen. Eine völlige Übereinstimmung mit der 'Hochlautung' ist also nicht zu erwarten (vgl. Boesch, 1957, Panizollo, 1982).

Auf Grund der Abbildungen 1 (Hochdt.) und 2 (Schweizerhd.) scheinen besonders folgende Einzelheiten interessant zu sein:

1. der Zug gespannt/ungespannt der langen/kurzen Vokale,
2. die Lage des F_1 in der Reihe /i, y:, u:/,
3. die Lage des F_1 in der Reihe /e:, ø:, o:/,
4. die Lage einiger einzelnen Vokaltypen, d.h. die des ɛ:/ in bezug auf /e:/, des /a:/ in bezug auf /a/, des /ø:/ in bezug auf /Y/ und des /ɔ/ in bezug auf die beiden a-Vokale.

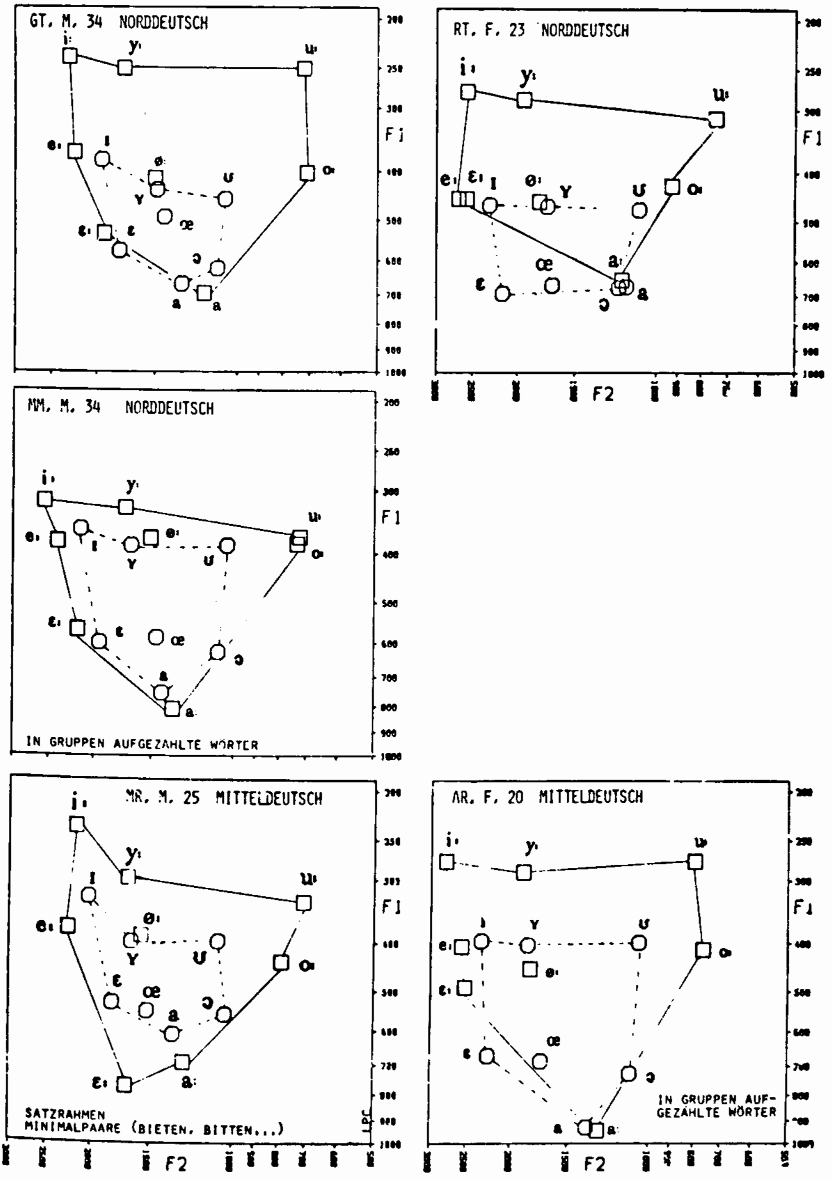


Abb. 1. F_1/F_2-Repräsentation der Vokale der hochdeutschen Sprecher. M = Mann; F = Frau; □ = lange Vokale; ○ = kurze Vokale.

Die norddeutschen Sprecher GT, RT und MM realisieren den Unterschied peripher/zentral sehr deutlich. Die gespannten (langen) Vokale bilden ein Peripheriesystem auf der Formantenkarte, während die ungespannten (kurzen) Vokale eine zentralisierte Position annehmen und deswegen ein geschrumpftes System bilden. Dasselbe gilt auch für die mitteldeutschen

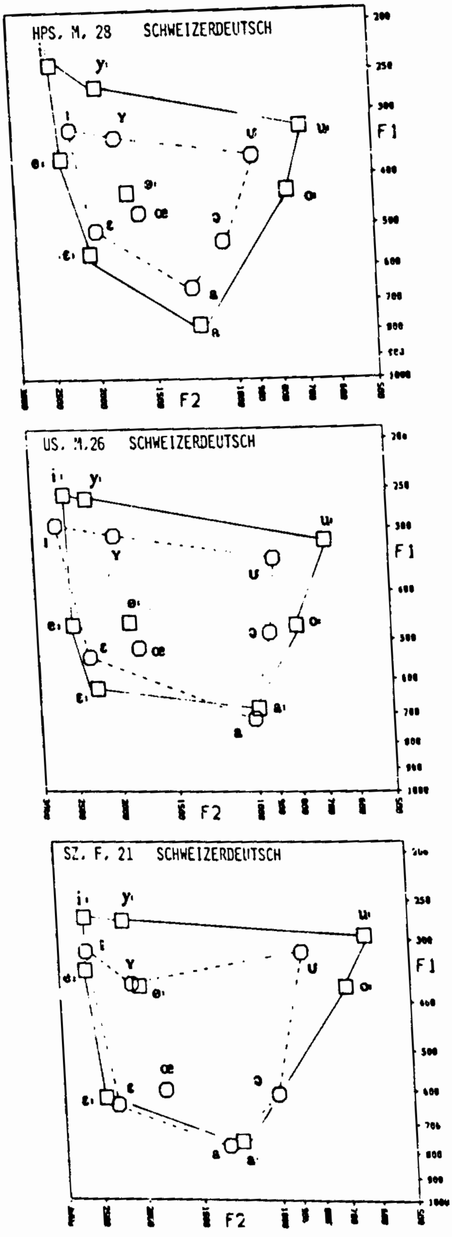


Abb. 2. F_1/F_2-Repräsentation der Vokale der schweizerischen Sprecher.

Sprecher MR und AR. Bei MM kommt der Unterschied in der geschlossenen Reihe nicht so deutlich zum Vorschein. Meine Interpretation ist jedoch, dass die ziemlich offene Lage der /iː, yː, uː/ bei MM durch eine Reduktion der Qualität bei hoher Sprechgeschwindigkeit während der Aufzählung der Einzelwörter erklärbar ist. AR hat dagegen die einzelnen Wörter viel langsamer und deutlicher gesprochen.

Die schweizerdeutschen Sprecher weichen voneinander und von den sonstigen Sprechern ab. US macht keinen grossen qualitativen Unterschied zwischen den langen und kurzen Vokalen, sondern scheint mehr nur die universelle Tendenz zu verfolgen: die kurzen Vokale sind etwas zentralisierter im Vergleich zu den langen Vokalen. In der Literatur wird behauptet, dies sei ein genereller Zug im Schweizerdeutschen (z.B. Keller, 1961:44). SZ produziert jedoch einen deutlichen Unterschied zwischen den Vokaltypen in den Paaren /eː, ɛ/, /øː œ/ und /oː, ɔ/, weil ihre Reihe /eː, øː, oː/ ziemlich geschlossen ist. Teilweise handelt es sich nicht um eine Zentralisierung, sondern um eine F_1-Distanzierung: der Öffnungsgrad der Vokale eines bezüglichen Paares (z.B. /yː, Y/) variiert. Teilweise wird der Öffnungsgrad nicht bedeutend getroffen, sondern der Formant F_2 (/uː, U/).

Bei HPS ist die Konstellation der Vokaltypen auf der F_1/F_2-Repräsenta tion sehr symmetrisch und er kommt nahe den hochdeutschen Sprechern. Die Zentralisierung ist bei den meisten kurzen Vokalen deutlich (sogar bei /a/). /øː/ und œ/ sind aber sehr nahe beieinander (weil /øː/ des HPS ziemlich offen ist).

Die Vokale /eː, øː, oː/ sind bei den schweizerdeutschen Sprechern US und HPS ziemlich offen, während die anderen Sprecher diese Reihe halbgeschlossen produzieren.

Was die Lage des /ɛː/ angeht, kann man vier verschiedene Möglichkeiten feststellen:

1. /eː/ und /ɛː/ sind qualitativ praktisch gleich (RT),
2. /eː/ und /ɛː/ unterscheiden sich voneinander nach der Standardnorm (GT, MM HPS, SZ),
3. /ɛ/ liegt zwischen dem /eː/ und dem normativen /ɛ/ (AR) oder
4. /ɛː/ wird 'zu offen' realisiert (MR und US), so dass eine Hyperartikulation schon vorliegt (MR, phonetisch [æ]).

Die Vokale /a/ und /aː/ sind bei den meisten Versuchspersonen qualitativ ähnlich. Nur US ist eine Ausnahme: seine a-Vokale liegen wesentlich weiter hinten (F_2 = ca. 1000 Hz). Auditiv haben seine a-Vokale ausserdem einen labialen Charakter, so dass sie besser durch das Zeichen [ɒ] angegeben werden könnten. Die exzeptionell hohen Werte des F2 seiner Vokale sind wahrscheinlich durch seine kleine Statur (und deswegen durch seinen kurzen Vokaltrakt) erklärbar.

Die Vokale /øː/ und /Y/ haben bei den Sprechern GT, RT, MM, MR und SZ beinahe dieselbe Position. Weiter voneinander liegen diese Vokale bei AR, HPS und US.

Besonders zu beachten ist, dass bei den nördlichsten Sprechern GT und RT die Lage des /ɔ/sehr nahe den a-Vokalen liegt. Bei RT ist die F_1/F_2-Repräsentation aller drei Vokale praktisch identisch. Die Kurzzeitspektra zeigen jedoch, dass die Intensität des F_2 in /ə/ systematisch dominierend ist, während F_2 /a/ und /a:/ dominiert. Auditiv (nach meiner eigenen Wahrnehmung) sind /a/ und /ɔ/ des Sprechers jedoch unterschiedlich. Dieser Umstand scheint nachzuweisen, dass die F_1/F_2-Repräsentation nicht völlig imstande ist, die systematische Vokalqualität zu zeigen. F_3 betrug in /a/ 2603 in /a:/ 2509 und in /ɔ/ 2564 Hz im Durchschnitt.

Dank

Ich bin folgenden Personen und Institutionen für ihren Beitrag zur vorliegenden Arbeit dankbar. Dr. Klaus Kohler und seinen Mitarbeitern Prof. Stefan Sonderegger, Frau Dr. Sonja Spörri-Bütler, Herrn Hanspeter Schifferle, Herrn Jouko Mikonsaari; dem Deutschen Akademischen Austauschdienst, der Stiftung Pro Helvetia; den Gewährsleuten.

References

Boesch, B. (1957). *Die Aussprache des Hochdeutschen in der Schweiz*. Zürich: Schweizer Spiegel.

Iivonen, A. (1982). Zur regionalen Variation der Vokalqualitäten des Hochdeutschen (dt. Zusammenfassung). In: *Publications of the Department of Phonetics, University of Helsinki*, **53**, 55-88.

Iivonen, A. (1983). Zur Frage der regionalen Variation der Hochdeutschen Vokale. *Neuphilologische Mitteilungen* **LXXXIV**, 45-52.

Karjalainen, M. (1980). Design of a microprocessor-based system for speech analysis. *Tutkimus raportti no. 9.* Tampereen Teknillinen Korkeakoulu.

Keller, R.E. (1961). *German Dialects*. Manchester: University Press.

Kohler, K. (1977). *Einführung in die Phonetik des Deutschen*. Berlin: Erich Schmidt.

Mangold, M. (mit P. Grebe) (1962). Duden-Aussprachewörterbuch. *Der Große Duden, Bd.* **6**. Mannheim.

Panizzolo, P. (1982). Die schweizerische Variante des Hochdeutschen. *Deutsche Dialektographie* **108**, Marburg: Elwert.

Temporal Control at the Utterance Level in German

K.J. Kohler
Kiel, FRG

1. Introduction

In German, the duration of a stressed syllable is inversely related to the number of the following unstressed syllables in a rhythmic foot structure (Kohler et al. 1982, Kohler 1982), not to the total number of following syllables in the sentence, as was suggested by Lindblom and Rapp (1973) for Swedish, This tendency is constrained by articulatory complexity (Kohler 1982) and by a ceiling effect in the number of syllables, more than two having little influence. Temporal adjustment within the foot is both left-to- right, as a function of the number of unstressed syllables, and right-to-left, as a function of the complexity of the stressed syllable (Kohler 1982). Further factors that might influence the stress-timing of German are (1) F_0 (e.g. in *EINe geZEIGT* vs. *EINe gezeigt*), (2) utterance-final position (e.g. in *er zeigt EINe/EINige* vs. *er wird EINe/EINige zeigen*), (3) complementary right-to-left adjustments between two feet, depending on the length of the first one (e.g. *zeigen* in *EINige zeigen* vs. *EINe zeigen*), (4) coarticulation (e.g. articulatory control in *-e* before *be-* vs. *-e* before *ge-*), (5) word structure (e.g. in *EINe ge-* vs. *EINige*) (About the use of capitals see below.)

2. Experiment

To investigate the hypotheses concerning these factors a production experiment with the following sentences was run:

a. er zeigt EINe/EINige
b. er wird EINe/EINige ZEIGen/beNENNen (zeigen/benennen)
c er hat EINe/EINige geZEIGT/beNANNT (gezeigt/benannt).

Capitals indicate sentence stresses, realised as falling pitch nuclei on the last capitalized word in each sentence and as level or rising pitches about the starting points of the falls on the preceding capitalized word. Each of the 18 sentences was typed on a separate card, and these cards were randomized. Two speakers (the author, KK and CW ♀, phonetically naive) produced the whole set 11 and 10 times, respectively. Microphone and laryngograph (Fourcin) signals were recorded simultaneously on a two-track Revox A-77 tape recorder (19 cm/s) in the studio of the Kiel Phonetics Department. The following sections of the audio time signals were determined and measured with the help of the DISPLAY programme (Barry et al. 1982): stem *ein*; first unstressed syllable *e/i*; second unstressed syllable *ge* in *einige*; prefix *be*, *ge*.

The end of each utterance was located either at the disappearance of periodicity (*zeigen, nennen, eine, einige*) or at the beginning of the stop release (*zeigt, nannt*). The data of each speaker were subjected to ANOVAs with fixed effects.

3. Results

1. F_0
   The items with double stress have longer durations of *EIN* than the corresponding single-stressed ones in both speakers and in all cases. 3-way ANOVAs of the *EIN* duration in the whole corpus (except for final *EINe/EINige*) with the factors 'word *eine/einige*', 'unstressed syllable after *eine/einige*' and 'stress' yield a high significance for 'stress' in both speakers. A rising F_0 (CW) or a high level one (KK) accompanies this timing difference.
2. Utterance final position
   Final *EINe/EINige* are considerably longer (in all their syllable components) than the non-final counterparts (i.e. single-stressed and with no prefixes following). Two-way ANOVAs with the factors 'word *eine/enige*' and 'sentence position' show high significance for the latter in all the dependent variables and in both speakers.
3. Right-to-left adjustments
   In KK's data there is a negative correlation between the duration of final *zeigen* or *zeigt* or *NENNen* or *NANNT* and the duration of the preceding foot as determined by the complexity of *EINe* vs. *EINige*. This is in accord with the findings from several speakers including KK, referred to in Kohler (1982). In this corpus the temporal adjustment between two feet does, however, show a combined influence of 'stress' and ' vowel quantity': a long vowel is more expandable under stress than a short one, and a short vowel is less reducible under lack of stress than a long one so that in these cases the timing control by the duration of *EINe/EINige* is less effective. CW has no such temporal regulation in this material although she showed it in the data of Kohler (1982). The mechanism behind this right-to-left adjustment of succeeding feet is obviously a tendency towards foot isochrony, subsequently correcting deviations caused by too complex articulations (cf. also Lehiste 1980).
4. Coarticulation
   The unstressed vowel before *be* is shorter than before *ge*. Two-way ANOVAs for the main effects 'stress' and 'following syllable *be/ge*' on the final vowels of *EINe* and *EINige*, respectively, prove '*be/ge*' to be highly significant in both speakers. This difference can be explained by coarticulation: the labial closing movement is initiated during the vowel because the lips are not involved in its production; this does, however, not apply to the dorsal gesture. On the other hand, *be* is always longer than *ge* because the distance the articulators have to move from the closure into the vowel is shorter for *ge*. In the sequence *EINige ge-* the first *ge* is longer than the second because the elevation of the dorsum is more sluggish than that of the apex. CW begins this slowing down of the movement even as early as the vowel preceding the first *ge*: *i* is longer than in *EINige be-*, which is in turn compensated for by a shortening of *EIN*, according to the negative correlation between stem and ending (cf. Kohler 1982). This adjustment is also found in KK's *EINe be-/ge-*, resulting in a highly significant duration difference of *EIN*, complementary to that of *e* which is in turn complementary to that of *be-/ge-*, with temporal equalization at the word level *EINe*.
5. Word structure
   In the data set 'EINe geZEIGT (gezeigt)/EINige ZEIGen (zeigen)' two-way ANOVAs for the dependent variable 'duration of the stem *ein*' show high significance in the main effects 'stress' and 'word boundary' for both speakers and no significant interactions for KK. The same applies to 'duration of *eine/eini-*',where CW, too, has no significant interactions. Neither speaker has a significant duration difference in the first unstressed syllable. There is thus a clear tendency in both speakers to signal the different word structures in durational patterns. This is further supported by a slower F_0 fall for *EINi-* of *EINige zeigen* vs. *EINe* of *EINe gezeigt* in both speakers.

4. Discussion

The timing of utterances in German is controlled at a number of levels. The tendency towards foot isochrony manifests itself not only in the well-known phenomenon of weak forms but also in a foot-internal negative duration adjustment according to the articulatory complexity of the beat syllable and to the number of unstressed syllables following it. Coarticulation further influences this temporal organization by producing duration effects which enter into the negative intra-foot correlations and may thus have repercussions over quite long signal stretches. Temporal compression is, however, limited by the complexity of the articulatory movements to be carried out in segmental types and in their combinations. If no elision occurs the regular foot timing cannot be maintained beyond these constraints. Isochrony is further disturbed by F_0, by utterance position and by word structuring. On the other hand, the principle still manifests itself above the foot level in negative inter-foot adjustments which compensate post hoc for rhythmic deviations. There is thus strong support for a stress-timing structure intervening in German between chains of syllables and utterance prosodies.

References

Barry, W.J., van Dommelen, W., Janßen, H., Kohler, K.J. Schäfer, K., Thon, W. and Timmermann, G. (1982). *Phonetic Data Processing at Kiel University. Developments an Applications.* Arbeidsberichte des Instituts für Phonetik der Unversität Kiel, **18**.

Kohler, K.J. (1982). Rhythmus im Deutschen. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel*, **19**, 89-105.

Kohler K.J., Krutzmann, U., Reetz, H. and Timmermann, G. (1982). Sprachliche Determinanten der signalphonetischen Dauer. *Arbeitsberichte des Instituts für Phonetik der Universität Kiel*, **17**, 1-35.

Lehiste, I. (1980). Interaction between test word duration and length of utterance. In: *The Melody of Language* (L.R. Waugh and C.H. van Schooneveld (eds.)), pp. 169-176. Baltimore: University Park Press.

Lindblom, B. and Rapp, K. (1973). Some temporal regularities of spoken Swedish. *Papers from the Institute of Linguistics, University of Stockholm*, **21**.

F_1 Locus and Place of Articulation

M. Pettorino and A. Giannini
Napels, Italy

1. Introduction

The aim of this research is to verify if the F_1 locus has a distinctive function in relation to the place of articulation of the consonant. The locus theory, formulated in the fifties by Cooper, Delattre, Liberman, Borst and Gerstman (1952) and Delattre, Liberman and Cooper (1955) on the basis of research with synthetic ṣpeech restricted to the labial, dental and velar stops, was formulated in order to account for the fact that in a CV context all the transitions of the second formants of the vowels tend to one point on the frequency scale as a function of the place of articulation of the consonant. The F_1 transition, on the contrary, is connected with the manner of articulation of the consonant, because it always shows the same rising trend for all the CV contexts depending on its voiced character. The onset of this transition is fixed at about 240 Hz but in order to get the best degree of perceptibility of the three consonants under consideration, this value can be lowered on the frequency scale. We must notice that notwithstanding the restriction to the labial, dental and velar places of articulation, Delattre (1955:773) and Stevens and House (1956:584) tend to generalize, the former affirming that the locus of the first formant is at 240 Hz 'for all the voiced stops', the latter fixing F_1 locus at zero Hz 'for all stop consonants'.

According to the acoustic theory developed by Chiba and Kajiyama (1958) and Fant (1960) these statements are explained by the fact that these places of articulation occur in proximity to an anti-node of velocity, all of them taking place in the oral cavity. In particular Fant (1960:210) finds that F_1 decreases if the oral constriction is narrowed, in accordance with the statements of the above mentioned works, but on the other hand, that F_1 increases if the constriction located in the posterior half of the vocal tract is narrowed.

From the mid-sixties a series of studies tend to criticize Delattre's observations about F_2 transitions, because in these studies no unique locus for every consonantal class is found. In particular Öhman (1966:167) underlines that 'the terminal frequencies of the formants in VCV utterances depend not only on the consonants but on the entire vowel context', so he concludes that 'the stop-consonant loci are, therfore, not unique'. In our opinion this conclusion is not justified because the apparent plurality of loci, found by spectrographic analysis of eal speech, is only the effect of the plurality of places of articulation within the same consonantal class. However, the distinctive function of F_1 transitions, as cues of manner of articulation has never been questioned. These experimental studies have been confined to the functional role that the oral articulations play inside the systems of western languages. However, a language system having a large variety of pharyngeal and glottal articulations, as for instance Arabic, allows the verification of the importance of F_1 transitions for the identification of the place of articulation of the consonant.

2. Analysis

The spectrographic and radiographic data gathered in our previous study Giannini and Pettorino (1982) on Iraqi Arabic showed that the shifting of place of articulation from the velarized dental /ḷ/ to the pharyngealized dentals /ṭḍṣẓ/ is reflected by the raising of F_1 locus from 250 Hz to 600 Hz, whereas F_2 locus remains in both cases at 1000 Hz. Furthermore, the difference between the uvular fricatives /ḫġ/ and the pharyngeal fricatives /h‘/ correlates with the raising of the F_1 locus from 500 Hz to 1000 Hz, whereas F_2 locus remains at 1400 Hz in both cases.

The spectrographic analysis shows that the articulation of /‘/ is realized as a voiced pharyngeal fricative only in an intervocalic context and as a voiceless stop in other phonological contexts. This is also noticed by Al Ani (1970) but, even though he associates /‘/ with the glottal stop /’/ (1970: 59), he purposely avoids to define the place of articulation of /‘/ (1970: 62). If we compare the spectrograms of /‘/ with those of the glottal stop, we can notice that they are articulated differently because even though both of them are voiceless stops, they show a very different trend in the formant transitions. In fact, the formants of a vowel next to a glottal stop, that is a vowel with an abrupt ending or beginning, have neither positive nor negative deviation from the target value of the vowel. As regards /‘/, however, both F_1 and F_2 show noticeable deviations, whose points of convergence have been spectrographically located at about 1100 Hz and 1500 Hz, respectively (Fig. 1). F_1 locus at about 1100 Hz is related to a constriction occurring in the proximity of the node of velocity that is at the glottal level; F_2 locus at about 1500 Hz is related to a constriction occurring at a point intermediate between a node and an anti-node of velocity. As shown in Fig. 2 these acoustic data, from an articulatory point of view, correspond to a constriction occurring at 2.5 cm above the glottis. The closure must be effected, therefore, by the epiglottis which, lowering on the arytenoid cartilages, closes the laryngeal opening; it is, in other words, a laryngeal stop. The acoustical and radiographic data of this study confirm the hypothesis formulated by Laufer and Condax (1979) who, from a fibroscopic analysis, assert that /‘/ is not characterized by the retraction of the root of the tongue towards the posterior wall of the pharynx but by the action of the epiglottis acting as an articulator.

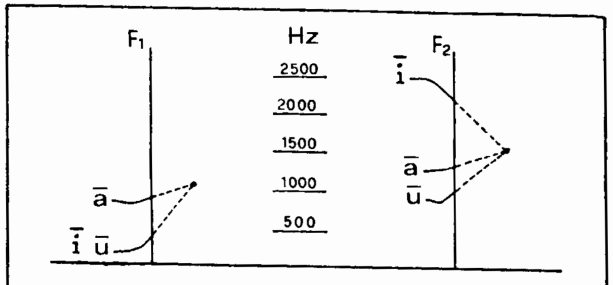


FIG.1 F_1 and F_2 loci of /‘/

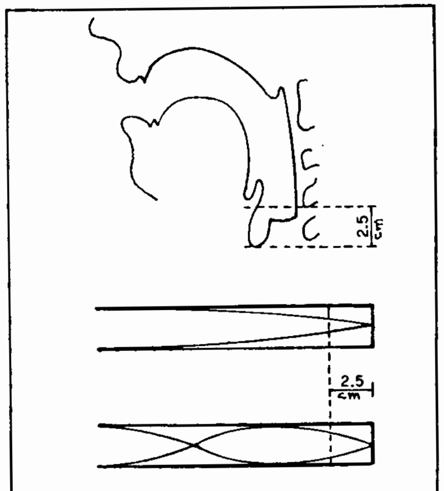


FIG.2

3. Conclusions

In conclusion, our data show that there is a fixed relation between F_1 locus and place of articulation but that is is arbitrary to give to this acoustical cue a specific distinctive function, as this function varies in relation to the pair under consideration. However, the case of /‘/ shows that if we limit our attention to a single articulation, avoiding any reference to pairs constituted a priori, both F_1 and F_2 loci come out to be indispensable elements for the identification of the place of articulation of the consonant.

References

Al Ani, S. (1970). *Arabic Phonology*. The Hague.

Chiba, T. and Kajiyama, M. (1958). *The Vowel. Its Nature and Structure*. Tokyo.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. and Gerstman, L.J. (1952). Some experiments on the Perception of Synthetic Speech Sounds. In: *JASA*, **24**, 597-606.

Delattre, P.C., Liberman, A.M. and Cooper, F.S. (1955). Acoustic Loci and Transitional Cues for Consonants. In: *JASA*, **27**, 769-773.

Fant, G. (1960). *Acoustic Theory of Speech Production*, The Hague.

Giannini, A. and Pettorino M. (1982). The Emphatic Consonants in Arabic. In: *SLR IV*, Ist. Univ. Orientale, Napoli.

Laufer, A. and Condax, I.D. (1979). The Epiglottis as an Articulator. In: *JIPA*, **9**, 50-56.

Öhman, S.E.G. (1966). Coarticulation in VCV Utterances: Spectrographic Measurements. In: *JASA*, **39**, 151-168.

Stevens, K.N. and House, A.S. (1956). Studies of Formant Transitions Using a Vocal Tract Analog. In: *JASA*, **28**, 578-585.

The Influence of Voice Sound Level on the Duration of French Long Vowels

D. Rostolland et C. Parant
Paris, France

1. Introduction

De nombreux travaux ont déjà été consacrés aux variations de la durée vocalique (revues bibl. in Fant et Tatham; 1975, Di Cristo; 1978, Rossi 1981). Cette durée est une grandeur parfois difficile à mesurer et souvent très difficile à interpréter, car les variations temporelles sont régies par des facteurs issus de domaines d'analyse variés: acoustiques (bruit ambiant), physiologiques (habitudes articulatoires, niveau, débit) et linguistiques (à l'échelle d'un phonème, d'une syllabe, d'un mot, etc.). Les travaux publiés concernent généralement la voix parlée au niveau normal et la langue anglaise. En présence de bruit le niveau de la voix peut être très supérieur, ce qui entraîne des modifications acoustiques et phonétiques (Rostolland, 1979 et 1982).

Dans une étude récente, nous avons cherché l'influence de l'effort vocal sur la durée des voyelles brèves (Rostolland et Parant 1983). Nous présentons ici les résultats concernant les voyelles suivies des constrictives les plus allongeantes. Nous cherchons dans quelle mesure la durée augmente, par hypothèse, avec le niveau vocal. L'ouverture de la voyelle et le point d'articulation de la consonne subséquente sont considérés comme 2 paramètres.

2. Méthode

Des séries de monosyllabes sont prononcés à 4 niveaux sonores: voix normale (N1), forte (N2), très forte (N3) et voix criée (N4). Ces niveaux de parole correspondent à des niveaux de bruit que le sujet entend au moyen d'écouteurs. Les mots sont du type C1 V C2 et placés en position accentuée dans une courte phrase. On a choise C1 = /k/, V = /a,i/ et C2 = /v z ʒ r/. A chaque niveau, on enregistre le sujet qui prononce en tout 20 fois chacune des 8 phrases mais qui ne répète à la suite que 2 phrases identiques. On mesure les durées sur des tracés logarithmiques, sonagraphiques et oscillographiques. La comparaison de ces 3 techniques permet de lever les ambiguités dans les cas de segmentations difficiles.

3. Résultats

Chez les 2 sujets S1 et S2, les différences de durée de la voyelle /a/, dues au niveau sonore d'émission, sont significatives entre N1N3 et N3N4. En forçant la voix, la durée de /a/ augmente jusqu'au niveau de la voix très forte et en criant la durée diminue, sauf avec /r/ chez S1. Mise à part cette consonne, la variation de durée est comparable chez S1 et S2 (les durées de /a/ sont plus longues chez S1, de 10 ms).

Les différences de durée de /i/, dues au niveau, sont significatives entre N2N3 et N3N4 mais pas entre N1N2 (sauf avec /r/ chez S1). Forcer la voix augmente la durée de /i/ entre la voix forte et la voix très forte avec, ensuite, une diminution très nette au niveau crié, comme avec /a/. La variation de durée est comparable chez S1S2, mais la diminution à N4 est plus marquée chez S2 (les durées de /i/ sont plus longues chez S1, de 35 ms).

La comparaison des durées de /a/ et /i/ a un niveau donné montre que chez S2 /a > i/ a NN2N3 et /a ≃ i/ a N4. Pour S2, on observe donc un regroupement des dureés vocaliques, en voix criée. Chez S1/a ≃ i/à N1N2 et /a < i/à N3N4: il n'y a pas de regroupement des dureés à N4.

En ce qui concerne les différences de durée dûes à la consonne finale, on observe que chez S_1 S_2 et pour /a,i/ :

- aux 4 niveaux, /v/ est la moins allongeante et les effets de /z ʒ/ sont très voisins sauf chez S1 avec /i/, à N1N2.
- à N4 les durées sont identiques devant /v z ʒ/. Elles sont plus longues devant /r/, sauf chez S1 avec /i/ où le regroupement a lieu pour l'ensemble /v z ʒ r/.
- à N3, les effects de /z ʒ r/ sont très voisins sauf chez S2 avec /i/ où /r/ est la plus allongeante.

Chez S1 et aux 4 niveaux, la consonne la plus allongeante est /z/, avec /a/ et /i/ sauf a N4 avec /a/ et les effets de /z,r/ sont identiques avec /i/ sauf à N1.

Chez S2 et aux 4 niveaux, la plus allongeante est /r/ avec /a/ et, surtout, avec /i/. Le tableau donne la durée moyenne en ms, de /a,i/ devant /v z ʒ r/ aux 4 niveaux d'émission. Les valeurs sont arrondies à l'unité la plus proche. Les dispersions sont < 10% dont 78% des cas pour S1 et dans 91% des cas pour S2.

Niveaux		N1		N2		N3		N4	
		/a/	/i/	/a/	/i/	/a/	/i/	/a/	/i/
S1	/v/	224	240	253	245	267	276	229	261
	/z/	254	275	281	295	298	318	242	283
	/ʒ/	259	258	288	264	292	302	241	272
	/r/	233	256	269	281	283	311	281	275
S2	/v/	236	205	249	210	257	231	209	193
	/z/	249	224	253	237	274	258	202	209
	/ʒ/	260	229	262	240	278	256	211	201
	/r/	269	257	272	267	290	283	225	227

4. Discussion

On sait que la durée intrinsèque d'une voyelle augmente avec son ouverture ou sa tension. La durée dépend aussi du mode phonatoire de la consonne suivante: par exemple, les voyelles accentuées sont en général plus longues de 25 à 35% devant les consonnes voisées (House et Fairbanks, 1953; Denes, 1955; Peterson et Lehiste, 1960; House, 1961; Klatt, 1973). Même en voix parlée normale, la durée est un indice difficile à interpréter car plusieurs causes agissent en sens contraire: une voyelle ouverte est plus ouverte lorsqu'elle est brève et une voyelle est d'autant plus tendue qu'elle est fermée et/ou longue (Durand, 1946; Delattre, 1966; Carton, 1974; Di Cristo, 1980; Rossi, 1981).

C'est à propos d'une étude de la force d'articulation consonantique, que Delattre s'est intéressé à la durée vocalique. Devant l'impossibilité de mesurer objectivement et directement cette force d'articulation, une méthode indirecte a été proposée par cet auteur. Etudiant les consonnes en fin de syllabes accentuées fermées (terminant des groupes rythmiques de longueurs voisines), Delattre a émis l'hypothèse que la durée de la voyelle précédent la consonne était un indice de la force d'articulation consonantique. Avec la voyelle /ɛ/, Delattre a mesuré des durées comprises entre 116 et 428 ms. Les voyelles brèves sont suivies de consonnes sourdes et les longues de consonnes sonores: /p t k/ < /b d g/ et /f s ʃ/ < /v z ʒ/. Le mode d'articulation influence la durée vocalique dans ce sens: /p t k/ < /f s ʃ/ et /b d g/ < /v z ʒ/ mais le lieu d'articulation ne semble pas jouer un role important (p ≃ t ≃ k).

Delattre aurait, le premier, établi une relation entre l'effort nécessaire pour articuler la consonne C2 dans des mots C1 V C2, et la durée de V: la voyelle s'abrège par un effet d'anticipation de l'effort articulatoire subséquent. (Delattre 1939, 1940, 1941, 1962). Ces travaux repris par Belasco (1953) et Malécot (1955) conduisent à remplacer l'ancienne division entre consonnes fortes, abrégeantes /p t k/ et faibles, allongeantes /v z ʒ r/ par une division en 3 et 5 groupes et même 8 (p t k, f l, n s m ʃ, b d g ŋ, j, v ʒ, z, r). La cause elle-même de l'allongement vocalique n'est pas absolument claire: s'agit-il d'un seul facteur, la force d'articulation consonantique? ou de trois facteurs, le voisement, le mode et le lieu d'articulation ? (Belasco 1953, House et Fairbanks 1953, Lehiste 1970, MacNeilage 1972).

En voix parlée normale, et en français, Delattre (1939) a mesuré avec la voyelle /ɛ/ devant /v z ʒ r/ des durées comprises entre 370 et 420 ms. Devant les spirantes sourdes et les occlusives sonores, /ɛ/ mesure 220 et 270 ms. Di Cristo (1980) a trouvé pour /a/ et /i/ devant /z/ des durées de 215 et 185 ms. Devant les obstruantes sourdes, il trouve avec /a/ des durees de 88, 121 et 155 ms selon le corpus utilisé; les valeurs correspondantes pour /i/ sont de 72, 96 et 122 ms (la durée de /i/ devant /v z ʒ/ serait de 108 ms). Pour la langue anglaise, nous avons effectué de nombreuses comparaisons mais nous n'indiquerons ici que les valeurs extrêmes du tableau comparatif: devant /v z ʒ r/ les durées de /ɒ, æ, ɔ/ vont de 304 a 400 ms, et celles de /i/ de 183 a 360 ms

(House et Fairbanks 1953, Peterson et Lehiste 1960, House 1961, Sharf 1964, Harris et Umeda 1974, Klatt 1973 et 1975, Umeda 1975).

Très peu de travaux concernent la durée des voyelles émises à des niveaux supérieurs a celui de la voix parlée et les méthodes expérimentales sont trop différentes pour permettre des comparaisons quantitatives. De manière qualitative, nos résultats confirment ceux de Hanley et Steer (1950) et de Fónagy (1980) qui ont trouvé un accroissement de la durée entre les niveaux N1N2N3. Fónagy avait déjà observé une limitation des durées à haut niveau (vers N3). En rapprochant la présente étude de celle que nous avons faite sur la durée des voyelles brèves, on voit apparaitre l'influence de facteurs liés soit à l'articulation de la voyelle elle-même, soit à la coarticulation consonantique subséquente.

4.1. Durée intrinsèque: niveau et ouverture vocalique (facteurs internes)

L'hypothèse de départ – allongement avec augmentation de niveau – est vérifiée entre N1N3, pour S1 et S2/a,i/ (l'augmentation relative de durée est plus importante pour les voyelles brèves – 28% de N1 – que pour les voyelles longues – 14%) mais pas entre N3N4, puisque la durée diminue au niveau extrême (de 7 a 16% de N3).

L'ouverture vocalique entraîne généralement une augmentation de durée (/a > i/), pour les voyelles brèves, chez S1 et S2. Entre N1N2N3, l'augmentation est de 22,5% chez S1 et de 27,5% chez S2. A N4, les durées de /a/ et /i/ sont identiques. Pour les voyelles longues, et entre N1N2N3, on note une augmentation plus faible chez S2 (8,5%) et même une diminution chez S1 (3,5%). A N4, on observe un regroupement des durées de /a/ et /i/ chez S2, mais cette tendance n'apparait qu'avec /r/ chez S1.

4.2. Durée extrinsèque: mode et lieu d'articulation consonantique (facteurs externes)

On considère ice l'effet cumulé du voisement et de l'ouverture consonantique. A N1, /a/ devant /v z ʒ, r/ est plus long que devant /p t k/ de 45,5% chez S1 et de 68% chez S2. Entre N1N2N3, les augmentations sont de 41 à 54%, respectivement. A N4, les augmentations sont moindres: 37 et 36%. Avec /i/ on trouve des augmentations plus grandes: à N1, 74, 5% chez S1 et 102% chez S2 et entre N1N2N3, 79 et 81% respectivement. A N4, les augmentations sont moindres: 48 à 34%.

L'influence du recul du lieu d'articulation, chez S1 et S2, est plus importante pour les longues que les brèves: a N1, les rapports 'ʒ/v' sont de 12,8% avec /a/ et de 13, 2% avec /i/ tandis que les rapports 'k/p' sont de 8 et 4,5%. A N1N2N3, les rapports moyens sont 10,5 et 14,5% (longues) et 8,5 et 7% (brèves). A N4, les rapports seraient plus petits pour les longues (3 et 6,5%) que pour les brèves (7,5 et 9%). Les valeurs maximums de ces pourcentages d'augmentation – en fonction du recul du lieu d'articulation – sont: 15,6% pour les longues (ʒ/v chez S1 à N1 avec /a/) et 17% pour les brèves (k/p chez S1 a N4 avec /i/). Si l'on considère le rapport 'r/v', le maximum est 27% (chez S2 a N2 avec /i/). Aux 4 niveaux, la consonne la moins allongeante est /v/, la plus allongeante étant soit /z/ (S1) soit /r/ (S2). Nous ne trouvons pas que l'influence du lieu d'articulation soit faible (House et Fairbanks 1953) ou négligeable (House 1961) mais trouvons que la durée vocalique est plutôt proportionelle au recul du lieu d'articulation (Lehiste, 1970; MacNeilage, 1972).

La présente étude confirme des résultats que nous avons obtenus avec les brèves, en particulier le fait que la durée diminue à haut niveau. Tout se passe comme si le sujet anticipait un effort articulatoire non plus à l'échelle des phonèmes mais à celles des mots. Des expérimentations sont en cours afin de vérifier certains résultats avec plus de locuteurs.

Références

Belasco, S. (1953). The influence of force of articulation of consonants on vowel duration. *J. Acoust. Soc. Am.* **25**, 1015-16.

Delattre, P. (1939). Durée vocalique et consonnes subséquentes. *Le Maitre Phonétique*. Londres 3e Série - N⁰ 67.

Delattre, P. (1966). *Studies in French and comparative phonetics*. The Hague: Mouton and Co.

Di Cristo, A. (1978). De la microprosodie à l'intonosyntaxe. Thèse d'Etat. Université de Provence. 1274 p.

Di Cristo, A. (1980). La durée intrinsèque des voyelles du francais. *Trav. Inst. Phon. Aix.* **7**, 211-235.

Durand, M. (1946). *Voyelles longues et voyelles brèves*. Paris: Klincksieck.

Fant, G. and Tatham, M. (1975). *Auditory analysis and perception of speech*. Londres: Academic Press.

Fónagy, I. (1980). Intensité et durée des sons. *Trav. Inst. Phon. Paris* **3**, 169-199.

Hanley, T.D. and Steer, M.D. (1950). Effect of level of distracting noise upon speaking rate, duration and intensity. *Journ. Speech and Hear. Dis.* **15**, 363-368.

House, A.S. and Fairbanks, G. (1953). The influence of consonants environment upon the secondary acoustical characteristics of vowels. *J. Acoust. Soc. Am.* **25**, 105-113.

House, A.S. (1961). On vowel duration in English. *J. Acoust. Soc. Am.* **33**, 1174-78.

Klatt, D.H. (1973). Interaction between two factors that influence vowel duration. *J. Acoust. Soc. Am.* **54**, 1102-04.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Mass.: M.I.T. Press.

MacNeilage, P.F. (1972). Speech physiology. In: *Speech and cortical functioning*, Chap. 1, 1-72. New York: Ac. Press.

Rossi, M. et coll. (1981). *L'intonation. De l'acoustique à la sémantique*. Paris: Klincksieck.

Rostolland, D. (1979). Contribution à l'étude de l'audition de la parole en présence de bruit: caractéristiques physiques, structure phonétique et intelligibilité de la voix criée. Thèse d'Etat. Faculté des Sciences de Paris. 281 p.

Rostolland D. (1982). Acoustic features of shouted voice. *Acustica* **50**,118-125.

Rostolland D. (1982). Phonetic structure of shouted voice. *Acustica* **51**, 80-89.

Rostolland, D. et Parant, C. (1983). Influence du niveau sonore de la voix sur la durée des voyelles brèves. *11e Congrès Int. d'Acoustique*, 4 p.

Umeda, N. (1975). Vowel duration in American English. *J. Acoust. Soc. Am.* **58**, 434-445.

Bavarian -a- Vowels: Experimental Investigations on the Comparability of the First Two Formants and of Phonetic Transcriptions

Klaus G. Schweisthal, Stephan Forstner, Kurt Kotten
Munich, FRG

1. Object

In contrast to the current Standard pronunciation of German it is a noticeable characteristic of several Southern Bavarian dialects that there are no central -a- sounds in monophthongs.

On the other hand one can postulate a front phoneme /a/ with at least two variants [a], [æ], and a back phoneme /ɒ/ lacking in Standard German which has about four allophones [ɒ], [ɔ], [o], [u]. This situation led us to hypothesize that the back variants in stressed syllable position occur only in words which have been in the language since the time of Old German or Middle High German, while the front variants in stressed syllable position occur only in later words e.g. loan words and foreign words in a way similar to the relatively late occurrence of the i-Umlaut in Bavarian dialects.

2. Method

2.1. Word List Used in the Tests

Three groups of twelve nouns all containing the letter -a- in stressed syllable position and also occurring in Standard German were selected, mainly from the 'Bayerisches Wörterbuch' (Schmeller^2 1872-1877, 1973). In the first group the twelve nouns were first recorded in the Old High German period, in the second in the Middle High German period, and in the third from Middle High German onwards. Each group had six examples of long -a- and six examples of short -a- (in the standard pronunciation). A further requirement on the list was that the individual words could be combined to form 18 compounds existing in Standard German. Furthermore it was necessary to distribute expected front variants (short and long) systematically over the first and the second position in the compounds, for example 'Nagellack' and 'Taxifahrer', 'Garagenwagen' and 'Lampenkabel'.

2.2. Recording of the Speech Data

In order to reduce the influence of Standard spoken and written language as far as possible we used picture puzzles as an elicitation technique. The

compound had to be guessed and spontaneously uttered from the pictures of the individual parts (e.g. the picture for 'Nagel' - nail, then the picture for 'Lack' - varnish, and then the compound 'Nagellack' - nail varnish). The compounds had to be uttered in the standardized context ('Das ist ein ... glaub' ich' - this is a .... I guess) in order to avoid list effects and to keep the accentuation reasonably constant.

Requirements for the Native Speakers: Born in the village where the recording was made; had spoken the dialect since childhood; no lengthy absences from the region; no higher education; dialect spoken as a matter of course at home; more than 30 years old.

Eight male speakers from Upper Bavaria; four speakers from the south and four from the north.

2.3. Digital Data Processing

The recordings were digitalized (sample rate 20 KHz) and segmented using a segmentation routine available on the Institute's PDP 11/50. Five consecutive pitch periods were segmented individually from the central portion of the -a-vowel in the individual words and in the compounds. An analysis program then calculated the fundamental frequency, the first three formants with their bandwidths and the length in milliseconds. For the production of the test material a waveform manipulation program was used to double every period because we imagined that a sequence only five periods long would be too short for a transcription to be possible.

2.4. Transcription

Each new sequence of ten periods was presented five times to the subjects over loudspeakers. The subjects (all trained listeners) were instructed to note tongue-height and tongue-position (front - back) on two scales with seven steps each (Fig. 1).

2.5. Combined scales

A system of coordinates was tried out. On the ordinate the seven steps of the transcription used for the evaluation of tongue-position and at the same time the measured values of F_2 are entered.

The abscissa shows the seven steps of the transcription of tongue-height and at the same time the measured values of F_1 (Fig. 2). For the definition of the formant scale we followed suggestions of Ungeheuer (1962) and Ladefoged (1971). We restricted ourselves to the vowel system forming the subject of our investigation. For each speaker the average of five F_1/F_2 measurements was entered on the system of coordinates, together with the average transcription results of ten trained listeners. The areas enclosed by the dashed lines were assigned the corresponding transcription symbols. In this way we also obtained a readily understandable aid for courses in experimental phonetics.

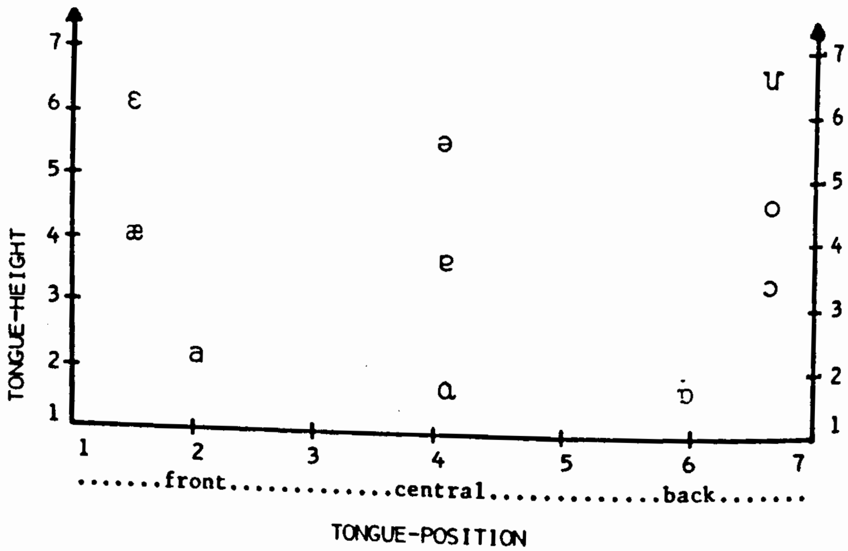


Fig. 1. Steps of transcription.

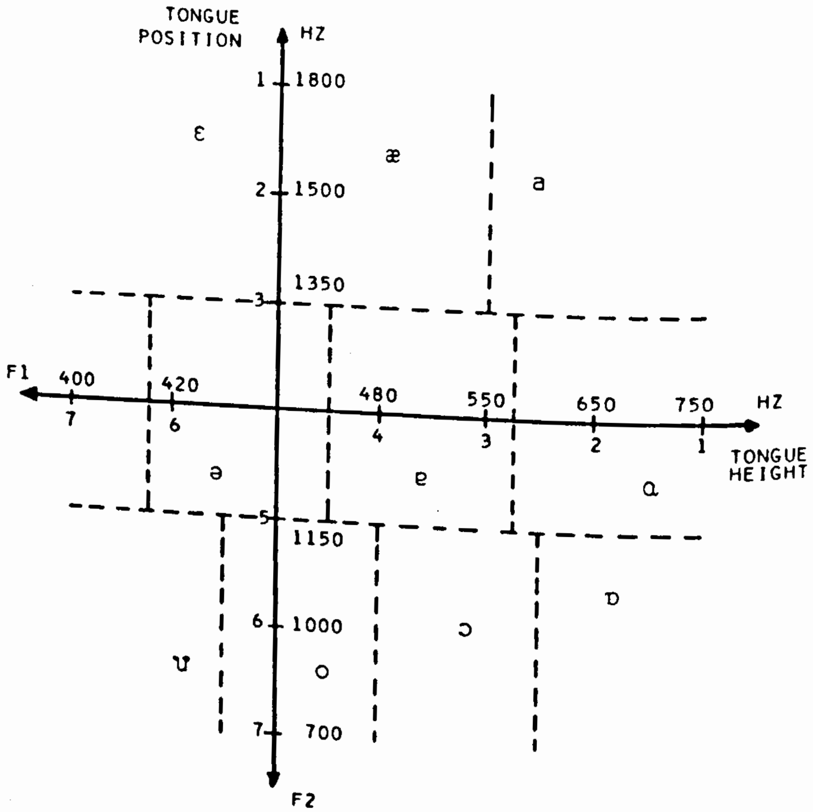


Fig. 2. Combined scales for F_1/F_2-measurements and steps of transcription.

3. Results and Discussion

3.0. To date a representative sample of 16 words from the total 36 (and thus eight of the 18 compounds) from five of the eight male speakers have been analysed. The speakers all had average fundamental frequencies between 100 and 120 Hz.

This gave a total of 160 -a-productions. The percentages given below relate to this total.

3.1. The following average formant frequencies were assigned to the transcription symbols on the basis of the combined scale:

	F_1 F_2		F_1 F_2		F_1 F_2
[æ]	550/1650;	[a]	680/1500;	[ɑ]	640/1250;
[a]	530/1250;	[ɒ]	650/1000;	[ɔ]	550/ 900;
[o]	450/ 850;	[u]	410/ 800.		

Comparison with Delattre's (1965) formant charts shows that the Bavarian F_2 values for front -a- realisations are much higher than the Standard German values, while the F_1 values of the back -a- productions are generally somewhat higher. The relatively low F_2 values for [u] can be explained by the influence of nasalization.

3.2. Complete agreement between measured and heard results was found in 78.2% of the cases. A difference of one step along the tongue-height scale was found in 18.3% and a one step difference in tongue-position in 3.2%. If one combines [a] and [æ] (Delattre 1965) the number of complete agreement rises to 88.5%. The rate of agreement could probably be improved still further by calibrating each speaker's formant scale on the basis of the formant values for his individual neutral vowel [ə].

3.3. In only 2.5% of the cases [ɑ] was measured and was transcribed in 4.3%. Most of the occurrences of [ɑ] were in the test word 'Harfe' - harp (three speakers) which is attested as early as the Old High German period. Centralized [ɐ] was measured and heard in 2.5% of the cases especially in the second position of the compounds (secondary stress).

A total of 95% of the test words were categorized on the basis of the measurements and transcription as either front or back.

Running counter to the hypothesis -a- in 'Kaffee' received a back realisation in all cases (18 × [ɔ] and 2 × [ɒ]). The hypothesis would predict an [a]-realisation since "Kaffee" has entered the language relatively late. The reason for the discrepancy is probably that it has been psychologically domesticated by daily use.

References

Delattre, P. (1965). *Comparing the Phonetic Features of English, German, Spanish and French.* Heidelberg.

Ladefoged, P. (1971). *Preliminaries to Linguistic Phonetics.* Chicago and London.

Schmeller, J. (1973). *Bayerisches Wörterbuch.* 3. Neudruck der von G. Karl Fromman bearbeiteten 2. Ausgabe, München 1872-1877. Aalen.

Ungeheuer, G. (1962). *Elemente einer akustischen Theorie der Vokal-artikulation.* Berlin, Göttingen, Heidelberg.

Über akustische Faktoren der Distinktiven Perzeption im Deutschen

L. Valaczkai
Szeged, Hungary

1. Einleitung

In der vorliegenden Studie wird die Perzeptionsrelevanz der Frequenz, der Intensität sowie der Bandbreite von Formanten der deutschen akzentuierten Vokale auf Grund der distinktiven Oppositionen erörtert. Die Untersuchung erstreckt sich auf die Bewegungstendenzen der Frequenzwerte von F_1, F_2 und F_3 als akustische Faktoren der distinktiven Opposition; ferner auf die Reaktionen der Grundfrequenz, der Intensität und der Bandbreite von Formanten auf diese Bewegungstendenzen.

2. Die Formanten.

Die Rolle der Formanten in der akustischen Perzeption wird unterschiedlich eingeschätzt. Ladefoged und Broadbent (1957: 98-104) bezweifeln die Identifizierung der Vokale lediglich auf Grund der absoluten Frequenzwerte. Otto von Essen (1966: 155) hält einen von zwei Formanten 'für die Erzielung einer vorgegebenen Vokalfarbe' für 'unbedingt notwendig', 'der andere rundet den Klang nur ab, sättigt ihn' Wängler (1976: 20) spricht zwar die Wichtigkeit der ersten beiden Formanten nicht expressis verbis aus, jedoch teilt er die Frequenzwerte der ersten beiden Formanten mit. Nach Lindner (1969: 172-173) 'haben nicht alle Formanten gleichen Informationswert.

2.1.Bewegungstendenzen von Frequenzwerten der Formanten (Hz)

2.1.1. Die Tendenzen von F_1

/i:/ 272 → /e:/ 398 → /ɛ:/ 623 → /ɑ:/ 850
/I/ 436 → /ɛ/ 509 → /ɑ/ 702
/y:/ 272 → /ø:/ 390
/u:/ 281 → /o:/ 445 → /ɑ:/ 850
/U/ 354 → /ɔ/ 500
/ɔ/ 500 → /ɑ/ 702

Die jeweiligen distinktiven Oppositionen sind auf Grund der senkrechten Bewegung der Zunge, und in Verbindung damit des Kiefers, von oben nach unten zusammengestellt worden. Die Bewegung des ersten Formanten zeigt in den Reihen distinktiver Oppositionen eine eindeutig steigende Tendenz.

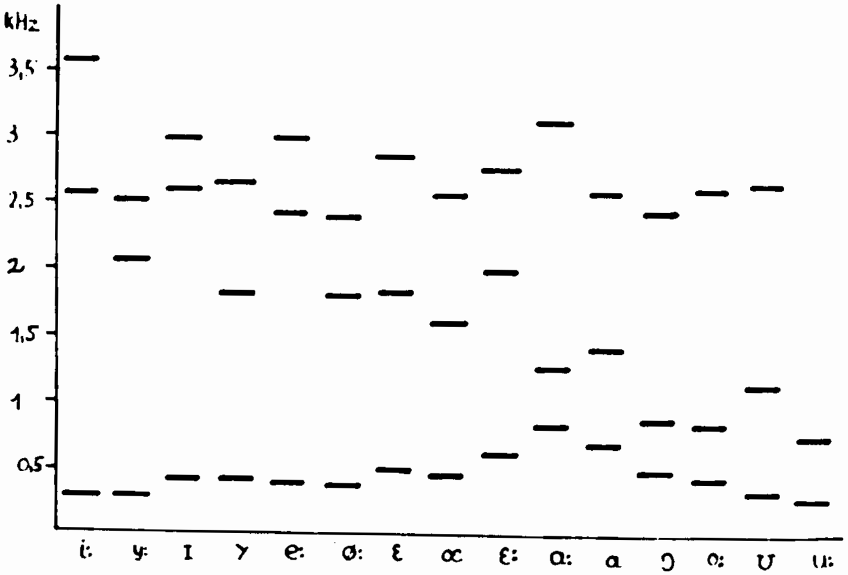


Abb. 1. Die spektrale Verteilung der Formanten.

2.1.2 Die Tendenzen von F_2
/i:/ 2245 → /e:/ 2425 → /ɛ:/ 2000 → /ɑ:/ 1286
/I/ 2590 → /ɛ/ 1838 → /ɑ/ 1409
/y:/ 2056 → /ø:/ 1797
/Y/ 1818 → /œ/ 1613
/u:/ 747 → /o:/ 852 → /ɑ:/ 1286

Die Tendenz der Werte von F_2 ist nicht so eindeutig wie die von F_1. In den Oppositionen der ersten vier Reihen ist sie fallend, in der fünften Reihe steigend.

2.1.3. Die Tendenzen von F_3
/i:/ 3575 → /e:/ 2975 → /ɛ:/ 2772
/I/ 2954 → /ɛ/ 2863 → /ɑ/ 2590
/y:/ 2511 → /ø:/ 2404
/U/ 2681 → /ɔ/ 2648
/o:/ 2618 → /ɑ:/ 3136

Die Tendenz der Werte ist in den ersten vier Reihen fallend, in der fünften steigend. Die Bewegungstendenzen der Werte von F_2 und F_3 stimmen in diesen Oppositionen überein.

Das proportionale Verhältnis der Werte der Formanten

Laut	F_1 : F_2	F_1 : F_3
/i:/	1 : 9,3	1 : 13
/e:/	6	7,4
/ɛ:/	3,2	4,4
/I/	5,9	6,7
/ɛ/	3,6	5,6
/ɑ/	2	3,6
/u:/	2,6	–
/o:/	1,9	5,8
/ɑ:/	1,5	3,6
/U/	3,2	7,5
/ɔ/	1,7	4,9
/y:/	7,5	9,2
/ø:/	4,6	6,1
/Y/	4,1	6
/œ/	3,7	5,9

Die Frequenzbereiche des ersten Formanten von /i:/ – /y:/ und /I/ – /Y/ sind gleich, und im Falle von /e:/ – /ø/ ist der zahlenmässige Unterschied so gering, dass er ausser acht gelassen werden kann. Trotzdem stehen diese Vokale in distinktiven Oppositionen, deren Grundlage man also in der Entfernung und im Unterschied der anderen Formanten sehen kann. Durch diese Feststellung wird die Auffassung bestätigt, die die oppositionelle Grundlage der akustischen Perzeption vor allem in der spektralen Verteiling, im proportionalen Verhältnis der Formanten sieht.

3. Intensität und Bandbreite

Der Zusammenhang zwischen der Frequenz und der Bandbreite der Formanten wird ebenfalls unterschiedlich beurteilt. Nach Tarnóczy und Vicsi (1975: 341-342) erhöhen sich die Werte der Bandbreite parallel zur Erhöhung der Frequenzwerte. Fujimura und Lindquist (1971: 541) haben dagegen eine solche Regelmässigkeit nicht gefunden. Aus unseren Untersuchungen geht folgendes hervor: Die Werte der relativen Amplitude sowie der Bandbreite reagieren zwar schwach auf die erörterten distinktiven Oppositionen, trotzdem kann ein Zusammenhang den Bewegungstendenzen der Werte der Frequenz, der Intensität und der Bandbreite der Formanten nachgewiesen werden, wenn auch die Tendenzen dei Werte der Intensität und der Bandbreite nicht so eindeutig sind wie die der Frequenzwerte der Formanten. Zur Ermittlung von klaren Tendenzen müssen Messungen auf einem umfangreicheren sprachlichen Korpus vorgenommen werden. In der folgenden Tabelle bezeichnen A_1 und A_2 die Intensität von F_1 und F_2, B_1 und B_2 die Bandbreite von F_1 und F_2.

Werte der Intensität (dB) und der Bandbreite (Hz)

	/ɑ:/	/o:/	/u:/	/ɛ:/	/e:/	/i:/
A_1	13,5	24	27	18	20	25
B_1	63	79	90	100	100	154
A_2	14	14	14	14	10	−0,5
B_2	138	–	–	93	100	63

Sowohl in der dunklen als auch in der hellen Oppositionsreihe zeigen die Werte von F_1 eine fallende, die Werte von A_1 und B_1 eine steigende Tendenz. Zwischen F_1 einerseits, A_1 und B_1 andererseits besteht also in den Bewegungstendenzen der Werte ein umgekehrtes Verhältnis. In der dunklen Vokalreihe ist die Tendenz von F_2 fallend, in der hellen Reihe steigend. Die Werte von A_2 sind in der dunklen Vokalreihe gleichbleibend, in der hellen Reihe fallend. Die Werte von B_2 ergeben in der hellen Vokalreihe eine leicht fallende Tendenz. Die zuverlässige Beurteilung von A_3 und B_3 bedarf noch einer weiteren Untersuchung.

Acknowledgements

Die zur Analyse notwendige Spektrographie wurde in der Phonetischen Abteilung des Institutes für Sprachwissenschaft der Ungarischen Akademie der Wissenschaften in Budapest unter der Leitung von Dr. K. Bolla vorgenommen.

References

Essen, O. von (1966). *Allgemeine und angewandte Phonetik*. Berlin.

Fujimura, O., Lindquist, J. (1971). Sweep-tone measurements of vocal tract characteristics. *J. Acoust. Soc. Amer.* **49**, 541-558.

Ladefoged, P., Broadbent, D.E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Amer.* **29**, 98-104.

Lindner, G. (1969). *Einführung in die experimentelle Phonetik*. Berlin.

Tarnóczy, T., Vicsi, K. (1975). Decay characteristics of vowel cavities recorded inside and outside the mouth. *Acustica* **33**, 340-343.

Wängler H. (1976). *Atlas deutscher Sprachlaute*. Berlin.

Section 2
Speech Synthesis

An Analysis Method for High Quality Formant Synthesis

P. Badin and G. Murillo
Grenoble, France

1. Introduction

Presently there are in France several laboratories cooperating in a Speech Communication Research Group (GRECO) that is supported by the French National Scientific Research Center (CNRS) and that is working on the constitution of a data base of French sounds. Within this framework, our aim is to carry out an analysis on a corpus of French sounds, in order to generate high quality synthetic speech.

Nowadays, a relatively large number of commercially available systems produce very intelligible speech. However, even if this speech is quite easy to understand, it is not very difficult to detect its synthetic nature. Moreover, it becomes more and more clear that the quality attained is not much improving anymore.

The great progress realized up to now in synthetic speech quality has been essentially based upon research using perception criteria: the acoustic cues contributing to sound perception have been determined exclusively from test verifications of a priori hypotheses. The famous 'locus' theory developed at Haskins Laboratories (Delattre, 1958), for example, was elaborated from perception tests of synthetic speech rather than from objective measures on the loci's values. Even if the synthetic sounds generated by the Pattern Play-Back were mediocre, this approach permitted considerable progress in phonetic knowledge. In fact, this work had been principally focused to determine the cues which allow us to perceive one sound as phonetically different from another one, regardless of its relation to production phenomena.

Since the elements contributing to speech quality are highly complex, and because the equipment employed was rather rudimentary, we consider that this approach is not adequate for research on speech quality and naturalness. Besides, the properties of sound perception are not yet mastered: because of this, perceptual compensation or masking effects caused that some acoustic cues are not important for intelligibility - but that might be important for quality - could be disregarded. Moreover, in order to easily manipulate the synthetic sounds, it is very important to possess reliable references: a production model and the results of an objective analysis.

We believe that, in order to generate very high quality synthetic speech, we must propose a more fundamental approach which takes into account the speech 'production' aspect, since this is the only way to point out all the acoustic details that might be important from a perceptual point of view. We define a production model and we go back to the classical Analysis-by-Synthesis method proposed for the first time by Bell et al. (1961), but we use more elaborate tools for the analysis and for the comparison as well as for the synthesizer.

2. Method

1. Analysis-Synthesis method principle

The method's goal is to carry out an accurate analysis of natural speech, using a production model. We shall distinguish two levels: the structure of the procedure (i.e. the algorithm surveying the set of operations) and the strategy for its use.

a. Analysis-Synthesis structure

We use the classical Analysis-by-Synthesis scheme: we aim to determine the evolution of the production model's control parameters that will permit us to obtain synthetic speech as close as possible to the original.

The algorithm is divided into two steps:

1. The first one consists of an automatic analysis of the original signal. Poles, bandwidths, F_0 and signal energy are computed. Besides that, sonagrams and DFT of the signal are drawn; finally a graphic display of the speech waveform is also made available;
2. The second step is a feedback procedure: starting from the data acquired in step 1, the evolution of the synthesizer control parameters is determined; the synthetic waveform is computed and compared to the original in order to edit again the control parameters. This procedure is repeated until a correct result is obtained. The comparison is threefold: time, spectrum and perception-wise. Parameter acquisition and editing are done by means of an interactive graphic program.

b. Strategy

It is clear that parameter acquisition and correction cannot be done for all parameters in one single step because of their large number. Once we have acquired the basic parameters (energy, F_0, formants and bandwidths), a first synthetic waveform is computed. This signal may be redrawn on outline, based on data in the literature, in order to get a first approximate result. Immediately after, the stationary zones are refined and verified using mostly the perceptual method (see below). After this, transitions are refined using all helpful analysis data, and proceeding by linear interpolation between the values of the zones surrounding the transition region, every time that analysis results are blurry. For each of these operations the basic parameters

(energy, F_0, formant frequencies) are adjusted first. The rest of the control parameters will serve to refine the results. At this moment, the whole utterance is checked by ear for verification: if the result is not satisfactory, the faulty segments are searched by the perceptual comparison method and readjusted until the whole utterance is considered correct. A long experience in applying this Analysis-Synthesis method shall permit us to increase the performance of the methodology we have described.

2. Tools employed

The production model we have chosen is a parallel type formant synthesizer with a mixed source (periodic signal and noise source) where a 19 parameter-updating is done every 5 ms (this structure is derived from the synthesizer by Klatt in 1980). We have opted for the formant configuration versus LPC synthesis because the former provides a direct acoustic interpretation of the control parameters. This technique, together with DFT spectrum analysis and sonagrams, makes is easier to edit the synthesis parameters by hand.

For the first evaluation of formants we decided to carry out an LPC analysis by the autocorrelation method (Markel and Gray, 1976). Pole and bandwidth values are obtained from the predictor coefficients of the analysis model. Even if this method is not highly accurate, particularly concerning bandwidths, it has the advantage of being fast and completely automatic. The quality of the results is good enough to provide the raw data for control parameter determination.

In an Analysis-by-Synthesis-like method, a most important point is the one dealing with the original versus synthetic comparison: we use a threefold criterion for this comparison. The first one is spectral matching. The second one is a comparison in the time domain : the waveform is displayed on a graphics screen and thus the transition zone boundaries of certain parameters such as noise source energy or voiced source spectrum are determined. The third criterion is the most important one: it takes care of the perceptual comparison ; as proposed by Holmes (1979), it consists of a 'repeated listening to natural and synthetic speech in immediate succession'. The synthetic sound may be composed of a complete synthetic utterance or of a 'synthetically patched natural utterance' (i.e. the natural utterance in which a certain section is replaced by the homologous synthetic one). In this way, it can be determined if there is an unsatisfactory section. This procedure enables one also to locate the different defects that might appear when an utterance is listened to as a whole.

3. First results - Discussion

We have begun to apply this method to the synthesis of a certain number of CVCVC sounds containing French voiced fricatives and stops. About twenty persons listened to a binary comparison-based preference test in which natural, formant synthesis and LPC synthesis homologous utterances are presented for comparison. The results confirm the high formant synthesis quality and the relevance and efficacy of the method.

The method's exploitation remains at present laborious for the operator: work is in progress in order to make a more interactive system and - more than anything else - to free the operator from jobs not requiring decision taking. The problem is to find a compromise between the operator's decision freedom and the system exploitation heaviness.

This method will further permit a build-up of a dictionary of sounds of a language, and will provide an efficient tool for the determination of acoustic cues of speech.

References

Bell, C.G., Fujisaki, H., Heinz, J.M., Stevens, K.N., and House, A.S. (1961). Reduction of Speech Spectra by Analysis-by-Synthesis Techniques. *J. Acoust. Soc. Am.*, **33**, 1725-1736.

Delattre, P.C. (1958). Les indices acoustiques de la parole: premier rapport. (Acoustic Cues of Speech: First Report.) *Phonetica* **2**, 108-118 and 226-251.

Holmes, J.N. (1979). Synthesis of Natural Sounding Speech using a Formant Synthesizer. In: *Frontiers of Speech Communication Research*, 275-285. Ed. by Lindblom, B. and Ohman, S. Academic Press, London.

Klatt, D.H. (1980). Software for a Cascade/Parallel Formant Synthesizer. *J. Acoust. Soc. Am.*, **67**, 971-996.

Markel, J.D. and Gray, A.H. (1976). *Linear Prediction of Speech*. Springer Verlag, Berlin.

Voxton, Russon: Systems Generating Impersonal Hungarian and Russian Speech by Rule

K. Bolla
Budapest, Hungary

1. The structure and use of Voxton

The physical constituents of the sound substance of speech are organised in two ways in linguistically relevant structures. One results in the segmental and the other in the suprasegmental structure of speech. A speech synthesizing system generating impersonal speech by rule, named Voxton, consists of three main parts : a) a collection of sound sections, i.e. the data base realizing the basic units of the acoustic structure; b) the operational computer programs; and c) a code which organizes the phonetic elements and structures of the sound stream.

A sound section can be defined as a homogeneous segment of the acoustic structure of the stream of speech sounds which can be isolated by considering changes in the acoustic constituents. The number of acoustic parameters determining the structure of a sound section is between 1 and 23. Four types of elements can be differentiated according to the acoustic quality of the sound section: pauses, elements with voiced structure, elements with noise structure and elements with mixed structure. The data of 550 sound sections are included in the Voxton speech synthesizing system. Of these elements 27 are used to synthesize the vowels and 89 are used to synthesize the consonants. Sound sequences are realized with the help of 'transitional sections', which are largest in number: 414. Not only do transitions have a role in forming the acoustic structure, they are also important in speech perception. Voxton can be used to deal with this question in depth. Temporal variation, i.e. the phonetic realization of long–short oppositions, is achieved by doubling one of the sound sections making up the relevant speech sound. According to its position in the word, each speech sound can be synthesized in three variants, with different qualities corresponding to word-initial, word-medial and word-final positions.

The phonetic code of Voxton consists of the identifiers of the sound sections and the 'call signals' of the speech sounds. The identifiers indicate a) which speech sound the section belongs to; b) its status in the structure of the speech sound; and c) its position in the sound sequence. The computer automatically selects the sound sections from the data base and combines them. The sound transitions are also built in automatically according to the structural code. The sound sequence called by the phonetic symbols corresponds to the segmental structure of Hungarian speech. The speech sounds in the sound sequence are realized on a monotone with their characteristic quantity and intensity. The appropriate suprasegmental structure is built up in a separate step. This can be carried out quickly after the data characterizing the intonation pattern, the dynamic structure, the tempo and the rhythm are fed in. It is possible to change the intonation pattern of a single segmental structure as needed; an infinite number of suprasegmental variants can be made (see Fig. 1). Each synthesized sentence or clause is given an identification, which makes repetitions, storage, repeated use and the production of longer texts more effective.

2. A brief description of Russon

Russon is a synthesizing system suitable for artificial production of the phonetic form of Russian speech. Its minimalized and optimalized data base contains the data of 265 sound sections from which 87 different speech sounds (35 vowels and 52 consonants) and moreover 4 pauses of different duration can be produced. The 87 speech sounds also include the positional variants of the sounds. Russon can be used in two variants: in a phonetic and a phonematic speech generating system. In the first case we use the phonetic characteristics of speech sounds in structuring the text, while in the second one, we describe the text or sequence of sounds to be synthesized with the phonemes of the Russian language. E.g.:

- phonetically: [SA“DY CV’IETUT V’IESNO’J’] # 00
- phonematically: /SADI’ CV’ETU’T V’ES+NO’J’+/ # 13

According to the phonetic code and phonotactic rules built into Russon the following steps take place: a) the building up of speech sounds from the proper sound sections, b) the selectiȯn of realizations/allophones of the phonemes used, c) the distinction of stressed and unstressed positions, d) the recognition of phonetic positions of vowels (apart from the word-stress, the modifications arising from the word-initial, word-medial and word-final positions; furthermore the patalized and pharyngealized consonantal surroundings are to be implemented, e) the assimilation of consonants according to their voiced/unvoiced quality, f) the assimilative palatalization of consonants, g) the alternation (lengthening) of duration of sounds, h) the recognition of focus in intonation in the phases and phrases of speech, i) transitions of the selected intonational constructions in the sound sequence. An F_0 matrix is used for the automatic synthesis of the intonation of Russian speech, (see Fig. 2).

3. The Common Features of the Two Speech Synthezising systems

We mention only the most important aspects. Both are built on the same hardware (a PDP 11/34 computer and an OVE III speech synthesizer).Its software consists of an RT-11 operating system and programmes written in

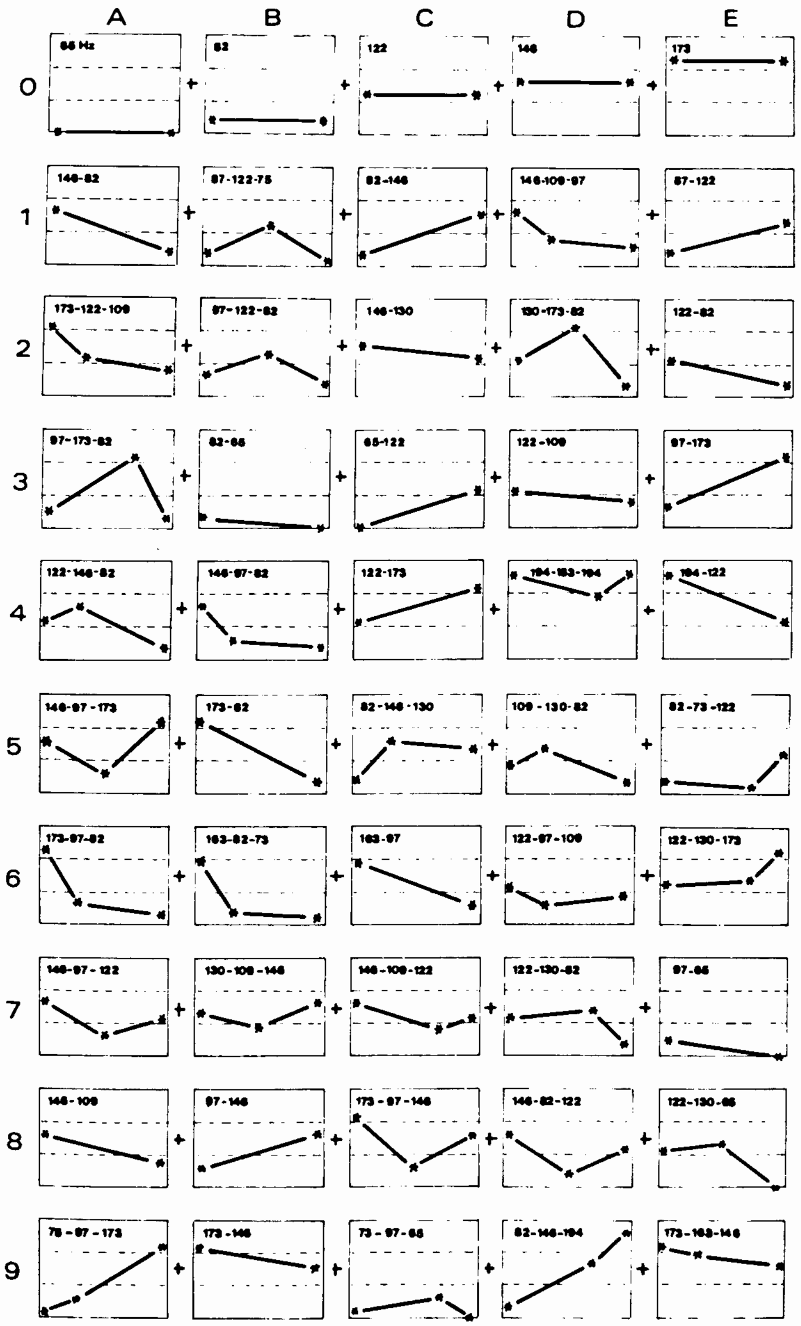


Fig. 1. F_0 patterns in Voxton used for automatic synthesis of Hungarian speech intonation.

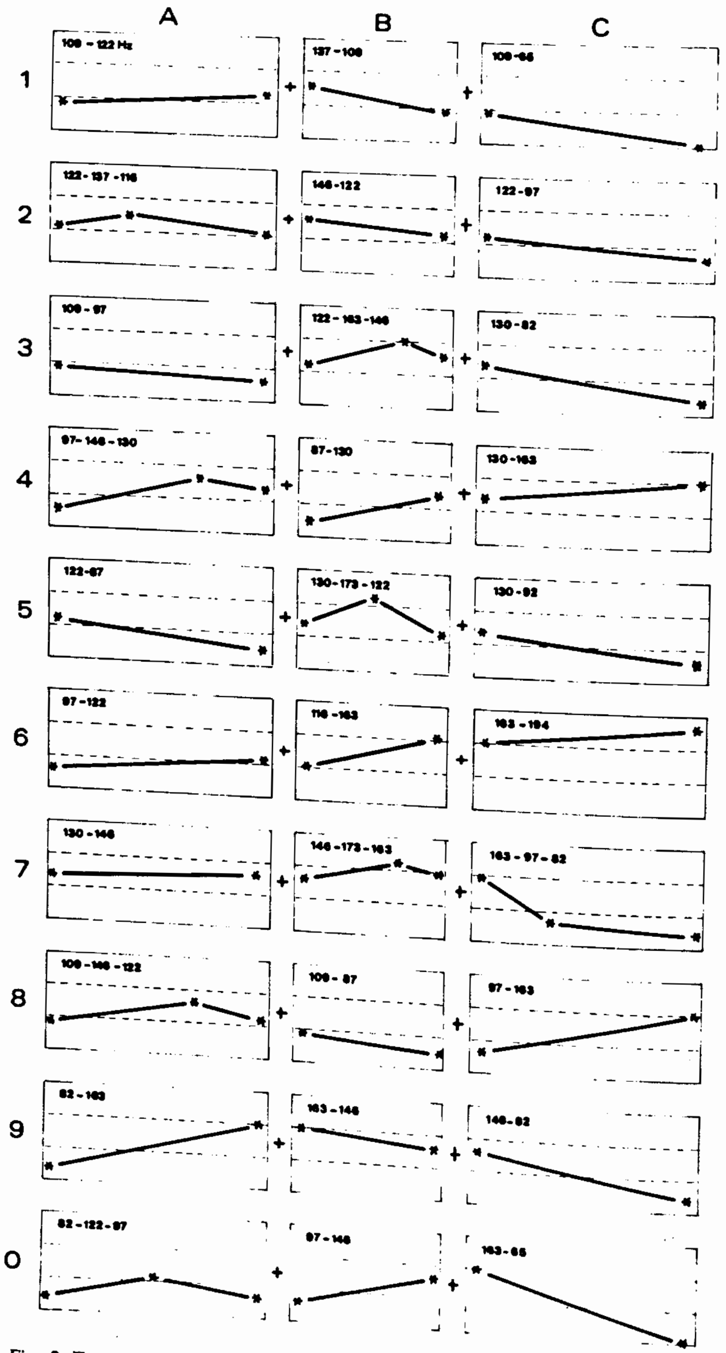


Fig. 2. F_0 patterns in Russon used for automatic synthesis of Russian speech intonation.

FORTRAN IV. The mode of producing speech based on rules is that of formant-synthesis. It is suitable for producing any kind of Hungarian or Russian text. The length of the texts which can be synthesized at one time is 5 s. The prescribed succession of sounds starts speaking with a delay of 30 s. The real speed of the speech is between 0.1 sounds/s to 25 sounds/s, but it can be altered between 6 sounds/s to 20 sounds/s. Voxton and Russon are phonetic systems i.e. their constructions and functions folow the phonetic—phonological systems of the Hungarian and Russian languages. There are three levels represented in its construction: a) the physical–structural level of the acoustic characteristics of speech, b) the so-called empiric phonetic level, c) the abstract phonological system-level of the language. These parts, which are well separable, easy to survey and stand in close connection with each other, constitute our phonetic synthesizing system as a structurally and functionally arranged whole. Any component of the acoustic structure can easily be changed within wide limits. The sound elements with their specific duration, pitch and intensity take part in the building up of the segmental structure. The formation of the suprasegmental structure of sounds is possible in three different ways: a) we manually give the F_0, A_0, Ac, T and tempo values by TON, ERO, IDO and IRA commands, b) by the automatic building of the intonational model chosen from the F_0 matrix and c) by the automatic transition of intonational structures from the patterns of intonation. The intelligibility of the speech produced by Voxton and Russon can be said to be good.

Reference

Bolla, K. (1982). Voxton: A system generating impersonal Hungarian speech by rule. *HPP* **10**.

Relative Importance of Parameters in Voice Similarity Judgment

R. Brown
Singapore

The field of auditory speaker recognition is concerned with the ability of human listeners to recognise a speaker's identity from hearing a sample of his speech. It involves a pattern-matching technique; on hearing the sample, the listener abstracts a representation of the voice which he then compares with an internalised reference pattern. Research in the field has concentrated on specifying those acoustic features which compose such voice patterns. Many experimenters have manipulated one feature in isolation, or isolated the glottal or vocal-tract contributions to voices (laryngograph, vocoder, inverse filtering, whispering, using an electrical larynx, etc.). However, results from such experiments indicate that each of the features investigated in isolation contributes something to speaker recognisability. Of greater relevance, therefore, to not only experimental, but also everyday speaker recognition is a statement of the *relative* importance of features.

The task in the present experiments is one not strictly of speaker recognition, but of voice similarity judgment, on the principle that the more similar a pair of voices are judged to be, the more difficult they will be to differentiate in a speaker recognition experiment, and vice versa. Synthetic voices were used, produced on a PAT synthesiser (Anthony and Lawrence, 1962). Stimulus samples consisted of various combinations of high and low values for the eight parameters below. The control sample, with which stimulus samples were paired for comparison, contained mid values for all eight parameters. With the exception of parameters 3, 6 and 7 below, these mid values were taken from a live utterance by the author.

1. Formant (F) range.
   High: approx. 30% increase in control value
   Low: approx. 30% decrease in control value
2. F mean.
   High: 15% increase in control value
   Low: 15% decrease in controle value
3. F bandwidth.
   High: 150 Hz. Control: 100z.
   Low: 50 Hz.
4. Fundamental frequency (F_0) mean.
   High: 20% increase in control value
   Low: 20% decrease in control value

5. F_0 range.
   High: approx. 45% increase in control value
   Low: approx. 45% decrease in control value
6. Larynx amplitude mean.
   Agreed auditory categorisations of loud, moderate and quiet (owing to lack of instrumentation)
7. Whisperiness.
   Agreed auditory categorisations of extreme, moderate and slight whisperiness
8. Tempo mean.
   High: 10% increase in control value
   Low: 10% decrease in control value

A homogeneous set of listeners were required to judge the similarity of pairs of voices (the control followed by a stimulus) on a 100-point scale ranging from SIMILAR (0) to DIFFERENT (100). A short-term memory task was set between the randomised trial presentations. Two replicates of a one-quarter replicate factorial design were employed, presenting 8 listeners with 16 trials each. A second experiment was carried out employing a full factorial design with 16 listeners, 16 trials per listener and the first four of the above factors, selected on the basis of the results of the first experiment. The results of the second experiment are therefore more reliable than those of the first.

Table I indicates the main effects and the 6 most important second-order interactions of the factors. These are expressed as shifts along the 100-point

Table I. Main effects and second-order interactions in the 2 voice similarity judgment experiments

	Experiment 1	Experiment 2
Main effects		
F range	1.87	-1.65
F mean	5.88**	-7.93**
F bandwidth	-4.48**	-10.07**
F_0 mean	-3.52*	-4.96**
F_0 range	-2.23	
Amplitude mean	1.76	
Whisperiness	0.54	
Tempo mean	-6.40**	
Second-order interactions		
F range/F mean	-4.99**	5.86**
F range/F bandwidth	-1.85	-5.54**
F range/F_0 mean	-1.46	1.22
F mean/F bandwidth	-0.37	-1.30
F mean/F_0 mean	0.80	-2.48*
F bandwidth/F_0 mean	0.13	-1.77

* Significant, $p < 0.05$.
** Significant, $p < 0.01$.

response scale (positively towards the DIFFERENT end, negatively towards the SIMILAR end). Main effects represent half the average difference in response between samples containing the factor at the high level against the low. Second-order interactions indicate the effect of having both factors at the same (high or low) level.

There are three main conclusions:

1. There is justification for the adoption of a design implying a linear model whereby a listener's response for a particular factorial combination is expressed as the sum of the mean response for that listener, the values of the appropriate main effects and interactions and an error factor. Although listeners differed in the average level of performance, their reactions to changes in the factors did not differ significantly.
2. F mean, F bandwidth and F_0 mean were consistently found to be significant. The reverse in polarity between the two experiments for F mean and for F mean/F range is, however, worrying.
3. Tempo mean, found to be significant in Experiment 1, deserves further investigation as a speaker-characterising feature.

Reference

Anthony, J. and Lawrence, W. (1962). A resonance analogue speech synthesizer. *Proceedings of the 4th International Congress on Acoustics, Copenhagen.* Paper G43.

The Analysis of Polish Intonation by Synthesis

E. Földi
Debrecen, Hungary

1. Introduction

Intonation as a means of expressing thoughts and emotions has an important role in understanding the message conveyed by speech. Therefore it is necessary to study its components, to describe them systematically, and to use results of the analysis in practice. Hungarian and foreign studies show a great diversity of opinions about intonation. Phoneticians do not have a uniform view of the term itself.

These different views can be divided into two large groups: the first one gives a "narrow" and the second one gives a "broad" interpretation of intonation. Intonation in the narrow sense refers to the tune as produced by changing the fundamental frequency. According to this view intonation is determined by a single component, and the interrelated acoustic parameters (i.e. the elements based on intensity and temporal factors) which accompany – or more exactly covary with – the changes of the fundamental frequency are examined independently, as secundary phonetic features, usually called suprasegmental, prosodic or sometimes phonetic features.

In the second interpretation – which has been recently gaining ground – intonation is a phonetic subsystem of more than one acoustic component which is determined by the changes of fundamental frequency, intensity and temporal components (quantity, tempo, rhythm) and elements of sound quality. However, it has to be noted that there is no agreement among phoneticians about the definition of the elements which constitute intonation – sometimes not even on opinions concerning the intonation system of one and the same language.

In my study the terms "intonation" and "speech intonation" are used synonymously with the term "phonetic construction" as defined by Kálmán Bolla. In this sense intonation – like speech sounds – is a complex unit of many components. Speech sounds can be characterized acoustically by determining their pitch, intensity, length and spectrum. Phonetic structure is also determined by these factors. Thus, as a result of the nature of its constituents, intonation is a linguistically relevant independent phonetic subsystem which can be viewed and described physically.

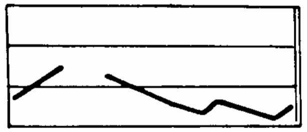
This paper presents the results of analysis and synthesis of Polish assertive, interrogative and negative sentences. The goal of the analysis was to study four elements of the intonation of speech (tempo, rhythm, tune, intensity). The synthesis was the verification of the data measured during analysis.

The recordings were made in the Acoustic Phonetics Laboratory of the Polish Academy of Sciences in Poznan. The corpus consisted of 150 sentences – simple, complex and incomplete – which were compiled as to fit the goals of the analysis. The recordings made with Barbara Klusinska were used in the experiment. The measurements were carried out with an FFM 650 Fundamental Frequency Meter, an IM 360 Intensity Meter, a 34 T four-channel mingograph, a PDP 11/34 computer and an OVE III speech synthesizer.

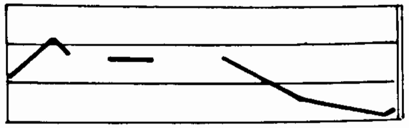
2. Results

The analysis of the intonation patterns of the sentences yielded the following results:

1. Polish affirmative sentences have a characteristic relatively steady rhythm, a relatively narrow register, a steady slightly falling and a slightly rising–falling tone, i.e. there is no abrupt change in the fundamental frequency.



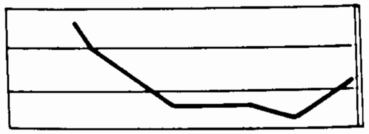
[na pra- vo]
(to the right)



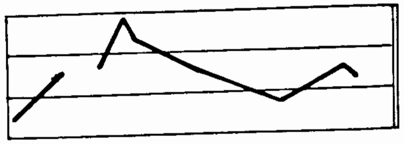
[on pra- t͜su- je]
(he works)

I measured 9.3 sounds/s as the average tempo of the examined Polish affirmative sentences. The tempo of the affirmative sentences was the slowest in relation to the types of interrogative and imperative sentences. The rhythm of the statements could be described as even, the duration data of the syllables compared to each other did not show any bigger change. The pitch of fundamental frequency fell on the first or second syllable of the sentence. The values of F_0 are less as compared to interrogative and imperative sentences, while no abrupt change of F_0 can be observed. The direction of change in intensity of affirmative sentences can be increasing–decreasing and increasing–equal–decreasing, their dynamics changed slightly according to the degree of change in intensity. Fundamental frequency and intensity usually occurred in parallel.

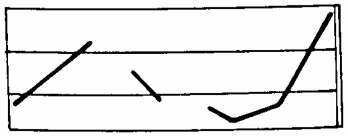
2. The question word questions and the yes-no questions examined are similar in the sense that compared to the imperative sentences their register is wider and their rhythm is usually accelerating–decelerating. The intonation pattern of the two types of interrogative sentences is different: the question word type is characterized by a rapidly falling–slightly falling–slightly rising tone or a rapidly rising–rapidly falling–slightly falling tone; yes-no questions, on the other hand, are realized on a slightly falling–rapidly rising tone or on a slightly rising–slightly falling–rapidly rising tone. Thus the abrupt change in the fundamental frequency is at the beginning of the sentence in the former case, while it is at the end of the sentence in the latter one.



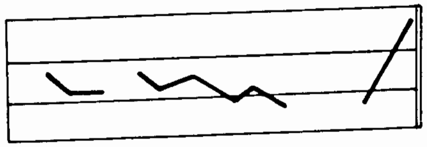
[fktu-rɨm do- mu]
(in which house?)



[do ko-go d͜ʑvo- n'iʃ]
(who are you calling up?)



[on pra-t͜su- je
(does he work?)

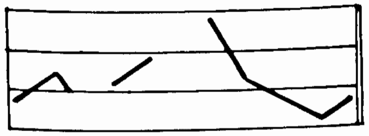


[t͡ʃu-jeʃ ruʒ-n'i- t͜sẽ]
(do you feel the difference?)

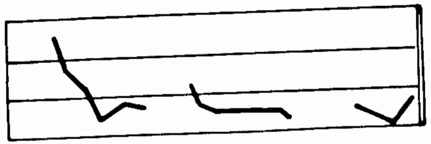
The tempo of yes-no questions was 9.94 sounds/s, which is quicker than the affirmative and imperative sentences, but it is slower than in question word questions. Their rhythms are variable, they are generally quickening and slowing down. The changes of fundamental frequency were always on the last syllable of the sentences. The changes of intensity occurred on the first and/or the second syllable, the minimum values fell on the last syllable. The dynamics of the sentences showed slight change.

Among the examined types of sentences the tempo of the question word questions was the quickest, it was 10.8 sounds/s. Their rhythms slowed down, more rarely equal/quickening–slowing. The F_0-pitch values – with the exception of the one word questions – always occurred on the first and/or on the second syllables of the sentences. The change of intensity and fundamental frequency took place in parallel. The direction of the change in intensity was mainly decreasing.

3. Imperative sentences are characterized by a wide register, a changing rhythm and a rapidly rising–slightly falling, rapidly falling–slightly falling or a level–rapidly rising–rapidly falling–slightly falling tone.



[on pra- t͜su- je]
(he does work!)



[t͜so za ʃt͡ʃẽ- ɕt͡ɕe]
(what lunch it is!)

The average tempo of the imperative sentences was 9.5 sounds/s, which is higher than that of affirmative sentences, but it is less considering the types of the two questions. Their rhythms are strongly variable, generally quickening–slowing. The place of changes of F_0 and intensity varied but they generally fell on the first syllable, and usually occurred in parallel. According to the direction of change in intensity the exclamations were decreasing, strengthening–decreasing, equal–decreasing or strengthening–equal–decreasing.

3. Conclusion

Intonation patterns characterizing the different sentence types are produced by considerable variation of the value changes of the components. Thus, the changing acoustic components produce different patterns which have linguistically different functions. Thus, as a result of the nature of its constituents, intonation is a linguistically relevant independent phonetic subsystem which can be viewed and described physically.

Comparing the intonation pattern of Polish and Hungarian sentences, I have found that the greatest difference is in the yes-no questions. The Hungarian intonation pattern in this case is usually slightly falling while it is rapidly rising in Polish. There is a greater similarity between the Hungarian and the Polish assertive and imperative sentences. The register of the Polish native speakers is wider, but there are no marked differences in the tempo of speech.

References

Bolla, K. (1978). Phonetic transcription. *Hungarian Papers in Phonetics (HPP)* **2**.

Bolla, K. (1979). The intonational analysis of speech·flow and the phonetic representation of intonation. *HPP* **3**.

Bolla, K. (1979). On the interlingual comparison of phonetic constructions (an outline of the problems involved). *HPP* **5**.

Dluska M. (1976). *Prozodia jezyka polskiego.* Warszawa.

Dukiewicz, L. (1978). *Intonacja wypowiedzi polskich.* Wroclaw.

Földi, E. (1979). Intonational means of expressing questionhood in Hungarian and Polish. *HPP* **5**.

Jassem W. (1962). *Akcent jezyka polskiego.* Wroclaw.

Jassem, W. (1973). *Podstawy fonetyki akustycznej.* Warszawa.

Mettas, O. (1981). *Les techniques de la phonétique instrumentale et l'intionation.* Bruxelles.

Wierzchowska, B. (1971). *Wymowa polska.* Warszawa.

Zinder, L.R. (1979). *Obscaja fonetika.* Moskva.

Recognition of a Spanish VV Sequence

J.A. Gurlekian and H.E. Franco
Buenos Aires, Argentina

1. Introduction

The fact that Spanish vowels are accurately identified in isolation and in context seems to present a suitable basis to explore the relevance of some dynamic information cues relative to steady state information cues. In this paper we will attempt to define some interactions of such types of acoustic information which consider listeners' responses to each of the speech-like sequences synthesized.

The Spanish language presents a relatively simple vocalic system with only five categories, while duration is phonologically not distinctive. The five vowels are phonetically monophthongs and they possess a known perceptual stability in running speech.

Recognition of natural and synthetic Spanish vowels in isolation is based on frequency bands sustained along preferred targets. Using mixed speakers, identification scores of 100% were obtained for natural sounds (Guirao and Manrique, 1972) and 100% for synthetic stimuli generated along the F_1-F_2 plane (Manrique and Gurlekian, 1976). In these experiments no type of training was necessary.

On the other hand changing spectral cues have been claimed to be responsible for high identification scores of vowels in other languages. English listeners in particular seem to improve recognition when this type of information is available. This occurs principally in vowels with consonant context (Strange et al., 1976; Gottfried and Strange, 1980) and it also occurs in diphthongized vowels, (Assman et al. , 1982). Strange and Gottfried (1980) obtained the lowest error rate in the identification of the isolated English vowels /o/ and /e/. Although recent experiments have shown that spectral changes, as provided by consonant context, are not an essential condition for vowel identification in the English language, higher identification scores are obtained for all isolated vowels, when listeners' training and task conditions are considered (Macchi, 1980; Diehl, 1981; Assmann, 1982). In a consonant context situation Spanish listeners preserve high identifiability of natural vowels, in spite of the fact that steady state cues are altered, showing that a dynamic spectral pattern, clearly a non-essential cue, may interact with stationary cues like those stated at the isolated emission level.

With this perspective, steady state cues which completely define the Spanish vowels in isolation will be presented for identification with additional dynamic information defined with relation to a context vowel. In the first experiment we expect to determine the effect of the spectral distances to different context vowels on the same initial frequency resonances of the vowels under study, and the effect of vowel duration, as a variation of the steady state cue, on the spectral distance to the context vowel. In the second experiment we explore the relevance of the temporal sequence on the identification of the vocalic pairs presented in Experiment 1, with the stimuli presented backwards. In a third experiment the effect of three different constant levels of fundamental frequency is analyzed in terms of the identification scores for the three different tested vowels.

2. Experimental Procedure

The sound /a/ with formant transitions is considered the 'context' or affecting vowel, based mainly on the hypothesis of perceptual stability of this central sound. This selection allows us a simpler and wider control of the formant transitions, resonances and durations of the vowel context in the synthesis process than the control required for a consonant environment where the perceptual stability of the affecting context might be lost and therefore confuse the listeners in their responses to the affected vowel. The vocalic segments under study consist of 'o1': a good example of an isolated vowel /o/ (100% identification) and two vowels used as control 'o2' (80% identification as /o/) and 'o3' (50% identification as /o/) obtained from a /u-o-a/ Spanish vocalic continuum presented for identification in a free choice condition (Manrique and Gurlekian, 1976). Two /a/ vowels are synthesized, both with formant values taken from the same continuum. These values can be considered as those produced in isolation by two different vocal tract lengths, one is called 'normal' /a/ or a(n), and the second corresponding to a shorter vocal tract (viz. higher resonances) is called 'elevated' /a/ or a(e). The vocalic segments are all combined as /o-a/ sequences through linear interpolation. Also, for each combination, we reduce the duration of the first vocalic segment systematically, thus imitating a speaking rate effect as it occurs in running speech.

In Experiment 1 the stimulus variations are: (1) duration of the first vocalic segment which ranged from 7.6 msec to 110 msec in steps of 25.6 msec, and (2) formant values, indicated in Fig. 1. The fundamental frequency was held constant at 130 Hz. The duration of the transition portion was also held constant at 70 msec for all the stimuli and the duration of the second vocalic segment was fixed to 110 msec. The stimuli were generated by a formant synthesizer designed by Klatt (1981). Three test tapes were generated. Each of them consisted basically of two VV sequences. One was made with a particular /o/ structure and the 'normal' /a/. The second VV sequence was composed of the same /o/ structure and the 'elevated' /a/. From these two sequences, five new stimuli were generated through digital segmentation of the first vocalic segment. This made a total of 10 stimuli which were then repeated at random five times for each identification test. Interstimulus intervals were always 5 sec. Ten native speakers, graduate students of the University of Buenos Aires, were asked to identify each sequence in the tests as the combination of two Spanish vowels.

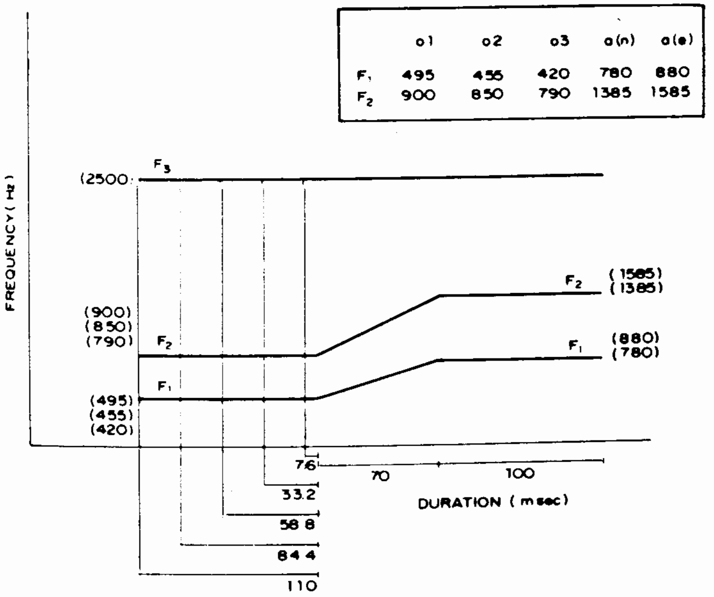


Figure 1. Acoustic patterns of the synthetic stimuli.

In Experiment 2 the synthetic waveforms of Exp. 1a, 1b, 1c were digitally reversed, and the same identification was requested.

In Experiment 3, the stimuli consisted of three basic VV sequences made with the three /o/ versions and the normal /a/ only. Each of these sequences have alternatively constant fundamental frequencies of 130 Hz, 230 Hz and 330 Hz. Three values of the duration of the vocalic segment /o-u/ was tested in different sessions. Again the subjects were asked to identify the stimuli as the concatenation of two Spanish vowels.

Both /oa/ and /ua/ are meaningless sequences in Spanish and they appear as syllables with similar frequencies of occurrence. Also, when they appear in CVV sequences before /a/, /o/ and /u/ seem to lose their oppositional character and to remain in free variation.

3. Results

The results of the experiments are presented as identification scores of the /oa/ sequence in every case. For each of the /o/ vowels tested in Exp. 1a, 1b and 1c there are two identification functions in Fig. 2, showing the effect of spectral distances (different slopes) to a(n) or to a(e) upon different dura-

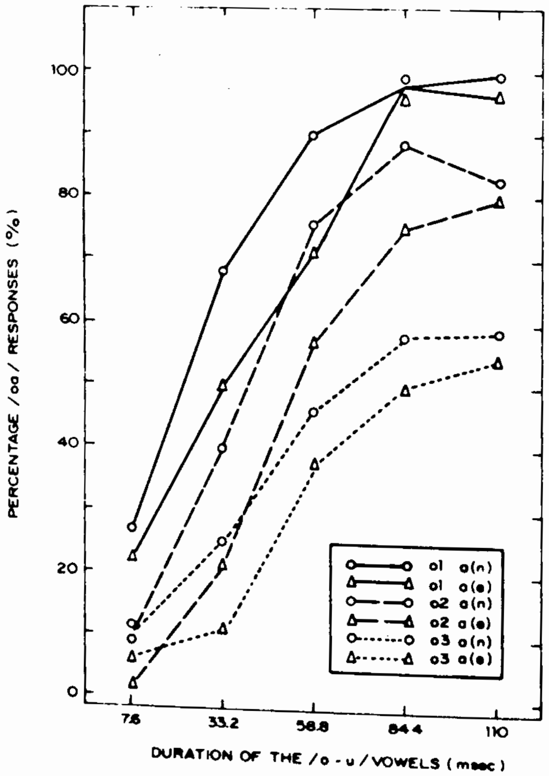


Figure 2. Average percent /oa/ identification for the vocalic sequences defined in Fig. 1. The differences to 100% correspond to /ua/.

tions of the /o/ segment. Higher identification as /u/ is observed when greater spectral distances occur between the vocalic segments in the identification functions, as shown in Fig. 2. Identification scores for o1, o2 and o3 at the maximum duration tested closely correspond to the percentages obtained in the isolated identification of this same segment in the previous work of Marique and Gurlekian (1976). At shorter durations of the /o-u/ segment, identification moves progressively to /u/.

These results indicate that: (1) recognition of a vowel like /o/ in isolation can be changed towards /u/ when affected by the spectral distance to a context vowel, despite the fact that formant values and the duration of the vocalic segment remain unchanged, (2) as a second and stronger effect, shorter durations reduce the information of the steady state cue reciprocally increasing the influence of the context vowel and the associated transitions. Trading relations can be established for both a higher spectral distance and a shorter duration of the vocalic segment and formant frequencies of the sounds tested.

Results from Experiment 2 (Fig. 3) show lower percentages of identification as /o/ for o1 and o2 versions than those obtained in Exp. 1. This is an indication that the temporal distribution of the acoustic cues: resonance frequencies of /a/ plus transitions that can be considered adequate to /u/ in running speech, and conflictive resonance frequencies for /o/ or /u/ may interact asymmetrically according to responses in Exp. 1. and Exp. 2 where all spectral and temporal parameters are the same. We have hypothesized that a preassignment occurs in the perceptual process as the listeners find some typical speech pattern and this first decision is highly responsible for our final phonological assignment. In Exp. 2 a clear acoustic pattern of /au/ is available at first, then a conflictive pattern of /u-o/ appears that diminishes the listeners' final responses. When the stimulus is in reverse order as in Exp. 1, a weaker preassigment occurs but is less influenced by the context vowel. In Exp. 3 it is observed that for each fixed duration the increase of fundamental frequency produces a significant increment of /u/ responses as shown in Table I. These responses may indicate a tendency for normalization of vocal tract resonances using fundamental frequency as a factor of normalization. In the Spanish vowel system, areas of /u/, /o/ and /a/ are aligned in a radial direction in the F_1-F_2 space. Due to the proximity of /u/ and /o/ areas, a partial overlapping occurs between them when considering emissions from

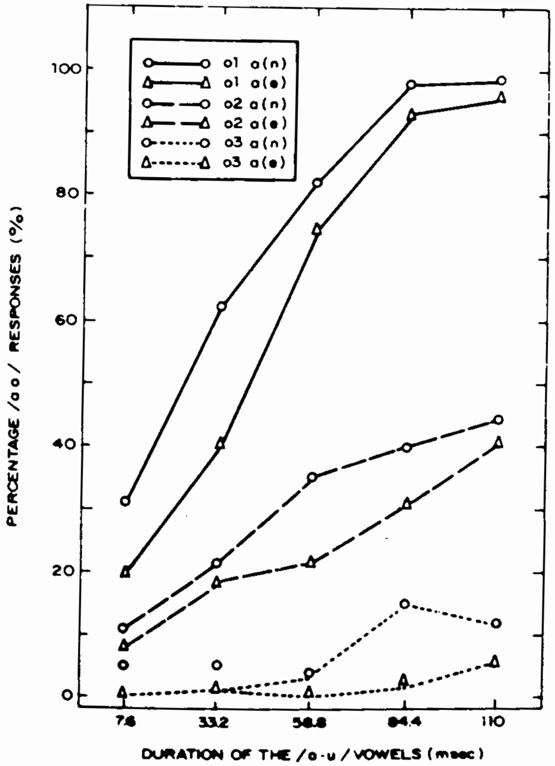


Figure 3. Average percent /ao/ identification for the vocalic sequences presented backwards. The differences to 100% correspond to /au/.

Table I. Percent of /oa/ responses

Duration of the /o-u/ segment	F_0	Vowel version		
	(msec)	(Hz)		
		o1	o2	o3
110	130	100	77.5	2.5
	230	100	42.5	5
	330	55	20	0
58.8	130	97.5	77.5	12.5
	230	77.5	12.5	2.5
	330	57.5	12.5	2.5
7.6	130	52.5	27.5	0
	230	42.5	0	0
	330	27.5	2.5	0

different speakers. Since there is a correlation between the fundamental frequency ranges and the different vocal tract lengths for children, women and men, the results suggest that listeners have used F_0 (in the absence of higher formant information) to assign the same formant pattern to /o/ or /u/.

References

Assman, P.F., Nearey, T.M., and Hogan, J.T. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *J. Acoust. Soc. Am.* **71**, 975-989.

Diehl R.L., McCusker, S.B., and Chapman, L.S. (1981). Perceiving vowels in isolation and in consonantal context. *J. Acoust. Soc. Am.* **68**, 239-248.

Gottfried, T.L., and Strange, W. (1980). Identification of coarticulated vowels. *J. Acoust. Soc. Am.* **68**, 1626-1635.

Guirao, M., and Manrique, A.M.B. (1972). Identification of Spanish vowels. In: Rigault, A., and Charbonneau, R. (eds.), *Proceedings of the Seventh International Congress of Phonetic Sciences*, 514-520. The Hague: Mouton.

Klatt, H.D. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* **67**, 971-995.

Macchi, M.H. (1981). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *J. Acoust. Soc. Am.* **68**, 1636-1642.

Manrique, A.M.B., and Gurlekian, J.A. (1976). Perception of a Spanish vowel continuum. *III World Congress of Phoneticians, Tokyo, Japan, Aug. 23-28*, 70.

Strange, W., Verbrugge, R.P., Shankweiler, D.P., and Edman, T.R. (1976). Consonantal environment specifies vowel identity. *J. Acoust. Soc. Am.* **60**, 213-224.

A Phonetically Based Data and Rule System for the Real-Time Text to Speech Synthesis of Hungarian

G. Olaszy
Budapest, Hungary

1. Introduction

Synthetic speech becomes more and more the focus of scientific, industrial and other applications. Speech synthesis by rule is a language-oriented task that means that the acoustical structure and the rules for building speech sounds, sound combinations and longer building elements have to be researched for every specific language.

The research of formant synthesis of Hungarian by rule has been done at the Institute of Linguistics in Hungary since 1979. During this research work we used a self developed analysis by synthesis method (Fig. 1.) to establish the data of the frequency, intensity and time structure of Hungarian speech sounds, sound combinations and longer sequences in such a form that the results could be used for the automatic building of words, and sentences by a computer and could be converted into speech by a formant synthesizer. For the analysis we used a Sound Spectrograph 700 and an Intensity Meter IM360, for the synthesis an OVE III Speech Synthesizer controlled by a PDP 11/34 computer.

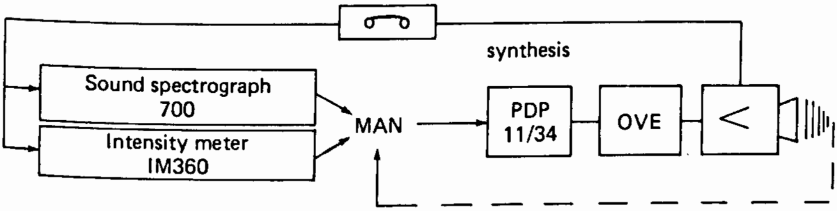


Fig. 1. Schematic diagram of the analysis-by-synthesis method used.

The software of the above mentioned synthesizing system was developed and made by Gabor Kiss (Kiss and Olaszy 1982).

2. Results

Our latest result (April 1982) is a *Hungarian speaking real-time text to speech synthesizing system* (named UNIVOICE), by which we can convert any kind of text into speech in real time. The system works without any vocabulary. It can be used from a typewriter keyboard. The typed text will be spoken by the synthesizer immediately after giving the full stop (or !, ?) at the end of the sound sequence. Keyboard operations can be replaced by an ASCII code stream of the letters of the required text. Using this form of operation the system 'utters' the previously coded text automatically. As far as we know this is the first real-time text to speech synthesizing system that accepts Hungarian spelling rules and converts written text immediately into speech.

In order to determine the intelligibility of the system's output, perception tests were carried out by the author (see Gósy and Olaszy 1983). On the basis of test material collected by the linguist-phonetician M. Gósy, 70 subjects were asked to test syllable, word and sentence size units generated by means of UNIVOICE system. According to the test results the synthetic speech proved to be well understandable.

In this paper I give a short description of the data base of the UNIVOICE system and some rules used to build up the speech from the elements of the data base.

Synthesis by rule demands a data base that contains the necessary building elements of a language to be synthesized, and a computer program that can handle the data base according to the rules given.

The aim of our research work was to create a real-time text to speech system for Hungarian. Practically we had to place the data base and the program as well in the central memory of the computer. It means that we had only limited place for both of them (our PDP has 28 K word memory). Thus attempts had to be made to find the minimal number of building elements of the Hungarian language and the optimal number of speech sound parts. We can do this minimalisation and optimalisation only when we know the exact acoustical structure of the language examined and the technical operation of the synthesizer we use. This implies that - on the one hand - research had to be done to discover the frequency, intensity and time structure of the speech sounds, inherent sound parts, sound combinations, transition phases etc. in Hungarian. On the other hand specific technical and physical knowledge was necessary for finding how the acoustical data of the human speech can be used by an electrical system of a limited scope. Man produces speech with a biological system, we have to do it with a technical one.

3. Discussion

For Hungarian we found that the minimalised and optimalised data base of the language for OVE III contains 370 speech sound elements. This data base was developed in 1980-81 (Olaszy 1981, 1982a). The 370 elements are not speech sounds or sound combinations but they are speech sound parts.

Minimalisation of this data base implies that one sound part (element) can have the function of representing not only one speech sound but provides information at all places in sound sequences where the acoustical structure of this sound part meets the phonetically necessary requirements. This results in sound parts that can be used at several places of the synthesized sound sequence and of course, there are those as well that can be used in the building process at only one position. Doing this process several times the number of building elements in the data base can be still further decreased. Finally we get the minimal number of building elements (in our case 370) that are necessary to build up any kind of words, sentences etc.

Optimalisation means that one has to find the optimal number of sound parts in the synthesis of a speech sound or sound combination. The more sound parts are used the better the sound quality is, and consequently the greater the memory demand. For example we can build a [b] sound from two, three, four or more sound elements. If we do it from two, the quality of the sound is bad, if we do it from three is becomes better and so on.

In the UNIVOICE system one speech sound and its transient phase consists of 3 or 4 sound parts. These sound parts are enough to realise the frequency, intensity and inherent time structure of the Hungarian speech sounds and sound combinations. One sound part is built from 1-5 microelements having the duration 4-50 ms. In one microelement the frequency and intensity data are constant. We can realise the formant and the intensity movements by making changes in the frequency and intensity data step by step in the microelements.

Using the UNIVOICE synthesizing system one can generate Hungarian speech sounds, syllables, words, words having no sense (for example for medical purposes), sentences and longer sequences as well. Non-Hungarian speech can be generated as well (English, German, Dutch, Finnish etc.) if we write the text phonetically using Hungarian letters. Of course the sound of any non-Hungarian language will be a little Hungarian-like because the UNIVOICE uses Hungarian phonemes only.

Hence by the synthesis process nearly every sound element of the data base can be linked to any other depending on the written text. This kind of operation demands that the 370 speech sound elements had to be planned in the way that if any of them comes into contact with an other - by the building of speech - the acoustical connection of them would be smooth without any transients, formant frequency shifts.

This data base for the real-time text to speech synthesis of Hungarian was developed for the OVE III speech synthesizer but it can be adapted for other formant synthesizers (for example MEA 8000) as well.

The program of the UNIVOICE system was developed and written by Gabor Kiss. How does it work? If we type a text on the keyboard of the terminal the program converts the letters to phonemes and phoneme combinations, after that it finds out which sound elements - from the 370 - are necessary to build the text, picks these elements from the data base and puts them into the appropriate order, determines the necessary melody pattern according to the punctuation marks typed at the end of the text and finally sends this data group to the input of the synthesizer.

The first demonstration of the UNIVOICE occurred at the 8th Colloquium on Acoustics, Budapest on the 6th of May 1982, where a 3 minute synthesized Hungarian speech was played to the audience (Olaszy 1982b)

A later version of UNIVOICE made it possible to make changes in the time structure of the typed text, melody patterns can be added at will and the intensity structure can be varied as well. For these changes the user only has to give some commands containing the data of the required time, melody or intensity structure.

4. Conclusion

Summarizing we can state that apart from a theoretical data base and rule system, a practical working model has been developed for synthesizing Hungarian of a good quality. The elasticity and speed of the system makes it useful for various industrial and other purposes.

References

Flanagan, J.L. and Rabiner, L.R. (1973). *Speech Synthesis*. Strassburg.

Gósy, M. and Olaszy, G. (1983). *The Perception of Machine Voice*. (Examination of the UNIVOICE, Hungarian speaking real-time text to speech synthesizing system). Nyelvtudományi kozlemények.

Kiss, G. and Olaszy, G. (1982). An Interactive Speech Synthesizing System with a Computer Controlled OVE III. *Hungarian Papers in Phonetics*, **10**, 21-45.

Olaszy, G. (1981). Preparation of Computer Formant Synthesis of Sound Sequences. *Hungarian Papers in Phonetics*, **8**, 147-59.

Olaszy, G. (1982*a*). The Analysis and Synthesis of the Hungarian Consonants and the Consonant-Vowel Combination Types. *Hungarian Papers in Phonetics*, **10**, 46-82.

Olaszy, G. (1982*b*). Some Rules for the Formant Synthesis of Hungarian. *8th Colloquium on Acoustics*, Budapest, Lectures, 204-10.

Semi-automatic Synthesis of Intonation for Dutch and British English

H.W. Zelle, J.R. de Pijper and J. 't Hart
Eindhoven, the Netherlands

1. Aim

The aim of the work reported here is to develop computer programs with which arbitrary utterances of restricted length in Dutch and English could be provided with acceptable pitch contours. The full melodic range of the language should be available. An additional requirement is that the amount of input data should be restricted to a minimum. This implies that the program should have built-in 'knowledge' of the rules and restrictions governing the internal structure of the pitch contours, and of accepted and forbidden combinations in each of the languages.

2. Background

We must make a distinction between Dutch and British English in view of differences in progress in our analysis of the intonation of the respective languages. Our earlier analysis of Dutch intonation has incorporated a corpus of considerable extension, which makes us confident enough to state that we have, in our 'grammar of Dutch intonation', incorporated practically all melodic possibilities. These results have been described in 't Hart and Collier (1975).

De Pijper's analysis of British English intonation has primarily concentrated on the most frequently occurring pattern of that language; in addition, it appeared possible to give explicit recipes for the construction of each of Halliday's (1970) seven Primary Tones (cf. 't Hart and de Pijper, 1983). There are reasons to doubt the adequacy of Halliday's analysis of the main features into precisely these seven categories. However, this is not the issue here. The examples recorded on Halliday's tape remain examples of fully acceptably intonated British English; this has been verified by de Pijper in listening experiments with native British subjects, (de Pijper, 1983).

The results of these analyses is a representation of both Dutch and English intonation in fully standardized form. In this only the perceptually relevant pitch movements come into play, in the shape of straight lines, with standard specifications of the size, slope and position of these movements in the syllable. On perceptual grounds, we have found no reason to apply a smoothing of these straight line contours.

3. Input data

It is assumed that an arbitrary speech file is stored on disk, in the form of formant, LPC, or channel vocoder parameters. These data may be derived from analysis of natural speech or obtained by diphone concatenation or allophone synthesis by rule. The experimenter has to provide the following additional data:

1. The temporal location of the vowel onsets of the syllables that should receive a pitch accent;
2. The location of syntactic boundaries to be marked intonationally;
   2a. In cases where this marking is required to be done by means of a continuation rise: the location of the last voiced frame before the boundary;
3. A code indicating the type of basis intonation pattern underlying the desired contour:
4. The choice made between a great number of optional variation possibilities.

Ad 1.
We do not have a sufficiently complete set of rules from which we can derive which words should be pitch-accented. Neither do we have rules for lexical stress of polysyllabic words. In the case of speech produced by diphone concatenation, the vowel onsets of all syllables are automatically available.

Ad 2.
In Dutch, the two most frequently occurring intonational markings for major syntactic boundaries are the postponed, or non-final fall, and the continuation rise followed by a (silent) fall. For the former, an indication of the frame number of the word boundary is sufficient. For the latter, the occurrence of a voiceless stretch immediately before the boundary would prevent its audibility. Therefore, the program also requires the location of the end of voicing.

The program for British English does not yet incorporate the feature of joining two sub-contours together by means of boundary markers. But since a continuation rise is one of the possibilities of that language as well, the same measure will have to be taken.

Ad 3.
For British English, there are Halliday's seven Primary Tones, numbered from 1 through 7, where 6 means Tone 13, and 7 means Tone 53 (in Halliday's terminology).

For Dutch, there are nine basic patterns in the program's menu. Pattern 1 is the 'hat-pattern' with choice of one to four pitch accents, but without continuation rise. Pattern 2 is the same, but with the option of one or more continuation rises at so-called primary boundaries. In no. 3 the final pitch accent is given by a rise, and in no. 4 the only one by a fall, the contour beginning high. No. 5 is the 'cap-pattern', with the typically 'scooped' rise, no. 6 is the 'valley-pattern' with gradual rise of 'inclination', nos. 7-9 are patterns with one or more half-falls ('call contours', 'terrace contours', or contours in which the half-fall serves as a boundary marker).

Ad 4.
Examples of optional variations are: postponing non-final falls to mark (secondary) boundaries, final rise, gradual fall, high onset. The program offers these options only if the intonation grammar allows these variations in the chosen pattern. Other options are: excursions other than standard (= 6 semitones), final frequency other than standard (= male 75 Hz, female 180 Hz).

4. Generating the pitch contour

All pitch movements called for by the choice of the pattern and by the options are to be superimposed on the declination line. The slope of this line is calculated from the automatically derived utterance duration with (D in semitones per second and t in seconds):

$D = -8.5/t$ for $t > 5$ s and $D = -11/(t+1.5)$ for $t \leqslant 5$ s (Cohen, Collier and 't Hart, 1982).

The program detects pauses within utterances (operationally defined as stretches of more than 250 ms during which the amplitude remains below a given threshold). During such pauses, declination is stopped and non-final falls are made 1 semitone smaller, such that a kind of 'declination resetting' takes place. Together with the standard (or chosen) final frequency, the slope defines the initial frequency. If a high onset is chosen, the initial frequency is 6 semitones higher (in the case of standard excursions). If pauses have been detected the initial frequency is adapted accordingly.

Figs. 1-3 show some examples. Fig. 1 refers to an English sentence (from Halliday's tape), Figs. 2 and 3 to a Dutch sentence, spoken without and with a pause, to show the effect of automatic declination resetting.

For each pattern, the grammar dictates the nature and order of the pitch movements, as well as their position with respect to the vowel onsets (or voicing offsets) of the syllables involved. As a consequence, the input data are sufficient for the program to generate the entire pitch contour.

5. Results and Applications

For Dutch intonation, the adequacy of the grammar has been tested on a large corpus earlier, and that of the fully standardized stylizations has been demonstrated on many occasions. For British intonation, the development

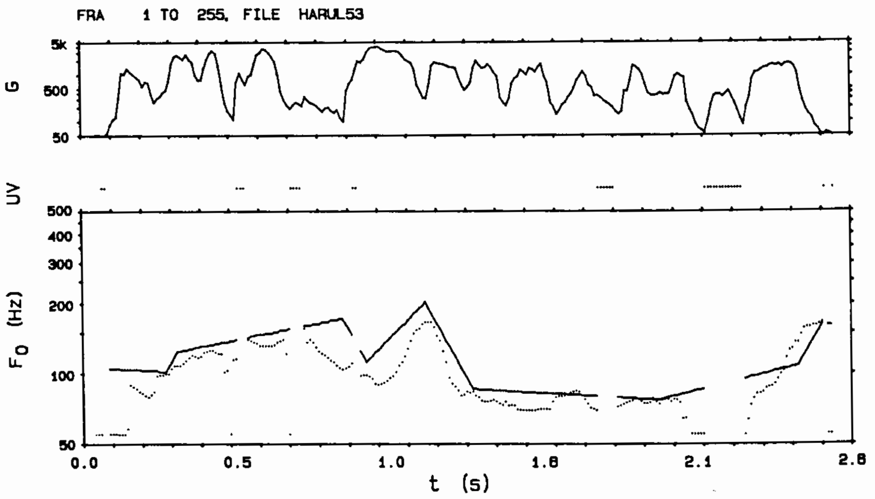


Fig. 1. Example of the sentence 'He's **never** taken **Jane** on any of his **visits** though', as spoken with Tone 53 (dashed), and as intonated artificially by means of the program (solid line). Necessary input data were: Tone: 7; Tonic (vowel onset of 'Jane'): 96; Pretonic (vowel onset of 'never'): 32; Secondary Tonic (vowel onset of 'visits'): 187; End of voicing: 250.

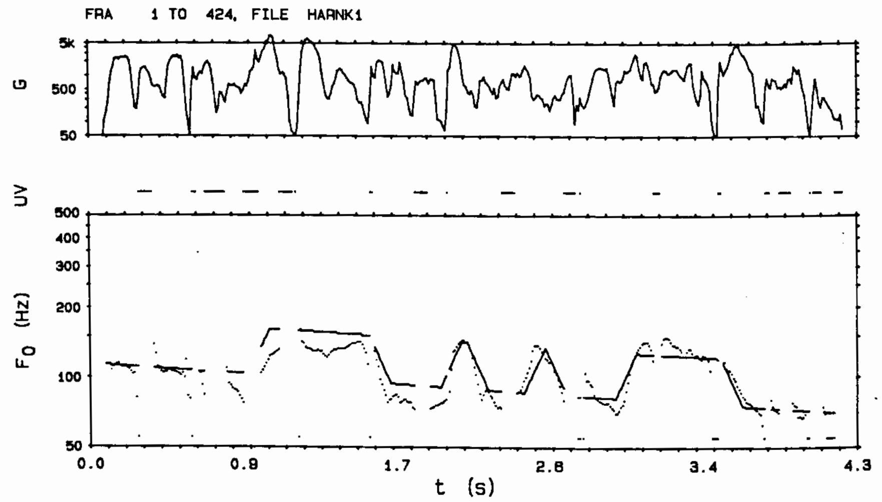


Fig. 2. Example of the Dutch sentence 'Wie het Fonetisch **Congres** bezoekt / en Amster**dam** wil zien // hoeft maar een **half** uur in de **trein** te zitten' (Who the Phonetic Congress attends / and Amsterdam wants (to) see // needs only a half hour in the train to sit) spoken (dashed) and intonated artificially (solid line). Input data were (free options omitted): Pattern: 2; No. of accents: 4; Flat hat between accents 1 and 2: no; Flat hat between accents 2 and 3: no; Flat hat between accents 3 and 4: yes; Frame no. accent 1: 96 (vowel onset of 'Congres'); Frame no. boundary: 157 (end of 'bijwoont'); Primary boundary?: no; Frame no. accent 2: 204 (vowel onset of '(Amster-)dam'); Frame no. boundary : 256 (end of voicing of 'zien'); Primary boundary?: yes; Frame no. accent 3: 302 (vowel onset of 'half'); Frame no. accent 4: 356 (vowel onset of 'trein').

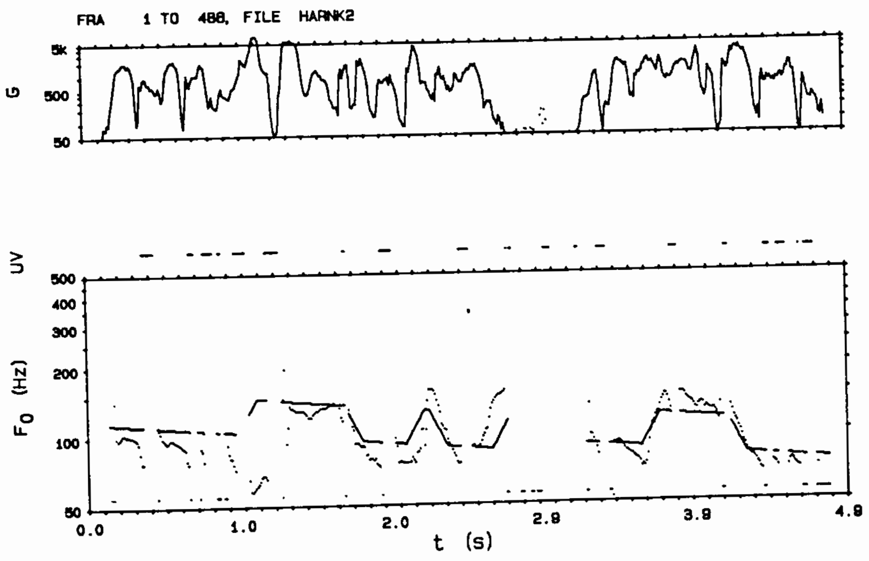


Fig. 3. The same sentence as in Fig. 2, now spoken with a pause, giving rise to automatic declination resetting.

of a grammar is still under way; on the other hand, testing of the acceptability of the stylized contours has been carried out very thoroughly with native users of the language ('t Hart and de Pijper, 1983).

Everyday application in the laboratory has been in speech (re-)synthesis activities where no original F_0 curves are available, such as in diphone synthesis, or in experiments in which the course of F_0 is a controlled experimental variable. The obvious advantage of the program is that experimenters no longer have to be instructed in the grammar of intonation before being able to actually produce acceptable synthetic pitch contours.

References

Cohen, A., Collier, R. and 't Hart, J. (1982). Declination: construct or intrinsic feature of speech pitch? *Phonetica* **39**, pp. 254-273.

Halliday, M.A.K. (1970). *A course in spoken English: intonation*, Oxford University Press, Oxford.

't Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis, *J. of Phonetics* **3**, pp. 235-255.

't Hart, J. and de Pijper, J.R. (1983). *Proc. Xth Int. Congress of Phonetic Sciences.*

de Pijper, J.R. (1983), *Modelling British English intonation. An analysis by resynthesis of British English intonation.* Foris Publications, Dordrecht.

Section 3
Acoustic Analysis and Coding of Speech

On the Acoustic Characterisation of the Oral and Nasal Vowels of French

M. Eskenazi and J.S. Liénard
Orsay, France

1. Introduction

The manner in which the vowels of any given language have been characterised up to the present has not fundamentally changed from the ideas put forth by Peterson and Barney (1952). It has depended on the quality of pnonologically and phonetically different vowels in that language as well as the relation of the vowels to one another in the acoustic dimension. Thus, the vowels of Japanese (there are five phonetically distinct entities (Komatsu et al., 1982)) could possibly be characterised by the type of information obtained from formant tracking. This implies that the phenomena of formants crossing one another, suddenly disappearing, or the apparition of a 'nasal' formant would be extremely unlikely to occur in Japanese. In French, there are generally considered to be ten oral vowels and three nasal vowels (for a typical Parisian speaker) – /a/, /ɔ/, /o/ /ɛ/, /e/, /i/ /y/, /u/, /œ/, /ø/, /ɛ̃/, /ɑ̃/, /ɔ̃/.

Despite this heavily populated vowel space, and due to the fact that there are no diphthongs or 'lax' vowels in French (so they may be considered to be 'steady state'), we shall show that these thirteen vowels may be sufficiently characterised by a relatively reduced amount of information in the frequency domain. This continues the work described in Liénard (1979). Articulatory representations, such as LPC, are not employed; rather an attempt is made to put to use the limited knowledge that presently exists on the manner in which the ear perceives sounds in the time and frequency domains.

After a description of the databases used, we explain the manner in which the speech signal was filtered and smoothed. A description of the simple statistics used to represent the dispersion of the vowels follows. The separation obtained, and the results of a first trial of unknown speaker vowel recognition are then presented as well as the explanation of a module that dynamically enlarges the base.

2. Databases

Several databases were made up: two to test the filtering, the spectral smoothings, and the statistical dispersions, and three to be used in the unknown speaker vowel recognition experiments.

A. Two test data bases

In order to explore different filtering and statistical approaches, and to confirm the dispersion results, two databases were constituted; one in context, and one of isolated vowels.

Both databases were recorded on a NAGRA IV S, with a BEYER M69N microphone. In the first there were 30 speakers, male and female, each recorded once (several repetitions of the list of sentences; the 'best' candidate for each vowel being retained). The frame sentence was, 'J'ai dit .... six fois' (/ʒɛdi .... sifwɑ/). The frame word always began with /t/, and ended either with one of the thirteen vowels mentioned above (example: thé /te/), or a consonant prolonging the duration of one of these vowels (example: thèse /tɛz/). The speech signal for each of the sentences was visualised, and the 50 ms portion to be used was indicated by hand (joystick).

The second, isolated vowel, database included ten speakers. These speakers pronounced the series of vowels three times at each recording session, and the 'best' candidate for each vowel was retained. Seven speakers (six male and one female) were recorded in two sessions (therefore two tokens of each vowel were present for each speaker); the other three speakers were recorded ten different times. The 50 ms portion of the vowel was obtained automatically : the starting point of the signal was detected, and the portion of interest was determined, after intelligibility tests, to begin at a fixed distance of 200 ms from this point.

B. Three recognition databases

For the unknown speaker recognition tests, three isolated vowel databases, alike in all ways except for the speakers included therein were constituted. Each included ten speakers, five male and five female, and were recorded on a REVOX B77, with the same microphone. There was only one recording session per person, but this time, of the five repetitions of the list that were requested, the three 'best' candidates for each vowel were retained. The 50 ms portions were excised in the same manner as above.

3. Filtering and Spectral Smoothing

Within 50 ms of speech and considering the possible use of a Hamming window, at least two, and as many as six, whole periods of the signal are present. This portion of vowel therefore seems to be correct as a base for obtaining a representative power spectrum. In all cases, the signal was preemphasized.

A. Filtering

In order to obtain the desired spectrum, two filtering methods were tested. For the first database, 32 fourth order filters (characterised according to a Bark scale) were used. Satisfactory vowel groupings were obtained; however, to confirm these groupings, and to lighten the computing load, an FFT (translated to a Bark scale) was also tried. Results on the same database were quite comparable. The two filtering methods were then tried on the second database, and after results were found to again be quite comparable, the FFT was chosen for use in further tests.

B. Spectral Smoothing

Directly after the FFT and the transformation of the resulting spectrum into 32 values separated according to a Bark scale, the linear values were transformed to quasi-logarithmic (base two segment approximation) ones. The statistical treatment described below was used on this original unsmoothed 32-point spectrum to determine whether the smoothed spectrum conveyed different information in this context. Recognition tests of the first of the three recognition databases on known speakers showed comparatively high error rates.

Spectral smoothing was carried out with two goals in mind. First, the original spectrum still conveys a considerable amount of information, part of which may be considered to be redundant for our needs. Smoothing eliminates accidental peaks and valleys, but respects the general distribution of energy in the spectrum. Second, inherent differences in amplitude variations from one vowel to another are not taken into account. Smoothing may be carried out in several stages, allowing us to subtract an extremely smoothed spectrum from a less smooth one (both coming from the same original vowel). The result is a series of 37 values for each vowel ($K = 1$ to 37) representing the degree of curvature of the spectrum over a wide (~1000 Hz) range. The values are independent of amplitude and therefore an /a/ pronounced very softly will not be confused with an /i/, nor a loud /i/ with an /a/.

The combinations of smoothings found to give an optimal characterisation of the spectrum was:

1. three-point averaging, with the central value weighted at 2
2. 24-point weighted smoothing
3. 9-point straight averaging
4. substraction of 3. from 2.

Figure 1 illustrates the evolution of three spectra from their original forms to the results of 4, where only an indication of the degree of curvature (c) of the spectrum remains.

The original vowels were also treated in another manner: normalising amplitude before FFT, and then proceeding up to 2. Results of known speaker vowel recognition tests were slightly less satisfying than the procedure described above. Further tests are being carried out using this approach.

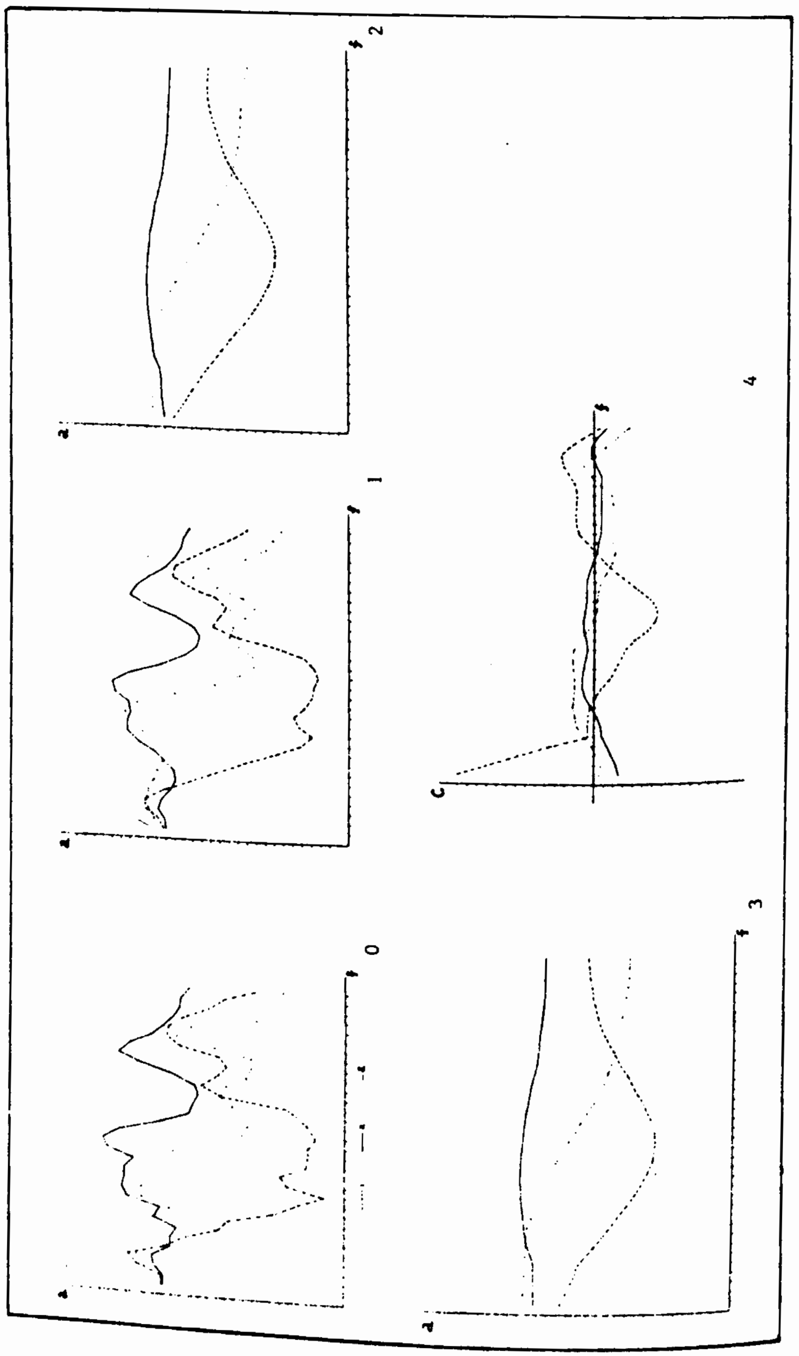


Fig. 1. Signal evolution from (0), the original power spectrum, to (4), amplitude-independent and smoothed.

4. Statistical Classification

The simple statistic tools chosen reflect the distance, at a given K, between the C values for different vowels (each 'vowel' now becoming a composite of all of the speakers in the given database).

First, the mean C value (m), and standard deviation (σ), for all the speakers for a given vowel were calculated at each K along the spectrum. We then calculated a dispersion value for all of the pairs of vowels at each K:

$$S\,(k, i, j) = \frac{|m_i - m_j|}{\sigma_i + \sigma_j}$$

A general indicator, (I(k)), of the dispersion of the vowels at each K, can be found by taking the mean value of the sum of the S(k,i,j) distances:

$$I\,(k) = 1/13^2\,[\sum_i \sum_j S(k, i, j)]$$

Figure 2 shows the Ik values for the second and third databases: the higher the value, the better the general dispersion.

Figure 3 shows the general dispersion of the vowels at K = 30 for database 3. It may be noted that such vowel pairs as /a/ and /i/, and /a/ and /ɛ/ have

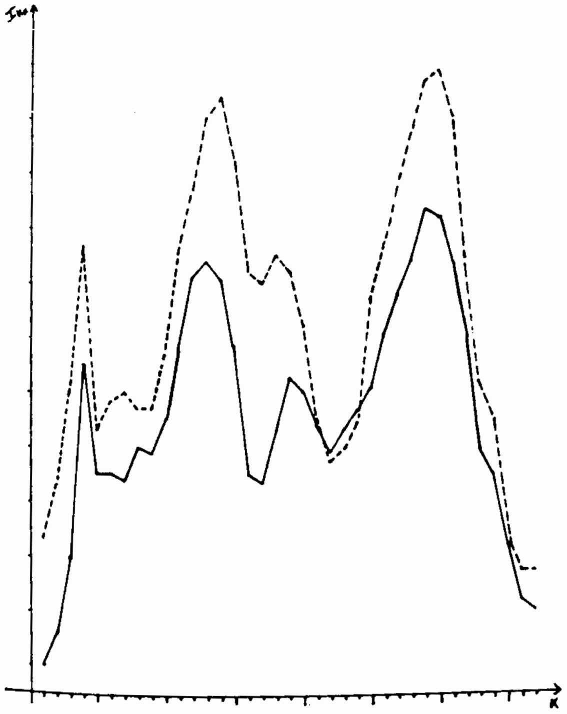


Fig. 2. Ik values for the second and third databases (dashes = database 2).

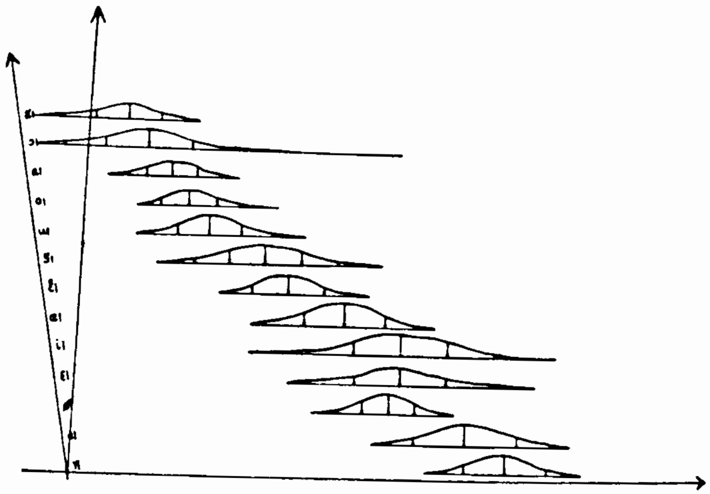


Fig. 3. General dispersion of the vowels at K = 30 (x axis = C, y Axis = vowels, z axis = N° of speakers).

well separated dispersions and cannot be confused with one another. At other values of K, the relations of the vowels among each other are different and, for example, at K = 14, /ɔ/ and /u/, and /u/ and /i/ are well separated.

The information contained in adjacent K's is quite similar, due to the smoothings. However, it would be interesting to see if an unknown vowel from an unknown speaker could be identified by superposition of certain Ks, or by trying to 'fit' the unknown spectrum against the 13 representative vowel curvature spectra obtained from the original database.

5. An Experiment in Vowel Recognition

Using the first of the three test databases to obtain the mean value and the standard deviation of each vowel at each K, we can filter and smooth a new unknown vowel from an unknown speaker (in the two other databases), and then compare its value at each K to the mean values of the 13 vowels in the base. The distance between the unknown, and a vowel mean value, at any given K is calculated as:

$$D_{uv} = \frac{|c_u - m_v|}{\sigma_v}$$

where u is the unknown vowel value at a given K, and v is the mean value of any one of the 13 base vowels. The shortest Duv distance should therefore indicate a likely vowel candidate. At K = 30, there is not enough information to make a decision on what the most likely vowel would be; however, a group of the three to four 'closest' candidates, almost always shows the vowel in question, and its closest 'neighbours' in the acoustic vowel space (for example, an unknown /a/ will often engender /o/, /ã/, /ɔ/, and /a/ as candidates).

We proceeded to test these distances on the whole of the spectrum. That is, taking the sum of the distances for all of the Ks for each of the 13 vowels, and seeing which vowel has the lowest cumulative distance from the unknown (the best 'total spectral fit'). Other tests were carried out to add up only half of the Ks (K = 2, K = 4, ... K = 36), one third, one fourth, one fifth, and one sixth of the Ks. Results are shown in Table I. Although so much of the information has been taken out of the signal, the correct vowel is closest to the unknown 45-50% of the time; it may be noted that when only half, one third, etc. of the Ks are used for comparison, the error rate does not change by more than a few points; one feature used for identifying oral and nasal vowels could only necessitate the retention of six m values and six σ values, New tests are presently being carried out on these same databases to see if the error rate decreases when only a few select Ks (the peaks shown in Figure 2, for example) are used for the cumulative distance calculations. Table II shows the confusion matrix for the total spectral fit.

Table I. Results of the vowel recognition experiment (as a percentage of 780 unknown vowels)

Test	Correct vowels proposed as place candidate (%)				
	1st	2nd	3rd	1st or 2nd	1st, 2nd, or 3rd
'Total Fit'	49.1	27.3	11.7	76.4	88.1
1/2 of the Ks	48.6	27.6	11.7	76.2	87.8
1/3 of the Ks	49.9	25.9	13.2	75.8	89.0
1/4 of the Ks	48.2	27.1	12.2	75.3	87.4
1/5 of the Ks	48.7	26.4	12.3	75.1	87.4
1/6 of the Ks	50.3	25.5	11.4	75.8	87.2

Table II. Confusion matrix for the total spectral fit

		confused with:												
		i	y	e	ɛ	ɛ̃	ø	œ	a	ã	ɔ	ɔ̃	o	u
unknown	i	72	6											
vowel	y	1	64	10	2		1							
	e	21	13	39	4							1		
	ɛ	1	4	17	45			1				10		
	ɛ̃				2	49		17	7		1	2		
	ø	1	3	1	6	1	46	13				7		
	œ		4		1	1	20	42	1	1	3	5		
	a					1			51	10	16			
	ã								5	32	38		3	
	ɔ								5	2	60	6	4	1
	ɔ̃						3	5			18	40	7	5
	o										24	4	46	4
	u									1	13	30	2	32

6. A Module to Dynamically Change the Statistics

The results of our recognition experiment may possibly be ameliorated in two ways; first, different manners of superposing the different Ks are being tried, second, the base of speakers is being enlarged to become more statistically representative, and therefore, used for unknown speakers with a decrease in error rates.

In order to enlarge the base for any vowel with any number of already known, or unknown speakers, we have developed a module that automatically changes the mean value, and the standard deviation value of the desired vowel. Information on the present status of the number of speakers in the base, the mean value, and the standard deviation, can be obtained at any time. Tests are now being carried out on the enlarged statistical base (now including all of the speakers in the second of the three test data bases).

7. Conclusions

We have shown that the ten oral and three nasal vowels of French may be characterized with quite a small amount of generalised information. This observation is in keeping with the work at LIMSI (Gauvain et al., 1983) which centers on the belief that the transitory parts of speech convey more information and therefore that less needs to be retained for the stable parts. Further work is being carried out on the enlarged statistical base (this base will soon be enlarged to encompass childrens' voices and whispered speech). Tests of the robustness of this representation are also being undertaken on vowel portions automatically taken from CV, VC, and C(C)VC syllables.

References

Cole, R., Alleva, F., Brill, S., Lasry, M., Phillips, M., Pilant, A., Specker, P., and Stern, R. (1982). FEATURE: Feature-based, speaker independent, isolated letter recognition. *Artificial Intelligence Conference*, Carnegie-Mellon University, Pittsburgh.

Gauvain, J.L., Liénard, J.S. and Mariani, J. (1983). On the use of time compression for word-based recognition. *IEEE-ICASSP*, Boston.

Komatsu, A., Ichikawa, A., Nakata, K., Asakawa, Y., Matsuzaka, H. (1982). Phoneme recognition in continuous speech. *IEEE-ICASSP*, Paris, pp. 883-86.

Liénard, J.S. (1979*a*). Sur quelques indices acoustiques des sons stables du français émis par plusieurs locuteurs. *9th International Congress of Phonetic Sciences*, Copenhagen.

Liénard, J.S. (1979*b*). Speech characterisation from a rough spectral analysis. *IEEE-ICASSP*, Washington, pp. 595-98.

Peterson, G., and Barney, H. (1952). Control Methods used in a study of the vowels. *J. Acoust. Soc. Am.*, **24**, 175-84.

Effective Implementation of Short-Term Analysis Pitch Determination Algorithms

W.J. Hess
Munich, FRG

1. Introduction

The problem of pitch determination counts among the most delicate problems in speech analysis. A multitude of pitch determination algorithms (PDAs) and devices have been developed; none of them works perfectly (Rabiner et al., 1976). A survey of the state of the art was presented in an earlier paper (Hess, 1982). In this survey the PDAs have been categorized into two gross categories: 1) time-domain PDAs, and 2) short-term analysis PDAs. The time-domain PDAs determine pitch (this term stands for fundamental period, fundamental frequency, or the elapsed time between two consecutive pulses of the voice source) directly from the signal as the elapsed time between consecutive laryngeal pulses. The short-term analysis PDAs, after subdividing the signal into a series of frames, leave the time-domain by a short-term transformation in favor of some spectral domain whose independent variable can be frequency or again time (in the latter case the independent spectral variable is called *lag* in order to avoid confusion).

The short-term analysis PDAs are further categorized according to the short-term transform they apply (Fig. 1). The main possibilities are *correlation, 'anticorrelation'* (i.e., the use of distance functions), *multiple spectral transform* (cepstrum), *harmonic analysis* (frequency-domain PDAs), and *maximum-likelihood* analysis. In the following we will deal only with three examples: 1) autocorrelation (pertaining to the the correlation PDAs), 2) maximum-likelihood, and 3) harmonic analysis.

2. Basic Computational Effort, Spectral Representation and Measurement Accuracy

In general the short-term analysis algorithms perform a short-term transformation of the form

$$X = W x. \tag{1}$$

In this equation, X is the spectral vector, x is the signal vector and W is the transformation matrix which represents the properties of the short-term transformation. For a frame of N samples (the transformation interval) the

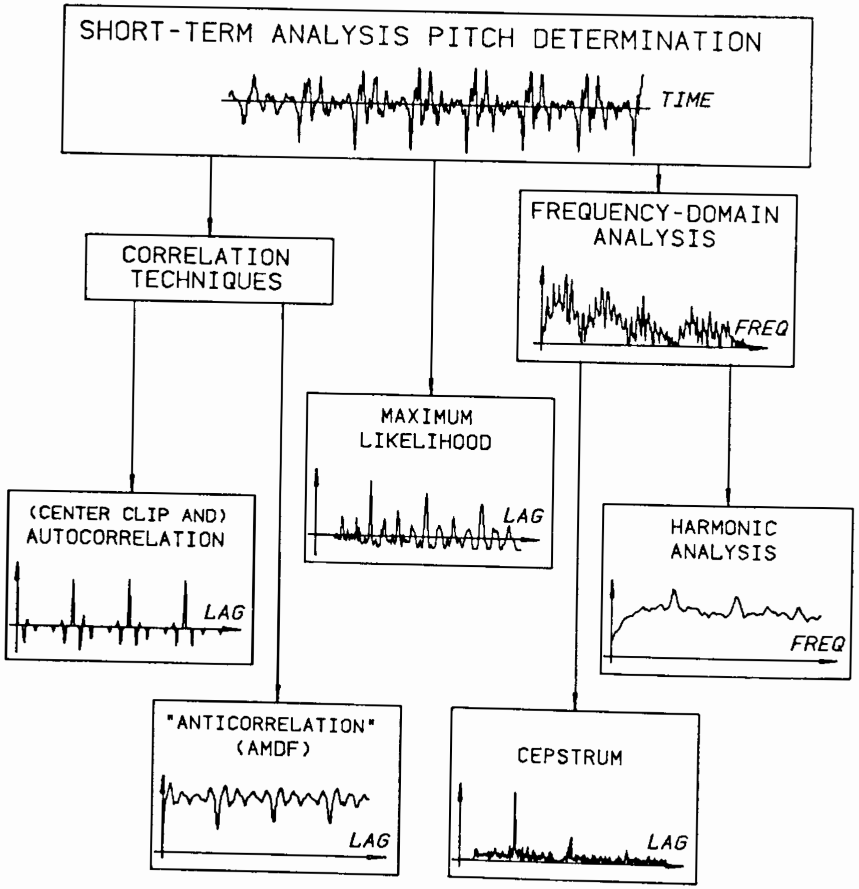


Fig. 1. Grouping of short-term analysis PDAs. Signal: part of the sustained vowel /ε/, male speaker, undistorted recording.

basic computational complexity of the short-term transformation is in the order of N^2 if the number of multiplications serves as the basic reference.

In principle the computational complexity given in (1) is valid for all three types of algorithm we are going to deal with. The discrete Fourier transform which is applied in frequency-domain PDAs and (not necessarily but possibly) in autocorrelation PDAs, clearly follows (1). Of course one always tries to use the fast Fourier transform (FFT) whose basic complexity is in the order of N IdN, where Id represents the dual logarithm. If the autocorrelation function is directly evaluated (without using the Fourier transform), it can also be brought into the form (1). The maximum-likelihood PDA involves squaring operations with the same basic complexity.

The question is now how to decrease the computational load when implementing these algorithms. With very complex hardware on-line performance can be obtained even if the algorithms have not been optimized. On the other hand, if the same results can be achieved with reduced effort, it is always worthwhile to think about such implementations. This shall be done in the following, first from a rather global point of view, later in more detail for the three algorithms cited.

The following actions to reduce the computing effort appear possible.

1. Replace multiplications and divisions by additions or table lookups;
2. Replace multiplications by logical operations due to sophisticated preprocessing;
3. Decrease the sampling rate in order to reduce the length N of the transformation interval;
4. Discard redundancies and irrelevance before the spectral transformation, again in order to reduce N;
5. Confine the operating range of the calculation to samples that are actually needed; and
6. Adaptively change the frame length K depending on the current fundamental frequency F_0 and/or the actual value of the trial period p within the short-term transformation.

Actions 5 and 6 are not possible when a FFT is applied; there all spectral samples are computed simultaneously from a constant-length transformation interval.

With respect to the representation of the relevant information on pitch in the spectral domain, the spectrum (in the following the output signal of the short-term transformation will always be labeled 'spectrum' regardless of whether it represents a Fourier spectrum, an autocorrelation function, or a maximum-likelihood estimate) is heavily oversampled. If we limit the measuring range to 50-1000 Hz, then for the PDAs which operate in the lag domain (autocorrelation, maximum-likelihood) a sampling rate of 2 kHz in the spectrum would be sufficient in order to correctly represent the information on pitch, i.e. to satisfy the sampling theorem. For a frequency-domain PDA a spectral resolution of less than 25 Hz would be necessary in order to correctly represent all the harmonics of a signal at the lower end of the measuring range as separate peaks. These crude sampling rates, if applied, would be sufficient to *represent* the relevant information on pitch in the spectral domain, but they are not at all sufficient to *measure* pitch accurately enough. The most critical judge with respect to measurement accuracy is the human ear; data by Flanagan and Saslow (1958) as well as the prevailing theories on pitch perception (e.g., Terhardt, 1979) suggest that errors of less than 0.5% are still perceived. To satisfy this requirement, we would thus need a spectral resolution of less than 1 Hz for a frequency-domain PDA and a sampling rate of more than 50 kHz for a lag-domain PDA. Since all these PDAs involve nonlinear processing (usually squaring and averaging), it is not sure whether interpolation in the spectral domain *after* the nonlinear step will yield correct results. Hence, in order to obtain a reasonable spectral resolution, frequency-domain algorithms usually perform voluminous FFTs on long transformation intervals (more than 200 ms) which consist of a short frame (30-40 ms) appended with zeros, and a number of autocorrelation PDAs compute the autocorrelation function with the full sampling rate of, say, 10 kHz although they compute it from a speech signal that has been low-pass filtered with a cutoff frequency of 1 kHz. So the basic problem is to find solutions that allow for the spectral resolution necessary for an accurate measurement, but cut down the computational effort as far as possible.

3. The Autocorrelation PDA

The breakthrough in the autocorrelation PDA came when Sondhi (1968) discovered that adaptive center clipping greatly improved the performance of this PDA which hitherto had suffered from a strong sensitivity to dominant formants. Dubnowski et al. (1976) then found that an adaptive three-level quantization did not significantly degrade the performance compared to the signal that was only center clipped. With this three-level quantization, however, it became possible to evaluate the ACF without any multiplications since the input signal of the so quantized signal can only take on values of +1, 0, and -1. It became even possible to replace the adder in the ACF evaluation logic by a simple up-down counter so that for this PDA the problem of computational complexity can be regarded as solved. Actions 4 and 6 from the above-mentioned list, which appear possible and promising, are no longer necessary under this aspect.

4. The Maximum Likelihood PDA

The maximum-likelihood PDA (Noll, 1970; Wise et al., 1976) emerged from the task to optimally separate a periodic component x(n) from Gaussian noise gn(n) in the signal a(n) = x(n) + gn(n) with the finite duration K. The mathematical formulation leads to a comb filter with the trial period p, and the best estimate of pitch is given when p optimally matches the harmonic structure; in this case the energy of the output signal of the comb filter is maximized. Computing the energy of this signal however, involves squaring, and the number of the pertinent multiplications is in the order of P^2_{max}, when P_{max} is the longest period possible within the measuring range. To reduce the computational effort, one can exploit the fact that the only multiplications needed are squaring operations, which can easily be implemented by a table-lookup procedure. Since the maximum-likelihood PDA is rather noise resistant, the input signal can be crudely quantized, and the table can be kept rather small. Another possibility of reducing the computational effort is obtained when one succeeds in replacing the squaring operations by other, less costly arithmic operations, such as the peak-to-peak amplitude of the output signal of the comb filter. This is indeed possible; taking the amplitude instead of the energy hardly affects the performance of the PDA (Ambikairajah et al., 1980).

5. Harmonic Analysis, Frequency-Domain PDAs

All frequency-domain PDAs (e.g. Schroeder, 1968; Martin, 1981) need a Fourier transform to enter the frequency domain. This is preferably done using the FFT although, for special applications, it might be profitable to use the conventional DFT and to compute only a few spectral samples (Duifhuis et al., 1982). In the following, the considerations will be confined to the case where the FFT is used. Let an arbitrary frequency-domain PDA need a spectral resolution of 5Hz. This usually meets the requirements with respect to accuracy since it is mostly possible to obtain the estimate of F_0 from a higher harmonic and thus to reduce the inaccuracy of the measurement due to quantization by the harmonic number of that harmonic. The easiest way to implement such a PDA is to take a segment of the input signal (30-40 ms), apply a suitable time-domain window, extend the segment by zeros, and apply the corresponding FFT to obtain the spectral resolution (a 2048-point FFT would be necessary when the sampling frequency of the input signal is 10 kHz). This procedure however, is in no way optimal with respect to the computational effort. Table I shows the different possibilities of optimization. The basic computational complexity of the FFT is in the order of N ld N (line 'no optimization' in Table I). Using sophisticated programming that avoids multiplication for such samples where the real and imaginary parts of the complex exponential values used in the transform are zero or have a magnitude of 1, the number of multiplications can be reduced by almost 30%. A reduction of 50% is achieved when one takes account of the fact that the input signal is real (and not complex). With programming alone one can thus save 64% of the multiplications.

Further reduction of the computing effort is only possible with additional digital filtering. First, the relevant information on pitch in the speech signal is contained in the frequency components below 2.5 kHz. In contrast to lag-domain PDAs the accuracy of the frequency-domain measurement is not influenced if the sampling frequency of the time-domain signal is reduced by a factor of 2. The computational complexity, however, is reduced by more than 50%.

A last possibility of optimization is given by interpolation in the frequency domain. There are two possibilities of increasing the spectral resolution: 1) appending many zeros to the input signal and applying a long FFT, or 2) appending few zeros to the input signal, applying the shortest FFT possible, and interpolate in the frequency domain using a digital filter with zero phase response (i.e. a linear-phase nonrecursive interpolation filter in a noncausal realization) until the required spectral resolution is achieved. These two possibilities are equivalent as long as the interpolation is performed on the complex spectrum, but they are rather different with respect to their computational complexity. Applying all these optimizations together, in our example, brings down the computational effort by as much as one order in magnitude.

Table I. Comparative evaluation of the effort necessary to compute an FFT spectrum for frequency-domain pitch determination under various aspects of possible algorithmic optimization. Assumed sampling rate: 10 kHz, frame rate: 100 Hz, required spectral resolution: 5 Hz

Optimizing operation	FFT Length	Number of multiplications			Saving (%)
		FFT	Other Operations	Total	
No optimization	2048	45056	–	45056	0
Optimized FFT Programming	2048	32776	–	32776	29
Exploit the fact that the input signal is real: perform spectral rotation, shift imaginary part of the signal against time, and decompose spectrum	1024	20480	2048	22528	50
All programming optimizations	1024	14344	2048	16392	64
Downsampling to 5 kHz	1024	20480	200	20680	56
Downsampling and programming optimizations	512	5942	1224	7166	84
Limit transformation interval to 51.2 ms and upsample spectrum by factor 4 in the frequency domain	512	9016	6144	15160	66
Time-domain downsampling, limitation of transformation interval, and frequency-domain upsampling	256	4096	3272	7368	83
All optimizations applied together	128	1032	3528	4660	89

6. Conclusions

A number of proposals to efficiently implement the short-term transformation in short-term analysis PDAs have been reviewed. The problem of computational effort arises from the fact that, for reasons of measurement accuracy, the spectral function (autocorrelation function, Fourier spectrum etc.) must be heavily oversampled. The proposals range from efficient preprocessing (combined center and peak clipping in an autocorrelation PDA), which avoids multiplications, to the use of signal amplitude instead of energy, and from the use of table-lookup procedures to the optimal combination of the FFT and interpolation by digital filters. If the PDA is carefully implemented, the gain in computing speed can be considerable.

References

Ambikairajah, E., Carey, M.J., Tattersall G. (1980). A method of estimating the pitch period of voiced speech. *Elektron. Lett.* **16**, 464-466.

Dubnowski, J.J., Schafer, R.W., Rabiner, L.R. (976). Real-time digital hardware pitch detector. *IEEE Trans. ASSP*-**24**, 2-8.

Duifhuis, H., Willems, L.F., Sluyter, R.L. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *J. Acoust. Soc. Am.* **71**, 1568-1580.

Flanagan, J.L., Saslow, M.G. (1958). Pitch discrimination for synthetic vowels. *J. Acoust. Soc. Am.* **30**, 435-442.

G-AE Subcommittee on Measurement Concepts (1967). What is the fast Fourier transform? *IEEE Trans.* AU-**15**, 45-55.

Hess, W.J. (1983). *Pitch determination of speech signals - algorithms and devices*. Springer, Berlin.

Martin, Ph. (1981). Détection de F_0 par intercorrelation avec une fonction peigne. In: *Actes, 12èmes Journées d'Etude sur la Parole*, Montréal, mai 1981. GALF, F-22301 Lannion.

Noll, A.M. (1970). Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In: *Symposium on Computer Processing in Communication*; ed. by the Microwave Institute; Vol. **19**, 779-797. University of Brooklyn Press, New York.

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, A. (1976). A comparative study of several pitch detection algorithms. *IEEE Trans.* ASSP-**24**, 399-413.

Schroeder, M.R. (1968). Period histogram and product spectrum: new methods for fundamental-frequency measurement. *J. Acoust. Soc. Am.* **43**, 829-834.

Sondhi, M.M. (1968). New methods of pitch extraction. *IEEE Transactions* AU-**16**, 262-266.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research* **1**, 155-182.

Wise, J.D., Caprio, J.R., Parks, T.W. (1976). Maximum likelihood pitch estimation. *IEEE Transactions* ASSP-**24**, 418-423.

Analysis and Validation of Higher Pole Correction Function

U.K. Laine
Stockholm, Sweden

1. Introduction

The concept of higher pole correction is connected to the all-pole modelling of speech production. Primarily it is used in the cascade terminal analog realizations where only the first few formants are included. The contribution of the higher formants to the lower frequency region is modelled by a higher pole correction (HPC) function to get the formant peaks in the spectrum to the proper levels. The modelling of HPC is usually (Fant 1959) determined by two independent parameters: the area of mouth opening and the physical length of vocal tract. The mouth opening determines the size of the radiation inductance which lowers the formant frequencies. This effect of the radiation inductance can be considered as a change in the physical length, 1, to give an 'effective' length $1_e = 1 + \Delta 1$, where $\Delta 1 = 0.8*SQRT(A_0/\pi)$ and A_0 is the area of the mouth opening. The term $\Delta 1$ is also called end correction. Henceforth the higher pole correction concept will be discussed only in terms of the effective length.

In the derivation of the higher pole correction function for any vowel sound it is assumed that the higher poles are located as they are in the lossless neutral vocal tract and that the higher poles have infinite Q-values (bandwidths of higher poles equal zero). Also the possible effects of cross-modes to the lower frequencies have not been discussed. The above derivation by Fant has been generally used. However, there have been no validating studies to verify the assumptions mentioned and the consequent effect on the accuracy of the HPC function.

In this study the true volume velocity transfer function is measured foı acoustic tubes of different shapes. Then the corresponding transfer functions are also calculated by a computer simulation of the transmission line (TL) model. By a comparison of the physical measurements with computer simulation results, the validity of the lip radiation impedance in particular, and the TL model in general, are studied. The computational TL model is then used as a reference system to study the HPC function in s- and z-domains.

2. Acoustical Measurements

To simulate vowel production, a straight uniform hard-walled tube of plexiglass was mounted into a spherical baffle with a radius of 9 cm. The tube was 17 cm in length and 2.95 cm in inner diameter (area 6.84 cm^2). The 'glottal' end of the tube was totally closed with a hard wall and the open end was mounted exactly on the surface of the baffle. Electrical sparks were used as excitation at the glottal end. The measurements were carried out in an anechoic room (Laboratory of Electroacoustics, Helsinki University of Technology). A real time spectrum analyzer was directly connected to the microphone amplifier to document the volume velocity transfer properties of the artficial vocal tract.

One measurement for the neutral tract is shown in Fig. 1. The corresponding transfer function calculated from the TL model (see sec. 3 for details) is superposed on the same figure. One can note that the match is fairly good up to about 7 kHz. One typical measurement for the nonuniform tube with the result from TL simulation is given in Fig. 2. In this case a three tube model was obtained by placing a block of 8.5 cm length into the tube, starting 2 cm from the 'lips', yielding a constriction area of 1.77 cm^2. This profile produces a response similar to the vowel /e/. Also in this case the TL simulation gives a good match to the physical measurements. These examples illustrate the validity of the TL model and the radiation impedance model used as a computational reference system in this study.

Since the TL model which describes only one-dimensional wave behavior in the tract, still gives a good match to the acoustical measurements, we can infer that the cross-modes in the measured physical model have only little effect on the lower frequencies. Thus it seems reasonable to leave out the cross-modes in the HPC derivation.

It can be noted from Fig. 2 that the formant frequencies have moved away

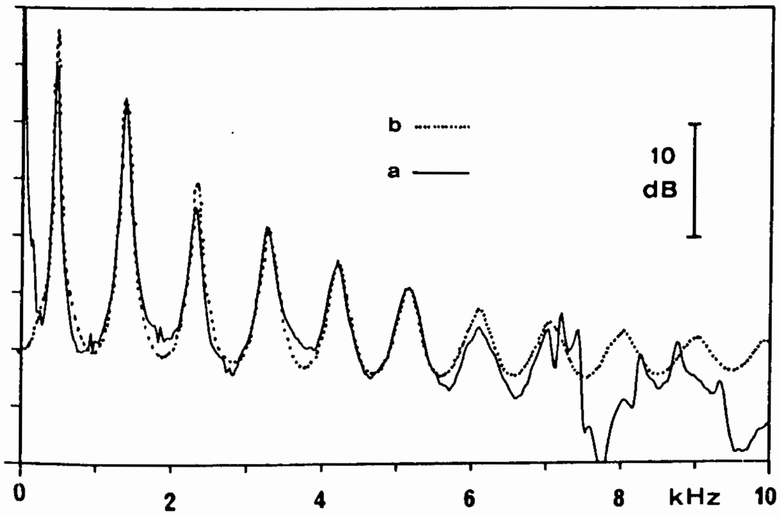


Fig. 1. Frequency response of the a) physical model b) TL model.

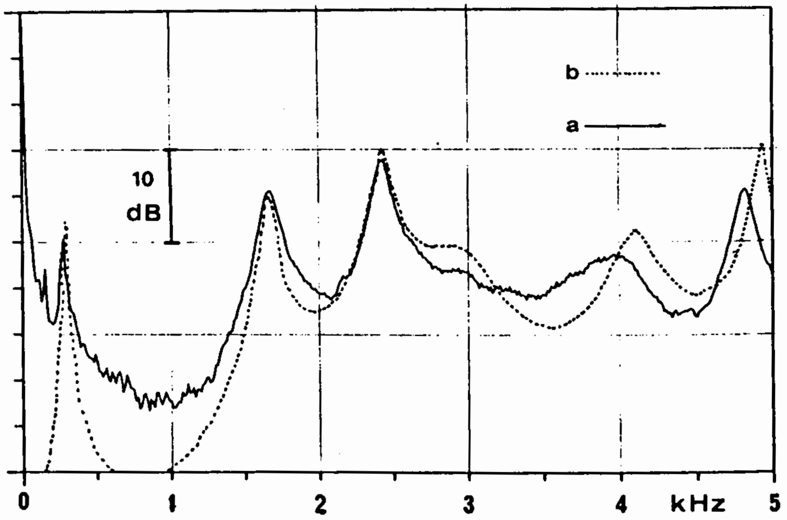


Fig. 2. Frequency response of the three tube model a) physical moαeι b) TL model.

from their neutral positions. Also, even in the case of a neutral tube (Fig. 1) the bandwidths are increasing towards the higher formants due to the larger radiation losses. How these factors affect the HPC function will be seen in the following analysis.

3. HPC Function in the S-Domain

As known, by assuming one-dimensional wave behavior in the vocal tract, its sound transmission properties can be described by applying the theories of transmission lines, The acoustic transfer functions are then given by the hyperbolic functions, which means that the model cannot directly be used for the time domain simulations. Therefore some approximations are needed to develop a corresponding lumped parameter model with rational transfer function (Skilling 1951, Flanagan 1972). However, in this study only the analysis in the frequency domain is needed and so the original TL model can be used in its full mathematical accuracy without any approximations. All frequency dependent losses (viscous friction, heat conduction, radiation etc.) can also be calculated from their original equations with the correct frequency dependency. The lip radiation impedance was approximated by a piston in a sphere model and calculated from the exact serial expansions given by Morse and Ingard (1968).

Vocal tract profile data for Russian vowels estimated from X-ray pictures by Fant (1960) were used in the TL model to calculate the corresponding volume velocity transfer functions. Values for the first five formant frequencies and bandwidths were estimated for the vowels /a/, /o/, /u/, /i/, and /e/.

Transfer functions for analog cascade resonant filters with the estimated formant parameters were calculated and substracted (in log. magnitude form) from those given by the TL model. Smoothly rising curves corresponding to the true HPC were achieved by these means. The HPC functions for the five formant model were then calculated by using Fant's formula and the known effective lengths. The error curves are shown in Fig. 3. The error is typically very small below 4.5 kHz, and then grows rapidly. This is mainly due to the sixth formant of the tract which usually has a lower frequency than assumed in the derivation of the HPC function.

The s-domain simulations confirm clearly the validity of the HPC function given by Fant. The function is seen to be an accurate and useful tool to be used in the all-pole modelling of the vowel tract transmission. In the analyzed five formant case the 3 dB point in the error curve is typically located between 4.2-4.6 kHz. Although the two assumptions made during the derivation of the HPC function are not confirmed by acoustical reality, the HPC function is seen to be quite insensitive to the absolute positions of the higher poles. Only their average density has to be known. The results are also in harmony with the fact that the bandwidth of the formant has primarily a local effect only. These general notes are also valid in the following analysis made in z-domain.

4. HPC Function in the Z-Domain

Gold and Rabiner (1968) have published a comparative study on analog and digital all-pole models. They compared the digital and analog transfer functions for ten English vowels. The models had three controllable for-

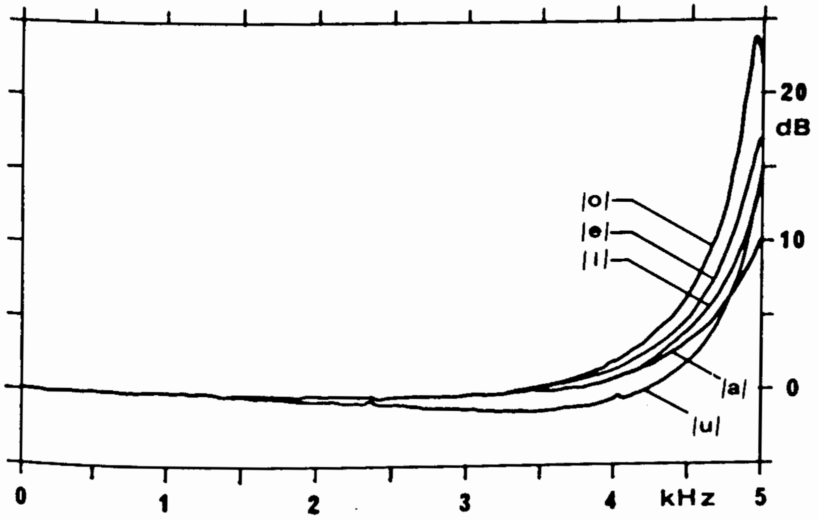


Fig. 3. Difference curves in the analog domain between the calculated, true HPC function and Fant's formula.

mants together with two or seven fixed higher poles corresponding to the 5- and 10-pole systems analyzed. As a result it was reported that the 10-pole analog system with HPC function and the 10-pole digital system without any HPC are 'extremely close' to each other, and it was also said that: 'This strongly indicates that higher poles of the vowel tract transfer function are automatically and more or less correctly taken into account by the repetitive nature of the digital formant frequency response. We also note that this intrinsic correction is actually more accurate than the quite good analog higher pole correction used in our computations. These results are generally valid for all the vowels'.

To have a closer look into this question a cascade all-pole model was compared with the TL model as in the previous s-domain study. Also five formants of the vowels were used and the sampling frequency was fixed at 10 kHz. This sampling frequency gives an optimum aliasing effect for a neutral tube of effective length 17.5 cm.

The results of the simulation are collected in Fig. 4. The difference between the TL model and the digital model is largest when the effective length of the tract differs most from the length fixed by the sampling frequency (17.5 cm). In this simulation one of the largest errors appeared for vowel /u/. This case is illustrated in more detail in Fig. 5, where the TL transfer function is shown together with the digital one. It can be noted that due to the large effective length of the actual vocal tract the digital model has approximately missed one formant (the 6th one). If we had the possibility to control the sampling frequency we could achieve a fairly good match between the models.

One can clearly see from the above results that the digital 5-pole model without any HPC does not give as accurate results as the corresponding analog model with HPC. Our result contradicts the conclusions made by Rabiner and Gold. However, their study was limited in the following sense: they did not compare the terminal analogs directly with the more realistic distributed parameter transmission line model because as they mentioned, it is 'quite difficult'. They have chosen some artificial values for the higher poles and compared the digital models directly to the analog one. The higher poles chosen led to vocal tract models with a fixed effective length of 17.5 cm, whereas in reality the effective length of the vocal tract varies.

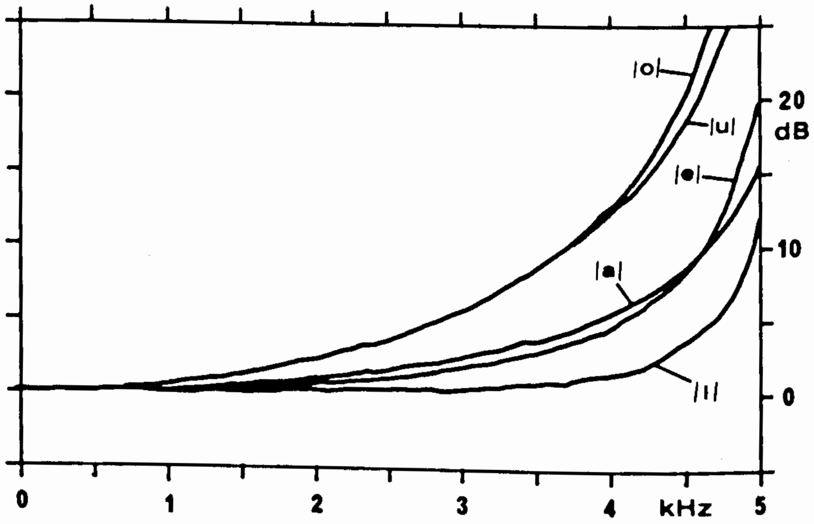


Fig. 4. Difference curves in the digital domain between the calculated TL model transfer function and the all-pole digital model.

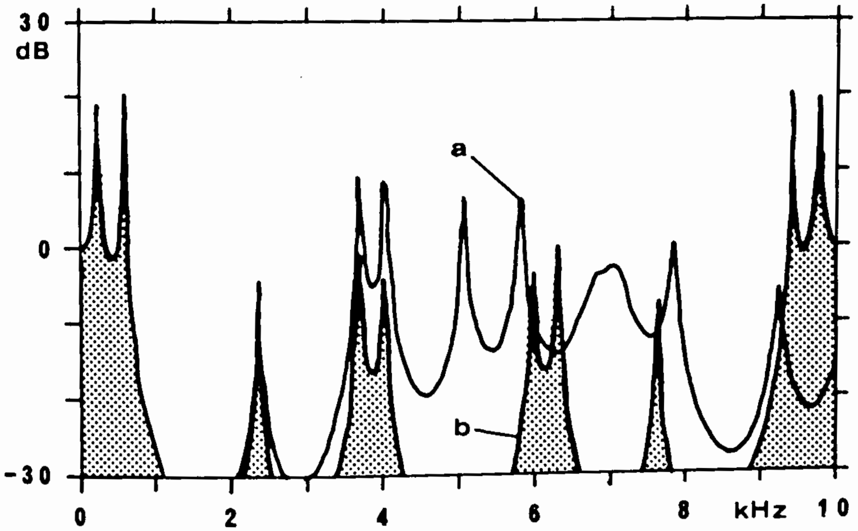


Fig. 5. Log. spectrum of /u/: a) TL model b) digital 5-pole model (aliased spectrum).

Acknowledgements

The author wishes to thank Dr. M. Karjalainen at Helsinki University of Technology for organizing the acoustic measurements used in this study and Prof. G. Fant for the comments upon the manuscript. The author is indebted especially to Dr. T.V. Ananthapadmanabha for many critical, constructive comments and discussions upon this work. This study has been supported by the Academy of Finland, Nordiska Forskarstipendier and the Cultural Foundation of Finland.

References

Fant, G. (1959). Acoustic Analysis and Synthesis of Speech with Applications to Swedish, *Ericsson Technics*, No. 1.

Fant. G. (1960). *Acoustic Theory of Speech Production*, Mouton, The Hague.

Flanagan, J. (1972). *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, NY.

Gold, B., Rabiner, L.R. (1968). Analysis of Digital and Analog Formant Synthesizers, *IEEE Tr. on Audio and Electoac.*, **16** (1), 81-94.

Morse, P.M., Ingard, K.U. (1968). *Theoretical Acoustics*, Mc Graw-Hill, NY.

Skilling, H. (1951). *Electric Transmission Lines*, Mc Graw-Hill, NY.

Estimating Vocal Tract Area Functions: A Progress Report

F. Lonchamp, J.P. Zerling, J.P. Lefèvre
Nancy, Strasbourg, Le Mesnil St-Denis, France

At the previous International Congress of Phonetic Sciences, Fant remarked that 'very little original data on area functions had accumulated. The 1960 Russian vowels have almost been overexploited'. A quantitative estimate of the variation in the cross-sectional area of the vocal tract is necessary to relate in a cogent way articulatory and acoustic data. Articulatory synthesis which promises high-quality elocution, requires accurate knowledge of area functions. The dearth of cross-dimensional, as opposed to sagittal, measurements is related to difficulties in using photographic data and to the potential hazards of tomographic X-ray exposure. Also, with any method, it is difficult to assess the accuracy of the recovered areas, as we lack normative values. We believe that several techniques should be used concurrently to provide a measure of consistency. We report preliminary results for one subject using two different procedures.

1. Area function determination from the tract impulse response

This first technique was originally devised by Sondhi and Gopinath (1971). The area function is determined from measurements of the vocal tract response to an impulsive acoustic pressure wave. The experimental set-up is shown in fig. 1. Following work mainly at the Electrical Engineering Dpt. of Laval University at Quebec, where the experimental data for this part of the study were gathered, area functions have been published for all French vowels, except the heavily labialized [u] and [y] (Tousignant et al. 1979; Lefèvre et al. 1981, 1983). The procedure is summarized in fig. 2 with numbers illustrating the following steps:

- 1. The acoustic impulse input signal e(t) and its response s(t) after reflection in the vocal tract are both sampled.
- 2. From the delay and magnitude of s(t), the major reflection coefficient (and hence the area value) is located and estimated.
- 3. Due to the bandlimited nature of the signals, smoothing of the area discontinuity with a transition is necessary.
- 4. Using the transmission line model, the tract response is obtained through convolution of the input signal with the impulse response of the vocal tract computed from the reflection coefficients.
- 5. Finally, the residual signal obtained by subtracting the calculated

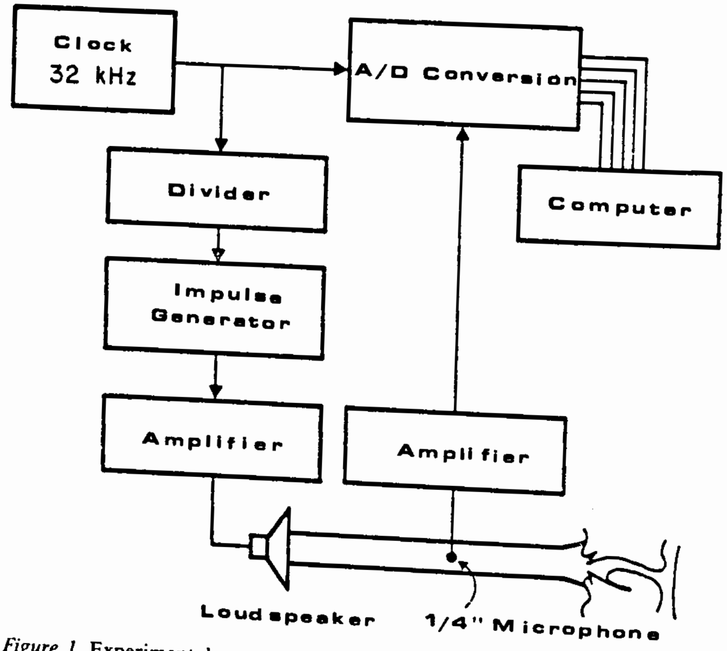


Figure 1. Experimental set-up for impulse response measurements.

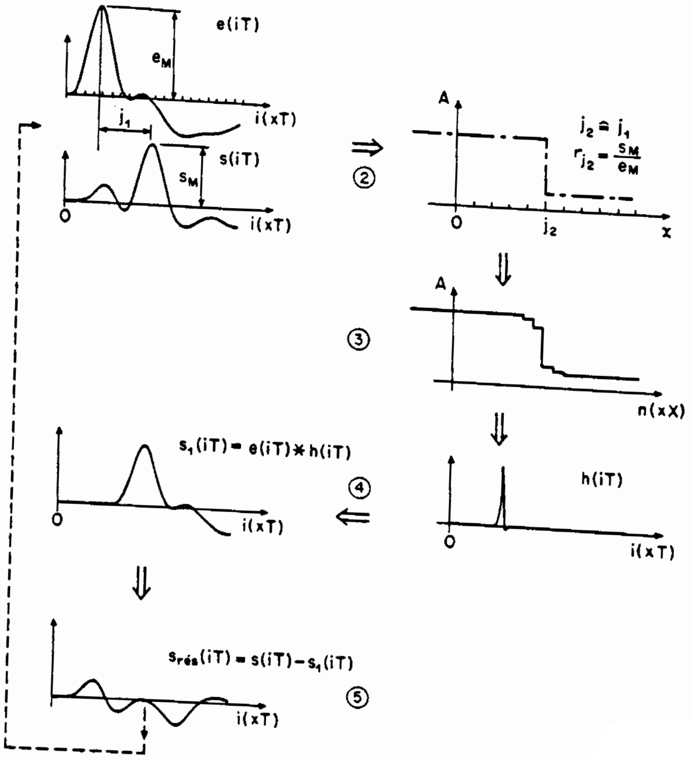


Figure 2. Area reconstruction algorithm.

response from the measured one can be further analyzed by repeating the procedure from step (2).

The area function is thus determined in 10 to 20 passes.

The known advantages of this method (Tousignant et al. 1979; Sondhi 1979) include absolute area estimates and recovery of true length. Drawbacks do exist however: exertion of suitable muscular tension; auditory feedback not available for the subject to check on his articulatory posture. Lip and jaw positions may not be natural while insuring airtight closure at the mouthpiece. To try to control for these last two effects, and to provide for a new test of accuracy, the subject was instructed to phonate as soon as the response measurements had been taken. All results reported here show a good correspondence between formant frequencies computed from the recovered area functions and measured from the tape recording of the session. In other cases, discrepancies were noted. But it is impossible to tell whether the recovered functions are incorrect or whether the subject changed his articulatory posture between measurement and phonation times. Although plausible area functions have been reported for 'constricted' vowels (i.e. [i] or [o]) for another subject with a seemingly larger and longer vocal tract (Lefèvre et al. 1983), only 'open' vowels such as [ɛ] or [œ] yielded reasonable area functions (cf. Sondhi and Resnick 1983). In all cases, total length and the location of the cavities seemed correct; only their volumes were not. Results for an [ɛ] token is shown in fig. 3. The dotted line refers to the area function for another [ɛ] by the same subject, derived from the second technique to be described below. Although discrepancies occur at the glottis and in the mouth region, the general shape is similar. Length values are within 0.4 cm. While the tokens are different, the formant frequencies are close (see caption). Measured and computed frequencies match satisfactorily. The area function is

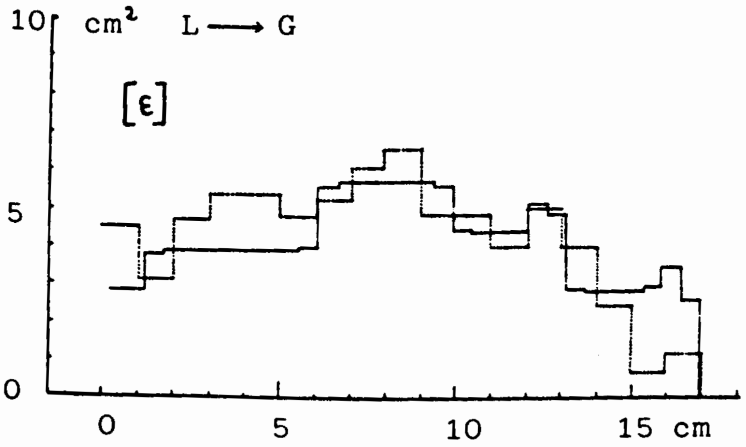


Figure 3. – Area functions for 2 [ɛ] tokens ___ acoustic impulse response : measured formants: 550-1670-2450-3430; computed formants: 545-1630-2500-3450; · · ·: sagittal to area conversion; measured formants: 580-1650-2512-3520; computed formants: 585-1650-2620-3360.

rather different from the one reported in Lefèvre et al. Fig. 4 illustrates 4 [œ] tokens with close spectral contents. In this case, the area function derived from the X-ray data for another token, shown in the insert, looks very different. The areas are [ɔ]-like in shape. Phonetic description of French has interestingly pointed out the acoustic similarity of [œ] and [ɔ] sounds. We cannot rule out the use of two different tract configurations for the same vowel sound. Clearly, simultaneous X-ray and impulse measurements would have to be made.

2. Conversion of X-ray sagittal measures to area functions

In this second approach, we sought to optimize the numerical coefficients of a set of functions relating vocal tract sagittal dimensions to area values, by simultaneously minimizing the discrepancy between measured and calculated formant frequencies on a set of vowels. A similar approach has been reported by Maeda (1971). Length and sagittal widths at 1 cm intervals were taken from an X-ray film for the vowels [i, ɛ, a, ɔ u] in an [əb-b] context (Zerling 1979). The first 4 formant frequencies were computed through autocorrelation LPC from the synchronous sound recording. The teeth, uvula hump and epiglottis were ignored. Being more distinct, the midline groove for [ɛ] and the side tongue outlines for [i] were traced in the upper pharynx zone. A crucial choice is the form and number of functions relating sagittal to area values. As reviewed in Wood (1982), most authors favour a power function for the mouth. The pharynx and larynx regions are modelled usually as a number of ellipses, the cross-dimensions of which (c_i) are set to a constant value. After a long series of pilot experiments, we selected 3 power functions (A (area in cm^2) = $a_i .x^{bi}$; x: sagittal width in cm) for the mouth, uvula and upper pharynx, and the lower pharynx (epiglottic region). Only the laryngeal tube was modelled as an ellipse. As the width of the pharynx never exceeded 2.7 cm, Wood's 'cosine' function was not used. Formant frequencies were computed from the estimated area functions using the transmission line approach (Liljencrants and Fant, 1975). Radiation losses were modelled as an extra section at the lips. Using recent simulation data by Maeda (1982), we were able to express Fw as a function of uncorrected F_1 values (F_1u) in the well-known correction formula ($F_1 = \sqrt{F\omega^2 + F_1u^2}$) for wall vibration effects: $F\omega = 0.04zF_1u + 187$. As no simple formula is available, no correction was made for the effect of open glottis. A random search procedure was used to simultaneously optimize the 7 coefficients for the 5 vowels. As [i, ɛ, a] gave consistently closer matches, they were more heavily weighed in the decision metric. As a final step, the first (lip and/or teeth) section was slightly adjusted to further improve the fit. The coefficients, error (%F_i) and absolute mismatch values in Hz (Δ F_i) are given below:
- mouth: a=1.95; b=1.53 - uvula and upper pharynx: a=3.00; b=1.40 - lower pharynx: a=2.40; b=1.23 – larynx: c=1.6.

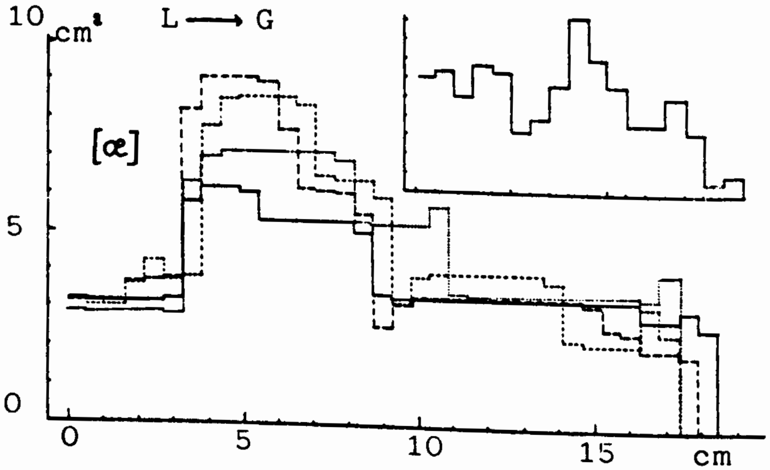


Figure 4. Area function for 4 [œ] tokens. Insert shows area from X-ray data. Measured and computed formant frequencies: —— 560-1275-2425-3310, 545-1296-2426-3330; — — —: 535-1230-2620-3425, 540-1220-2520-3395; ----: 525-1355-2420-3670, 535-1340-2500-3650; . . .: 530-1345-2590-3450, 530-1360-2530-3300.

	%F_1	ΔF_1	%F_2	ΔF_2	%F_3	ΔF_3	%F_4	ΔF_4
[i]	–3.0	10	–0.9	19	–1.5	44	–1.3	45
[ɛ]	0.5	3	0.0	1	4.5	96	–4.6	163
[a]	–0.5	3	2.0	26	–0.9	22	1.3	45
[ɔ]	4.9	29	4.7	52	–8.1	211	1.4	47
[u]	–4.9	18	8.1	70	–7.0	177	–0.7	25

Figs. 5 and 6 show the area functions for the 5 vowels. The mouth coefficients are almost identical to the values reported by several authors (see Wood, 1982). They also check with values derived from a plaster cast made on the subject. The upper pharynx values are slightly above those for 'level 1' reported by Gauffin and Sundberg (1978) especially for large sagittal values (i.e. 8.5 cm^2 vs 6.8 cm^2 for a 2.1 cm width). The lower pharynx values ('level 2') are also larger (i.e. 5.3 cm^2 vs 3.2 cm^2 for a 1.9 cm width). The larynx tube appears rather narrow. It may be worth noting that the laryngeal tube mainly controls the frequency of F_4, which was not corrected for the effect of the open glottis phase during a period.

It is not clear whether 'the relationship of lateral to cross-dimensions is affected by the phonetic nature of each vowel' and whether a tongue gesture factor, computed from the sagittal shape and related to the high/low and front/back features, is necessary to convert sagittal to area values, as suggested by Maeda (1971). While it is true that the formant fit is sometimes poor, the mean error is only 2.5 %. We have recently tested the 5 following vowels [i, e, oy, ø, œ], from a new X-ray film, using the coefficients given above. Mean percentage errors for the first to the fourth formants are 7.3, 9,7, 6.1, and 7.5 % respectively. These larger errors indicate that the set of coefficients

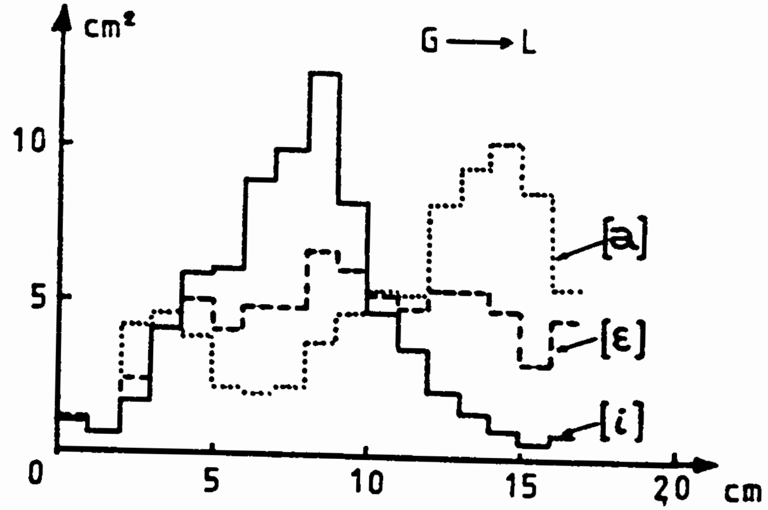


Figure 5. Area functions from the conversion of sagittal widths of x-ray data: vowels [i, ε, a].

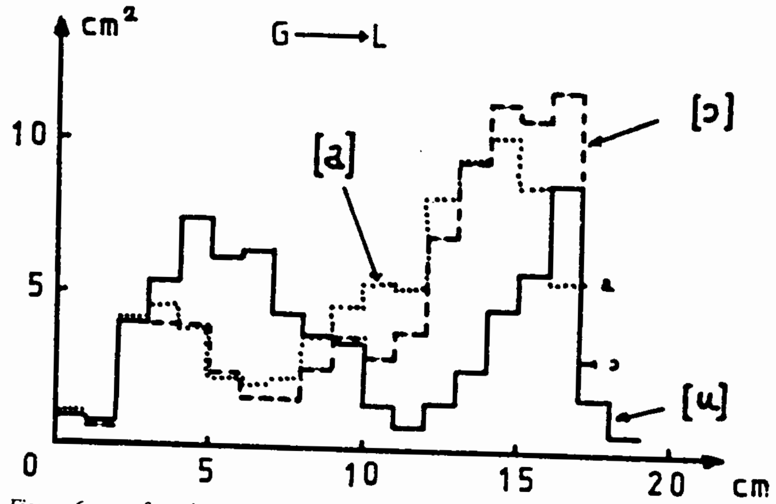


Figure 6. area functions from the conversion of sagittal widths of x-ray data: vowels [a, ɔ, u].

is not optimal. But the possibility exists that forcing area values to be a monotonically increasing function of the sagittal widths generates 'pseudo-area functions' with too strong a constriction in the larynx area. A multistage optimization procedure as developed by Charpentier (1982), combining table look-up for initial estimates and an optimization algorithm, might seem a better approach. But in this case, only one vowel can be dealt with at a time, as there are no constraints on the respective shapes of a set of vowels. Also, in order to have more acoustic variables than articulatory ones, one has to resort to measures of formant bandwidths or amplitudes, the relevance of which is doubtful. In our approach it can be checked that no sharp discontinuity exists for contiguous sections computed with different functions (i.e. at the junction between the mouth and upper pharynx functions) and that the area variation is physically compatible with the variation in sagittal width at most or all points along the vocal tract.

Further work, including simultaneous X-ray and acoustic impulse measurements as well as sagittal to area values conversion for a larger set of vowels, is definitely needed to better estimate area functions.

References

Charpentier, F. (1982). Application of an optimization technique to the inversion of an articulatory speech production model. *IEEE-ICASSP* 1982-1987.

Gauffin, J. and Sundberg, J. (1978). Pharyngeal constrictions. *Phonetica* **35**, 157-168.

Lefèvre, J.P., Tousignant, B. and Lecours, M. (1981). Utilisation de méthodes acoustiques pour l'évaluation des fonctions d'aires du conduit vocal. *XII^e Journées d'Etude du GALF*, Montréal, 26-41.

Lefèvre, J.P., Tousignant, B. and Lecours, M. (1983). Etude des configurations vocaliques des voyelles françaises à partir de mesures acoustiques. *Acustica* **52(4)**, 227-231.

Liljencrants, J. and Fant, G. (1975). Computer program for V.T. resonance frequency calculations. *STL-QPSR* **4** 1975, 15-20.

Maeda, S. (1971). Conversion of midsagittal dimensions to vocal-tract area function. *JASA* **51**, 88A.

Maeda, S. (1982). A digital simulation method of the vocal tract system. *Speech Communication* **1(3/4)**, 199-229.

Sondhi, M.M. (1979). Estimation of vocal tract areas: the need for acoustical measurements. *IEEE-ASSP* **27/3**, 268-276.

Sondhi, M.M. and Gopinath, B. (1971). Determination of vocal tract shape from impulse response at the lips. *JASA* **49**, 1867-1873.

Sondhi, M.M. and Resnick, J.R. (1983). The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis. *JASA* **73(3)**, 985-1002.

Tousignant, B., Lefèvre, J.P. and Lecours, M. (1979). Speech synthesis from vocal tract area function acoustical measurements. *IEEE-ICASSP*, 921-924.

Wood, S. (1982). X-ray and model studies of vowel articulation. *Lund University Working Papers n°* **23**.

Zerling, J.P. (1979). *Articulation et coarticulation dans les groupes occlusive-voyelle en français.* Thèse de 3ème cycle, Nancy.

Real Time Fundamental Frequency Analysis Using the Spectral Comb Method

P.J. Martin
Toronto, Canada

1. Introduction

Although reasonably good real-time fundamental frequency visualizers are now commercially available, it is difficult to conduct a phonetic analysis of prosodic parameters when speech material is recorded in a noisy environment. Indeed, most Fo analyzers are designed to work with a relatively high signal to noise ratio, which rules out their use in certain cases such as the analysis of spontaneous speech recorded in a noisy street. This is due to the fact that these instruments rely on methods of analysis which take into account fundamental frequency as the only parameter.

More reliable methods based on the analysis of the short-time spectrum of the speech signal are now available due to the development of inexpensive highspeed hardware.

2. The Spectral Comb Method

Most methods of pitch extraction from the short- time spectrum are based on detecting the periodicity of F_0 harmonics in the spectrum. The cepstrum (Noll, 1969), for instance, computes the Fourier transform of the logarithm of the power spectrum. Other methods are based on a more direct search for periodicity in the spectrum. Schroeder (1968) uses a histogram of subharmonics derived from spectral peaks, and F_0 is considered as being the smallest common multiple of the periods of its harmonic components. Harris and Weiss (1963) examine a high resolution Fourier spectrum and retain as fundamental frequency the most numerous equal spacings of adjacent peaks. Sreenivas and Rao (1972) use only selected high quality peaks, and compute their approximate highest common factor to obtain the pitch value. Sluyter, Kotmans and Leuwaarden (1980), in order to account for possible phase distortion of the peak harmonics in the spectrum, use a minimum distance criterion to recognize harmonic patterns, and derive pitch value from these patterns.

The spectral comb method (Martin, 1981) is also based on short-time spectral analysis. Although it is similar to other spectral compression techniques, reliability is due essentially to the fact that both the frequency and the amplitude of each harmonic component are considered in the evaluation of

F_0. More precisely, the comb method uses the cross correlation between the short-time spectrum /F (ω)/ and a spectral comb function P (ω_p,ω), with teeth of decreasing amplitude and variable intervals. The maximum of this crosscorrelation function is obtained when the comb's teeth coincide with the harmonic peaks of the spectrum, i.e. to the harmonic pattern which has the largest sum of its harmonic amplitude.

Although presenting some similarities with the method based on Goldstein's work (Goldstein, 1973), by Duifhuis et al. (1978), the spectral comb method differs from the harmonic sieve approach by using information pertaining to the amplitudes of the harmonics rather than using a minimum distance criterion in order to select the appropriate Fo.

3. Hardware Implementation

After analog-to-digital conversion, (4 kHz sampling frequency), the speech signal is split up into overlapping segments of 32 ms duration (128 samples). The speech samples of each segment are multiplied by 64 × 128 complex Fourier coefficients stored in a 16 k byte PROM, and accumulated into a 125 ns cycle time 12 × 12 multiplier-accumulator chip. (Gaussian windowing is performed on the trigonometric functions themselves, rather than on the speech signal). The module of the real and imaginary frequency values is computed from a table, and the logarithm of the module stored in the memory of a small microprocessor system (8085). These values range from 0 to 1024 Hz, with a 16 Hz resolution. An algorithm determines the spectral peaks above a threshold situated 40 dB below the maximum amplitude of the spectrum. These spectral peaks are 'rewritten' as parabolas whose peaks coincide exactly with the interpolated spectral peaks in order to obtain a modified power spectrum /F′ (ω)/ with reduced energy between the harmonic components. The modified spectrum is then crosscorrelated with a comb function with teeth of decreasing amplitude (law $n^{-1/8}$)

$$C (\omega_p,\omega) = \Sigma\, n^{-1/8} (n\omega_p - \omega)$$

the crosscorrelation function

$$I (\omega_p) = \Sigma\, n^{-1/8} \,/F' (\omega_p)/$$

is thus the sum of equally sampled values of the modified power spectrum /F′ (ω)/ which are weighted according to the amplitude of the comb's teeth.

The computation of I (ω_p) is performed by the microprocessor using a large table of coefficients containing all the possible parabola sampled values of 128 possible amplitude levels. Fast computation can thus be realized performing only sums and no multiplications. The final F_0 value is obtained by looking at the maximum of the crosscorrelation function I (ω_p).

The voiced-unvoiced distinction conducted by using an index compares the maximum value of I (ω_p) to the value obtained with a comb function shifted by $\omega_p/2$. This ratio indicates the relative energy of the detected harmonic structure compared to the weighted sum of noise components.

4. Visualization

The F_0 values are displayed on a color monitor, together woth the intensity curve (which is measured independently). The display converter generates a color video signal containing both curves and alphanumeric information pertaining to display duration and frequency scale. Two cursors allow alphanumeric readout of frequency, intensity and duration values at any point along the curve.

Pitch visualization would be of even greater usefulness in phonetic research if the user were able to segment the speech signal on the screen and identify each segment. Although the speech signal can sometimes be segmented by simply looking at the information given by the intensity and fundamental frequency curves, one needs still more information in order to perform an accurate segmentation.

Usually, this information is provided by the speech wave. Since the resolution of the display (256 dots horizontally) does not allow for displaying the speech wave itself, the envelope of the signal is displayed, instead. F_0 analysis is performed in real time, and any length of speech (up to 9 s.) may be displayed on the screen. Two cursors can be moved along the time axis in order to segment the speech signal according to the cues provided by the envelope. Numerical information pertaining to the segment defined by the two cursors includes the amplitude of F_0 (in Hz), and intensity (in dB) variations, the time duration of the segment, as well as the average and the standard deviation of both the pitch and the intensity between the cursors. It is also possible to obtain absolute values of F_0 and intensity at any point of the curves. With this equipment, an accurate analysis of speech data can be realized quite rapidly, and can even be performed interactively in the presence of an informant if desired.

5. Conclusion

A new reliable real-time pitch-visualizer has been presented, using a noise resistant F_0 tracking algorithn. The instrument allows real-time display of both intensity and fundamental frequency, together with the envelope of the speech signal.. Numerical readout of F_0 and intensity is provided, which makes the instrument easy to use for practical phonetic analysis.

References

Duifhuis, H., Willems, L.F. and Sluyter, R.J. (1978). Measuring Pitch in Speech. *Inst. Perceptie Onderzoek,* Annual Progress Report no **13**, 24-30.

Goldstein, J.L. (1973). An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones. *JASA,* Vol. **54**, 1496-1516.

Harris, C.M. and Weiss, M.R. (1963). Pitch Extraction by Computer Processing of High Resolution Fourier Analysis Data. *JASA,* Vol. **35**, 339-343.

Martin, P. (1981). Comparison of Pitch Detection by Cepstrum and Spectral Comb Analysis. *Proceedings of the Int. Conf. on Acoustic, Speech and Signal Processing,* Vol. 1, 180-183.

Noll A.M. (1969). Short-time Pitch Spectrum and 'Cepstrum' Techniques for Vocal-Pitch Detection. *JASA,* Vol. **36**, 296-302.

Schroeder, M.R. (1968). Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurement. *JASA,* Vol. **43**, 829-834.

Sluyter, R.J., Kotmans, H.J. and Leuwaarden, A.V. (1980). A Novel Method for Pitch Extraction from Speech and a Hardware Model Applicable to Vocoder Systems. *Proceedings of the Int. Conf. on Acoustic, Speech and Signal Processing* 80, Vol. **1**, 45-48.

Sreenivas, T.V. and Rao, P.V.S. (1979). Pitch Extraction from corrupted Harmonics of the Power Spectrum. *JASA,* Vol. **65**, 223-228.

Jitter in the Singing Voice

R.A. Rasch
Utrecht, the Netherlands

1. Introduction

In the musical tones produced by the singing voice, frequency variations are always present. They can be grouped into three categories: (1) *trend* (slow, gradual changes), (2) *vibrato* (periodical variations with a frequency of 5 to 6 Hz mostly), and (3) *jitter*, to be defined as small-scale period-to-period fluctuations with a random or pseudo-random character. The vibrato of the singing voice has been studied extensively, beginning with Seashore (1932), but this is not the case with jitter. Jitter has been studied in connection with pathological speech (Lieberman, 1963) and from a point of view of hearing theory (i.e. Pollack, 1968; Cardozo and Ritsma, 1968), but only very few studies are available that deal with the jitter of any musical tones, and they are mostly concerned with the tones of stringed instruments (e.g., Cremer, 1973 and McIntyre et al., 1981). A few data about the singing voice are given by Bennett (1981). As far as my knowledge goes, the jitter of the singing voice has not been studied systematically. Still, jitter is present in the singing voice to a significant extent, as will be shown in this paper.

Jitter (s_j) is quantitatively defined as the standard deviation of period duration (s_p) when there is no vibrato or trend:

$$[\; s_j = s_p = (\Sigma (p_i - \bar{p})^2 \; / \; n)^{1/2},$$

in which p_i is the duration of the i-th period, p the mean period duration, and n the number of periods. This measure can be made more meaningful in a musical context by interpreting the standard deviation as a musical interval which can be converted into cents*:

$$[\; J = 1200 \log_2 (1+s_j/\bar{p}) \text{ cents},$$

which measure we will use throughout this paper.

When vibrato and trends are also present, the situation is a bit more complex. The amount of jitter (s_j, now defined as the standard deviation of period durations as far as due to jitter) can be derived from the variances of period durations (s_p^2) and of the time differences of successive periods (s_d^2) in the following way:

$$[\; s_j^2 = \frac{(1-r_v)s_p^2 - \tfrac{1}{2}s_d^2}{r_j - r_v},$$

in which r_j and r_v are the autocorrelations of successive period durations due to jitter (r_j) and to vibrato/trend (r_v), respectively. For a vibrato, this autocorrelation equals cos $(2\,\pi f_v/\bar{f})$, in which F_v is vibrato frequency and $\bar{f}$ is mean signal frequency. When vibrato/trend is sufficiently slow compared to the period durations and the variance due to vibrato/trend is not too large, then the amount of jitter is well approximated by:

$$[\; s_j = \frac{s_d}{(2 - 2r_j)^{1/2}},$$

which means that, when r_j can be assumed to be constant, jitter is about proportional to the standard deviation of the time differences of successive periods. For the singing voice, r_j can be assumed to be about zero.

2. Measurement Procedure

Musical tones sung on Dutch syllables in an anechoic room were recorded on tape. Period durations were measured be feeding the filtered fundamental component to a Schmitt trigger that produced pulses at every positive zero crossing. The onset times of the pulses were measured by a 10 MHz clock and stored in computer memory. Further analyses were done with help of computer programs. Diagrams with instantaneous frequency (as the reciprocal of period duration) could be plotted on a graphic terminal (see Fig. 1).

3. Results

We will present here the results of measurements of the jitter in *h*-vowel-*t* syllables sung by two professional singers, a bariton and a soprano. They sang nine different vowels on various fundamental frequencies in three modes: straight (no vibrato), neutral (small vibrato), and strong vibrato. Mean jitter of the nine vowels (i.e. averaged per vowel over the three singing

* The *cent* scale is a logarithmic transformation of the interval as frequency ratio:

$$[\; I = 1200 \log_2 q/p \text{ cents},$$

where q/p is the frequency ratio, and I the interval size in cents. One cent is 1.000578/1, the octave (2/1) is 1200 cents, a semitone in equal temperament ($2^{1/12}/1 = 1.059$) is 100 cent.

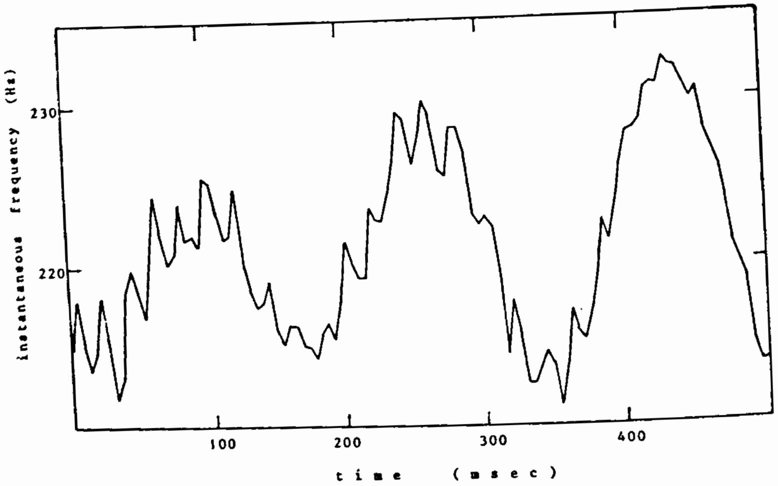


Figure 1. Frequency per period as a function of time. The tone is A2 (220 Hz), sung with vibrato by the bariton. The duration of the fragment is 500 msec. The number of periods is 110. The vowel is /u/ in Dutch *hoet*, with an amount of jitter of about 10 cents and a vibrato depth of about 60 cents.

modes) has been presented as a function of vowel in Figure 2, with signal frequency as parameter. Generally spoken, there is a frequency effect. Low tones have more jitter than high tones. But there is a vowel effect as well. Vowels with a low first formant frequency F_1 (like Dutch *hiet* and *huut*) show relatively small amounts of jitter, while vowels with a high F_1 (like Dutch *haat* and *hat*) show relatively large amounts of jitter.

I want to propose here the hypothesis that the jitter differences between vowels are due to the interaction between the vocal cords and the vocal tract. The vocal cords and the vocal tract may be seen as two coupled oscillators. In principle, the vocal cords are driving, while the vocal tract is driven, which implies a certain phase relation between the oscillations of the two systems. However, because of the coupling the vocal tract may also have a driving impact on the vocal cords, which implies a phase relation that is surely not the reverse of the one just mentioned. It seems plausible that these conflicting phase relations can be the cause of a small phase instability in the transmitted wave from the vocal cords to the vocal tract, which, as a matter of fact, becomes apparent in the produced sound signal as a small frequency instability.

The phase lag ϑ of a driven oscillator relative to a driving one is most simply given by $\cot \vartheta = Q(F_1/F_0 - F_0/F_1)$, in which Q refers to the damping of the driven oscillator, F_0 is the frequency of the driving oscillator and F_1 is the resonance frequency of the driven oscillator. When we use the F_1 values given by Pols et al. (1973) and Nierop et al. (1973), the correlation between the amount of jitter and the quantity $F_1/F_0 - F_0/F_1$ appeared to be 0.89 for the 131 and 220 Hz bariton tones and 0.85 for the 220 Hz soprano tones. For the highest soprano tones, the calculation of such a correlation makes less sense, since the F_0 then is in the neighbourhood of the F_1, while it is possible that the singer adapted the F_1 in one way or another to the F_0. The data show that, as long as F_0 is around or above F_1 the amount of jitter is at some minimum base-line level of about 2 to 3 cents*. The vowels with an F_1 substantially (more than 100 to 200 Hz) higher than the F_0 have more jitter. For the 392 Hz tones, this is beginning with *hot* (F_1 578 Hz), for the 660 Hz tones only with *haat* (F_1 986 Hz), for the 880 Hz tones not at all.

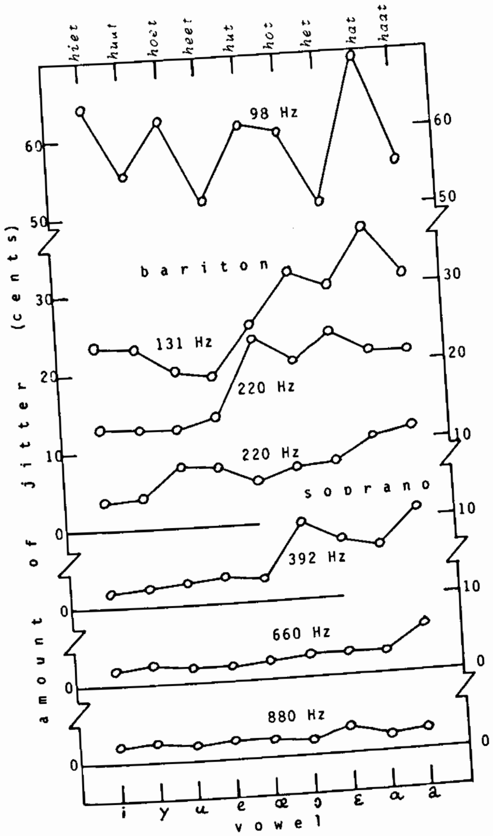


Figure 2. Amount of jitter as a function of vowel. Parameter is signal frequency. The vowels are ordered after their first formant frequencies, from low to high. We used data given by Pols & al. (1973) and Nierop & al. (1973).

The large, vowel-independent amounts of jitter of the 98 Hz bariton tones may be due to the surpassing of some critical frequency separation between F_0 and F_1.

4. Conclusion

The data given about the amounts of jitter present in the singing voice must be seen as preliminary. However, they are encouraging for further research. Jitter is a 'noisy' phenomenon, which can vary with singer, mode of singing, frequency and vowel, but also with attention and concentration, fitness and fatigue, exercise and mood. Really smooth, nice data should never be expected. However, jitter seems to be a rather fundamental property of musical tones in general. Couplings like the one between the vocal cords and the vocal tract are present in almost all acoustical musical instruments, like the strings and the top plate or sound board of stringed instruments, the reed and tube in the reed wind instruments, the lips and tube in brass wind instrument, etc. Very often one of the oscillators is vibrating with another than its resonance frequency. When jitter is indeed caused by the interaction between these coupled systems, it should be expected to be present in almost every musical tone. It may well be one of the features of tones that make possible a distinction between the dead sounds of electronic frequency generators and the live sounds of musical instruments.

Acknowledgements

My thanks go to Gerrit Bloothooft who lent me his recordings of singing-voice tones, and to the Audiology Department of the Institute for Perception TNO, Soesterberg, the Netherlands, where I could make use of computing and audio equipment.

References

Bennett, G. (1981). Singing synthesis in electronic music. In: *Research aspects on singing*. Royal Swedish Academy of Music, Stockholm. 34-50.

Cardozo, B.L., and Ritsma, R.J. (1968). On the perception of imperfect periodicity. *IEEE Trans. Audio Electroacoust.* **16**, 159-164.

Cremer, L. (1974). Der Einfluss des "Bogendrucks" auf die selbsterregten Schwingungen der gestrichenen Saite. *Acustica* **30**, 119-136.

Lieberman, Ph. (1963). Some acoustic measures of the fundamental periodicity of normal and pathological larynges. *J. Acoust. Soc. Amer.* **35**, 344-353.

McIntyre, M.E., Schumacher, R.T., and Woodhouse, J. (1981). Aperiodicity in bowed-string motion. *Acustica* **49**, 13-32. Corrigendum in *Acustica* **50** (1982), 294-295.

Nierop, D.J.P. van, Pols, L.C.W., and Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica* **29**, 110-118.

Pollack, I. (1968). Detection and relative discrimination of auditory jitter. *J. Acoust. Soc. Amer.* **43**, 308-315.

Pols, L.C.W., Tromp, H.R.C. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Amer.* **53**, 1093-1101.

Rasch, R. (1983). Jitter in violin tones. Submitted to the Proceedings of the Stockholm Music Acoustics Conference 1983.

Seashore, C.E. (ed.) (1932). The vibrato. *Studies Psychol. Music* **1**. Iowa Univ., Iowa City.

A Perceptual Evaluation of Two V/U Detectors

N. van Rossum and A. Rietveld
Nijmegen, the Netherlands

1. Introduction

In many analog pitch meters, however different they may be, the pitch detection circuitry is controlled by a voiced/unvoiced (V/U)-detector. This is to say that pitch will only be determined in those segments which have been labelled 'voiced' in a previous stage.

Both parts of such pitch meters, the voiced/unvoiced decision system and the pitch detector itself, have one thing in common: the evaluation problem. In both cases it is difficult to find clearly operationable correlates of the features to be detected in the acoustical and perceptual domains.

In this contribution we will focus on the description and perceptual evaluation of two V/U-detectors which are parts of two analog pitch meters described elsewhere (van Rossum, 1982). We tried to find an answer to the following questions:

1. Is it possible to obtain reliable 'voiced/unvoiced' judgments from a panel of listeners?
2. Do judges agree equally well on the onset and offset of voiced segments?
3. Which of the two V/U detectors (to be described below) corresponds best with the decisions of the listeners?

2. Short Description of the two V/U-detectors

We tested two different detectors which are integral parts of two analog pitch processors developed in our laboratory. Fig. 1 gives a blockdiagram of these V/U detectors; their main characteristics will be summarized below.

a. A classical V/U detector, in which the energy in a low frequency band (20 Hz - 1 K Hz) is compared with a predetermined criterion. This detector is based on a principle already applied by Dudley (1939); it was found to be very reliable by Wiren and Stubbs (1956).

b. A V/U detector which measures the spectral balance in the speech signal. To this aim the energy difference in the bands 20 Hz - 1 kHz and 5 kHz - 14 kHz is determined, whereafter the result is compared with a criterion value. Voiced segments are assumed to have predominantly low frequency energy and voiceless segments high frequency energy.

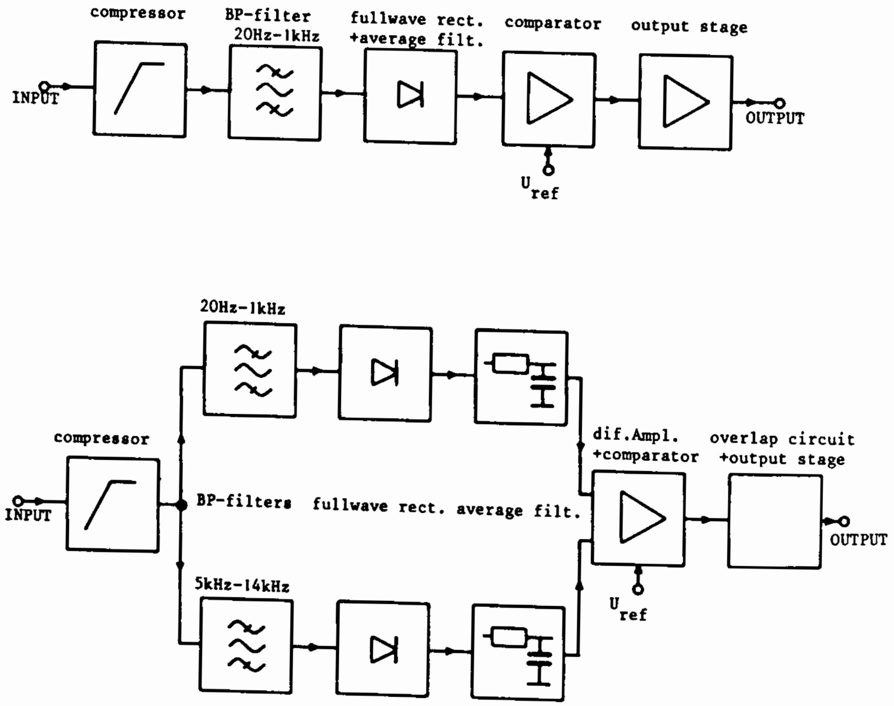


Figure 1. Blockdiagrams of a 'classical' V/U-detector (A) and of an energy difference V/U-detector (B).

This type of detector was originally designed by Knorr (1979) and has been adapted by us. The detector does not only differ from the previous one in the frequency bands which are used, but also in an 'overlap circuit' applied to the output of the detector. In order to avoid 'jittering' in the V/U-output, this circuit was integrated in the design; the resulting delay time amounts to about 10 ms, depending on the input waveform.

Knorr's evaluation of the detector was based on the comparison of the V/U-decisions and the waveform; the result of this evaluation was reported to be very satisfactory. In our opinion, however, it is rather difficult to determine whether a signal is semiperiodic - especially at the boundaries of voiced segments - and should consequently be labelled as 'voiced'.

For that reason we designed another evaluation test, a perceptual one.

3. The perceptual evaluation of V/U-detectors

The evaluation of V/U-detectors is notoriously difficult for many reasons; the main problem is the absence of a one-to-one relation in the three domains involved. Dolansky (1968) showed that the vibration of the vocal cords does not always result in a periodic signal, whereas Glave (1973) found that stochastic signals without a clear periodicity may have a 'tonal' quality and will consequently be judged as voiced. Furthermore, the distinction tone/no tone appeared not to be a categorical one.

A supplementary problem resides in the fact that different evaluation procedures may lead to different results. The scores obtained in a perceptual evaluation of a V/U-detector will depend on the size of the signal units which are to be judged. Segments of phone size will be judged in a totally different manner than segments of say 30 ms. In the former case a complex of cues will be used (Slis and Cohen, 1969), in the latter case the judgments will be based on spectral features.

If a perceptual evaluation is realized by means of speech resynthesis in which voice-onsets and offsets are controlled by a V/U-detector, the two possible biases of the detector, voiced and unvoiced, will differently influence the acceptability of the synthesized speech signal.

An 'acoustic' evaluation is also rather hard to perform. Proper periodic signals hardly occur in real speech; at the end of voiced segments semiperiodic signals cannot easily be distinguished from noise.

We chose a perceptual criterion in the evaluation of the V/U-detectors involved, one reason being the difficulties which can be expected in an acoustic evaluation. Our experiment was designed in such a way that subjects had to judge short successive speech segments of 30 ms as 'voiced' or 'unvoiced'. We chose this perceptual scanning procedure because it simulates to a certain extent the functioning of the V/U-detectors; these detectors determine whether short speech segments (in fact indefinitively short) will be further processed by the pitch detector or not.

4. Procedure

Before the real trials started, the subjects were given a set of 'anchoring' trials in which short segments (30 ms) of clearly voiced or unvoiced speech were presented.

The experimental trials consisted of segments from six sentences each spoken by two speakers (one male, one female).

By means of a variable gate successive segments of 30 ms were presented over earphones to ten listeners (5 male, 5 female). By means of a thumbwheel switch the subjects could shift a segment by incremental steps of 10 ms. The subjects had to mark the transitions from 'voiced' to 'unvoiced' segments. The thumbwheel switch indicated the onset of the trapezoidal window, as is shown in Fig. 2.

In this way a perceptual scanning of the twelve sentences took place; the resulting data were the perceived onsets and offsets of the voiced segments in milliseconds.

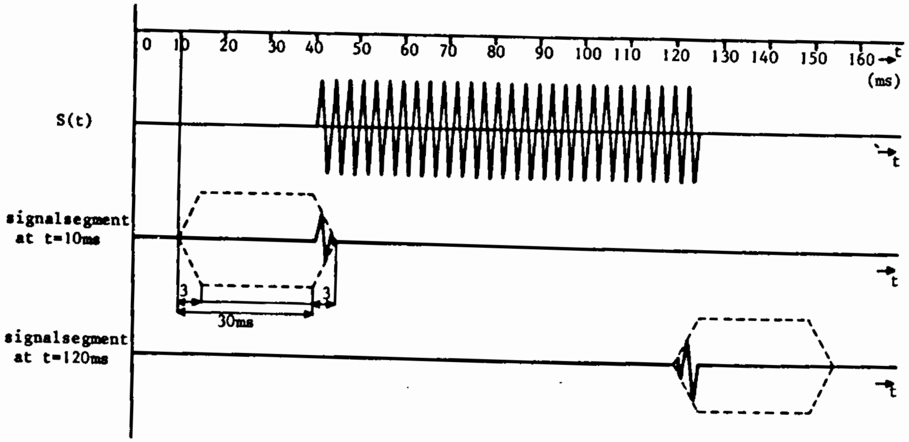


Figure 2. Position of the trapezoidal window and the perceived onset and offset.

5. Results

The perceptual scanning resulted in 78 segments which were labelled 'voiced' by at least two of the ten listeners. For further processing we only used the registered voice onsets and offsets of those segments which were judged 'voiced' by at least 7 of the ten listeners (62 segments).

Subjects were found to be reliable in their judgements. Ebel's reliability coefficient was .99 for both onset and offset times. Therefore, we feel justified to make use of the mean scores in the subsequent analyses.

In spite of the high reliability, an F-test showed differences in agreement between the perceived onset and offset times: $F = 18.77$ df1/2=9, $p < 0.001$. It appeared that subjects agreed more on the onset of voiced segments than on the offset; this difference is probably due to the often observed asymmetry of the waveform of voiced segments.

As for the agreement between the decisions of the listeners and both detectors, we performed separate analyses for onset and offset times. In Table I we present the mean onset and offset times perceived by the listeners and those found by detector 1 and detector 2, respectively.

A Multiple Range Test (significance level: 0.05) showed significant differences between the onset times determined by the detectors and the perceived onset times; the difference between the detectors was not significant.

Table I. Onset and offset times of voiced segments in milliseconds; mean values of the listener's judgements arbitrarily set to zero

	listeners	Detector 1	Detector 2
onset	0	33	32
offset	0	2	6

The same test was applied to the offset times. The only significant difference was that between detector 2 and the listeners, a difference of not more than 6 milliseconds.

6. Discussion and Conclusion

The difference between the performance of the detectors and the decisions of the listeners on the onset of voiced segments can be explained by the method we used in the perception experiment. The registered onset times of the listeners equal the opening time of the variable gate. If a subject labels a segment 'voiced' as soon as the last part of the window is voiced, the registered onset time will be 30 ms ahead of the 'real' onset. An interval of about thirty milliseconds happens to be the difference we found between the decisions of the listeners and those of the detectors. As in most cases a close agreement was observed between the onset of semiperiodicity and 'voiced' labels of the detectors, we may conclude that the observed difference is for a great part due to the window we used in the experiment.

As for the offset times, the fact that the 'voiceless' decisions of detector 2 were significantly later than those of the subjects can be explained by the overlap circuit which is part of the system. This circuit appeared to be rather important as without it (detector 1) much more and longer jitterings - about 100% - were found. If we take into account the above mentioned effects, we may conclude that the two V/U detectors performed equally well and in close agreement with the judgments of the listeners. This finding should not obscure the existing differences between both detectors. In particular the overall-amplitude of the signal has a noticeable influence on the functioning of the detector which only operates on the LP-frequency band (detector 1) and much less influence on the performance of detector 2. For that reason, the latter should be preferred to the former.

As is well known, voiced/unvoiced decisions have a strong influence on the quality of synthesized speech. It is not yet clear whether parameter estimation in which our V/U detectors are used, will lead to acceptable resynthesized speech. Experiments in that direction are planned.

References

Dolansky, L. and Tjernlund, P. (1968). On certain irregularities of voiced speech waveforms. *IEEE Trans. Audio and Electroacoust., Vol. AU-***16**, 51-56.

Dudley, H. (1939). The Automatic Synthesis of Speech. *Bell Teleph. Syst. Techn. Publ. Monograph B-1169.*

Glave, R.D. (1973). *Untersuchungen zur Tonhohenwahrnehmung Stochastischer Schallsignale.* Hamburg: Helmut Buske Verlag.

Knorr, S.G. (1979). Reliable Voiced/Unvoiced Decision. *IEEE Trans. on Acoustics Speech and Signal Processing, Vol. ASSP-***27**, **3**, 263-267.

Rossum, N. van and Boves, L. (1978). *An analog pitch-period extractor.* IFN-proceedings, **2**, 1-17.

Rossum, N. van (1982). Technische beschrijving van de F_0-processor. Interne Publicatie IFN.

Slis, J.H. and Cohen, A. (1969). On the complex regulating the voiced-voiceless distinction. Part I. Part II. *Language and Speech* **12**, 80-102, 137-155.

Wiren, J. and Stubbs, H.L. (1956). Electronic Selection System for Phoneme Classification. *JASA* **28** (6), 1082-1091.

F_0 Behaviour in Mandarin and French: An Instrumental Comparison

R. Sneppe and V. Wei
Brussels, Belgium

1. Introduction

Learning Mandarin Chinese is almost proverbially difficult for adults and this is not restricted to 'foreigners'. Native Chinese raised in a different dialect, e.g. Cantonese, have a problem in speaking Mandarin, see Tse (1974). What are the reasons for this difficulty? Three types of explanations have been advanced.

First Tse (1974) argues for the simple interference hypothesis. Cantonese have problems when learning Mandarin because two tone contours that are in free variation in Cantonese, have a lexical status in Mandarin. This is the main source of confusion.

A second hypothesis has been advanced by Chen (1974). Comparing the native speech of Chinese and Americans, he reaches the following conclusion: 'When the test subjects spoke their native language, (...) the average (pitch) range of the four Chinese subjects was 154% wider than that of the four English-speaking subjects... an English speaking person... should widen his normal pitch range at least 1.5 times if he wants to successfully learn to speak Chinese.

Finally Eady (1982), comparing American English and Mandarin arrives at a very different conclusion. '... for both the Mandarin and English speakers, the compass[1] of the voice (which is directly related to standard deviation) was a good indication of the range of F_0 values produced. The lack of significant difference between the two language groups for the variable SDFF (Stand. Dev. of Fund. Freq.) indicates that the F_0 values for the Mandarin speakers varied over a range that was approximately equal to that of the English.' Instead of a difference in pitch range, Eady records a statistically significant increase of F_0 fluctuation (as a function of time and of the number of syllables per second) for Mandarin speakers.

These quite different and sometimes conflicting hypotheses call for more investigation. Yet another comparison is presented here. Using French, a language known for its relatively 'flat' intonation, and Mandarin, stretches of continuous speech were analyzed with respect to F_0 behaviour.

Cross-language comparisons using different speakers are difficult to interpret. It is indeed a very intricate business to disentangle the various factors and their influences. Hence the usefulness of bilingual speakers. In this study a single bilingual French-Chinese speaker was used. If there are differences between Mandarin and French, these should be obvious in his speech.

2. Method

2.1. Subject

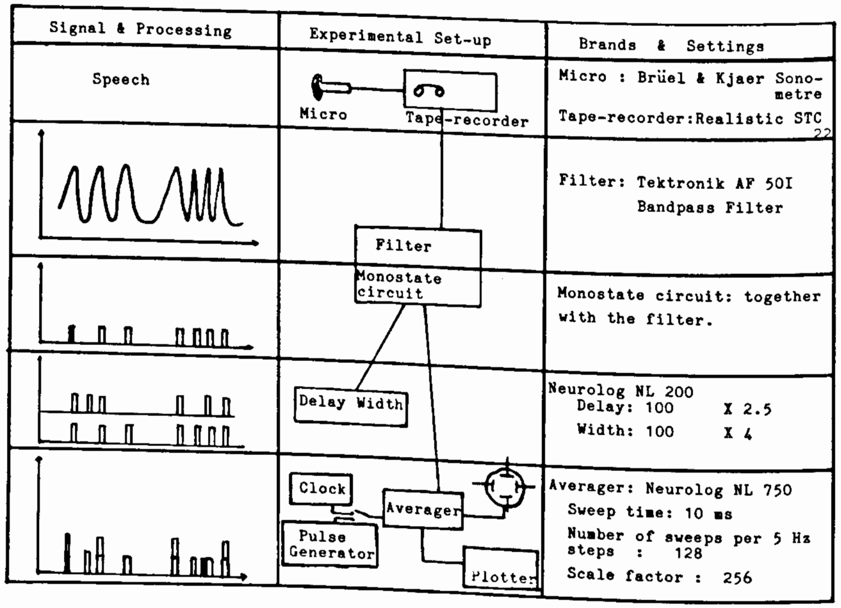
One of us being Chinese-French bilingual, the choice of the subject was obvious. The definition of bilingualism adopted here is: on the phone, no native speaker of Chinese or French would mistake the bilingual as a foreigner. Off the phone, any accent should be attributed to intra-dialectal variation rather than to foreign accent.

2.2. Recording procedure

Two texts were read. In French it was a fairy tale: 'La Reine des Neiges.' (The Queen of Snow). In Chinese it was a monologue taken from De Francis' Intermediate Course (Lesson 2). Both texts were read unemotionally and at normal pace. The recording was done under low level background noise.

2.3. Analysis procedure (see figure 1)

The filter is set at a given frequency. Any voiced segment coming out of it starts a time window (here 10 ms.). During this time, each period, transform-



Micro : Brüel & Kjaer Sonometre
Tape-recorder:Realistic STC 22

Filter: Tektronik AF 50I Bandpass Filter

Monostate circuit: together with the filter.

Neurolog NL 200
Delay: 100 X 2.5
Width: 100 X 4

Averager: Neurolog NL 750
Sweep time: 10 ms
Number of sweeps per 5 Hz steps : 128
Scale factor : 256

Figure 1. Description of the experimental set-up.

ed as a normalized pulse, is recorded in one of the 256 memory blocks of the averager. When the window has opened 128 times the averaging stops and the process is restarted, with the filter 5 Hz lower or higher depending on whether the filter is sliding towards 120 Hz or 200 Hz in 5 Hz steps. This gives two curves for each language. One called the increasing curve (i.e. when the filter is sweeping upwards to 200 Hz) and the other is, of course, the decreasing curve. Motivation for the upsweep downsweep is discussed in 4.2.

3. Results (see figure 2)

From the curves the following information can be deduced:

1. Weighted mean F_0 for Chinese upsweep: 173 Hz Standard Deviation: 29 Hz
   Weighted mean F_0 for Chinese downsweep: 173 Hz Standard Deviation: 31 Hz
   Weighted mean F_0 for French upsweep: 166 Hz Standard Deviation: 22 Hz
   Weighted mean F_0 for French downsweep: 166 Hz Standard Deviation: 21 Hz
2. Chi-square tests were run, giving the following results: the two Chinese curves were similar with a probability under .1; the two French curves

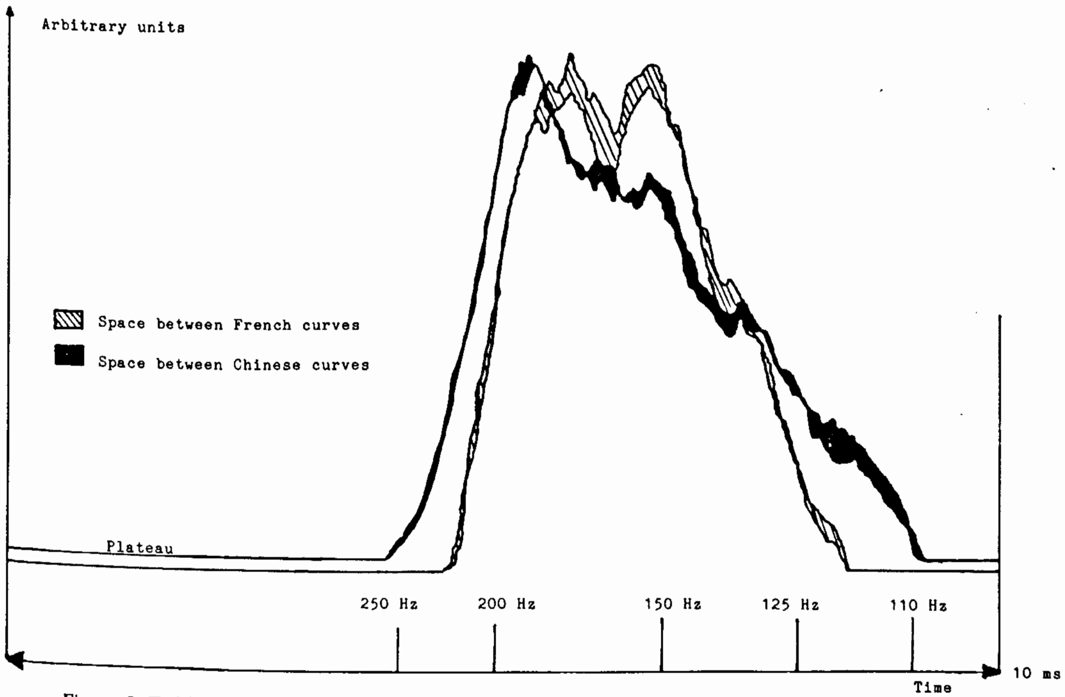


Figure 2. F_0 histogram, expressed in arbitrary units vs. frequency (time).

were also similar with a probability under .10; finally each Chinese curve was different from each of the two French curves with a probability over .99 . From these results, it is clear that Chinese F_0 behaviour is different from the French one.

4. Discussion

Various factors influence the curves presented here. They will be

1. Choice of the window width: 10 ms was chosen because the speaker did not have a significant contribution of pitch under 110 Hz.
2. Filter: It was found that changing the filter value resulted in substantial curve modification. In order to reach a curve closer to reality, it was decided to resort to a 5 Hz upsweep and downsweep from 120 Hz to 200 Hz for each language. This coupled with the same number of averaging gave a more balanced picture.
3. Segmentation: The procedure adopted here cuts up the speech in groups of 128 windows of 10 ms. This segmentation is not the same for the upsweep and the downsweep because first we did not make sure to start at exactly the place on the tape; secondly, frequency repartition in the spoken stretches are not identical, hence for some frequencies (especially the ones close to the mean) the 128 windows are recorded in less time than for other frequencies. This means that the upsweep and downsweep do not yield the same segmentation. It is as if two different texts were read, one for the upsweep and another for the downsweep. This fact diminishes the influence of the particular texts.
4. Pulse width: this was chosen here to give a reasonably detailed picture.
5. Plateau: Pulses are created by a voltage, however the absence of pulse does not correspond to an absence of voltage; even in the absence of pulse there is a slight continuous voltage. This explains the presence of the plateau. The plateau height is proportional to the number of averaging done.
6. Number of averaging was set at 128 pulse trains. (i.e. the window opened 128 times). We found this to be a convenient value.

The factors above mentioned should be put in the context of this important argument: the subject's reading of French and Chinese texts was submitted to the same treatment. If any difference can be detected, it has to be attributed to language difference.

5. Conclusion

From the results obtained, it is clear that French and Chinese have different F_0 behaviour. Pitch range is greater in Chinese than in French. It can be tentatively inferred that French speakers ought to increase their pitch range by 1/3.

The procedure used here does not allow inferring any information as to the rate of fluctuation on Eady's sense (Eady 1982). This will be a topic for further investigation.

[1] voice compass: the stretch of frequencies to standard deviations on either side of the mean F_0.

References

Chen, Gwang-tsai (1974). The pitch range of English and Chinese speakers. *J. of Chinese Linguistics* **2** 2, pp. 159-171.

De Francis (1967). *Intermediate Chinese.* Yale University Press.

Eady, S.J. (1982). Differences in the F_0 Patterns of Speech: Tone Language vs Stress Language. *Language and Speech,* **25**, 1, pp. 29-42.

Tse, J.K.P. (1974). An Instrumental Study on the Difficulties with Respect to Tones encountered by Native Speakers of Cantonese when learning Mandarin Chinese. *Education and Culture Monthly*, pp. 62-66, Taipei.

On Difference Operation in Linear Prediction

M. Yanagida and O. Kakusho
Osaka, Japan

Abstract

The relationship between the predictors obtained on differenced data and those on original data is derived for both the covariance method and the autocorrelation method. The physical interpretation of the derived relationship is discussed in connection with spectral enhancement.

1. Difference Operation in the Covariance Method

The linear prediction model for a sampled $\{y_n\}$ is expressed in the form

$$y_n \doteqdot \sum_{i=1}^{p} \alpha_i y_{n-i} \tag{1}$$

where α_i denotes the ith predictor and p is the prediction order. In matrix form, eq.(1) can be written as

$$\underset{m\times 1}{y} \doteqdot \underset{m\times p}{Y}\ \underset{p\times 1}{\alpha} \quad \text{or} \quad \underset{m\times(p+1)}{[\,y \,\vdots\, Y\,]}\ \underset{(p+1)\times 1}{\alpha} \doteqdot \underset{m\times 1}{0} \tag{2}$$

where

$$y \bar{\mp} \begin{bmatrix} y_n \\ \vdots \\ \text{j} \\ \vdots \\ y_{n-m+1} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{n-1} \cdots\cdot y_{n-p} \\ \vdots \qquad \vdots \\ \cdot \qquad \cdot \\ \vdots \qquad \vdots \\ y_{n-m} \cdots\cdot y_{n-m-p+1} \end{bmatrix} \text{ and } \alpha = \begin{bmatrix} -1 \\ \\ \alpha \end{bmatrix} = \begin{bmatrix} -1 \\ \alpha_1 \\ \cdot \\ \vdots \\ \alpha_p \end{bmatrix} \tag{3}$$

Vector *a* is called the augmented predictor vector for α. The normal equation for the covariance method is obtained by premultiplying both sides of eq. (2) by Y^T.

$$Y^T y = Y^T Y \hat{\alpha} \tag{4}$$

where the product matrix Y^TY represents the covariance matrix of $\{y_n\}$.

The least-squares solution for α or ɑ is derived from the normal equation (4) and is expressed as

$$\hat{\alpha} = Y^+ y \quad \text{or} \quad \hat{\alpha} = \begin{bmatrix} -1 \\ \hat{\alpha} \end{bmatrix} \begin{bmatrix} -1 \\ Y^+ y \end{bmatrix} \tag{5}$$

where Y^+ denotes the generalized inverse of Y and is usually identical to $(Y^TY)^{-1}y^T$ except for rank deficient cases. That was the formulation of the covariance method by the generalized inverse of matrices.

In the same way, the linear prediction model for the differenced sequence of the form $\{ y_n - wy_{n-1}\}$ is expressed as

$$y_n - w y_{n-1} \doteqdot \sum_{i=1}^{p} \beta_i (y_{n-i} - w y_{n-i-1}) \tag{6}$$

where β_i denotes the ith predictor for the differenced data. Equation 6 is written in matrix form as

$$\underset{m\times(p+1)}{[\,\Delta y \,\vdots\, \Delta Y\,]}\ \underset{(p+1)\times 1}{b} \doteqdot \underset{m\times 1}{0} \tag{7}$$

where

$$\Delta y \begin{bmatrix} y_n - w y_{n-1} \\ \vdots \\ \vdots \\ y_{n-m+1} - w y_{n-m} \end{bmatrix}, \qquad b = \begin{bmatrix} -1 \\ \beta \end{bmatrix} = \begin{bmatrix} -1 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

and

$$\Delta Y = \begin{bmatrix} y_{n-1} - w y_{n-2} & \cdots\cdots\cdots & y_{n-p} - w y_{n-p-1} \\ \vdots & & \vdots \\ y_{n-m} - w y_{n-m-1} & \cdots\cdots\cdots & y_{n-m-p+\bar{1}}\ w y_{n-m-p} \end{bmatrix} \tag{8}$$

Here again, **b** is the augmented predictor vector for **β**.

In order to investigate the relationship between $\hat{ɑ}$ and $\hat{b}$, it will be reasonable to start with the same number of prediction equations on the same number of data samples. Standing on that point, we assume β_p=0 with the intention of preparing the same number of differenced data that can provide the same number of prediction equations as those on the original data. Under this assumption the pth column of ΔY is arbitrary and eq.(7) can be modified as follows:

$$[y \vdots Y] W b \doteq 0 \tag{9}$$

where

$$\underset{(p+1)\times(p+1)}{W} = \begin{bmatrix} 1 & & & & 0 & \\ -w & 1 & & & & \\ & -w & 1 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & 1 & \\ & 0 & & & -w & 1 \end{bmatrix} \tag{10}$$

Comparing eq.(9) with eq.(2), we can derive the least-squares solution for **Wb** as

$$W \hat{b} = \begin{bmatrix} -1 \\ \\ Y^{+}y \end{bmatrix} = \hat{a} \tag{11}$$

and we get

$$\hat{b} = W^{-1} \hat{a} \tag{12}$$

Equation (12) represents the relationship between the predictors obtained by the covariance method on differenced data and those on original data.

2. Difference Operation in the Auto-Correlation Method

The normal equation of the auto-correlation method for original data sequence is expressed as

$$R \hat{a} = r \qquad \text{or} \qquad [r \vdots R] \hat{a} = 0 \tag{13}$$

where

$$\mathbf{R} \begin{bmatrix} r_0 & r_1 & \cdots\cdots & r_{p-1} \\ r_1 & r_0 & \cdots\cdots & r_{p-2} \\ \vdots & \vdots & & \vdots \\ r_{p-1} & r_{p-2} & \cdots\cdots & r_0 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} r_1 \\ \vdots \\ \vdots \\ r_p \end{bmatrix} \tag{14}$$

and

$$r_i = \sum_n y_n \, y_{n+i} \, . \tag{15}$$

The least-squares solution for α or *a* is obtained as

$$\hat{\alpha} = R^{-1} r \qquad \text{or} \qquad \hat{a} = \begin{bmatrix} -1 \\ \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} -1 \\ \\ R^{-1} r \end{bmatrix} \tag{16}$$

On the other hand, the normal equation for the differenced data is expressed as

$$P \hat{\beta} = \rho \qquad\qquad [\mathrm{p} \vdots P] \hat{b} = 0 \tag{17}$$

where

$$P = \begin{bmatrix} \rho_0 & \rho_1 & \cdots\cdots & \rho_{p-1} \\ \rho_1 & \rho_0 & \cdots\cdots & \rho_{p-2} \\ \vdots & \vdots & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots\cdots & \rho_0 \end{bmatrix}, \quad \rho = \begin{bmatrix} \rho_1 \\ \vdots \\ \vdots \\ \rho_p \end{bmatrix} \tag{18}$$

and

$$\rho_i = \sum_n (y_n - w\, y_{n-1})\,(y_{n+i} - w\, y_{n+i-1}). \tag{19}$$

Since ρ_i is rewritten as

$$\rho_i = -w\, r_{|i-1|} + (1 + w^2)\, r_i - w\, r_{i+1}, \tag{20}$$

we can rewrite $[p \vdots P]$ as

$$[\rho \vdots P] = W^{\mathrm{T}} [r \vdots R] W \tag{21}$$

and eq. (17) as

$$W^{\mathrm{T}} [r \vdots R] W \hat{b} = 0. \tag{22}$$

As $|W^{\mathrm{T}}| \neq 0$, we get

$$[r \vdots R] W \hat{b} = 0. \tag{23}$$

The least-squares solution for *W b* is obtained as

$$W \hat{b} = \begin{bmatrix} -1 \\ \\ R^{-1} r \end{bmatrix} = \hat{a} \tag{24}$$

and we get

$$\hat{b} = W^{-1} \hat{a} . \tag{25}$$

Equation (25) represents the relationship between the predictors obtained by the auto-correlation method on differenced data and those on original data.

3. Physical Interpretation of the Relation between $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$

The relations between $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ for the covariance method and that for the auto-correlation method are identical to each other as formulated in eqs. (12) and (25) or eqs. (11) and (24). The latter two equations express the following relation between the two sets of predictors:

$$\hat{\beta}_i - w \hat{\beta}_{i-1} = \hat{\alpha}_i , \quad .i = 1,2,\ldots,p , \quad \hat{\beta}_0 = -1 . \tag{26}$$

Modifying this equation, we get the following successive equation:

$$\hat{\beta}_i = \hat{\alpha}_i + w \hat{\beta}_{i--1} \tag{27}$$

In closed form, it is written as

$$\hat{\beta}_i = \sum_{j=1}^{i} w^{i-j} \hat{\alpha}_j - w^i \tag{28}$$

Equation (28) is another expression of eqs. (12) or (25), because

$$W^{-1} = \begin{bmatrix} 1 & & & & 0 \\ w & 1 & & & \\ w^2 & w & 1 & & \\ \vdots & \vdots & & \ddots & \\ w^p & w^{p-1} & \cdots & w & 1 \end{bmatrix} . \tag{29}$$

Equation (26) proclaims that $\{ \hat{\alpha}_i \}$ is the differenced sequence of $\{ \hat{\beta}_i \}$. Both the difference operations on sampled data and that on the predictor sequence are interpreted to have the same effects of spectral enhancement in higher frequency region as depicted in Fig. 1, where w is assumed to be unity for simplicity.

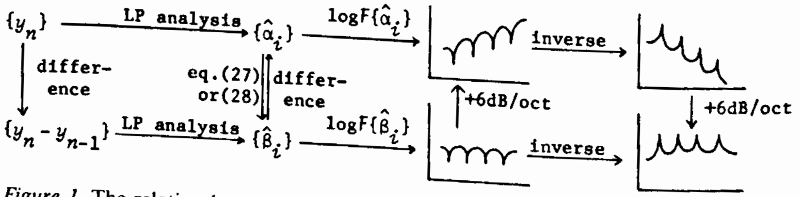


Figure 1. The relation between the difference operation on data and that on predictors.

4. Conclusions

The relationship between the predictors obtained on differenced data and those on original data has been derived. Although the way of derivation employed here is rather rough, the authors have already shown two ways of strict derivations: one (Yanagida et al., 1982a), equating the prediction errors at each sampling point for both sequences, and the other (Yanagida et al., 1982b), employing several theorems concerning the generalized inverse of matrices. Those, however, were only for the covariance method. This paper has discussed the difference operations in a linear prediction analysis aiming at a unified description for both the covariance method and the auto-correlation method.

In this paper, the discussion has been limited only to the first-order differencing, but the derived results are easily expanded to general higher order difference operations (Yanagida et al., 1982b).

Our present interest is to develop an efficient method to replace the sample differencing of fixed pre-emphasis factor with an adaptive inverse differencing on predictors, that is to replace the fixed pre-processing with an adaptive post-processing.

Acknowledgement

This work has been supported by Grant-in-Aid for Scientific Research for 1982 by the Ministry of Education, Science and Culture (No. 56550038).

References

Yanagida, M. et al. (1982a). Effects of the difference operation on linear predictive coefficients. *Proc. Autumn Meet. ASJ.* 3-6-6.

Yanagida, M. et al. (1982b). Analysis of the differencing operation in the linear prediction model by the generalized inverse matrix. *Trans. of the Committee on Speech Research, Acoust. Soc. Jpn.* S82-51.

Section 4
Automatic Speech Recognition

Outline of an Auditory Theory of Speaker Normalization

R.A.W. Bladon, C.G. Henton and J.B. Pickering
Oxford, United Kingdom

Research towards systems of speaker-independent speech recognition and theoretical research on speech perception are both confronted by the problem of speaker normalization, perhaps most forcibly illustrated by the speaker sex problem in acoustic phonetics. This problem can be summed up as follows: two vowels (cf. the data of Peterson and Barney, 1952) or two fricatives (cf. the data of Schwartz, 1968), one of which is spoken by a male and one by a female can be judged 'the same' even by a trained phonetician, yet our analysis equipment reveals great differences. The differences are exemplified by - but not exhausted by - the well known ones of formant frequency.

According to Fant (1975) the problem is compounded, because the measured differences (his 'k-factors') are not consistent from vowel to vowel, nor from formant to formant. One is reminded of his dictum that (ibid.), 'In terms of the acoustic code, female speech remains an obscure dialect.'

Hitherto, approaches to the speaker sex problem (see Disner, 1980 for a review) have mostly concentrated on scaling the acoustic data according to inferred or observed differences in speaker physiology such as vocal tract length.

Our approach is different: it is listener-orientated, and it draws on current knowledge about human auditory analysis, in the belief that this forms an important building-block in the modelling of speech perception. As our point of departure we take Potter and Steinberg's old (1950) idea that, 'a certain pattern of stimulation along the basilar membrane may be identified as a given sound, regardless of position along the membrane'.

That their idea was attractive can be deduced from Figure 1. Now that a good estimate of basilar membrane frequency analysis is available to us (in the psychophysical form of the critical-band scale of Bark units) it is instructive to plot on a Bark scale the same heterogeneous cross-language data from Fant (1975) whose apparent non-uniformity disturbed us at first sight. The plot in Figure 1 is of inter-formant distance F_1-F_2 (in Bark), female against male. The emergent signs of correlation are encouraging.

Our auditory theory of normalization goes somewhat further. As reported at length elsewhere (Bladon and Lindblom, 1981), we postulate a series of acoustic-to-auditory spectral transforms which contain not only a conversion to the Bark scale, but also an auditory filter designed to reflect aspects of

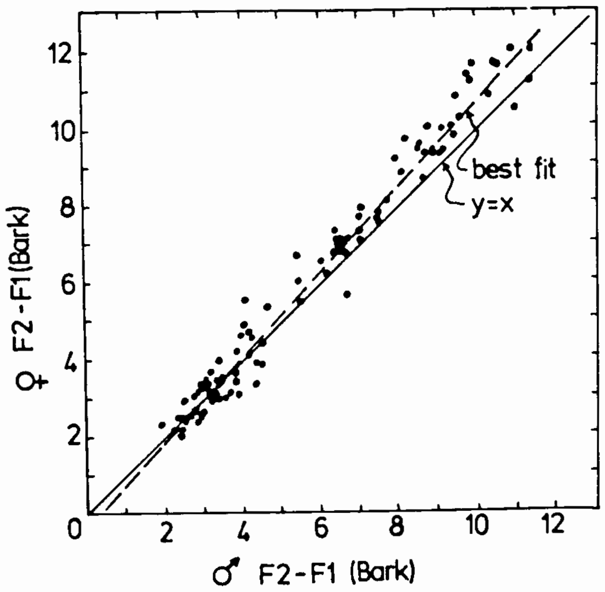


Figure 1. Inter-formant distance F2-F1 (in Bark), females against males. Data from six languages taken from Fant (1975).

masking, together with transforms of intensity level in terms of equal loudness curves and of total loudness calibrated in terms of loudness density per Bark. Further details of this 'auditory model' need not be rehearsed here: suffice it to say that the output is a quasi-auditory spectrum (of e.g. a vowel), which is meant to correspond to that vowel's excitation pattern on the auditory nerve.

Next, take two vowels represented as auditory spectra in this sense: a male vowel and a female equivalent. Suppose that we follow the Potter and Steinberg idea and, analogically speaking, preserve the two excitation patterns in the auditory system but displace the position of one of them. In our terms, we effect a simple linear Bark scale shift of the female pattern. Figure 2 illustrates this procedure applied to several vowels in our data, using a shift of 1 Bark. The coincidence of the resultant spectral shapes is not complete, but it is encouraging as a first approximation. Some progress has in fact been made upon the modifications which are apparently needed, and these are being reported elsewhere. Foremost among these modifications is a warping of the spectrum, especially in the F_1 region, owing to interference from F_0 (see Bladon, 1982).

This auditory theory of normalization has to date been tested on seven sets of male/female vowel data from four languages. Preliminary indications from this (far from satisfactory) database are that the optimal male/female normalization (expressed on average for the vowels of a particular language or dialect) may not necessarily be by 1 Bark. The optimal normalization displacement varies considerably across speech communities. Figure 3 demonstrates this finding quantitatively, insofar as present data permit. The suggestion is, rather as observed by Labov (1978) in Martha's Vineyard speakers, or as concluded by Goldstein (1980) from her vocal tract modelling, that males and females may in some speech communities speak more unlike (or, more like) each other than their vocal tract physiology would predict. In other words, the data of Figure 3 appear to implicate a learned, socially motivated factor for part of a model of speaker normalization.

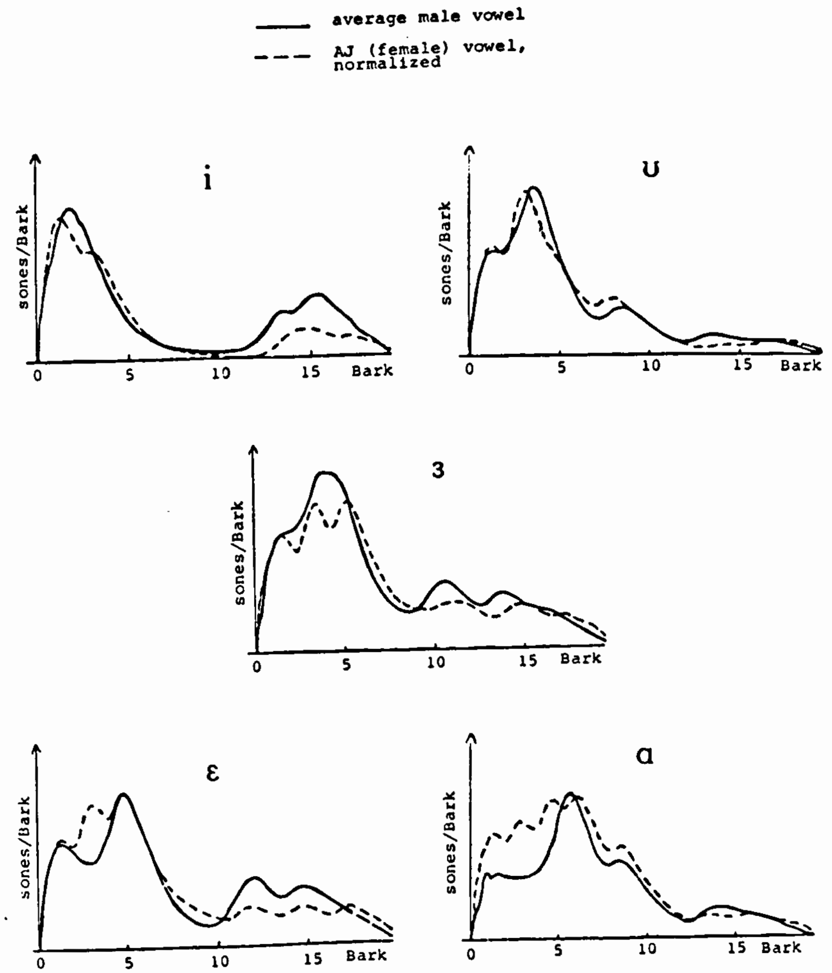


Figure 2. Vowel pairs, male and female, after normalization of the female (dotted vowel) by a downward shift of 1 Bark. The male vowel spectra are averaged over 5 speakers; the female spectra are from one speaker AJ; all are speakers of Middle Northern British English.

However, in order to establish with any certainty these tentative suggestions of phonetic role-stereotyping in vowels, the investigator must be conscious of the need to control a large number of variables. These include

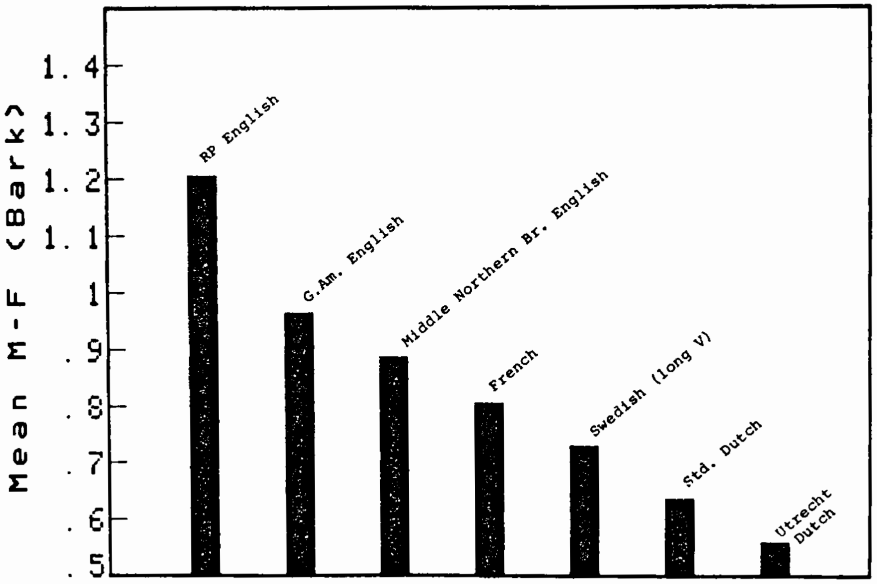


Figure 3. Optimum mean male/female vowel normalization (in Bark), for seven languages/dialects. The data for RP and Middle Northern British English are our own. The remaining sources are: General American (Peterson and Barney 1952), French (Carton 1974 and Mettas 1979), Swedish (Fant 1979), Standard Dutch (Pols et al. 1973 and van Nierop et al. 1973), Utrecht Dutch (Koopmans-van Beinum 1973).

speaker variables such as physique, age, socio-economic background and state of health, as well as experimental variables such as linguistic context, F_0 used and recording conditions. For a discussion of these problems see Henton (1983). It is unfortunate that a good many important controls are missing from most of the data examined in this paper; consequently the conclusions should be treated as no more than suggestive at this stage.

Acknowledgement

This work was supported in part by a Scientific Investigations Grant from the Royal Society.

References

Bladon, R.A.W. (1982). Problems of normalizing the spectral effects of variations in the fundamental. *Proc. Inst. Acoust. 1982*, A51-A55.

Bladon, R.A.W. and Lindblom, B. (1981). Modelling the judgement of vowel quality differences. *J. Acoust. Soc. Am.* **69**, 1414-1422.

Carton, F. (1974). *Introduction à la Phonétique du Français.* Paris: Bordas.

Disner, S.F. (1980). Evaluation of normalizations. *J. Acoust. Soc. Am.* **67**, 253-261.

Fant, G. (1975). Non-uniform vowel normalization. *RIT Stockholm Qu. Prog. Stat. Rep.* 2-3/1975, 28-52.

Fant, G. (1979). *Speech Sounds and Features.* MIT Press.

Goldstein, U. (1980). An Articulatory Model for the Vocal Tracts of Growing Children. Doctor of Science Thesis, MIT.

Henton, C.G. (1983). Changes in the vowels of Received Pronunciation. *J. Phon.* **11** (forthcoming).

Koopmans-van Beinum, F.J. (1973). Comparative Phonetic vowel analysis. *J. Phon.* **1**, 249-261.

Labov, W. (1978). *Sociolinguistic Patterns.* Oxford: Blackwell.

Mettas, O. (1979). French oral vowels analysed from recording spontaneous conversations. In: B. Lindblom and S. Öhman (eds.), *Frontiers of Speech Communication Research.* London: Academic Press.

Peterson, G.E. and Barney, H. (1952). Control methods used in a study of the vowels. *J. Acoust Soc. Am.* **24**, 175-184.

Pols, L., Tromp, R. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 50 male speakers. *J. Acoust. Soc. Am.* **53**, 1093-1101.

Potter, R.K. and Steinberg, J.C. (1950). Towards the specification of speech. *J. Acoust. Soc. Am.* **22**, 807-820.

Schwartz, M.F. (1968). Identification of speaker sex from isolated voiceless fricatives. *J. Acoust. Soc. Am.* **43**, 1171-1179.

Van Nierop, Pols, L. and Plomp, R. (1973). Frequency analysis of Dutch vowels from 25 female speakers. *Acustica* **29**, 110-118.

Automatic Segmentation of the Speech Signal into Phone-length Elements

W. Jassem
Poznań, Poland

For the purposes of automatic speech recognition, the quasi-continuous signal has to be split up into fragments correlated with linguistic units such as words, morphs or phones. Various considerations of intended applications, technological feasibility as well as the theoretical approach decide what these elements are and how they are discovered, in particular systems. Phoneticians, linguists and acousticians are divided on the subject of segmentability of the speech signal into succesive phonetic entities. Studdert-Kennedy (1981) and Hammarberg (1982), for instance, maintain that phonetic segmentation is only performed at the level of perception and human recognition and that the acoustic speech signal is not segmentable. Such assertion ignores the unshakeable evidence produced, e.g. by Fant (1964) and Reddy (1967) showing that both the original waveform and its transformation into a dynamic spectrogram display unmistakeable points, or brief moments, along the time axis, at which some quite definite changes take place and that these points, or moments, correlate almost perfectly with boundaries between phone-related segments.

A system has been developed in the Acoustic Phonetics Research Unit which consists of a bank of 63 analogue band-pass filters, an interface including an A-to-D converter, and a very primitive 8-K-byte, 8-bit-word minicomputer. This system enables four different kinds of digital spectrograms to be made. For each spectrogram, the signal level is averaged in 4 contiguous frequency bands with a 3-band overlap the effective bandwidth being 320 Hz up to 3560 Hz and progressively larger up to 8310 Hz, resulting in 60 analysis channels. The width of the non-overlapping time windows is 23 ms. (1) The basic spectrogram indicates the signal level in each time-frequency cell above a pre-selected threshold in units of 0.6 dB. (2) The differential spectrogram shows the difference, in positive or negative numbers, between the levels in successive cells in each channel. (3) The difference sign spectrogram only indicates the sign of the difference calculated for (2). Finally, (4), a binary spectrogram signals whether or not the level in each cell exceeds a dynamically varied threshold level. Technical details of the method of obtaining binary spectrograms with the MERA-303 minicomputer are obtained in Kubzdela (1980). The segmentation into phone-length elements here proposed is based on type-3 spectrograms, an example of which appears in Fig. 1. A light sign stands for a minus and a heavy sign for a plus, while unmarked dot indicates defaults corresponding to below-threshold cells. Note that some of the columns in Fig. 1 contain only minusses or only plusses, e.g. column 016, 034, 036, others contain only minusses in the lower frequencies and only plusses in the higher frequences, or vice versa, e.g. 046, 048, while still others are irregular, e.g. 012, 060, etc. in that they consist of more than two same-sign vertical sequences. Basically, segmental boundaries are assumed to occur in columns with one same-sign vertical sequence and, in specific cases, in columns with two same-sign sequences. The formal definitions of the boundaries are as follows:

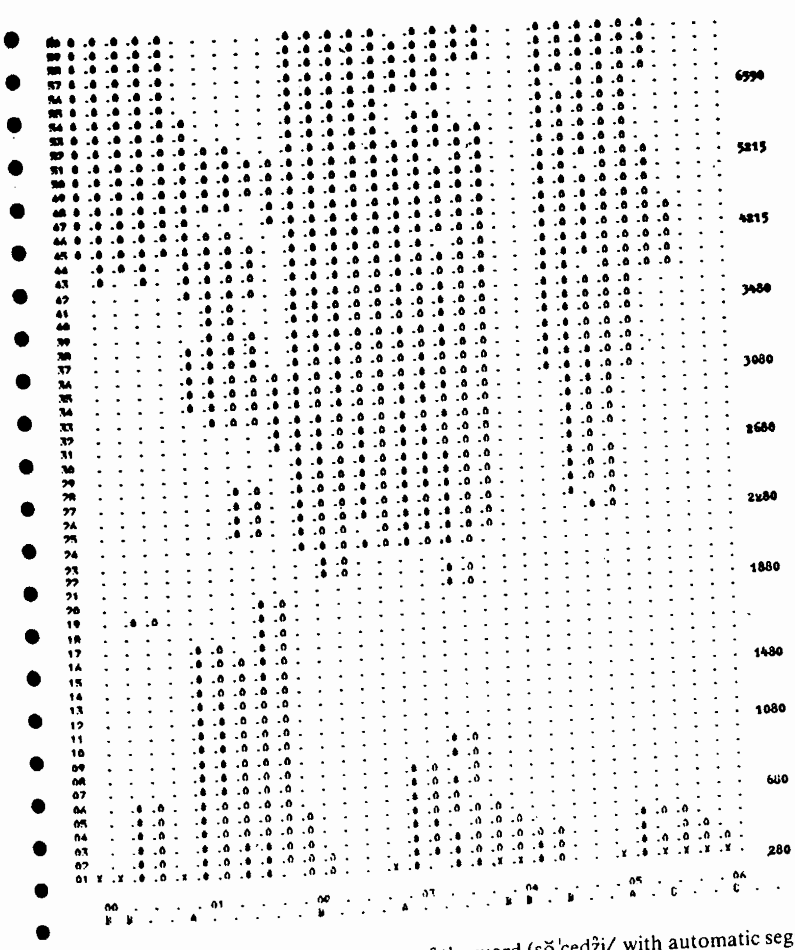


Figure 1. A differential sign spectrogram of the word (sŏˈçedẑi/ with automatic segmentation.

We denote c(k) -- number of same-sign vertical sequences in the k-th column; z(k) -- sign of the highest-frequency vertical same-sign sequence; z(k) equals 0 for a negative sequence and 1 for a positive sequence.

An A boundary appears in the k-th column if
$c(k) = 1_{\wedge}\{c(k+1) \neq 1_{\vee}[c(k+1) = 1_{\wedge}z(k+1) \neq z(k)]\}.$

A BB boundary appears in the k-th and the (k+1)-th column if $c(k) = c(k+1) = 1 \wedge z(k) = z(k+1) \wedge \{c(k+2) \neq 1 \vee [c(k+2) = 1 \wedge z(k+2) \neq z(k)]\}$.

A C boundary appears in the k-th and the k+ n-1)th column if $c(k) = c(k+i) \wedge z(k) = z(k+i) \wedge \{c(k+n) \neq 1 \vee [c(k+n) = 1 \wedge z(k) \neq z(k+n)]\}$, with $i = 1,2,\ldots n-1$ and $n \geq -3$.

A D boundary appears in the k-th column if $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) \neq 1 \wedge c(k+2) \neq 2$.

An F boundary appears in the k-th column if $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge z(k) \neq z(k+n) \wedge c(k+n) = 2$, with $i = 0,1,2\ldots, n-1$, and $n \geq 1$.

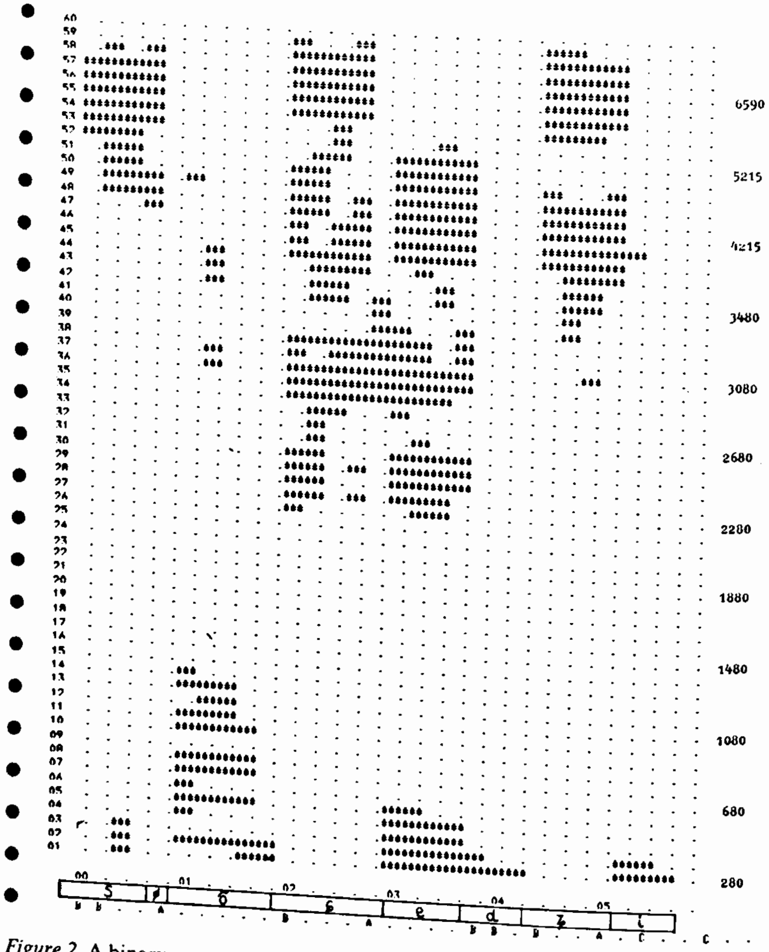


Figure 2. A binary spectrogram of the word /sŏ çedẑi/ with visual and automatic segmentation.

A G boundary appears in the k-th column if $c(k) = 2 \wedge c(k-1) \neq 1 \wedge c(k+1) = 2 \wedge c(k+i) = 2 \wedge z(k) = z(k+i) \wedge c(k+n) \neq 1 \wedge c(k+n) \neq 2$, with $i = C, 1, \ldots, n-1$, and $n \geq 2$.

The detailed algorithm of segmentation is presented in Jassem, Kubzdela and Domagata (in press). A formal test of the algorithm was performed with one male and one female voice pronouncing, in isolation, 17 Polish words containing sequences of phone types known to be particularly difficult to segment. Overall results were as follows:

	Correct	Misses	False alarms
Male voice	129	25	13
Female voice	122	23	9

Fig. 2 represents a binary spectrogram with visual and automatic segmentation.

The results obtained here compare favorably with those obtained in other ASR systems, particularly in view to its extreme simplicity which enables it to be implemented in a microprocessor. Meanwhile improvements of our system are in progress.

References

Fant, G. (1964). Phonetics and speech research. In: *Research Potentials in Voice Physiology*. New York, 159-239.

Hammarberg, R. (1982). On re-defining co-articulation. *J. Phonetics* **10**, 123-137.

Jassem, W., Kubzdela, H., and Domagata, P. (in press). Automatic acoustic-phonetic segmentation of the speech signal. In: *From Sounds to Words*, Umeå Studies in the Humanities.

Kubzdela, H. (1980). A method of word recognition based on binary spectrograms (in Polish). *IFTP Reports* 15/80.

Reddy, D.R. (1967). Computer recognition of continuous speech. *Journ. Acoust. Soc. Am.* **41(5)**, 1295-1300.

Studdert-Kennedy, M. (1981). Perceiving phonetic segments. In: *The Cognitive Representation of Speech*. T. Myers et al. (eds.), North-Holland Publ. Co., Amsterdam, 3-10.

Speaker Recognition in Open Sets

W. Majewski and C. Basztura
Wrocław, Poland

1. Introduction

The development of methods and techniques for automatic speaker recognition creates a real promise for their future practical applications in a variety of social and business activities. These include such fields as criminology, protection of information stored in computer memory, voice control of machines by authorized persons. In all these applications a problem of speaker verification or identification may occur. The term verification refers to a decision-making process leading to an acceptance or rejection of claimed identity of a speaker (Rosenberg, 1976). Similarly, the term identification – in most common understanding (Rosenberg, 1976) – refers to a process that assigns a speech sample of an unknown speaker to one speaker from a given set of speakers. This classical approach to speaker identification makes sense only in case of closed sets of speakers, i.e. such sets in which it is a priori assumed that the unknown speaker belongs to the given set. Such presumption is, however, an oversimplification of the speaker identification problem. If the set of speakers is open, the unknown speaker may not necessarily belong to the set of known speakers, and the identification task has to be preceded by a corroboration that the unknown speaker does in fact belong to the set of known speakers. Thus, the question arises how the recognition system should react and what decision criteria should be adopted to find out whether the speech pattern of an unknown speaker belongs to the given set of M speakers or is a representation of a new, M + 1 class of voices. The present study, which is part of a substantial series of investigations focused on speaker identification (e.g. Majewski et al. 1981; Basztura and Majewski, 1981), is an attempt to answer these important questions.

2. Recognition System — Preliminary Assumptions

A model of a recognition system consisting of two basic operational blocks, for measurement and for classification, was adopted for this study. It is a simple system with programmed learning and statistical decision criteria (Skripkin and Gorielik, 1977) Reference patterns W_m are developed by means of learning sequences representing I_m repetitions of speech samples produced by each speaker belonging to the set of M classes. The assumption

of a system with programmed learning fits the majority of known, practically applied recognition systems (Rosenberg, 1976; Atal, 1976) and decision criteria based on statistics (Skripkin and Gorielik, 1977).

The main goal of speaker recognition procedure is to assure a minimum of incorrect decisions, which may be classified as α and β errors. The first kind of error is called a *false alarm* and it defines the chance of incorrect rejection. β errors, or *misses*, define the probability of incorrect identifications. It should be emphasised that the efficiency of a recognition system, i.e. minimization of both α and β errors, depends on the selection of effective parameters representing the individual voice features in suitably chosen utterances (Wolf, 1972). The expected value of correct decisions is also a function of population size M (Rosenberg, 1976) and representativeness of learning sequence. The error rates depend also on the classification stage, which is the main concern of this study. Thus, the present discussion concerns a fixed measuring procedure and is aimed at a classification procedure that permits avoiding an incidental increase of β errors in cases of open sets of test speech samples.

The classical approach to speaker identification does not assure such features of classification, since the pattern X_n of unknown speaker is assigned to the particular class m* from the set of M classes, if the adopted similarity measure between this pattern and reference patterns of all classes reaches a maximum for the class m*.

In terms of distance measures:

$$\left[\begin{array}{l} X_n \in m^{*} \ \text{if} \ \underset{m=1}{\overset{M}{\text{Min}}}\left[d(X_n,W_m)\right] = d(X_n,W_{m^{*}}) \\ \text{where } m = 1,2,\ldots M \quad M \text{ - number of classes} \end{array}\right.$$

It is clear that the application of the classical approach to open sets of speakers also assigns the pattern of unknown voice to some class from the set M in case where the test sample does not belong to any of the known classes.

One possible conception that permits avoiding the stated inconvenience of the classical approach seems to be the idea of reducing the problem of speaker recognition in open sets to the known and successfully solved problem of speaker verification. A speaker recognition system is presented in the next section of the paper based on this idea.

3. Proposed Model of the Recogniton System

The proposed system is based on a dichotomous classification (Cypkin, 1973) with the utilization of statistical decision theory, in one case followed by the verification task.

The operation of the system, which enables us to reduce the problem of speaker recognition in open sets to the problem of speaker verification, may be presented in the following points.

A. Learning cycle

- For a number of speakers M, and learning sequenc I_m, establish the kind of parameters extracted from the speech signal, and resulting from them patterns X of individual voice features, the probability P_m of appearances of patterns from a given class, and adopt a particular similarity measure between speech patterns as well as preliminary assumptions related to expected values of α and β errors, which should be facilitated by the introductory experiments.
- Derive estimators of the probability distributions $p_m(X)$ for each of all the known M classes of voices.
- On the basis of assumed loss functions c and probabilities P_m derive threshold values h for each pair of classes from the set M.

$$\left[h_{m,m+1} = \frac{c_{m+1,m} - c_{m+1,m+1} \quad P_{m+1}}{c_{m,m+1} - c_{m,m} \quad P_m} \right.$$

- For each class from the set M derive the reference pattern W_m and decision threshold H_m as in the case of a simple discrimination task.

B. Classification cycle

- Read in the parameters of a pattern X_n from an unknown speaker.
- For an arbitrary chosen pair of classes compute the likelihood ratio

$$\left[l_{m,m+1}\left(X_n\right) = \frac{p_m \ X_n}{p_{m+1} \ X_n} \right.$$

and compare it with the corresponding threshold value derived in the learning cycle. If $l_{m,m+1}$ (Xn) > $h_{m,m+1}$, the pattern X_n from the unknown speaker is 'closer' to the class m than to class m + 1, which is rejected. In the opposite case the pattern X_n is closer to the class m + 1 and class m is rejected.
- Replace the rejected class with the values $P_{m+2}(X_n)$ from the next class and compute a subsequent likelihood ratio and compare it with the corresponding threshold value. In the course of checking all known M classes the index m* of the class closest to the pattern X_n is established.
- Carry out the verification procedure, i.e. compare the value of the similarity measure between the test pattern X_n and reference pattern W_{m*} with the decision threshold H_m* derived in the learning cycle. Thus:

$$\left| \begin{array}{l} X_n \in m^* \quad \text{if} \quad d\left(X_n, W_m *\right) < H_m * \\ X_n \notin m^* \quad \text{if} \quad d\left(X_n, W_m *\right) \geq H_m * \end{array} \right.$$

In the first case, the pattern X_n with the assumed probability of error is assigned to the class m*, in the second – the pattern X_n does not belong to any class from the set M and may be stored as a pattern generating a new, M + 1 class of voices.

4. Concluding Remarks

In comparison to the classical approach to speaker identification the presented procedure is more complicated and requires more computations. However, considering the actual state of the art in computer hardware and software, this should not be a serious obstacle in practical applications of this procedure. On the other hand, the proposed system is more universal and it enables realization of a speaker identification procedure both in closed and open sets of speakers, as well as realization of a speaker verification procedure. The application of the present system permits performing the speaker recognition task with an established, for any given input data, probability of correct decision, which is impossible to achieve when a classical decision rule of speaker identification is applied to open sets of speakers.

References

Atal, B.S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, **64**, 460-475.

Basztura, C. and Majewski W. (1981). The effect of chosen parameters of a telephone channel on voice identification. *Archives of Acoustics*, **6**, 359-370.

Cypkin, J.Z. (1973). *Podstawy teorii ukladów uczacych sie.* Warszawa: Wydawnictwa Naukowo-Techniczne.

Majewski, W., Tosi, O. and Jurkiewicz, J. (1981). Speaker identification over telephone lines with unknown frequency response. *Proceedings of the Fourth FASE Symposium on Acoustics and Speech, Venezia*, **I**, 313-316.

Rosenberg A.E. (1976). Automatic speaker verification: a revue. *Proceedings of the IEEE*, **64**, 475-487.

Skripkin, W.A. and Gorielik, A.L. (1977). *Mietody rozpoznavanija.* Moskva: Visshaia Shkola.

Wolf, J.J. (1972). Efficient acoustic parameters for speaker recognition. *Journal of the Acoustical Society of America*, **51**, 2044-2056.

The Machine as an Addressee: When Paralinguistics Fails

M. Ohala
San Jose, USA

1. Introduction

Speech normally encodes both linguistic and non-linguistic information. The latter conveys, among other things, the following:

a. The speaker's state, e.g., relaxed, angry, interested, bored, nervous, sad, etc.,
b. The speaker's attitude towards the addressee, e.g., formal, informal, cooperative, aggressive, condescending, etc.,
c. The speaker's attitude toward the content or referent of the message, or, more generally, the way the speaker is reacting to the informational context in which the speech is uttered, especially whether it represents knowledge shared with the addressee or not.

Although such 'paralinguistic' signals are always present in speech, and humans have learned to adapt to them, there is one very new use of speech where such non-linguistic features impair communication, namely when the listener is a machine.

Machines which will recognize spoken commands are now commercially available and are being used in a number of applications: mail sorting, assembly lines, control of wheelchairs by quadraplegics, etc. These automatic speech recognition (ASR) devices are generally speaker-dependent and work on words spoken in isolation. To use one, a given speaker must train the device by first giving it samples of his/her pronunciation of all the words it will be required to recognize (typically ≤ 200). The acoustic pattern of these words (called 'templates') are stored in the device's memory; the incoming 'unknown' word is then compared with all the stored templates and is identified as the word corresponding to the template which produced the closest match. (If no template produces a sufficiently close match the device may prompt the speaker to repeat the word.) Once a speaker gets used to the device, accuracy rates of 97% or better are not uncommon. But the problem is that some users take a long time to 'get used' to such devices (2 to 6 weeks) due to a high degree of initial variability in pronunciation (as Doddington and Schalk (1981) remark: 'Speech recognizers commercially available today are effective only within narrow limits. They have relatively small vocabularies and frequently confuse words. Users must develop the skill to talk to the

recognizer, and the machine's performance often varies widely from speaker to speaker'. p. 26). There would be considerable practical advantage if the source(s) of within-speaker variation could be identified so that strategies to control it/them could be developed. A number of determinants of within-speaker variation in language use have already been identified and described in some detail. Conceivably, at least some of the causes of speaker variability which plagues ASR are one or more of these previously-identified factors.

Among the things that need to be considered in trying to find the causes of speaker variability are the following:

Does the speaker cast himself/herself in some well-defined social role vis-a-vis the ASR device, e.g., as a superior, as an adult speaking to a child, an owner to a dog, an English teacher to a foreigner? Does the perception of role persist or, worse, does it change depending on the type of feedback (or lack of it) received from the device? Very possibly the user finds him/herself in a totally new 'social' situation and discovers that the old and familiar sociolinguistic roles do not apply. It does not pay, however, to be alternately helpful, exasperated, condescending, etc., to an ASR device. To the extent that the speaker 'tries out' various socially-dictated modes of speaking, the ASR device is more likely to fail.

A more subtle source of variation, mentioned in (c), is the speaker's presumption of shared knowledge with the addressee. To oversimplify, whatever the speaker thinks the listener knows, or *should* know, can be weakly articulated. Conversely, whatever represents new information must be pronounced carefully, or at the speaker's option emphatically. Both weakly-articulated and emphatic pronunciation may differ from context-neutral pronunciation, thus creating problems for an ASR device.

2. Experiment

I hypothesized that a significant part of within-speaker variation stems from a speaker using the familiar emotional and attitudinal qualifiers which, though appropriate when communicating with other humans, are inappropriate when speaking to machines. Thus, the more 'emotional' the speaker becomes the more he will vary the way he speaks in order to express that emotion, and such pronunciation variability will lead to degradation of ASR performance.

3. Experimental design

To test this I observed and recorded 20 subjects' interaction with an ASR device under circumstances where their emotional arousal could be controlled. Subjects were randomly assigned to 4 groups of 5 each in an experimental design whereby two binary factors were varied independently, high vs. low subject involvement in (or anxiety towards) the task, and high vs. low confusability of the vocabulary used (such that error rate would be high vs.

low). Involvement in the task varied by paying half the subjects by the number of words recognized correctly and paying the other half by the hour. The vocabularies, one with many phonetically similar words and one with highly dissimilar words provided variation in the inherent difficulty of the recognition task. The assignment of the 4 groups according to these experimental variables is shown in Table I. It was hypothesized that subjects who were paid by the word and assigned the confusable vocabulary would have the highest error rate.

Table I. Experimental design.

Variable 1 Subject interest	No involvement (paid by the hour)	High involvement (paid by the word)
Variable 2 Complexity of voc.	Group I	Group II
Low error rate (distinct voc.)		
High error rate (confusable voc.)	Group III	Group IV

With one exception all subjects were students at the University of California, Berkeley, from various disciplines. Each subject first trained the ASR device by pronouncing a single time each of the 30 words of the vocabulary assigned to him/her. These samples constituted the stored templates. Then, with the computer prompting them via printed words (randomized and in blocks of 30) on the CRT of the terminal, they repeated the words for a total of 450 trials in one session and 450 trials in a second session on another day. In all, thus, each subject had 900 trials. During the recognition session there was an interval of 2 seconds between responses and the next prompt. After 90 trials subjects were given short breaks. All sessions were audiotape recorded for later acoustic analysis. A record of correct/incorrect recognition was automatically maintained by the computer. There was a constant threshold for 'rejection', i.e., when a noise or utterance was judged to be so dissimilar as to be unlike any of the stored templates. Rejections were not counted as errors. (For details on the algorithm used in the recognition see Murvelt, Lowy and Brodersen (1981).

4. Results

Figure 1 shows the average error rate by blocks of 750 words for all 4 groups. As would be expected subjects using the distinct vocabulary (Groups I and II) made fewer errors than those with the confusable one (Groups III and IV). Counter to expectation, those paid by the word (Groups II and IV) did not have more errors than those paid by the hour, however the effect of subject's anxiety on performance is reflected in another interesting way. Groups III and IV (in contrast to the other two groups) give evidence of experiencing an *increasing* error rate on successive blocks of trials. This is unusual because speakers normally adapt to the constraints of these tasks, i.e., manage to lower the error rate. Evidently errors beyond a certain level trigger an emotional reaction in speakers which in turn causes them to express this emotion in their speech thus leading to more errors.

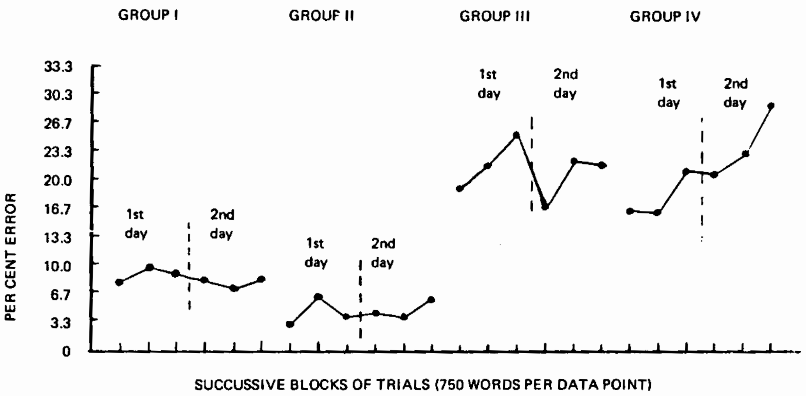


Figure 1. Percent error for the 4 experimental groups plotted as a function of successive blocks of 750 words (see text for further details).

5. Conclusions

The results of this study support the hypothesis that one source of error in ASR is the variation in speakers' pronunciation which encode their changing emotional state. Further studies are underway to identify the precise acoustic features which manifest these emotions.

Acknowledgments

This research was funded by an Affirmative Action Faculty Development Program grant awarded by the Chancellor's Office of the California State University system. The research was conducted using the experimental ASR facility at the Electrical Engineering and Computer Science Department at the University of California, Berkeley. I am extremely grateful to R.W. Brodersen for making this facility available to me and to M. Murveit for his time and guidance in the use of the ASR device and also for writing the computer program. I am also grateful to Margit Peet for her help in conducting the sessions with the subjects and collecting the data, and to John Ohala for comments on the experimental design.

The vocabulary was designed by John Ohala. The distinct vocabulary included words such as *lunch, claud, point, lake* and the confusable vocabulary included words such as *bad, dad, bead, deed.*

References

Doddington, G.R. and Schalk, T.B. (1981). Speech recognition: turning theory to practice. *IEEE Spectrum* **18 (9)**: 26-32.

Murveit, H., Lowy, M., and Brodersen, R.W. (1981). An architecture of an MOS-LSI speech recognition system using dynamic programming. *JASA* **69**: 42.

Extensions to the Locus Theory

M. O'Kane
Canberra, Australia

1. Introduction

The Locus Theory originally put forward by Delattre, Liberman and Cooper (1955) proved useful in explaining formant transitions between vowels and consonants. According to this theory, which was developed from experiments with synthetic speech, the formant transitions from the steady state of a vowel to a neighbouring plosive consonant point to a particular frequency or locus depending on the place of articulation of the consonant. Öhman (1966) showed that in real speech the situation was somewhat more complex. He showed that for /b/ and /d/ the position to which the transitions pointed was dependent not only on the place of articulation of the consonant but also on the second formant of the vowel preceding the consonant. Thus, the second formant /b/ focus could be anywhere in the range 500-1400 Hz; being low if the second formant of the preceding vowel was low and high if the second formant vowel was high. A similar situation occurred for /d/ with the locus occurring in the range 1400-1700 Hz. For /g/ the formants pointed to a locus which was determined by the vowel following the /g/ if the vowel preceding the /g/ was a front or central vowel, and to a low locus if the vowel preceding the consonant was a back vowel.

We decided to use formant locus as one of the components in an algorithm to automatically differentiate between plosive consonants occurring in a VCV context in continuous speech. We found, however, that the Locus Theory even with the Öhman modifications was somewhat inadequate for our purposes. Unanswered questions included the following:

How did the Locus Theory apply to Australian English, particularly continuous, conversational Australian English?

Were locus ranges the same for all speakers? In particular were male and female locus range the same?

Did the presence of a word boundary within a VCV combination affect the locus range? VCV combinations in English occur more frequently across word boundaries than within words.

The answers to these questions were sought through an experiment which is described below.

2. Method

The experiment was actually designed to investigate coarticulation, juncture, plosive consonant phenomena and the interaction between all three continuous speech examples of Australian English. Only the results pertaining to the Locus Theory are presented here; results concerning other aspects of the experiment are given in O'Kane (1981).

Lists of two-word sequences were prepared. Each of these two-word sequences was of either of two forms:

1. The first word ended in a VC combination and the second word began with a V, where the vowels could be either /i/ or /ɔ/ (the foremost and backmost vowels in Australian English), and the consonant was one of the six plosives e.g. 'heat ought'.
2. The first word ended in V and the second word began with a CV combination where the vowels could be either /i/ or /ɔ/ and the consonant was one of the six plosives e.g. 'he taught'.

Thus, with the two vowels and six plosive consonants and two juncture positions there were a total of forty-eight two-word combinations. The five male and five female speakers all spoke general Australian English. They were from a geographically diverse area of Eastern Australia and ranged in age from twenty-four to fifty. It should perhaps be pointed out that Australian English is remarkably homogeneous geographically as regards pronunciation, the greatest variation occurring between people from different social and employment groups (Mitchell and Delbridge, 1965).

Each speaker was presented with the list of two-word sequences and instructed not to study the list but to immediately begin saying sentences containing the word sequences. It was impressed on the subjects that the sentences they produced were to be spoken at a conversational speed and that the semantic content of the sentences was not of particular importance. It was even suggested that slightly silly sentences would be perfectly acceptable. If a subject was having particular difficulty in producing a sentence containing any given two-word sequence he was told not to waste time over it but to say something such as 'I can't think of a sentence with ——— in it'. All this was to keep the subject speaking at as conversational a rate as possible. This aim was largely achieved.

Experimental studies on conversational speech are difficult as it is hard to exercise any control over the many interacting parameters and still obtain the characteristic spontaneity of conversational speech. In the experiment described here phonetic and junctural contexts were controlled and almost always the speech of the subjects sounded (in the author's subjective opinion) to be at conversational rate. It is perhaps indicative of the relative newness of continuous speech recognition research that paradigms such as the one used in the experiment outlined above generally have not been systematically developed as a means of controlled study of conversational speech phenomena.

The sentences containing the two-word sequences were recorded on a Nakamichi 550 cassette recorder, using a Bayer microphone. The required VCV tokens were excised from the sentences using a waveform editing routine. These VCV combinations were then analysed using the Interactive Laboratory System waveform analysis package. For the male voices the speech was sampled at 10 kHz and for the female voices the sampling rate was 16 kHz. For each token a linear prediction analysis was done using the autocorrelation technique. Plots of the spectral peaks derived from this analysis as a function of time provided a means of measuring the formants.

3. Results

The data examined in the experiment described above revealed the following results (which can be seen in Figures 1 and 2).

1. That in continuous speech (at least for Australian English) the position of the second formant locus is primarily determined for *all* consonants by the vowel preceding the consonant. *This is more general than Öhman's result.*
2. That the position of the F_2 /b/ locus can range for male voices from 500 Hz for the case when the preceding vowel is /ɔ/ to 1300 Hz for the case where the preceding vowel is /i/.
3. That the position of the F_2 /d/ locus ranges for male voices from 1350 Hz for the case where the preceding vowel is /ɔ/ to 1900 Hz when the preceding vowel is /i/.
4. That there are high and low loci for /g/. If the preceding vowel is /i/ the /g/ locus for male voices is in the range 2000-2300 Hz. If the preceding vowel is /ɔ/ the /g/ locus for male voices is in the range 800-1500 Hz.
5. That, at least for the case where the consonant is /g/, the exact position of the locus is also determined by the vowel following the consonant. Thus if the preceding vowel is a back vowel and the following vowel is a front vowel then the locus position will be at the high end of the low /g/-locus range. But if the preceding vowel is a front vowel and the following vowel is a back vowel then the locus position will be at the low end of the high /g/-locus range.
6. That the higher the locus position is in frequency, the greater is the difference between male and female locus values for a given VCV production. The average locus range for various consonants are given in Table I for both male and female voices.

Table I. Locus ranges of plosive consonants at the three places of articulation for male and female voices. All measurements in Hz

	Labial locus range	Alveolar locus range	Low velar locus range	High velar locus range
Male voices	500-1300	1350-1900	800-1500	2000-2300
Female voices	500-1500	1550-2400	800-1800	2400-2900

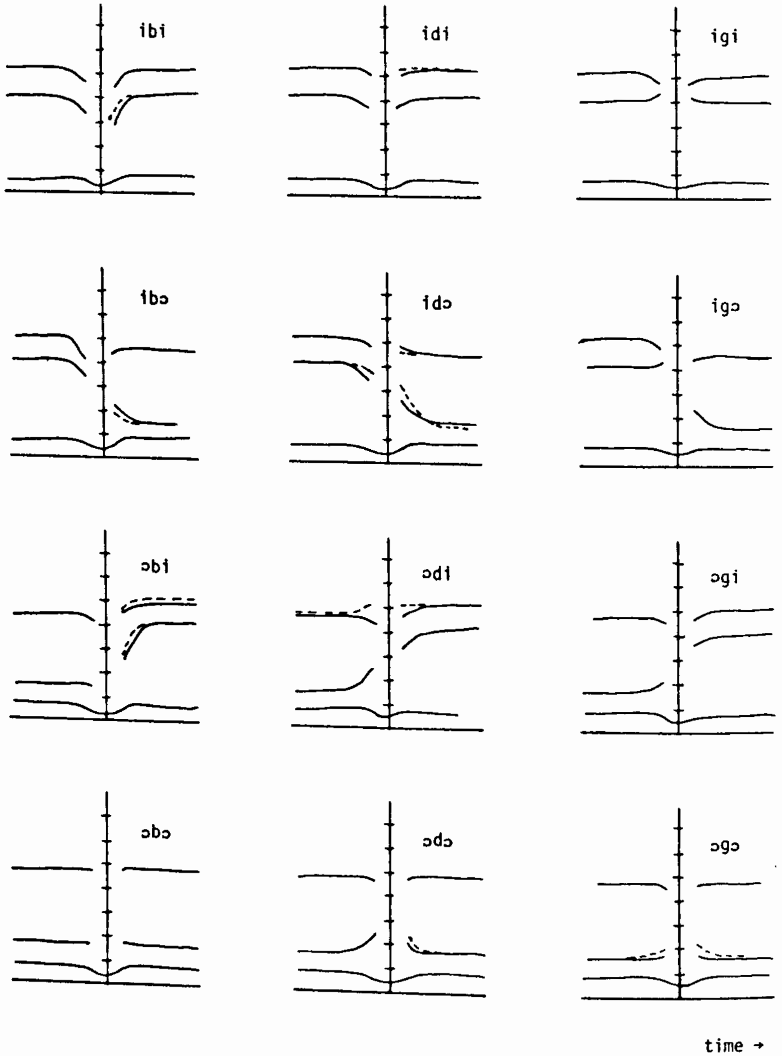


Figure 1. Average, first, second and third formants for male voices VCV sequences in which C is a lax plosive. Where dotted lines are shown they represent the /VCV/ case in contradistinction to the /VCV/ case (full lines). Where only a full line is shown both cases essentially coincide. Frequency marks are 0.5 kHz apart. Time is not strictly to scale but is of the order of 450 msec, in total.

7. That the difference between male and female locus positions are reflections of the differences between formant positions for vowels for male and female voices. This can be seen by reference to Figure 3. Figure 3 shows the average first and second formants of several vowels in Australian English for male and female speakers based on data from Oasa (1980) and O'Kane (1981). The differences in male and female voices is slight for the transitions from /ɔ/ to the /d/ but are considerable for the transition from /d/ to /i/.

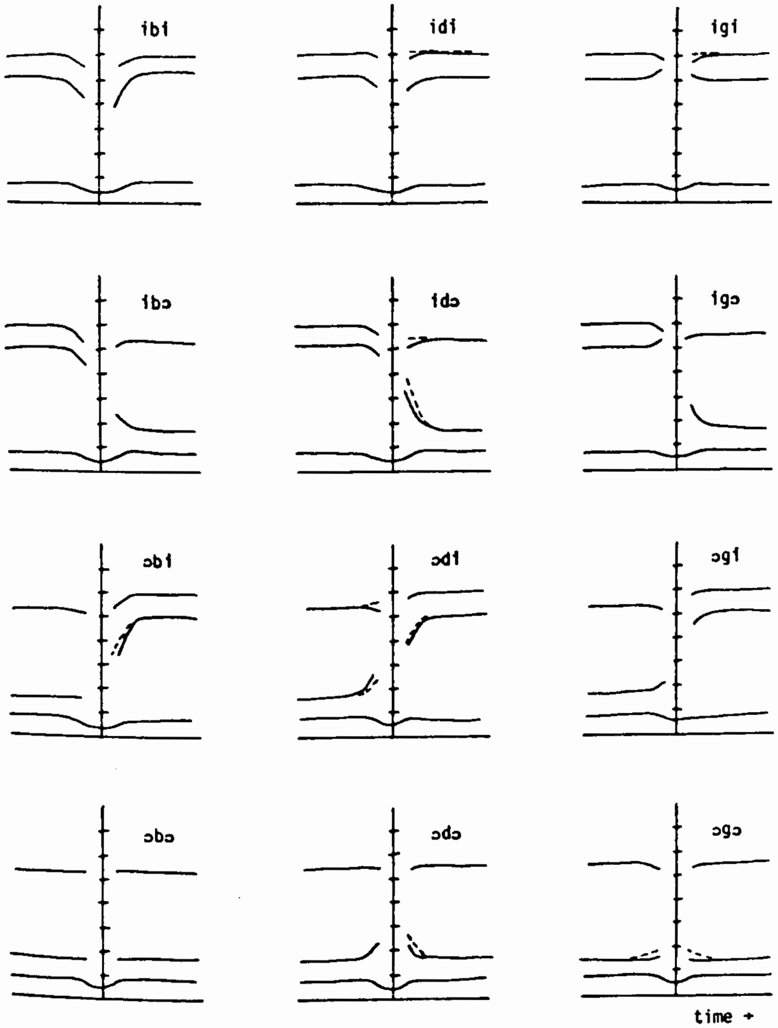


Figure 2. Average, first, second and third formants for female voices for VCV sequences in which C is a lax plosive. Where dotted lines are shown they represent the /VCV/ case in contradistinction to the /VCV/ case (full lines). Where only a full line is shown both cases essentially coincide. Frequency marks are 0.5 kHz apart. Time is not strictly to scale but is of the order of 450 msec, in total.

8. Inter-speaker differences connected with locus effects within each sex grouping are slight particularly for the labial and velar consonants. In the production of alveolar consonants, however, some speakers produce barely noticeable formant transitions which makes the finding of the locus position difficult.

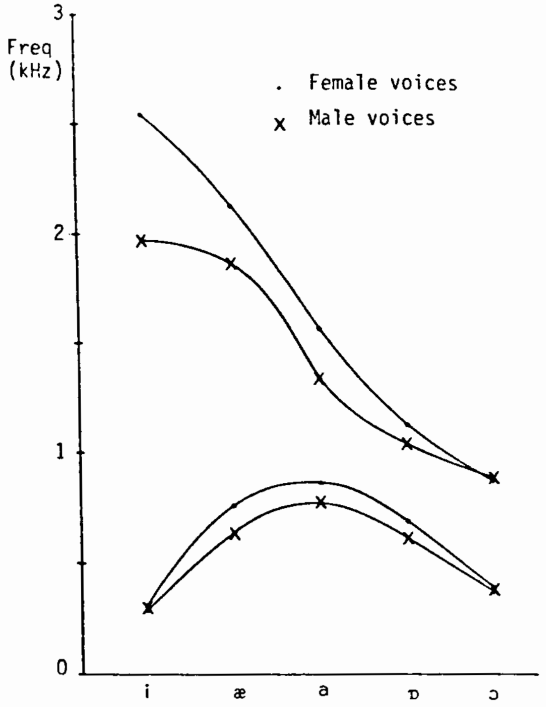


Figure 3. Average first and second formants of the vowels /i/, /æ/, /a/, /ɒ/, /ɔ/ for male and female speakers of Australian English.

9. That differences in the locus positions due to the word-final or word-initial nature of the consonant are only noticeable in a few cases. If an effect is present at all it is often only noticeable in the CV transition. As can be seen in Figures 1 and 2 the most noticeable cases in which juncture effects are present are /ɔbi/, /idɔ/, /ɔdɔ/ and ɔgɔ/.

4. Conclusion

An expanded version of the Locus Theory for plosive consonant transitions has been presented. The results given here can be incorporated as a partial cue for the automatic recognition of plosive consonants. This cue is particularly helpful in cases (5% in the experiment described here) where no burst is produced in the articulation of the consonant. It should also be noted (see Table 1) that the locus cue for plosive recognition is slightly more useful for female voices than for male voices.

References

Delattre P.C., Liberman, A.M. and Cooper, F.S., (1955). Acoustic loci and transitional cues for consonants. *J. Acoust. Soc. Amer.*, **27**, 769-773.

Mitchel, A.G. and Delbridge, A. (1965). *The Speech of Australian Adolescents: A Survey.* Angus and Robertson, Sydney.

Oasa, H. (1980). Regional Variations in Australian English. M.A. Thesis, Australian National University, Canberra.

Öhman, S.E.G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *J. Acoust. Soc Amer.*, **39**, 151-168.

O'Kane, M. (1981). Acoustic-Phonetic Processing for Continuous Speech Recognition. Ph.D. Thesis, Australian National University, Canberra.

Syllable-based Analysis of Spectral and Temporal Features for Automatic Speech Recognition

G. Ruske
Munich, FRG

1. Introduction

In systems for automatic recognition of continuous speech an initial segmentation is required in order to obtain appropriate segments which can be used as basic units for the subsequent analysis and classification procedures. The syllable structure of the speech signal provides segments which include the essential coarticulation effects. A substantial reduction in the number of different syllables is achieved by a further subdivision of each syllable segment into two parts: an initial demisyllable ranging from the starting point of the syllable to the syllable nucleus, and a final demisyllable ranging from the nucleus to the end point of this syllable. In the German language we only have to discriminate about 50 initial consonant clusters, 19 vowels (short and long vowels, and 3 diphthongs). and there is a maximum of about 160 final consonant clusters. The efficiency of syllabic segmentation has been demonstrated in previous experiments (Ruske and Schotola, 1981).

2. Demisyllable segmentation

A syllable is defined here as an 'acoustic syllable'. According to this definition the localization of syllable nuclei as well as the determination of a suitable syllable boundary between two consecutive syllable nuclei can be based on an evaluation of the loudness contour and on spectral information from the speech signal. In our experiments we examined several methods for syllabic segmentation of spoken utterances:

a. The maxima of a smoothed loudness function have proved to be suitable candidates for syllable nuclei. An additional vowel/non-vowel classification of the spectrum at the maximum rejects maxima produced by high energy consonants.

b. A demisyllable boundary is indicated by a loudness minimum.

c. A crude vowel classification is performed at intervals of 10 ms using several vowel spectra as prototypes. High estimates of this classifier indicate syllable nuclei.

d. A set of demisyllable templates is applied consisting of spectral-temporal patterns and which are representative for all possible demisyllable shapes. The comparison is performed by dynamic programming methods (2-level DP matching). This method provides syllable nuclei as well as the syllable boundaries.

Methods a. and b. were used for the following experiments, since they yield reliable segmentation results and in addition save computation time.

3. Recognition of demisyllables

As a first approach, recognition of German demisyllables was carried out using spectral-temporal templates of complete consonant clusters and spectral templates of vowels. Time normalization was performed by a so-called 'dynamic interpolation' procedure. After normalization a city-block metric was applied for the calculation of similarity. However, template matching needs a lot of storage and computation time since an unknown consonant cluster has to be compared with all reference templates regardless of their phonetic structure.

Since knowledge about the gross phonetic structure could considerably reduce the number of templates to be actually compared a second method was developed starting from a description of the relevant acoustic events within each segment by evaluating spectral and temporal features or 'cues' which can be objectively measured in the signal. These features have been defined in some analogy to the classical perception experiments with synthesized speech sounds which have been reported in the literature (Delattre, 1968). The cues describe: the 'loci' of the first 3 formants of the syllable vowel, the formant transitions, formant-like 'links' for nasals and liquids, duration and spectral distribution of bursts and turbulences, pauses, and voicing during pauses or turbulences.

A main problem is the depency of most of the features on phonetic context. In the present paper the context dependencies are taken into consideration by collating the results of feature extraction within each demisyllable segment. This enables the contextual dependencies between the acoustic features to be determined statistically from representative speech material. The feature vector for an initial consonant cluster as well as for a final demisyllable have a fixed number of components. In syllable-initial position 1 nasal, liquid or glide, and up to 2 fricatives or plosives are possible; in syllable-final position the maximum number of plosives or fricatives can be limited to 3. Therefore, initial consonant clusters are completely described by 24 feature components and final consonant clusters by 31 components, see Table Ia and Table Ib.

4. Feature extraction method

Feature extraction starts from a spectral representation calculated by linear prediction analysis. From the LPC-coefficients power spectra are computed with a frequency resolution of about 78 Hz. Depending on a voicing parameter each spectral frame is labelled voiced, unvoiced, or silent, respectively.

Table 1a. Feature vector for initial consonant clusters

Component number	Feature	Dimension
1	First turbulence (or burst)	Yes/No
2	Center of gravity	Hz
3	Lower cut-off frequency	Hz
4	Upper cut-off frequency	Hz
5	Duration	ms
6	First pause	Yes/No
7	Duration of pause	ms
8	Second turbulence (or burst)	Yes/No
9	Center of gravity	Hz
10	Lower cut-off frequency	Hz
11	Upper cut-off frequency	Hz
12	Duration	ms
13	Voice-bar	Yes/No
14	Duration of voice-bar	ms
15	Nasal or liquid links	Yes/No
16	Low link	Hz
17	High link	Hz
18	Transition of links	Hz/ms
19	Transition of the formant F_1	Hz/ms
20	Transition of the formant F_2	Hz/ms
21	Transition of the formant F_3	Hz/ms
22	Locus of the formant F_1	Hz
23	Locus of the formant F_2	Hz
24	Locus of the formant F_3	Hz

Finally, formant tracking is performed within voiced parts. Acoustic parameters are then derived from the energy in selected frequency bands which allow a gross characterization of the spectral shape; the feature extraction procedures are in some accordance with the methods proposed by Weinstein et al. (1975). Based on these parameters, a set of rules has been established in order to detect voiced and unvoiced turbulences and bursts, pauses, and liquid and nasal links. After detection the individual features are characterized by gross measurements of their spectral and temporal distribution (e.g. center of gravity, upper and lower cut-off frequency for turbulences, and spectral peaks for links) and used as components of the common demisyllable feature vector.

5. Classification procedure

Classification of the feature vectors is based on Euclidean distance measurements within the feature space. All components are normalized to equal variance. However, when comparing the unknown vector with a reference vector, the single components are handled in different ways. For all binary components which indicate the presence or absence of a single feature, the distance is calculated in each case. The distances of the remaining components describing the temporal or spectral characteristics of a certain feature are only taken into account in those cases where the corresponding acoustic feature is present in both vectors. Finally the accumulated distance is divided by the number of all feature components evaluated during the distance measurement. This results in the calculation of an 'average normalized distance' which allows the comparison of all feature vectors even if they differ quite considerably as to their current composition.

Table 1b. Feature vector for final consonant clusters

Component number	Feature	Dimension
1	First turbulence (or burst)	Yes/No
2	Center of gravity	Hz
3	Lower cut-off frequency	Hz
4	Upper cut-off frequency	Hz
5	Duration	ms
6	First pause	Yes/No
7	Duration of pause	ms
8	Second turbulence (or burst)	Yes/No
9	Center of gravity	Hz
10	Lower cut-off frequency	Hz
11	Upper cut-off frequency	Hz
12	Duration	ms
13	Second pause	Yes/No
14	Duration of pause	ms
15	Third turbulence (or burst)	Yes/No
16	Center of gravity	Hz
17	Lower cut-off frequency	Hz
18	Upper cut-off frequency	Hz
19	Duration	ms
20	Third pause	Yes/No
21	Duration of pause	ms
22	Nasal or liquid links	Yes/No
23	Low link	Hz
24	High link	Hz
25	Transition of links	Hz/ms
26	Transition of the formant F_1	Hz/ms
27	Transition of the formant F_2	Hz/ms
28	Transition of the formant F_3	Hz/ms
29	Locus of the formant F_1	Hz
30	Locus of the formant F_2	Hz
31	Locus of the formant F_3	Hz

6. Recognition experiments

The speech test material consisted of several hundred initial and final demisyllables which were automatically extracted from German words spoken by one male speaker. The set of demisyllables contained 45 initial consonant clusters and 48 important final consonant clusters, all combined with 8 vowels. Syllabic segmentation, feature extraction and classification was applied to this material. As expected some difficulties arise in the detection of the glide /r/ and the liquid /l/. The consonant /r/ often cannot be discriminated from the vowel, whereas in the case of /l/ often parts of the vowel /o/ or /u/ were indicated erroneously as liquid links as e.g. in /ʃu:/.

Some of the confusions observed in the feature classification experiments can be explained by inspection of the mean values and standard deviations. The data display the typical order of the F_2-loci for the plosives: low for the labial /p/, mid for the dental /t/ and high for the velar /k/. However, the standard deviations are rather large so that the corresponding confusions are to be expected. A special problem is the discrimination between /m/ and /n/ which were often confused. On the other hand, the calculation of the gradient of formant-to-link transitions enabled nasals to be discriminated from liquids. From the recognized consonant clusters the recognition scores of the single consonants were computed; the average recognition rate for the single consonants was about 62% for initial and 68% for final consonants.

For comparison, the same speech material was processed by *template matching* methods using complete spectral-temporal templates for each consonant cluster. Here, the average recognition score was about 4-7% better and on average amounted 66% for initial and 75% for final consonants. Again confusions occurred between the unvoiced and voiced plosives, and between the nasals and /l/ and /v/. The fricatives received the best recognition scores. It is worth noticing that, roughly speaking, the distribution of confusions obtained by template matching is very similar to or even identical to that obtained by feature extraction. While the feature extraction approach could not yet reach the recognition accuracy of template matching, it has to be borne in mind that the feature vector for a consonant cluster has only 24 or 31 components whereas a corresponding template constructed from a series of consecutive spectra needs on average more than 500 components; this results in about 20 times more storage and computation time. Thus the features components can be seen as an efficient representation of the units. In both experiments the recognition scores were not very high; they have to be seen as pilot experiments. But the main goal of this investigation was only to compare the efficiency of the two methods. Our own previous investigations (with template matching) showed that an 85-90% consonant recognition score can be reached with a large training set. This encourages us to believe that the recognition scores of the feature approach can be considerably improved by further optimizing the feature extraction procedures.

References

Delattre, P.C. (1968). From acoustic cues to distinctive features. *Phonetica* **18**, 198-230.

Ruske, G. and Schotola, T. (1981). The efficiency of demisyllable segmentation in the recognition of spoken words. *IEEE ICASSP* 1981, 971-974.

Weinstein, C.J., et al. (1975). A system for acoustic-phonetic analysis of continuous speech. *IEEE Trans. Vol. ASSP-23*, 54-67.

Section 5
Physiology and Acoustics of Speech Production

Between Formant Space and Articulation Space

L.J. Bonder
Amsterdam, the Netherlands

1. Introduction

One of the problems we meet in the study of speech production models is how we can describe the relation between vocal tract and generated sound. More specifically, for vowellike sounds, the question is: given a vocal tract shape, what are the formant values, and conversely, given a certain set of formant values, which shapes can produce these formants. The latter problem of determining shapes from formants will be referred to as the inverse problem.

In this paper we will sketch the relation between formants and shape of a model of the vowel tract. Finally, we will give an outlook on further research on this issue.

2. The n-tube model

The starting point in the description of the relation between formants and articulation is the modelling of the vocal tract as a lossless n-tube, i.e. a concatenation of n cylindrical tubes of equal length l, but different cross-sectional area (Dunn, 1950). The cross-sectional area of segment i is denoted by S_i (i=1,....,n). Further, we define the so-called k-parameters k_i, by putting $k_i=Si/S_{i+1}$ (i=1,...,n-1).
The propagation of sound through such a tube is described mathematically by the one-dimensional wave equation. The pressure and the volume velocity are considered continuous at the junctions of the segments. If we put together the n-tube model, the one-dimensional wave equation, and the continuity conditions, we get the so-called n-tube formula, a closed form expression relating shape to formants, and vice versa (Bonder, 1983a).

3. Properties of the n-tube model

As we have seen, the model has three important features: continuity of pressure and volume velocity, one-dimensionality, and its lossless nature. Beside these features there are some interesting consequences of the model (cf. Bonder, 1983a).

First, there is the modelling of the vocal tract in a non-continuous way as an n-tube. The consequence of this non-continuity is that, from an acoustic

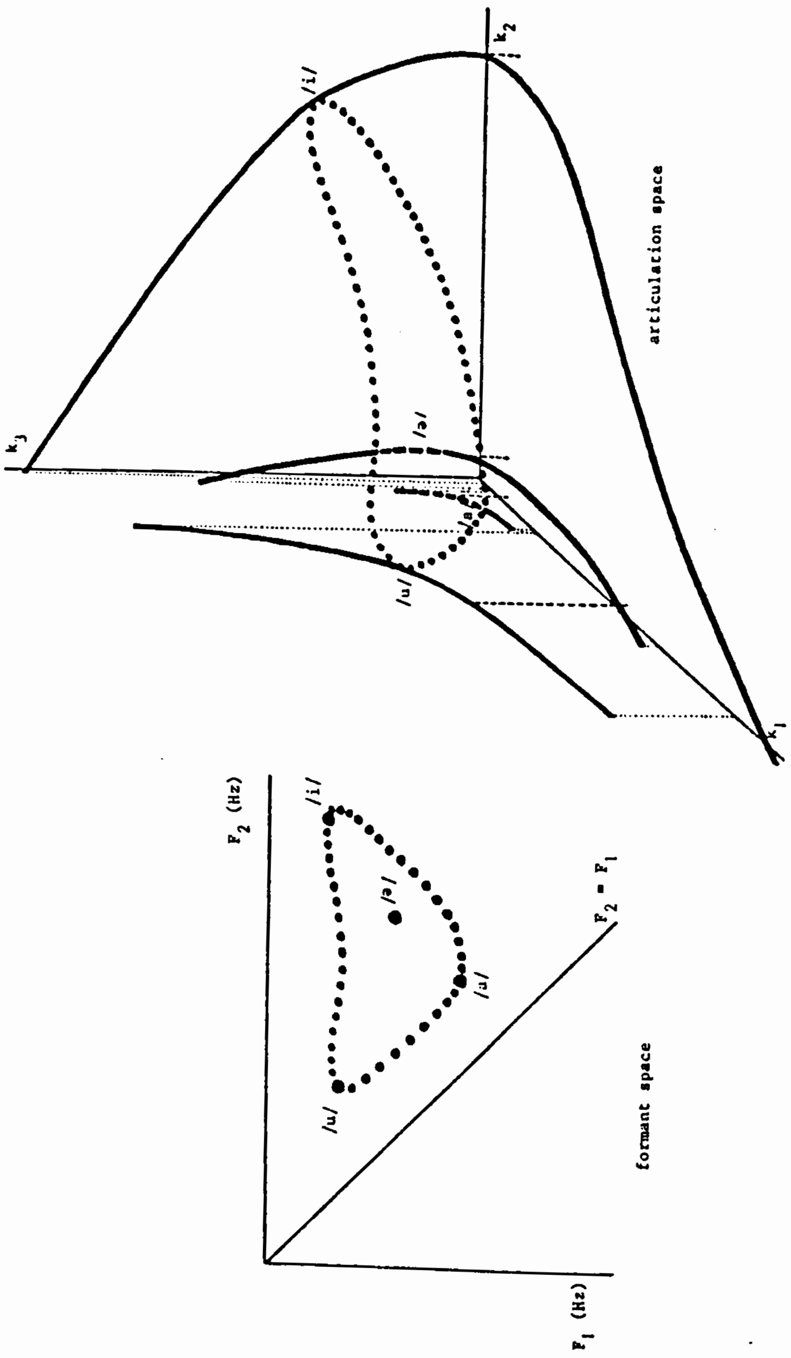


Figure 1. An n-tube model of the vocal tract. The k-parameters are $k_i = S_i/S_{i+1}$ (i=1, . . ., n-1).

point of view, only the first [½n]* formants of an n-tube can be taken seriously; the pattern of the higher formants is merely a repetition of the lower [½n] formants.

Secondly, the formant frequencies of an n-tube will not change if the n cross-sectional areas S_i are multiplied by the same factor. So one of the S_i can be taken as a reference for the other S_i. This means that an n-tube can be fully described by the n-1 parameters $k_i = S_i/S_{i+1}$ (i=1,..., n-1). The description of an n-tube in terms of its n-1 parameters k_i enables us to view such a tube as a point in the (n-1)-dimensional space spanned by these parameters. An n-tube P with k-parameters $k_1, ..., k_{n-1}$ is denoted by $P = (k_1 ..., k_{n-1})$. The space spanned by the k-parameters will be called 'articulation space'. In Fig. 2 the location of the straight 4-tube in the 3-dimensional space of k_1, k_2, k_3 is shown. As all S_i have the same value, it follows that $k_1 = k_2 = k_3 = 1$. So, the straight 4-tube is denoted by (1,1,1).

Thirdly, when calculating n-tube shapes from formant frequencies there are [½(n-1)] degrees of freedom. This means that we can choose freely the values of [½(n-1)] parameters k_i. For example, 4-tubes have one degree of

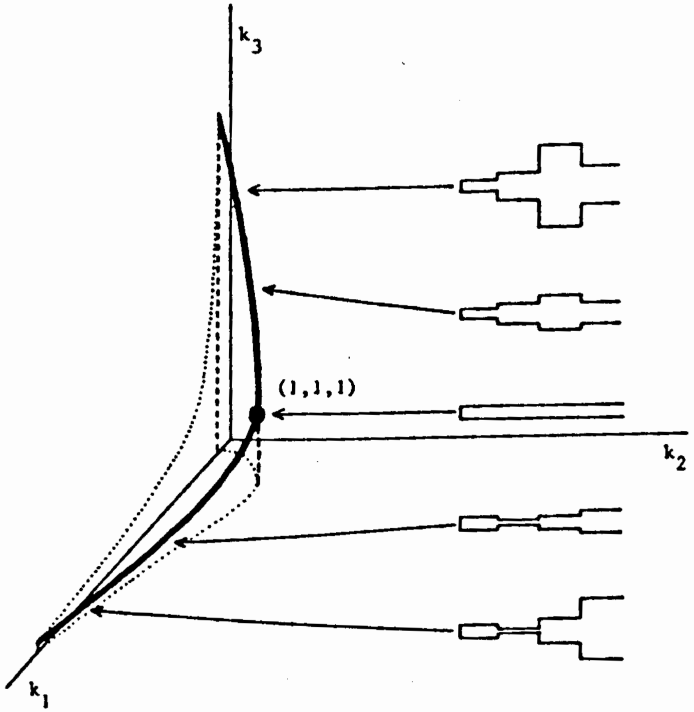


Figure 2. The 3-dimensional articulation space of 4-tubes spanned by the parameters k_1, k_2, k_3, and the location of the straight 4-tube (1,1,1) in it.

* [½n] is defined as the greatest integer number not greater than ½n.

freedom: $[\frac{1}{2}(n-1)]=1$. We observe that the inverse problem does not have a unique solution for n-tubes with more than two segments.

4. The Inverse Problem

The greatest contribution to the research on the inverse problem is from Atal, Chang, Mathews and Tukey (1978). They treated the subject numerically, and showed that there are many vocal tract shapes having the same formant frequencies. A disadvantage of their numerical inversion is that it does not show the structure of the relation between shape and corresponding formant pattern. We will briefly indicate when and how we can handle the inversion analytically (Bonder, 1983b).

The n-tube formula is the starting point for the attack on the inverse problem. The main step in our method of inversion is the decomposition of the n-tube formula, i.e. the replacement of the n-tube formula, which is an equation of degree n, by a set of $\frac{1}{2}n$ equations relating explicitly formants to shape of the tube. By means of these $\frac{1}{2}n$ equations we can solve the inverse problem analytically up to 10-tubes, but, the more segments, the more involved the calculus. For n-tubes consisting of more than 10 segments the inverse problem is no longer analytically solvable, in which case the problem has to solved numerically.

For 4-tubes the inversion is rather simple. The analytical inversion yields the following expressions from which the k-parameters can be determined if the formants F_1 and F_2 are known:

$$\begin{aligned} k_2 &= (-C_2k_1^2+C_1k_1-1) / ((1+k_1)(1+C_2k_1)) \\ k_3 &= 1 / C_2k_1 \end{aligned} \qquad (1)$$

where

$$\begin{aligned} C_1 &= \tan^2\tau F_1 + \tan^2\tau F_2 \\ C_2 &= \tan^2\tau F_1 \cdot \tan^2\tau F_2 \\ \tau &= 2\pi L/4c \end{aligned}$$

c being the velocity of sound, L (=4.1) the overall length of the 4-tube. It is obvious from equations (1) that we have one degree of freedom for 4-tubes: one parameter, k_1, has to be given a value in order to be able to compute the other two parameters k_2 and k_3. Expressions (1) can be used to calculate equivalent 4-tubes, i.e. tubes with the same formant frequencies, to a given length L. Each equivalence class consists of an infinite number of 4-tubes, all of them having the same length L. In the articulation space, an equivalence class turns out to be a continuous trace. In Fig. 3 we show the equivalence class of the straight 4-tube with formants F_1=500 Hz and F_2=1500 Hz. All equivalent tubes have the same overall length L=17.5 cm (which is about the average male vocal tract length). Of course,we cannot show the whole class but this finite subset gives us a good view of it.

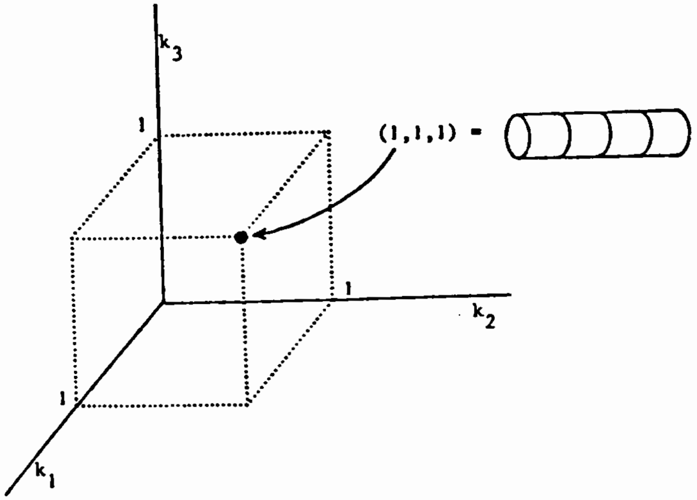


Figure 3. The equivalent class of the straight 4-tube (1,1,1) with F_1=500 Hz and F_2=1500 Hz and length L=17.5 cm (solid curve). At the right hand side the corresponding 4-tube shapes are shown.

5. Inversion Applied to the Vowel Triangle

In the way mentioned above we can calculate a trace in the articulation space for each point in the formant space. All traces in the articulation space look very much the same as the one in Fig. 3. In Fig. 4 we have sketched the traces in the articulation space corresponding to the three ‘vertices‘ of the vowel triangle /u/, /i/, /a/. As we may see from Fig. 4, the structure of the vowel triangle is rather alike in both spaces. From this we conclude that our choice of the k-parameters as parameters of articulation seems to be adequate. The structure of the articulation space of Atal et al. (1978) is more complex. Besides, one of their dimensions is not contained in our space, namely the length L of the tube. From an acoustical point of view, the length L is not essential, as it is no more than a scaling factor in the formant space (for, if all the segment lengths are multiplied by the same factor α, the cross-sectional areas being unchanged, the overall length will change to αL and the corresponding formant frequencies F_i to F_i/α).

6. Outlook on Further Research

After this rough sketch of the relation between formant space and articulation space by means of the n-tube model of the vocal tract we might come to the question if there are preferential areas in the articulation space, and how

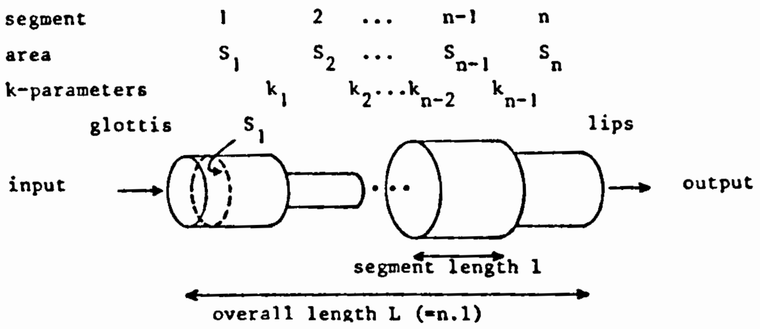


Figure 4. The vowel triangle in both formant space and articulation space. The solid curves in the articulation space are the equivalence classes of the vowels /u/, /i/, /a/, /ə/. In the articulation space we have drawn an arbitrary cross-section (solid dots) of the bundle formed by the equivalence class curves of vowel sounds. The curve corresponding to the /ə/ intersects this cross-section.

we can describe this phenomenon in terms of our k-parameters k_i. As a starting point in this direction we use the paper by Lindblom and Sundberg (1971). They suggested, on the basis of numerical experiments, that a principle of minimal articulatory antagonism between tongue and jaw might play an important role in the realization of isolated vowels. If we want to translate this mechanism into the language of our model we obviously have to define some measure with which we can quantatively indicate the resemblance of two n-tube shapes. As a measure of comparison between two n-tubes.

$$P=(k_1^{(p)},\ldots,k_{n=1}^{(p)}) \text{ and } Q=(k_1^{(q)},\ldots,k_{n-1}^{(q)})$$

we introduce in a forthcoming paper (Bonder: the MAD model)

$$d_{P,Q} = \sum_{i=1}^{n-1} (k_i^{(p)} - k_i^{(q)})^2 \}^{1/2}$$

which is the euclidean distance between the two tubes P and Q in the (n-1)-dimensional articulation space.

The translation of the suggestion by Lindblom and Sundberg (1971) into our model seems to be that we have to look, in the articulation space, for the point on the trace corresponding to a vowellike sound that has minimum distance to the straight tube, the point with coordinates (1,1,1) in the 3-dimensional articulation space of k_1, k_2, k_3. This is the so-called MAD model, where MAD stands for Minimal Articulatory Difference.

Acknowledgements

This paper is based on research which was supported by the Foundation for Linguistic Research, funded by the Netherlands Organization for the Advancement of Pure Research Z.W.O. under Contract 17-21-08.

References

Atal, B.S. et al. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America* **63**, 1535-1555.

Bonder, L.J. (1983*a*). The n-tube formula and some of its consequences. *Acustica* **52**, 216-226.

Bonder, L.J. (1983*b*). Equivalency of lossless n-tubes. *Acustica* **52**, (to appear).

Dunn, H.K. (1950). The calculation of vowel resonances, and an electrical vocal tract. *Journal of the Acoustical Society of America* **22**, 740-753.

Lindblom, B.E.F. and Sundberg, J.E.F. (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America* **50**, 1166-1179.

Physiological Explanations of F_0 Declination

R. Collier[1] and C.E. Gelfer[2]
[1]*Antwerp, Belgium; Eindhoven, the Netherlands and* [2]*New York; New Haven, U.S.A.*

Declination is the tendency of F_0 contours to exhibit a tilted overall pattern: the major rises and falls appear to be superimposed on an imaginary baseline that drifts down over the entire course of the utterance. This baseline becomes actually visible during longer stretches of speech in which no major F_0 changes occur, for instance, when there is only one pitch accent in a fairly long utterance.

Cohen, Collier and 't Hart (1982) point out that the notion of declination has developed from an operational construct, useful in the interpretation of F_0 recordings, to a theoretical concept of phonetic and linguistic importance. Since declination is now being considered an intrinsic feature of speech pitch, the question regarding its physiological origin and, hence, its programming and control, becomes increasingly relevant.

1. A simple model

A simple model of pitch control in speech may take the following form:

(a) all consciously intended pitch rises and falls are effected by appropriate variations in the tension of the vocal folds; this tension is regulated by the laryngeal muscles (mainly the cricotheroid muscle and some strap muscles of the neck, such as the sternohyoid or the thyrohyoid).

(b) the gradual downdrift of the overall pitch level, i.e. declination, is caused by slowly decreasing subglottal pressure over the course of the utterance.

Part (a) of this model is well supported by quite a number of physiological experiments, which also indicate that short term variations in subglottal pressure do not significantly contribute to the production of momentary pitch inflections (see survey in Atkinson 1978 and Ohala 1978). However, these experiments are not directly relevant to part (b) of the model, which concerns long term variations in the pitch parameter. Only a subset of the data presented in Collier (1975) suggest that part (b) of the model may be a plausible account of declination. Therefore an experiment was set up in which declination could be studied in greater detail.

2. Experimental procedure

In order to observe declination in its pristine form of appearance, utterances were constructed containing no more than two pitch accents, implemented by rise-falls, while the rest of their contours shows only declination pitch (see the stylized F_0 contours in Table I). The length of the utterances, more particularly the length of the actual stretch of declination, was varied in five steps: from 5 to 18 syllables (roughly 0.7 to 3 seconds). In order to factor out the effects of segmental perturbations, the (Dutch) utterances were also mimicked in reiterant speech with /ma/ and /fa/ syllables. In all there were 45 utterance types, each read five times in succession by one subject, the first author.

Simultaneous recordings were made of, among others, the following physiological variables: subglottal pressure (P_s), recorded directly through a tracheal puncture above the first tracheal ring, and the electromyographic activity in the crico-thyroid (CT) and sternohyoid (SH) muscles, recorded with hooked-wire electrodes. The sampling and processing techniques for these parameters have been described by Harris (1981).

F_0 was measured with the algorithm designed by Duifhuis, Willems and Sluyter (1982).

3. Results

3.1. F_0 declination

After having established in a sample of data that the inter-token variability was negligibly small, only one exemplar of the five repetitions of each utterance type was selected for F_0 analysis. The F_0 values, measured at the points indicated in Table I, were similar across the conditions of length variations and normal versus reiterant speech. Therefore the F_0 values at these points were averaged and only the three accent conditions were kept separate.

Since the declination stretches vary in length, a relatively fixed ΔF_0 over a variably amount of time leads to systematic differences in the declination rate, as can be seen in Table I. This variable rate has been observed before, among others by 't Hart (1979).

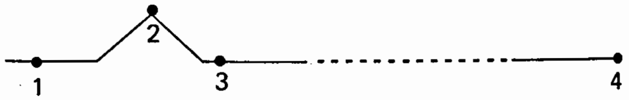
3.2. P_s declination

The P_s values, measured at the same points as the F_0 values, and averaged over all five repetitions of each utterance type, exhibit the same tendencies as the F_0 data (see Table I): they are very stable across conditions and therefore pooled together, keeping only the three accent conditions apart. It is clear that the rate of P_s declination varies with utterance length the same way as the F_0 declination does. This gross correspondence between the two variables

Table I.

A. Early pitch accent

Stylized F_0 contour and measuring points

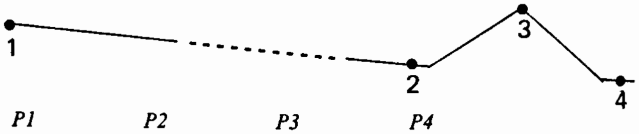


Average values (and standard deviations) at point P_i	*P1*	*P2*	*P3*	*P4*
F_0 (Hz) n = 15	109 (7)	150 (14)	99 (6)	84 (4)
P_s (cm aq) n = 75	8.3 (0.4)	9.2 (0.3)	7.7 (0.8)	4.1 (0.4)

Declination rate in length category L_i	*L1*	*L2*	*L3*	*L4*	*L5*
F_0 (Hz/sec)	-14	-12	-9	-8	-7
P_s (cm aq/sec)	-3.5	-3	-2.7	-1.9	-1.4

B. Late pitch accent

Stylized F_0 contour and measuring points

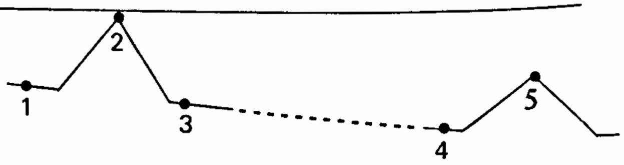


Average values (and standard deviations) at point P_i	*P1*	*P2*	*P3*	*P4*
F_0 (Hz) n = 15	115 (3)	90 (4)	117 (3)	77 (3)
P_s (cm aq) n = 75	8.4 (0.8)	5 (0.8)	7 (0.8)	2.6 (0.8)

Declination rate in length category L_i	*L1*	*L2*	*L3*	*L4*	*L5*
F_0 (Hz/sec)	-46	-23	-16	-12	-12
P_s (cm aq/sec)	-4.3	-3	-1.8	-1.4	-1.3

C. Double pitch accent

Stylized F_0 contour and measuring points



Average values (and standard deviations) at point P_i	*P1*	*P2*	*P3*	*P4*	*P5*
F_0 (Hz) n = 15	111 (5)	154 (12)	109 (6)	91 (6)	120 (4)
P_s (cm aq) n = 75	8.4 (1.5)	10.7 (0.9)	7.9 (0.9)	5.3 (0.8)	7.3 (0.6)

Declination rate in length category L_i	*L1*	*L2*	*L3*	*L4*	*L5*
F_0 (Hz/sec)	-75	-30	-11	-11	-7
P_s (cm aq/sec)	-6.4	-3	-1.9	-1.5	-1.2

suggest that F_0 declination may indeed be caused by the gradual decrease of P_s.

3.3. A causal relationship?

The extent to which P_s variations can effect F_0 changes has been studied mostly in a number of 'push in the stomach' experiments (see a survey in Bear 1979). In the chest register the P_s/F_0 ratio appears to vary between 1/3 and 1/7, which means that a ΔP_s of 1 cm aq results in a ΔF_0 of 3 to 7 Hz. Therefore a necessary condition for P_s declination to be the (sole) cause of F_0 downdrift is, that their ratio remain within these established limits.

In the 'early accent' condition of our data his is invariably the case : the P_s/F_0 ratio varies exactly between 1/3 and 1/7 in the 15 utterance types that exemplify this condition. The average ratio is 1/4.

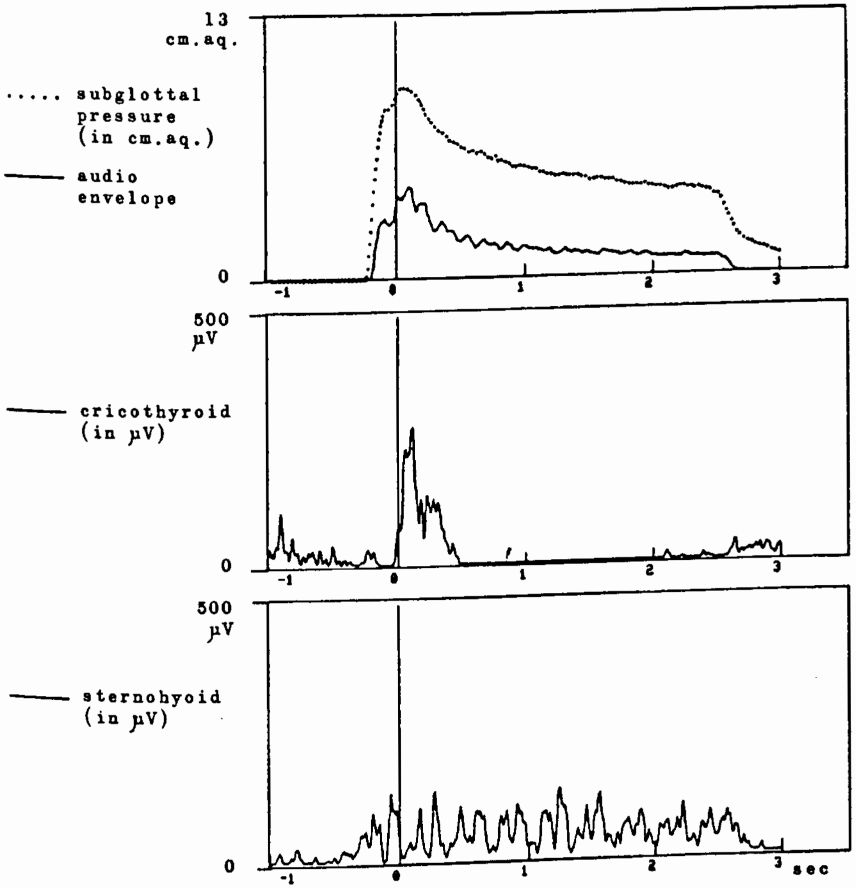


Fig. 1a. P_s and EMG data, averaged over 5 repetitions of a reiterant speech utterance with /ma/ syllables. 'Early' pitch accent at line-up point 0.

In the 'double accent' case, only 9 out of 12 utterance types comply with the established ratio. Overall the ratio varies between 1/4 and 1/11, with an average of 1/7.

In the 'late accent' condition no more than 2 utterance types have a P_s/F_0 ratio greater than 1/7. For the ensemble of this condition the ratio ranges between 1/6 and 1/16, with an average of 1/11.

Clearly, the P_s/F_0 ratio is significantly different in the three accent conditions. In the 'early accent' situation P_s declination can in itself explain F_0 declination. In the other two accent conditions the gradual decrease of P_s is often too small to account for the full extent of the F_0 downdrift.

3.4. Other factors?

In the 'early accent' condition, illustrated in Figure 1a, there is typically no CT activity during the declination stretch and SH shows nearly equal peaks of activity, mainly associated with segmental speech gestures such as jaw lowering. In the 'late accent' condition F_0 starts at a relatively high level and this is preceded by a fairly large amount of CT activity. In many instances CT relaxes gradually (over a period of up to one second) and its relaxation is then sometimes accompanied by an increasing amount of SH contraction (see Figure 1b). Thus, the combined patterns of activity in these two muscles may account for some fraction of the F_0 lowering, in cooperation with decreasing P_s. However, this picture of combined laryngeal and respiratory action does not emerge systematically enough to explain F_0 declination whenever P_s alone cannot account for it. Moreover, the same pattern also emerges in some of the utterance types in which declination can in principle be explained by references to P_s only.

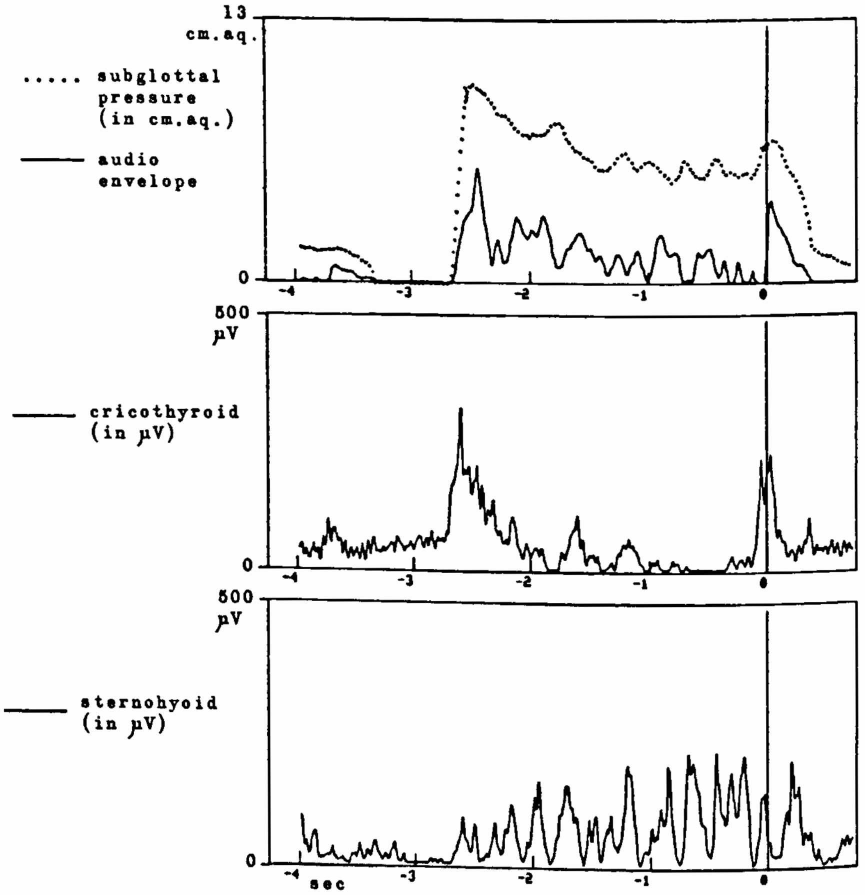


Fig. 1b. P_s and EMG data, averaged over 5 repetitions of the Dutch utterance 'Je weet dat Jan erover nadenkt ons hiervoor met genoegen te betalen'. 'Late' pitch accent at line-up point 0.

Roughly the same state of affairs holds for the 'double accent' condition.

4. Discussion and conclusion

Evidently, part (b) of our simple model accounts for the situation in which the last or only rising-falling accent occurs early in the utterance. In such a case there is no interaction of P_s with the (inactive) CT muscle or with SH (which then shows no pitch related activity). But whenever CT and (sometimes) SH are involved in pitch control, their activity is not limited to bringing about momentary pitch inflections; it can also assist in the continuous gradual pitch lowering, called declination. This means that, in certain cases, F_0 declination is not the mere byproduct of respiratory regulation but is partly controlled by laryngeal action. However our data do not show any transparant trading relationship between P_s variation and CT or SH activity. Therefore, other muscular or mechanical factors that affect vocal fold tension and glottal resistance may also be involved.

Acknowledgement

The physiological data were sampled and processed at Haskins Laboratories (New Haven, CT, USA). The F_0 measurements were made at IPO (Eindhoven, the Netherlands). We thank both institutions for their hospitality.

References

Atkinson, J.E. (1978). Correlation analysis of the physiological features controlling fundamental voice frequency. *J. Acoust. Soc. Am.* **63**, 211-222.

Baer, T. (1979). Reflex activation of laryngeal muscles by sudden induced subglottal pressure changes. *J. Acoust. Soc. Am.* **65**, 1271-1275.

Cohen, A., Collier, R. and 't Hart, J. (1982). Declination: construct or intrinsic feature of speech pitch? *Phonetica* **39**, 254-273.

Collier, R. (1975). Physiological correlates of intonation patterns. *J. Acoust. Soc. Am.* **58**, 249-255.

Duifhuis, H., Willems, L.F. and Sluyter, R.J. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. *J. Acoust. Soc. Am.* **71**, 1568-1580.

Harris, K.S. (1981). Electromyography as a technique for laryngeal investigation. In: C.L. Ludlow and M.O. Hart (Eds.), *ASHA Reports: Proceedings of the Conference on the Assessment of Vocal Pathology*, 1981.

Ohala, J.J. (1978). Production of tone. In: V. Fromkin (Ed.), *Tone: a linguistic survey*. New York, Academic Press.

't Hart, J. (1979). Explorations in automatic stylization of F_0 contours. *IPO Annual Progress Report* **14**, 61-65.

Spectral Consequences of a Time-varying Glottal Impedance

B. Cranen and L. Boves
Nijmegen, the Netherlands

1. Introduction

Virtually all research in modern acoustic phonetics relies on the linear source-system model of acoustic speech production (Fant, 1960; Flanagan, 1972). The alleged linearity of the source-system cascade permits the separate development of accurate models of both the source and the system. The assumption that the glottal impedance is very high enables one to treat the vocal tract as an acoustic tube that is closed at the glottal end. Since all articulators have an appreciable mass they are constrained to relatively slow motions, and consequently the vocal tract resonances can be supposed to vary so slowly that they may be considered as constant during time intervals of 10-30 ms. This assumption is at the basis of all present day speech technology, notably including all LPC techniques.

The ubiquitous experience of the machine-like quality of much synthetic speech, however, has motivated a re-examination of the customary source-system model. However unimpeachable the assumption of an infinite glottal impedance may be during the closed glottis interval of a fundamental period, during the open glottis interval the termination impedance at the glottis may well assume values of the same order of magnitude as the input impedance of the vocal tract.

Although it is recognized that changes in the glottal termination impedance may affect the formants of speech signals, little is known about the way in which specific changes in glottal termination affect the formants of various vowels. The present study atttempts to derive techniques with which the effects of the termination impedance can be determined so that it becomes possible to estimate their contribution to the naturalness of synthetic speech. To this end we combined measurements on real speech with model simulations.

2. The Wakita-Fant model

Unlike the customary models of speech production, the Wakita-Fant approach (henceforth WF-model) includes a model of the subglottal system. In the WF-model it is dimensioned so as to represent the measurements of Ishizaka, Matsudaira, and Kaneko (1976) on the input impedance of the subglottal system, who found resonances at 640 Hz, 1400 Hz, 2150 Hz, and 2850 Hz, with bandwidths between 200 and 400 Hz. In our implementation the supraglottal system consisted of six sections of unequal elliptical cross-sections and unequal lengths. We have studied the model for seven vocal tract configurations representative of the vowels /a,o,u,i,ɨ,e,ə/ Cross-sectional areas and lengths of the sections were adapted from the data in Fant (1960). Subglottal pressure was varied from 4 cm water up to 10 cm water in steps of 2 cm, and glottal area(Ag) was varied between 0 and .2 cm^2 in steps of .01 cm^2. In order to assess the importance of the subglottal system, all calculations have been repeated with a model in which the glottal termination consisted of an R-L circuit modeled after the properties of the air plug in the glottis.

The results of these model calculations can be summarized as follows: Formant frequencies and bandwidths appear to behave as monotonic (though not necessarily linear) functions of subglottal pressure and glottal area. The effect of changes in Ag appear to outweigh those of variations in subglottal pressure by far. This permits us to restrict our discussion to the results obtained for the 'typical' condition of 8 cm water for Psg.

For the model with the subglottal system the frequency of F1 increases by approximately 10% as Ag increases from 0 to .2 cm^2. Although the actual rate of change is quite vowel dependent, all vowels show variations which are clearly greater than 3%, the JND for formant frequencies according to Flanagan (1972). Except for the /a/-vowel, variations in F2 do not exceed the critical value of 3%, nor do the higher formants of any vowels.

Formant bandwidths in the complete model increase by approximately 400% in F1, 150% in F2 and 50% in F3 as Ag increases from 0 to .2 cm^2. The variations are strongly vowel dependent: F2 of /i/, for instance, increases by no more than 18%, which is clearly less than the JND of 50% (Flanagan, 1972). F2 of /o/ and /ə/ on the other hand, increase by more than 230%.

No simple conclusions can be drawn with respect to the results obtained for the model without a subglottal system, except that the dependence on vowel type is much greater than was the case with the complete model. Interestingly F1 of /i/ decreases slightly with increasing Ag in the model with the R-L glottis. The rate of change of formant bandwidth as a function of Ag is between two and three times as great for the simplified model.

Since the WF-model assumes stationary conditions and since the glottal impedance changes with a frequency which cannot be considered as very low with respect to the frequencies of the formants, a verification of the results of the model is in order. This has been done by studying real speech and the output of another model.

3. Formants measurements on real speech

Five subjects participated in an experiment in which simultaneous recordings were made of the electroglottogram, subglottal pressure, and the acoustic speech wave (Boves and Cranen, 1982). The electroglottogram is a reliable indicator of the moment of glottal opening and closure. Thus the signal provides the information needed to obtain the open and closed glottis intervals from each glottal cycle. The subglottal pressure signal enables us to determine the resonance of the subglottal system. Using a linear (time domain) averaging technique combined with cepstral smoothing spectra of the subglottal pressure waves were computed. The results consistently show resonances around 500, 1200, and 1900 Hz. Thus, we appear to find resonance frequencies of the subglottal system which are consistently lower than those reported by Ishizaka et al. More often than not clear spectral dips are present between these resonances.

Estimation of the parameters of the supraglottal system was attempted using a covariance LP analysis. In view of the fact that a (rectangular) analysis window -the left boundary of which is positioned at the beginning of the closed (or alternatively open) glottis interval- is not allowed to extend into the following open (or closed) interval, window lengths had to be constrained to at most some 30 samples. The extremely short-time LP analysis appears to yield stable and credible results for the closed glottis intervals. The open glottis intervals, on the other hand, consistently give very unstable results in the sense that variations from period to period are very great. Also, more often than not F1 and F2, or F3 and F4 are no longer resolved.

4. Formant measurements made on the Ishizaka-Flanagan model

The speech production model as described in Ishizaka and Flanagan (1972) (henceforward IF-model) was implemented using the 1130 Continuous System Modelling Program (developed for use on IBM 1130 and 1800 computers but adapted to run on a DG Eclipse S/200). Seven vowels were generated using the same vocal tract configurations as in the experiments with the WF-model described above. The parameters of the two-mass model of the vocal folds were given 'typical' values. Our implementation of the IF-model did not include a subglottal system. Outputs of the model comprise the speech signal radiated at the lips and the volume flow at the glottis. From the latter signal the boundaries of the open and closed glottis intervals can be established unequivocally.

A covariance LP analysis carried out on the synthetic speech confirmed the results obtained with real speech. Formant estimates during closed glottis intervals are stable and in very good agreement with the results of the WF-model, whereas estimates derived from the open glottis intervals often fail to resolve all formants.

5. The failures explained

Merging of the formants can be accounted for by an increased damping in the open glottis interval. Formant bandwidths may become excessively large and may thereby prevent neighbouring peaks from being resolved.

This increased damping may be due to a decrease in glottal impedance during the open glottis interval. This effect is apparent in the (stationary) WF-model for e.g. the vowel /a/.

A time-varying glottal area might, however, contribute to a merging of formants in yet another way: the time derivative of a uniformly growing conductance acts like an inertance. Ananthapadmanabha and Fant (1982) have shown that the source filter interaction during the open glottis interval will give rise to both an upward shift in frequency and a broadening of the first formant, in addition to the introduction of a number of spectral zeros.

Fourier analysis of glottal flow signals obtained by inverse filtering of real speech and as an additional output of the IF-model shows that the source signals indeed tend to contain clear and sharp spectral dips. As can be seen from figures 1 and 2 one dip invariantly lies very close to the F1-frequency of the closed glottis interval as predicted by Ananthapadmanabha et al. (1982).

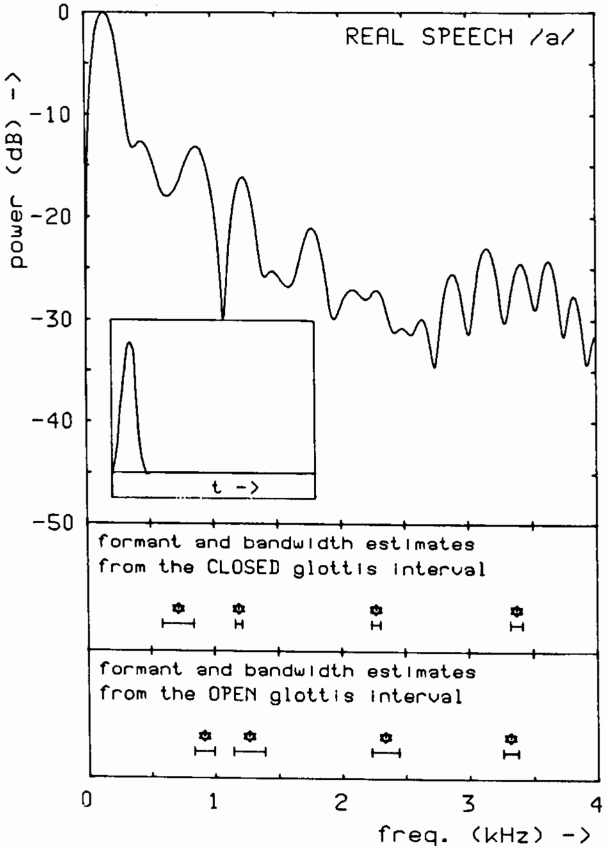


Figure 1. 1024 point FFT-spectrum of (the time-derivative of) a reconstructed glottal pulse (REAL SPEECH, vowel /a/) and formant/bandwidth estimates of the speech signal in the closed, respectively open glottis interval. The pulse of which a 25 ms time segment is shown in the inset, has been obtained by inverse filtering of the speech signal. The inverse filter has been estimated in the preceding closed glottis interval by means of an LP-analysis (covariance method, rectangular analysis window = 24 (closed)/36 (open) samples, prediction order = 10 (closed)/12)open)).

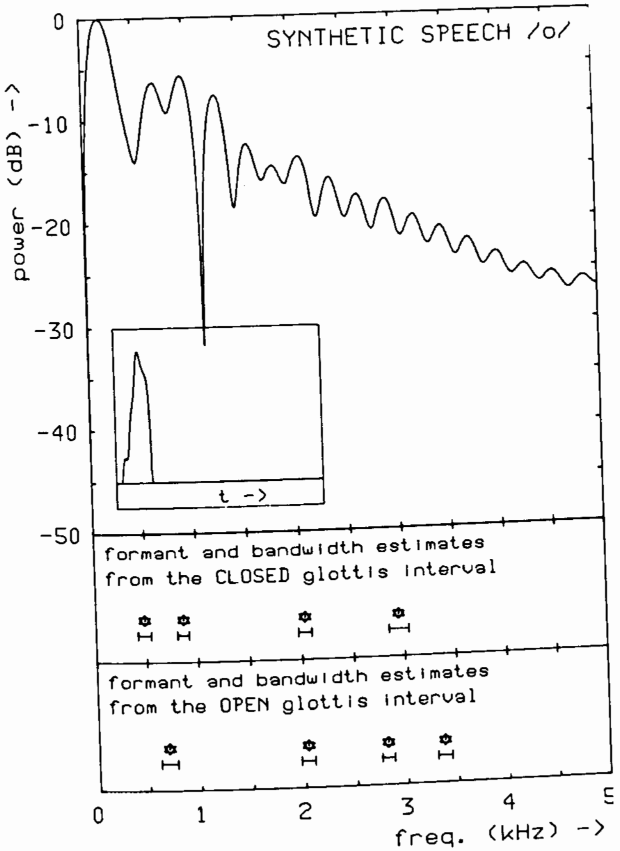


Figure 2. 1024-point FFT-spectrum of (the time-derivative of) a glottal pulse (IF-MODEL, vowel /o/) and formant/bandwidth estimates of the speech signal in the closed, respectively open glottis interval. The pulse of which a 25 ms time segment is shown in the inset, has been generated by the IF-model. The formant estimates have been obtained by LP-analysis of the corresponding speech segments (covariance method, rectangular analysis window = 24 (closed)/36 (open) samples, prediction order = 10 (closed)/12 (open)).

6. Conclusions

From the experiments with the WF-model it is clear that within-cycle variations in the termination impedance at the glottis may well have effects on formant parameters of such an order of magnitude as to be perceptually relevant. Thus, such variations may account for (part of) the machine-like character of much synthetic speech.

Attempts to verify the predictions of the WF-model by measurements on real speech and on the output of the IF-model confirmed the importance of within-cycle variations, but failed to yield numerical results since formant estimations in open glottis intervals appeared to be virtually impossible. To a large extent this failure can be explained by the presence of clear spectral zeros in the voice source.

The outcomes of our measurements, combined with the theoretical work of Fant and his coworkers, raise the question whether it is at all possible and sensible to use the formant concept in describing the speech signal during the open glottis interval.

Acknowledgement

This research was supported by the Foundation for Linguistic Research which is funded by the Netherlands Organization for the advancement of pure research, ZWO.

References

Boves, L. and Cranen, B. (1982). Evaluation of glottal inverse filtering by means of physiological registrations. *Conference Records on the International Conference on Acoustics, Speech, and Signal Processing ICASSP-82,* 1988-1991.

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Ananthapadmanabha, T.V. and Fant, G. (1982). Calculation of true glottal flow and its components. *Speech Communication* **1**, 167-184.

Flanagan, J.L. (1972). *Speech analysis, synthesis, and perception.* Berlin: Springer.

Ishizaka K. and Flanagan, J.L. (1972). Synthesis of voice sounds from a two mass model of the vocal cords. *Bell System Technical Journal,* **51**, 1233-1268.

Ishizaka, K., Matsudaira, M. and Kaneko, T. (1976). Input acoustic impedance measurement of the subglottal system. *Journal of the Acoustical Society of America,* **60**, 190-197.

Wakita, H. and Fant, G. (1978). Toward a better vocal tract model. *Speech Transmission Laboratory-QPSR,* **1**, 9-29.

On the Reliability of the Intraoral Measuring of Subglottal Pressure

G.L.J. Nieboer, H.K. Schutte and T. de Graaf
Groningen, the Netherlands

1. Introduction

Measuring subglottal pressure is a useful tool in diagnosing patients with voice problems. And of course it is a necessary part of a complete physiological description of speech events.

But up to now none of the established methods to measure this subglottal pressure is simple enough to be applied in every day practice. These established methods are:

- first, the transcutaneous needle;
- secondly, the oesophageal balloon;
- thirdly, the tip-catheter or the small tube-catheter through the glottis.

None of these methods is easy-to-do, and the help of a medical doctor is nearly always needed. So the search for easier methods deserves our attention, not in the last place because research in the Groningen ENT Clinic has revealed that the subglottal pressure is an important parameter in establishing vocal dysfunction.

2. The intraoral method

An interesting and easy-to-do method proposed by Rothenberg in 1973 and used again by Smitheran and Hixon in 1981 is: to measure the subglottal pressure during phonation indirectly, by estimating it from the intraoral pressure during the occlusion phase of adjacent voiceless plosive consonants. This is what makes this method simple: the intraoral pressure can be measured in a comparatively uncomplicated way by inserting a small catheter into the mouth through the lips.

The method is based on two assumptions: the first is that the peak oral pressure obtained in this way corresponds to the subglottal pressure during phonation. The second assumption, though it is never mentioned as such as far as we know, is that the glottis is open during the occlusion phase of voiceless plosive consonants. It is known for French and Russian speakers that the glottis is not open during this occlusion phase. For Dutch speakers we are not sure at all, because, like in French, there is no aspiration after voiceless plosive consonants; this may point to a closed glottis before the phonation starts.

3. Comparative measurements

We carried out measurements on two Dutch-speaking subjects in order to test the reliability of the method. The most obvious way to do this is to compare data from simultaneous measurements of first the intraoral pressure and secondly the subglottal pressure, assessed by one of the afore-mentioned established methods. In our study we used the oesophageal balloon.

Figure 1 shows the parameters we recorded in our measurements.

The oesophageal pressure is measured with an oesophageal balloon, fixed to a long thin catheter passing through the nose. The oral/nasal flow is measured with a Lilly flowhead, fixed to a rubber mask. The flow curve provides information on the moments of lip opening of the plosive consonant /p/. A signal related to the lung volume is computed by integration from the flow curve. We need to know this lung volume because it is related to the oesophageal pressure by lung mechanical factors. The microphone signal is used to assess the moments of voice onset and offset. The intraoral pressure was measured with a small catheter through the mask, with a pressure transducer fixed directly to the mask.

The way in which the intraoral and oesophageal pressure are compared is shown in Fig. 2.

In order to assess the value of the subglottal pressure we must have a reference point. This is necessary because the intraoesophageal and intratracheal pressures are different to a varying degree, depending upon lung mechanical factors. We have chosen the reference point to be the moment of minimal lung volume (on top of the curve) during quiet breathing. The real intraoesophageal pressure at that moment is about -0.4 kPa. We will call the pressure at this moment the reference pressure. When the lung volume curve during the repeated phonations indicates that the lung volume is the same as at the reference moment, we can consider the amount of pressure in the oesophageal pressure curve above the reference pressure to be equal to the driving pressure of the lungs, that is the mean subglottal pressure. So at this moment of the curves this mean subglottal pressure can be directly compared to the simultaneously measured intraoral pressure.

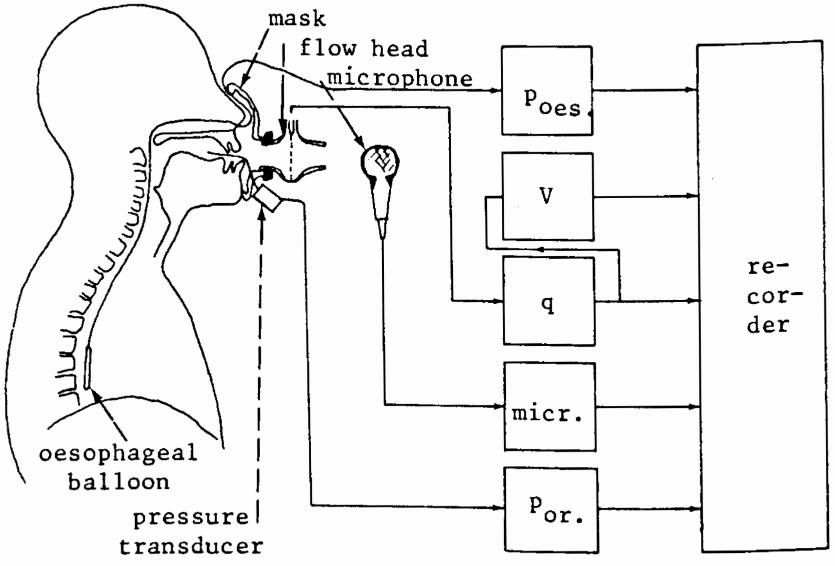


Figure 1. Diagram of the experimental set up. The curves are recorded on a Mingograph.

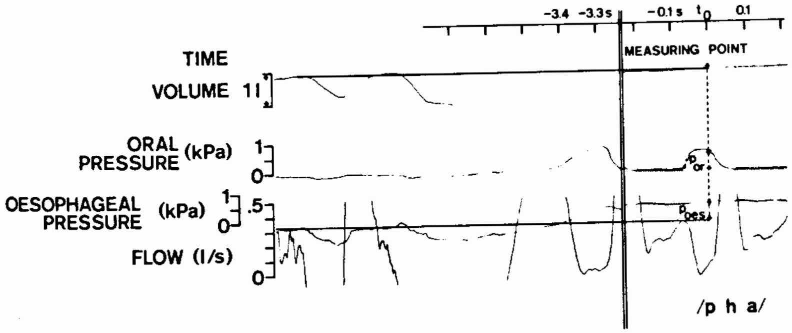


Figure 2. The way in which the subglottal pressure is derived from the oesophageal pressure and compared to the intraoral pressure. The subject repeats /pha/ at a rate of ca. 4 per second.

Schutte (1980) gives a more complete description of the intraoesophageal measuring of subglottal pressure, and of the registration of flow and volume curves.

4. Results of the measurements

In Figures 3 and 4, the reference pressure, which was always assessed in the way we described, is indicated by a solid horizontal line; an arrow points to the moment of measuring.

In Figure 3 the subject repeats /pa/ at a rate of about 4 per second. The oesophageal pressure above the reference pressure and the intraoral pressure have both the value of 0.6 kPa. When /pa/ utterances are repeated at a slower rate, the two pressures are equally well comparable.

But as soon as aspiration occurs, differences show up as is shown in Figure 4, representing a /pha/ utterance, aspirated. We see that the oesophageal pressure curve shows different values when we compare the moments of occlusion and of phonation.

The speech rate is 2 per second and the pressure difference is 0.2 kPa. At a speech rate of 4 per second this difference is 0.1 kPa.

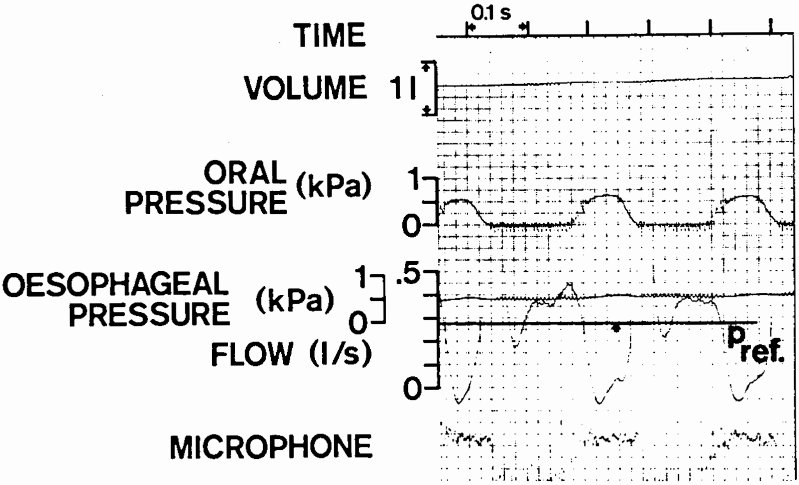


Figure 3. Lung volume, oral pressure, oesophageal pressure, flow and microphone curves of one of our subjects repeating /pa/ at a rate of ca. 4 per second. The arrow points to the moment of measuring.

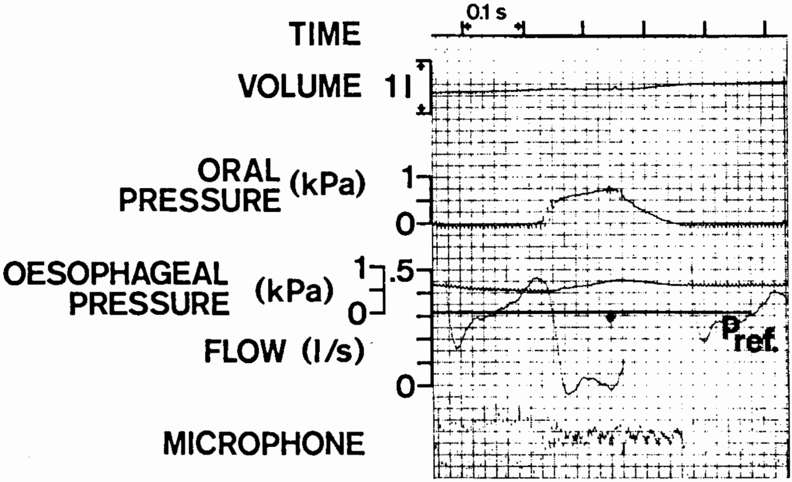


Figure 4. Lung volume, oral pressure, oesophageal pressure, flow and microphone curves of one of our subjects repeating /pha/ at a rate of ca. 2 per second. The arrow points to the moment of measuring.

5. Conclusions

Our first conclusion from these experiments is that the reliability of the method described does not depend on the rate of the utterances, when it varies between 2 and 4 per second. Of course, this and the following conclusions are valid for Dutch speakers.

A second conclusion is that when aspiration occurs the curve of the oesophageal pressure shows fluctuations, indicating that the moment of intraoral measuring provides a deviant estimation of the mean subglottal pressure. In this case, as can be seen in Figure 4, the estimate is too high. The differences amounted to 0.2 kPa, which means a measuring error of about 25%.

A third conclusion is that the intraoral pressure curve must have a horizontal part. This can be obtained by using a pressure measuring system with a response time within 30 ms, as was pointed out by Rothenberg in 1981. The need of such a short response time is illustrated by the fact that we found the pressure build-up in the mouth to take place in 40-100 ms for /pa/ and /pha/ utterances.

Our last conclusion is that on the basis of these and other measurements, the intraoral method to estimate the mean subglottal pressure is not, or not yet, a reliable tool when patients are involved with articulatory or vocal dysfunction. When for example a patient aspirates his /p/ consonants, or closes his glottis more or less violently during the occlusion phase, we can be almost sure that the estimates become unreliable. So we think that this method, as far as it has been developed up to now, is a reliable tool to measure subglottal pressure in trained phoneticians; but it needs more research in order to be a reliable method for the untrained voice patient.

References

Rothenberg, M. (1981). Interpolating subglottal pressure from oral pressure. *Journal of Speech and Hearing Disorders,* **47**, 218-219.

Schutte, H.K. (1980). *The efficiency of voice production.* Doct. thesis, Groningen.

Smitheran, J.R., Hixon, T.J. (1981). A clinical method for estimating laryngeal airway resistance during vowel production. *Journal of Speech and Hearing Disorders,* **46**, 138-146.

Lung and Larynx Coordination in a Composite Model of Speech Production

C. Scully and E. Allwood
Leeds, United Kingdom

1. Introduction

From different combinations of a few rather simple articulatory actions, a variety of quite complex aerodynamic conditions and acoustic outputs can be created. The most basic requirement of all for speech is the creation of voice; this is easily achieved by new-born babies. What is examined here is the building up of a repertoire of lung and larynx actions appropriate for *controlled* operation of the voice source. Even apparently simple speech sounds demand correct coordination. The auditory goal of the simulation described here was an [i] vowel quality with 'modal' as opposed to 'breathy' or 'pressed' ('laryngealised') phonation type and with falling pitch. The tasks of speech production are by no means clear, but one basic aim is to achieve a subglottal pressure suitable for the onset and maintenance of voice.

2. The model

A model of speech production processes implemented on a VAX 11/780 computer was used. The stages modelled are shown in Figure 1. Inputs to the model define speaker dimensions, initial conditions, larynx type for a functional model of voicing and articulatory transitions. Eight quasi-independent articulators are used, as controllers of the geometry rather than as anatomical structures. Most articulatory actions are represented by changes in cross-section area of a few constrictions of the vocal tract. Articulations of the lung walls are represented either by air pressure in the lungs Pl, or as in the study described here, by the rate of change of lung volume DVLU. Vocal fold articulations are represented by the slowly changing (d.c.) component of glottal area Ag and by a variable called Q, for the effective stiffness and mass of the vocal folds. Vertical movements of the vocal folds are not modelled at present. The bases for the modelling have been described (Scully, 1975; Allwood and Scully, 1982).

Timing and coordination in the articulatory block determine aerodynamic conditions throughout the respiratory tract. Articulatory states and aerodynamic conditions combine to determine the magnitude of turbulence noise sources for aspiration and frication. A pulse source, derived from rate of pressure change in the oral cavity, has been introduced recently, but was not used in this study. A parametric description of the voice source is used as shown in Figure 2 (Fant, 1980). A minimum Δ P of 2 cm H_2O was assumed for the onset and offset of voicing. Fundamental frequency F_0 was derived from $F_0 = \Phi + 4.\Delta$ P. A voicing 'plateau' region was defined between AG= 0.04 cm^2 and Ag = 0.08 cm^2. F_0 decreased for Ag less than 0.04 cm^2. K varied inversely with Ag. TCR was constant at 0.1. Aspiration and frication sources were weakened and modulated when voicing was present. In an alternative form of the voicing model the wave parameters VOIA, K and TCR can all be made to vary as linear functions of three controlling physiological variables: Ag, Δ P and Q. Using the model interdependence of vowel and consonant durations have been demonstrated for voiced and voiceless fricatives having constant supraglottal articulation and for open and close vowel contexts. The effects were similar to those of real speech and the model's outputs were intelligible and speech-like (Allwood and Scully, 1982).

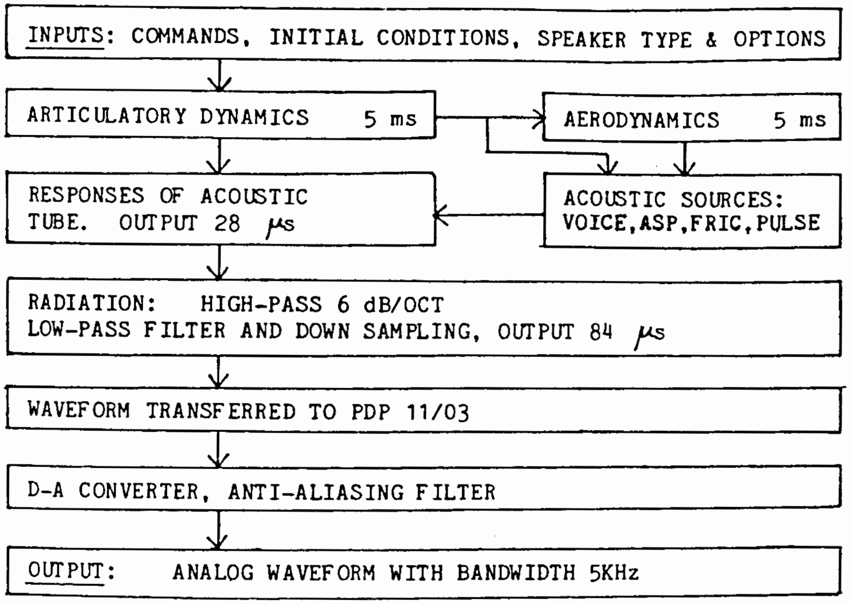

INPUTS: COMMANDS, INITIAL CONDITIONS, SPEAKER TYPE & OPTIONS
ARTICULATORY DYNAMICS 5 ms
AERODYNAMICS 5 ms
RESPONSES OF ACOUSTIC TUBE. OUTPUT 28 μs
ACOUSTIC SOURCES: VOICE, ASP, FRIC, PULSE
RADIATION: HIGH-PASS 6 dB/OCT
LOW-PASS FILTER AND DOWN SAMPLING, OUTPUT 84 μs
WAVEFORM TRANSFERRED TO PDP 11/03
D-A CONVERTER, ANTI-ALIASING FILTER
OUTPUT: ANALOG WAVEFORM WITH BANDWIDTH 5KHz

Figure 1. Block diagram of the model.

3. Modelling of aerodynamic processes

The system in Figure 3. A set of first order differential equations expresses the assumptions made and the physical principles invoked in the model, which are as follows:

1. The compliance of the lung walls need not be included. It is assumed that the speaker takes the nett compliance (recoil) into account when adjusting muscle pressures at different lung volumes so as to give a pre-planned rate of lung volume decrement. Passive changes in rate of lung volume decrease are not modelled at present.

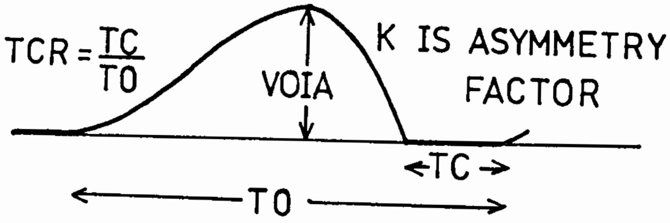


Figure 2. The parametric description for the voice source.

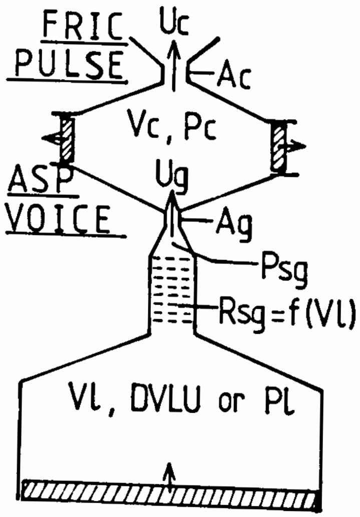


Figure 3. The aerodynamic system.

2. The walls of the subglottal airways are taken as rigid, with flow rates in speech well below limiting flow rate.
3. The supraglottal cavity has an active component of volume change due to articulatory actions, added to a passive component associated with wall compliance (Rothenberg, 1968).
4. All but 4% of the subglottal volume is located in the respiratory zone of small airways, with generations higher than 16. Subglottal flow resistance is almost totally confined, on the contrary, to the large tubes of generation less than 10. This striking separation of subglottal volume and flow resistance justifies a model with one lumped lung volume and a separate single flow resistance linking it to the glottal orifice. This constrasts with the more complex representation in the model of Rothenberg (1968)
5. Subglottal flow resistance is an 'ohmic' conductance which increases linearly with lung volume, up to a maximum value of about 2 L/cm H_2O.
6. Inertance of air and tissues may be neglected.
7. The air in the respiratory tract is assumed to be an ideal gas and to be compressible. Departures from atmospheric pressure are small. Isothermal conditions are assumed. The flow is taken as one-dimensional. There is continuity of mass flow for each of the two cavities.
8. For each of the two orifices (constrictions) there is conservation of energy at the inlet (the Bernoulli effect), but energy is lost in the turbulent mixing region at the outlet. This gives a turbulent, flow-dependent component of pressure drop. A laminar 'ohmic' component of pressure drop is added to this. The same empirical constants are used for both orifices.
(Space does not permit reference to the relevant respiratory literature).

Parameter values are chosen to define cavity wall compliance, subglottal properties and initial conditions, Lung volume Vl, lung and supraglottal air pressures Pl and Pc are integrated at each time step to obtain values for the next sample. Merson's method (NAG library, 1981) was used here. There were problems with numerial instabilities in the aerodynamic variables, especially when oral pressure Pc was very low, in vowel-like segments. Other methods for the integration, including Gear's method for dealing with 'stiff' equations (NAG library, 1981), have recently given improved stability and much reduced computation time for the aerodynamics.

4. The modelling of lung and larynx coordination

Some articulatory plans yielded inappropriate pitches, voice qualities or vowel lengths. Two of a series of articulatory plans are shown in Figure 4

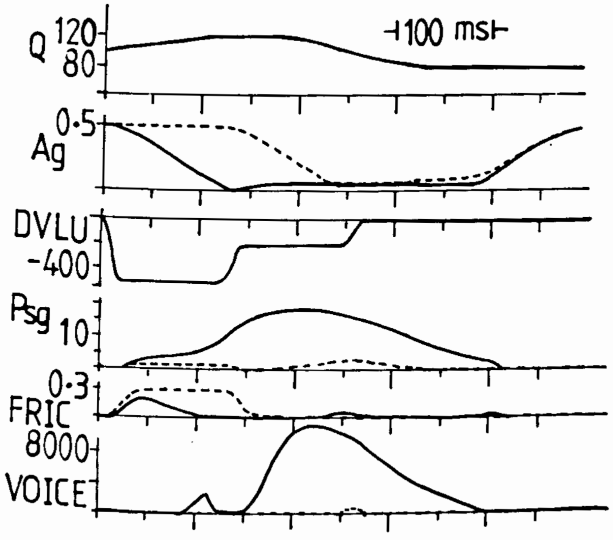


Figure 4. Two contrasting coordinations for the lung walls (DVLU in cm^3/s) and the larynx (Ag in cm^2 and Q in Hz). Also shown: a computed aerodynamic variable Psg in cm H_2O and the envelopes of acoustic sources voice and frication noise (FRIC) in arbitrary units. (a) --- (b) ———.

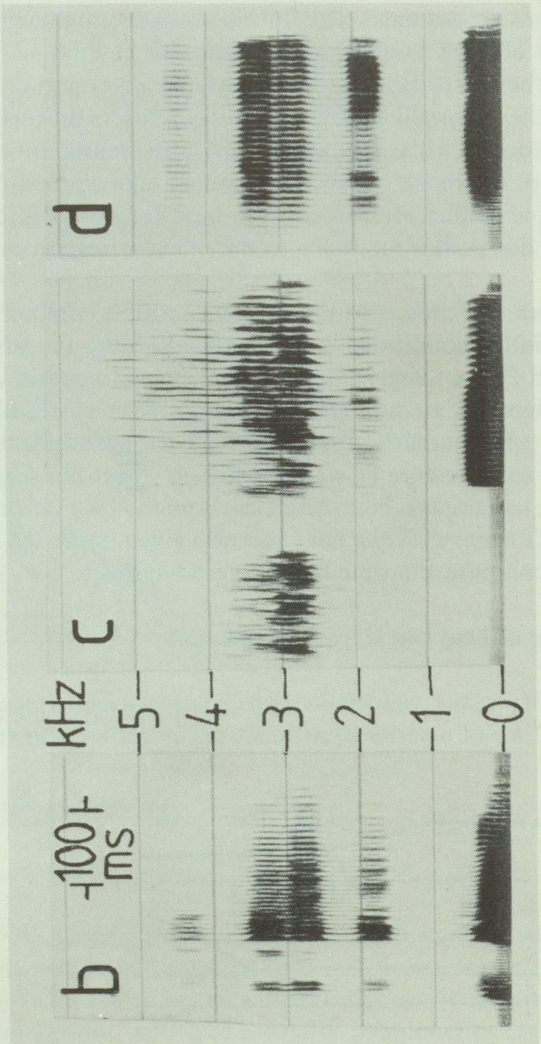


Figure 5. Spectrograms for (a) and (b) in Figure 4 and for additional runs (c) and (d).

together with some of the aerodynamic and acoustic results. Unwanted sounds were generated in both cases. (a) was an attempt at 'braethy' attack. It was transcribed auditorily as [çʰi̥ʔi] with falling pitch. (b) was an attempt at 'hard' (or 'glottalised') attack and was transcribed as [hiʔi·] with 'gulp' effect, sudden onset and falling pitch. Spectrograms for (a) and (b) are shown in Figure 5. Two other unsuccessful attempts at the auditory goal are shown as (c) and (d) in Figure 5. (c) gave [breath drawn in sharply] then [i] falling pitch. (d) gave a 'strong' [i] sound with no audible noise, but not a falling pitch. In another set of syntheses for target words 'purse' and 'purrs', unwanted vowel-like segments were often generated at the speech offset. By trial and error, combinations of lung and larynx actions could be found which avoided unwanted onset and offsets. It is suggested that auditory feedback must be of overwhelming importance for the acquisition of speech, as in our modelling. The onset and offset of speech present speakers with specific problems. The options selected by a particular speaker for the achievement of rather broadly defined auditory goals will be reflected in the details of acoustic structure. Modelling of the kind outlined here may be able to assist in defining the probable acoustic variations within one accent, with potential applications in automatic recognition of speech.

Acknowledgement

This work is supported by the Science and Engineering Research Council, Grant Gr/B/34874.

References

Allwood, E. and Scully, C. (1982). A composite model of speech production. In: *Proceedings of the IEEE International Congress on Acoustics, Speech and Signal Processing, ICASSP 82,* Paris, 932-5.

Fant, G. (1980). Voice source dynamics. STL-QPSR 2-3/80. Stockholm: Department of Speech Communication and Music Acoustics, RIT. 27-37.

Numerical Algorithms Group (1981). Numerical Algorithms Group FORTRAN Library Manual, Mark 8, Vol. 1.

Rothenberg, M. (1968). *The Breath-Stream Dynamics of Simple-Released-Plosive Production.* Bibliotheca Phonetica, No. 6. Basel: Karger.

Scully, C. (1975). A synthesizer study of speech segment durations. In: *Speech Communication* (G. Fant, ed.), Vol. 2, pp. 227-234. Stockholm: Almqvist and Wiksell.

Use of an Optical Position-Sensitive Device for Recording Lip and Jaw Movements in Speech

Y. Sonoda
Kumamoto, Japan

1. Introduction

For the study of the speech production process it is highly important to observe articulatory movements during speech. Most of the studies on articulatory behavior have been done by using x-ray techniques. Obvious disadvantages of x-ray observations are the exposure of the subject to harmful radiation and the time needed to extract data to obtain displays of motion in time. A computer-controlled x-ray technique has been developed to provide time-motion signals on articulators (Kiritani, Itoh and Fujimura, 1975). However, its use still must be under the constraints that the dose be minimized, and any scheme without the use of x-rays is useful both independently and as a supplement to an x-ray method.

This paper introduces a new optical recording system for speech research (Sonoda and Wanishi, 1982) and presents some early data on lip and jaw movements as examples.

2. Measuring Principle

The recording system consists of an optoelectric position-sensitive device (PSD) and light-emitting diodes (LEDs).The PSD provides two electrical analog outputs, specifying the individual X and Y positions of input light spots from LEDs, which are attached to several selected points of articulators. The system provides simultaneous recordings of the upper lip, lower lip, and jaw movements, continuously and in real time, with a high accuracy and a very fast response.

The PSD, developed and manufactured by Hamamatsu Corp. of Japan, is composed of a planar-type PN silicon photodiode with very uniform resistive layers formed on both the top and bottom surfaces as shown in Fig. 1. Pairs of electrodes (X_1-X_2 and Y_1-Y_2) are provided on each surface of the P and N layers, to sense X- and Y-axis signals, respectively. When a light spot from an LED is focused via a lens system on the active surface of the PSD, the photo current, which is proportional to the light intensity, is induced in the depletion layer at PN junction. This current is separated into two groups of output current, which flow through P and N layers to and from the opposite electrodes on the X and Y axis in the figure. Each of these currents is divided

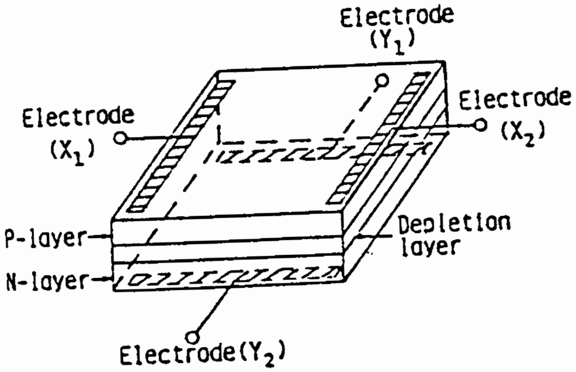


Fig. 1. Schematic construction of the PSD.

in accordance with the resistances of the corresponding layer between the light spot position and each electrode. Measuring each current and taking the ratio of their difference to their sum, the spot point on the PSD is converted into X and Y distance.

A circuit of the arithmetic operation utilized for a PSD is easily designed by the use of the analog IC arithmetic modules as shown in Fig. 2. The calibration patterns are slightly deformed around the edge. The maximum deviation in the useful area (5 × 9 cm) is estimated at 0.19 mm and the resolution on the position of an LED is estimated at 0.05 mm at a distance of 50 cm from the LEDs.

3. Data Collection

Preliminary experiments for recording lip and jaw movements have been conducted. Five of the eight LEDs were attached to these articulators within the sagittal plane as shown in Fig. 3. Two of the five LEDs were for the monitoring the lip movements; upper lip (UL) and lower lip (LL). For the

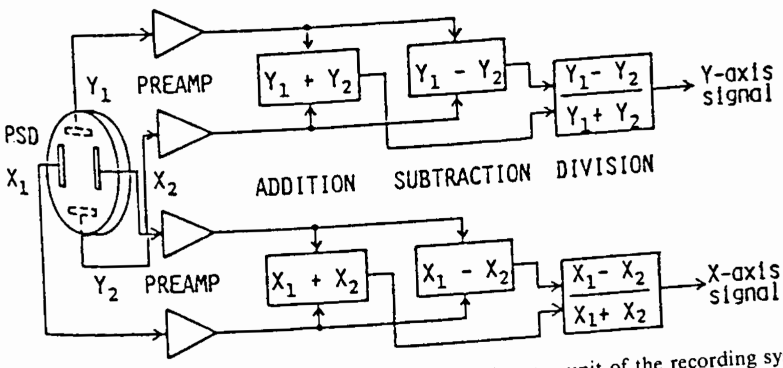


Fig. 2. Block diagram of the arithmetic positional detector unit of the recording system.

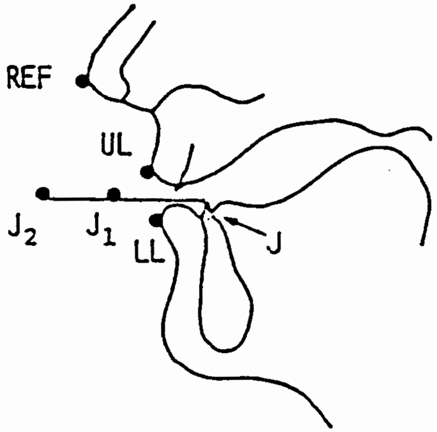


Fig. 3. Typical arrangement of LEDs used to observe articulatory movements.

observation of jaw movements, two LEDs (J_1 and J_2) were fixed on the rigid wire protruding from an artificial teethridge fixed at the canines. The movements of the jaw, the edge of the lower incisor J, were determined by these movements. The reference LED, REF, was used for monitoring the head displacement during experiments.

Fig. 4 shows an explanatory diagram for calculating the coordinates of the points J and C which correspond to the edge of the lower incisor and the point of the head of condyle, respectively. On these points hidden from view, the following calculations are made at each sample interval:

Jaw : $J(X_0,Y_0)$

$$X_0 = X_2+(X_1 - X_2)\,\frac{l_1+l_2}{l_1}\,,\quad Y_0 = Y_2+(Y_1 - Y_2)\,\frac{l_1+l_2}{l_1}$$

Condyle : $C(X_3,Y_3)$

$$X_3 = X_2 + \frac{1}{l_1}\,[(X_1 - X_2)\,l_0 + (Y_1 - Y_2)\,l_4]$$

$$Y_3 = Y_2 + \frac{1}{l_1}\,[(Y_1 - Y_2)\,l_0 + (X_1 - X_2)\,l_4],\quad \text{where } l_0=l_1+l_2+l_3$$

The parameters of l_3 and l_4 were obtained from the x-ray cephalograms for individual subjects.

In order to minimize gross movements during the experiments, the subject's head was fastened onto the headholder. However, small displacements of the head were observed and therefore subjects were instructed to take the jaw-closed position (rest position) keeping the lips relaxed after every tenth utterance of the test words. The data obtained under this condition were used for correction in terms of the translational and rotational perturbation in the head position during the course of an experiment. After the correction of the point REF to the origin, variations of the upper lip (UL) and jaw (J) were estimated. Standard deviation of the three subjects were very small; the individual values were 0.14, 0.52 and 0.43 (mm) at the point of upper lip, and 0.24, 0.44 and 0.41 (mm) at the jaw, correspondingly. These results show that the relative geometric relations among these points remained almost unchanged regardless of the small movements of the head. The lower lip is biomechanically linked with the jaw, the observed movement of the point LL is the sum of the individual lip and jaw components. The net movement of the lower lip was obtained by subtracting the movement of the point J from that of LL.

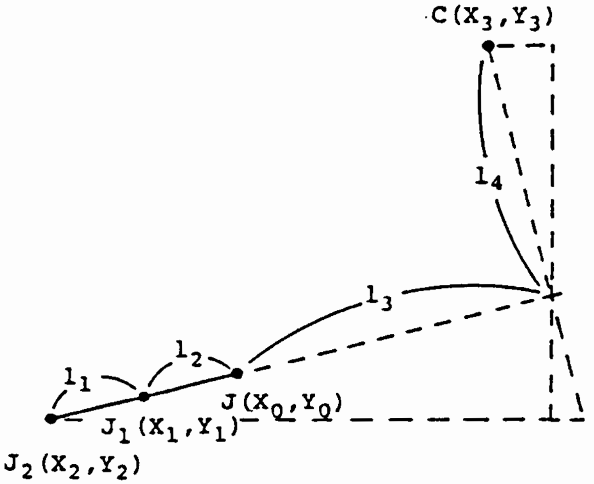


Fig. 4. Explanatory diagram for calculating the coordinates of the points J and C.

4. Experiments

All data were recorded once on an analog tape recorder and then stored in the computer at a frame rate of 200 Hz. Fig.5 shows typical time patterns (a) and two-dimensional trajectories (b) during an utterance of /epepe/ spoken by one subject. The curves in the figure are alternatively horizontal (H) and vertical (V) displacements of each point on the articulators. The curves denoted as L-J and CNDYL are the movement patterns of the lower lip itself and the head of the condyle respectively. The lowering of the upper lip for /p/ occurred synchronously with the closing (raising) of the lower lip. During the closure intervals for /p/, these were clipped and sustained movements on their position were observed in the figure. The jaw, however, gradually moved upward (closing) from the first vowel /e/ through the

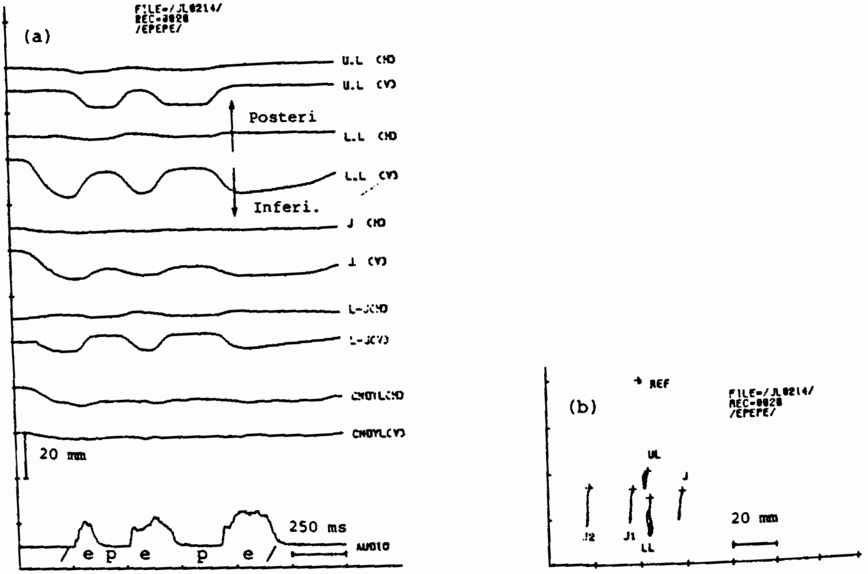


Fig. 5. Time patterns (a) and two-dimensional trajectories (b) of the articulatory movements in the utterance of /epepe/.

consonant /p/ and downward (opening) to the next /e/. During the first phase of jaw opening for /e/ from the closed (rest) position, the head of the condyle glided forward and somewhat downward. Since the consecutive movements of this point are very small compared with the corresponding jaw movements, the rotational movement of the mandible will apparently occur around the area on the head of the condyle.

The recording system presented in this paper may prove to be one of the most useful measuring techniques for the study of articulatory movements. By using this system, we are now conducting experiments to study the articulatory dynamics for several subjects.

References

Kiritani, S., Itoh, K., and Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *J. Acoust. Soc. Am.* **57**, 1516-1520.

Sonoda, Y., and Wanishi, S. (1982). New optical method for recording lip and jaw movements. *J. Acoust. Soc. Am.* **72**, 700-704.

Section 6
Physiological Manifestations of Speech Sounds

Tafkhim in Arabic: the Acoustic and Psychological Parameters

S.H. Al-Ani and M.S. El-Dalee
Bloomington, U.S.A.

1. Introduction

The phonetic phenomenon of 'tafkhim' (retraction) has been intensively investigated. As a categorical feature, tafkhim has been found to color any consonant or vowel, depending on the specific phonetic context.

The domain of distinctive features is considered one of the most important elements in the study of both phonetics and phonology. Accordingly, to describe a phonetic feature properly, is to delineate phonetic representations and behavior of linguistic sounds. It is apparent that the feature *retraction* is marked in the vowels by a definite raising of Formant one (F_1) and the lowering of Formant two (F_2). As for Formant three (F_3), it is used as a separate acoustic parameter for distinguishing some phonemes in the language.

2. Method

The method for obtaining the acoustic-phonetic results was determined through an experiment which was designed and carried out at the Phonetics Laboratory at Indiana University. A sequence of 220 utterances were read by a native speaker of Arabic from Alexandria, Egypt. The focus of the experiment was to investigate the unifying acoustic parameters of the Arabic consonants and vowels. The structure of each utterance of the data is the phonetic patterns /ʔ,VC,V/. Spectrograms were prepared of each utterance on a wide-band sound spectrogram machine (Voice Identification, Series 700). Measurements of the components of each utterance were made. The acoustic-phonetic cues of the vowels and consonants were measured and identified in the following manner.

3. Results

Vowel variable

The physical correlate of the retraction is manifested in the formants of the vowels with varying degrees. The vowel which shows the feature of retraction most significantly is /a/, and the least significant are the back vowels (See Fig. 1).

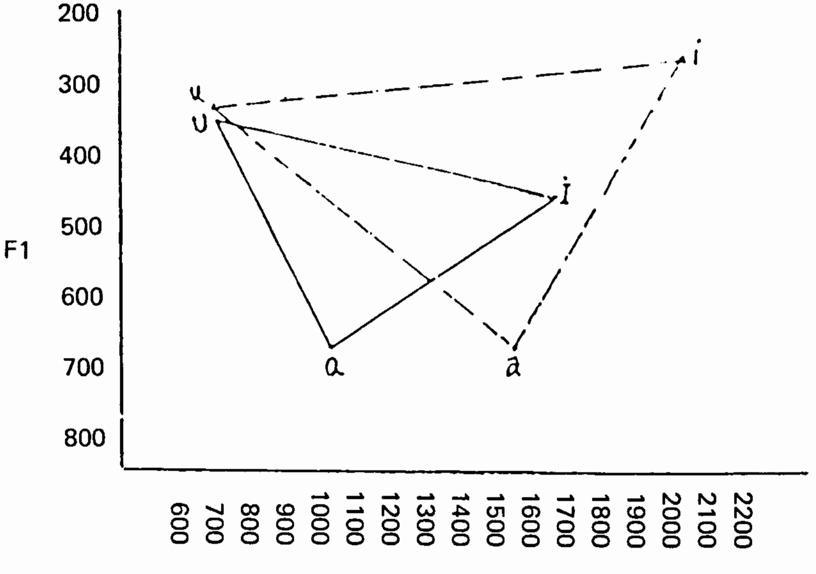


Fig. 1. Vowel diagram showing retraction, straight lines.

Formant variable

Both F_1 and F_2 possess a significant clue for the feature of retraction. But F_2 is not a consistent criterion since its significance fluctuates so that in some cases it is highly insignificant. (Consider the F-pattern in some phonetic classes of consonants in Figure 2).

Distance between formants

The F_2-F_1 difference is more significant than the F_3-F_2 difference. This is consistent with the weak significance of F_3 (See Tables I and II).

Table I. The mean measurements of vowels

	Onset		Steady		Offset			Onset		Steady		Offset	
	F_1	F_2	F_1	F_2	F_1	F_2		F_1	F_2	F_1	F_2	F_1	F_2
							V_1						
a	724	1581	743	1614	705	1655	ɑ	729	1029	736	1062	683	1060
i	269	2069	314	2086	314	2055	ɪ	457	1712	495	1667	469	1586
u	333	698	357	850	338	925	ᴜ	348	706	371	867	338	945
							V_2						
a	681	1629	721	1602	693	1571	ɑ	598	1081	707	1038	710	962
i	302	2017	279	2081	224	2131	ɪ	426	1498	433	1631	360	1874
u	343	995	317	760	336	652	ᴜ	305	960	310	771	350	625

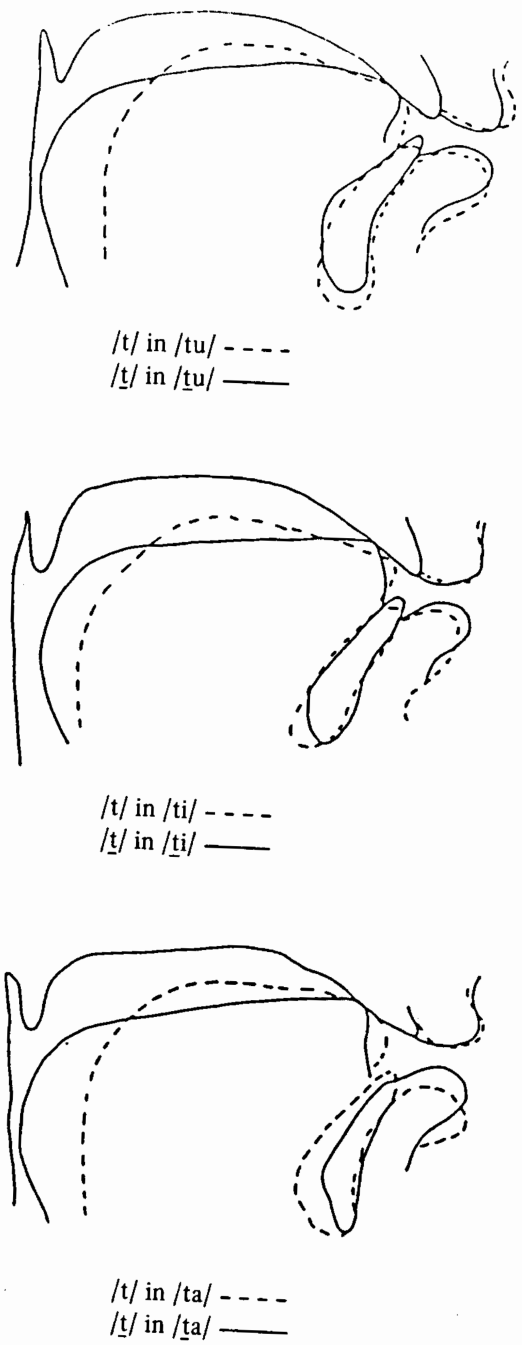
Table II.

					FD$_1$ (F_2 - F_1)						
V_1						V_2					
	Mean	SD		Mean	SD		Mean	SD		Mean	SD
a	871	(114)	ɑ	326	(240)	a	881	(161)	ɑ	331	(224)
i	1771	(64)	ɪ	1171	(171)	i	1802	(150)	ɪ	1198	(246)
u	493	(121)	ᴜ	495	(122)	u	443	(99)	ᴜ	462	(112)
F = 246.67						F = 24.82					
					FD$_2$ (F_3 - F_2)						
a	724	(971)	ɑ	1424	(189)	a	871	(781)	ɑ	1064	(947)
i	529	(68)	ɪ	757	(242)	i	526	(90)	ɪ	650	(524)
u	–283	(1122)	ᴜ	–512	(788)	u	–62	(1095)	ᴜ	2	(1292)
F = 10.42						F = 7.43					

SD = Standard deviation; F = Fisher test.

With regard to the manifestation of retraction on segments, the vowels used in this study fall into two groups. The front vowel /i/ shows significant changes in quality as measured by first and second formant frequencies. The back vowel /u/ shows insignificant changes between the plain and retracted allophones. The compact vowel /a/ is the most significant as it enhances and demonstrates clearly the color of the feature of retraction. The acute consonants are the consonantal category which shows most significantly the color of the feature under discussion (compare Table I with Table II).

The physiological phonetic parameters of retraction have been examined. As they became evident through the observation of the production of sequences of utterances, it was noted that the utterances contained the four independent emphatic consonants. Their conditioning of the neighboring sounds, specifically the vowels, was the focus of investigation. An experiment was designed and its utterances were recorded on a sound X-ray film at the University of Iowa in the Department of Otolaryngology and Maxillofacial Surgery. Tracings of certain selected frames were made. These tracings show the various configurations of the tongue movements associated with the emphatic sounds. It was indicated that clear proof exists of the lowering of the body of the tongue which is simultaneously correlated with the retraction movement of the whole body of the tongue, especially its back (See Figs. 2a, 2b, 2c).



/t/ in /tu/ - - - -
/t̠/ in /t̠u/ ——

/t/ in /ti/ - - - -
/t̠/ in /t̠i/ ——

/t/ in /ta/ - - - -
/t̠/ in /t̠a/ ——

Fig. 2. Tongue movements compared in normal vs. emphatic articulation of /t/.

4. Conclusion

The acoustic phonetic results correlate and agree with the physiological findings concerning the description of the characteristics of the phenomena of retraction. The combined findings of the acoustical and physiological investigations gave clear proof of how the phenomena of 'tafkhim' operates in Arabic.

References

Al Ani, Salman H. (1970). *Arabic Phonology: An Acoustical and Physiological Investigation.* Mouton and Co., The Hague, the Netherlands. Janua Linguarum, 61 (Series practica). Translated into Arabic, 1980.

Fant, G. (1970). Analysis and Synthesis of Speech Processes. B. Malmberg, ed. *Manual of Phonetics*, North Holland, Amsterdam.

Jakobson, R., Fant, G. and Halle, M. (1969). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates.* The MIT Press.

Jakobson, R. and Waugh, R. (1979). *The Sound Shape of the Language.* Indiana University Press.

Potter, K.P., Kopp G.A. and Green, H.C. (1947). *Visible Speech.* D. Van Nostrand Company, Inc., New York.

Coordinated Tongue Muscle Activity During /əpVp/ Utterances

P.J. Alfonso, K. Honda and T. Baer
New Haven and Storrs, USA

1. Introduction

There are many studies in the phonetic literature that describe the EMG activity of various tongue muscles during the production of vowels. However, there are relatively few studies that have simultaneously recorded muscle activity with hooked-wire electrodes from the complete set of tongue muscles. Thus, phoneticians have been forced to rely on different experiments, which use different speakers, with different dialects, often producing different speech utterances, to make assumptions about tongue control during speech.

2. Method

The purpose of the present experiment was to study control during vowel production by simultaneously recording EMG potentials from the complete set of muscles responsible for tongue movements. We recorded from the extrinsic tongue muscles: anterior and posterior genioglossus, (GGA and GGP), hypoglossus (HG), and styloglossus (SG). To our knowledge, EMG potentials from the HG muscle have never been recorded with hooked-wire electrodes before. In addition, we recorded from the accessory tongue muscles, the geniohyoid (GH) and mylohyoid (MG), and from orbicularis oris superior (OOS). We also monitored jaw position using measurement techniques similar to those described by Sonoda and Wanishi (1982) and we extracted formant frequency trajectories from the acoustic signal with the aid of an LPC-based, interactive computer system. The data were obtained while a single speaker of a New York dialect of American English produced ten repetitions of a randomized list containing eleven vowels in a /əpVp/ environment. Ensemble averages of the ten tokens were created for the EMG, acoustic, and jaw movement channels.

3. Results and Discussion

Figure 1 shows the acoustic vowel space plotted in average F_1-F_2 values for the eleven vowels produced in this experiment. Formant values were obtained by averaging across the ten repetitions per utterance type and then pooling the data points within an average formant trajectory. The acoustic data suggest that this subject produced the vowels /i/ and /e/ with a high and front tongue position, when compared to the remaining vowels. The vowels /I/ and /ε/ are lower and more centralized. The back group appears more tightly clustered.

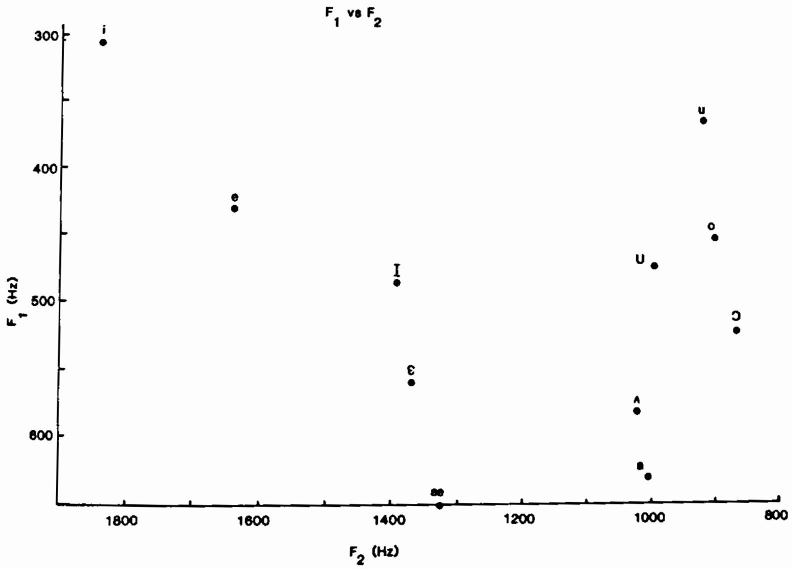


Fig. 1. Average F_1 - F_2 values for the eleven vowels produced in this experiment. Each data point represents the F_1 and F_2 values obtained by averaging both over time, during the vocalic segment, and over the ensemble of ten repetitions.

With the acoustic data in mind, we turn to the results of the EMG analysis. Figure 2 shows the ensemble averages for the EMG, acoustic, and jaw movement channels. The various channels are labelled across the vertical axis. Vowel type is shown across the top. For convenience, we have grouped front and back vowels, and show /Λ/ separately. The units along the abscissa represent 100 ms intervals. The heavy vertical lines represent the acoustic onset of the vowel, which served as the line-up point for each of the tokens. The EMG signals were smoothed with a 70 ms window. Audio amplitude is shown on the top row.

The extrinsic tongue muscles are shown on rows two through five. Considering these muscles as a group, note that vowel-related EMG potentials for GGA, HG and SG are more clearly differentiated in front versus back vowel height in both the front and back group. GGP activity appears to be inconstant, that is, EMG potentials for this muscle vary with vowel height in both the front and back group. These muscles are differentially organized to horizontally and vertically position the tonguė. Jaw movement contributes primarily to vertical tongue position. With these comments in mind, we next consider the function of each of the extrinsic and accessory muscles on

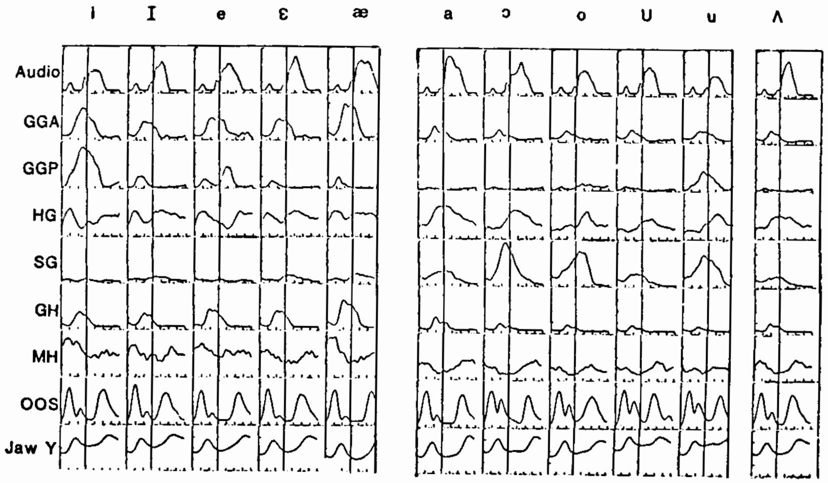


Fig. 2. Ensemble averages of the audio envelopes, electromyographic signals, and jaw vertical position signals for each of the eleven vowels. Electromyographic signals are shown for the anterior genioglossus (GGA), posterior genioglossus (GGP), hyoglossus (HG), styloglossus (SG), geniohyoid (GH), mylohyoid (MH), and orbicularis oris superior (OOS) muscles. Tick marks arbitrary. EMG data were smoothed using a 70 ms trianglar window and were averaged over ten repetitions. The vertical line through each set of plots indicates voice onset for the vowel, the temporal line-up point for ensemble averaging.

tongue positioning without taking into account the contribution of the jaw to tongue elevation.

Although the anterior and posterior fibers of the genioglossus are part of the same muscle anatomically, they can and do function independently in different vowel environments. GGA, shown on the second row, is much more active for front vowel than for back vowel production. GGP activity, shown on the third row, is most active during the production of the high vowels /i/, /e/, and /u/ than for the low vowels. Note that during the production of the high front vowel /i/, both GGA and GGP are active. Given the anatomical configuration of GGA fibers, this muscle should pull the tongue dorsum downward and forward. However, in the vowel /i/, the lowering component of GGA is reduced by activity of GGP, resulting in forward and upward movement of the tongue. During the production of the low front vowel /æ/, GGA, but not GGP, is active and during the production of the high back vowel /u/, GGP, but not GGA, is active. Thus, GGA contributes to tongue fronting and lowering, whereas GGP contributes primarily to tongue raising.

Next we consider the HG and SG muscles, shown on the fourth and fifth rows respectively. Note that these muscles show greater vowel-related activity for back vowels than for front vowels. During front vowel production, a peak in HG activity occurs at about the time of lip closure for the initial stop. This pre-vocalic activity possibly serves to lower the tongue during bilabial closure. Peak activation is followed by vowel-related suppression of the HG. The magnitude of HG suppression is directly related to the magnitude of GGP activity. SG activity is negligible in all front vowel environments. The activity of the HG and SG for back vowels is clearly different from that observed for front vowels. First, note the absence of the pre-vocalic HG activity. The HG is active during the production of all back vowels, maximum activity occurring during the production of /a/. The SG is also active for all back vowels, the greatest activity occurring for the high back vowels /u/, /o/ and /ɔ/. Thus, while HG acting alone pulls downward and backward on the tongue body , and SG acting alone pulls upward and backward on the tongue body, these two muscles in fact act synergistically to vertically and horizontally position the tongue during the production of all back vowels by varying the timing and magnitude of their contraction.

Finally we consider the GH and MH muscles, shown on the sixth and seventh rows respectively. These muscles exert a mechanical influence on the hyoid-larynx complex: However, we still discuss here the contribution of GH and MH to tongue positioning. Note first that GH and MH EMG activity is greater for front vowels than for back vowels. For front vowels, these muscles serve to raise the tongue base by stiffening the oral floor. For back vowels, activity of these muscles is suppressed. Thus, GH and MH assist in lowering and backing tongue gestures by relaxing and lowering the oral floor. The patterns of activity for these two muscles are clearly different. The GH muscle has greater effect on horizontal tongue movement than does the MH. Notice that the peak in GH occurs at the same time as the peak in GGA activity, indicating that GH and GGA assist in tongue fronting. The MH has a greater effect on vertical tongue displacement, especially for the posterior part of the tongue, than does the GH. Although we recognize that MH function is complex, we note that during EMG activity temporally associated with the vowel, MH suppression is related to the HG activity. Greater MH suppression occurs when HG activity is high, as in the back vowel group. On the other hand, overall MH activity is high when HG activity is lower, as in the front vowel group.

In summary, we have presented the preliminary analysis of simultaneously recorded tongue EMG data collected from a single speaker of American English. The purpose of the study was to investigate the muscular control of the tongue during vowel production. We note that there is EMG activity associated with tongue movements prior to and following the period associated with the vowel. This activity shows some systematic variation with vowel identity, and thus bears further consideration. Tongue movements are dependent upon the complex interdependencies among extrinsic and intrinsic tongue muscles, the accessory tongue muscles, the hyroid-larynx complex, and jaw position. Before we can ultimately understand the complex control of the tongue, we must be able to specify the individual function of these various parameters, The study reported here is an attempt to do that by examining the relationship of some of these muscles in the same speaker.

Acknowledgement

This research was supported by NIH grant NS13617 to Haskins Laboratories.

Reference

Sonoda, Y. and Wanishi, S. (1982). New optical method for recording lip and jaw movements. *J. Acoust. Soc. Am.* **72**, 700-704.

An Electromyographic Study of Laryngeal Adjustments for the Korean Stops in Syllable-Initial and Final Positions

H. Hirose, H.S. Park and M. Sawashima
Tokyo, Japan

1. Introduction

It is known that in Korean there is a three-member distinction in both manner and place of articulation that serves to differentiate nine stop consonant phonemes. For classification, the three types are generally referred to as ' forced' (Type I), 'lax' (Type II) and 'aspirated' (Type III). All stop types may occur in the syllable-initial position to be realized as voiceless, while in the medial position, the lax stops are usually manifested by voiced allophones. In syllable-final position, the three stop types are phonetically realized as voiceless 'applosives', being characterized by the absence of oral release.

The purpose of the present study is to investigate electromyographically the laryngeal adjustment for Korean stops both in the syllable-initial and syllable-final positions in various phonological conditions.

2. Procedures

Two native Korean speakers of the Seoul dialect served as the subjects. They read meaningful test words which were selected so as to place the stop consonants in different phonological environments, where (1) stop consonants are placed at the syllable-initial position and (2) they are placed at the syllable-final position and followed by an appropriate word starting with a syllable-initial stop. Electro-myographic (EMG) recordings were made using hooked-wire electrodes inserted percutaneously into the thyroarytenoid (VOC) muscle of the subjects. The recordings were made when the subjects read the utterance samples ten to twelve times each repeatedly. The recorded EMG signals were then reproduced and computer-processed after appropriate rectification and integration and average indication of the muscle activity of VOC was obtained over more than ten selected tokens of each test utterance type.

3. Results

1. EMG findings on syllable-initial stops

For both subjects, it was revealed that VOC activity was suppressed for each type of the stop consonants examined, the degree of which was slightest for the forced type and most marked for the aspirated type, while it was moderate for the lax type. VOC activity increased (reactivated) again toward the voice onset of the postconsonantal vowel after the suppression. The timing of the reactivation was earliest for the forced type followed by the aspirated and the lax types in that order. The peak of the reactivation was higher for the forced and the aspirated than for the lax type.

2. EMG findings on syllable-final stops

It was revealed that the pattern of VOC activity for those utterance types containing the syllable-final stop followed by the syllable-initial lax or forced stop was quite similar to that for the initial forced stop in terms of the degree and timing of VOC suppression for consonant segments, and of the degree and timing of VOC reactivation for postconsonantal vowels. EMG patterns were found to be similar even when the place of articulation of syllable-final applosives was different from that of the following syllable-initial stop. In contrast, when the syllable-final stop was followed by the aspirated stop, the pattern resembled that for the single syllable-initial aspirated stop, suggesting the occurrence of assimilation.

4. Discussion

The present study reveals that the three types of Korean stops in the syllable-initial position are characterized by different patterns of VOC activity. For the production of the aspirated type, VOC is markedly suppressed and then reactivated for the following vowel. Such activity patterns seem to correspond to a marked abduction gesture of the glottis for the aspirated type, which has been confirmed by fiberoptic observation (Kagaya, 1974).

EMG patterns for the lax type can also be considered to correspond to the glottal abduction gesture for this type of stop, in which a moderate degree of glottal opening is always observed by fiberoptic observation.

It has been observed by fiberoptic study that the glottal opening is smallest for the forced type among the three stop types of Korean, and the glottis tends to close earlier relative to the voice onset of the following vowel (Kagaya, 1974). Minimum suppression and early reactivation of VOC activity found in the present study for the syllable-initial forced stop seem to correspond at least to the temporal feature of the glottal dynamics for the forced type. However, the increase in VOC activity before the voice onset of the vowel following the forced type, the degree of which is relatively higher than that for the lax type and comparable to that for the aspirated type, cannot be explained by a simple dimension of glottal abduction-adduction. Rather, as already suggested by Hirose, Lee and Ushijima (1974), the relatively steep increase in VOC activity for the forced type must be taken as a characteristic feature of this type of Korean stop. This activity pattern may correspond to the acoustic feature of 'laryngealization' described by Abramson and Lisker (1972) and Ladefoged (1973), and can be a physiological correlate of the rapid intensity build-up after stop release which was found to be characteristic for the forced type by Han and Weitzman (1970).

Results of the present study also indicate that the patterns of VOC activity for the consonant clusters consisting of a syllable-final applosive and a syllable-initial forced or lax stop is quite similar to that of the syllable-initial forced type, regardless of the type of the syllable- final stop. It is claimed that when the syllable- final applosive is followed by an initial lax stop, the initial lax stop first changes into the forced type and then assimilates the preceding stop (Sawashima and Park, 1979). It is also shown that the pattern of VOC activity for the cluster type consisting of a syllable-final applosive followed by a syllable-initial aspirated stop resembles that of the syllable-initial spirated type. These EMG results are comparable to those of fiberoptic observations reported by Sawashima, Park, Honda and Hirose (1980), who claimed that for consonant clusters with an intervening word boundary the laryngeal feature of the final stop is assimilated to the following syllable-initial stop, regardless of the difference in the place of articulation.

It is reasonable to consider that the dynamics of glottal configuration observed by a fiberscope is under the muscle control of the larynx. The present study suggests that the pattern of VOC activity is an important physiological correlate for differentiating the three stop types of Korean as well as for phonetic realization of consonant clusters across an intervening word boundary in Korean with reference to the dynamic control of glottal configuration.

5. Summary

An EMG study was conducted to investigate the laryngeal control in Korean stop production. The results are summarized as follows:

1. The pattern of VOC activity appeared to characterize the three different types of Korean stops.
2. Specifically, for the production of the forced stop, VOC showed marked increase of activity with relatively earlier timing before the onset of the following vowel.
3. When the final stop was followed by the syllable- initial lax or forced stop, the EMG pattern of VOC for the cluster resembled that for the initial forced stop. In contrast, when it was followed by the aspirated stop, the pattern resembled that for the single syllable-initial aspirated stop.

References

Abramson, A.S. and Lisker, L. (1972); Voice timing in Korean stops. *Proceedings of the 7th International Congress of Phonetic Sciences, Montreal, 1971.* The Hague: Mouton, pp. 439-446.

Han, M.S. and Weitzman, R.W. (1970); Acoustic features of Korean /P,T,K/, /p t,k/ and /p^h t^h, k^h/. *Phonetica* **22**, 112-128.

Hirose, H., Lee, C.Y. and Ushijima, T. (1974); Laryngeal control in Korean stop production. *J. Phonetics* **2**, 145-152.

Kagaya, R. (1974); A fiberscopic and acoustic study of the Korean stops, affricates and fricatives. *J. Phonetics* **2**, 161-180.

Ladefoged, P. (1973); The features of the larynx. *J. Phonetics* **1**, 73-83.

Sawashima, M., and Park, H.S. (1979); Laryngeal adjustments for syllable-final stops in Korean. *Ann. Bull. Research Institute of Logopedics and Phoniatrics, University of Tokyo*, **13**, 83-89.

Sawashima, M., Park, H.S., Honda, K. and Hirose, H. (1980); Fiberscopic study on laryngeal adjustments for syllable-final applosives in Korean. *Ann. Bull. Research Institute of Logopedics and Phoniatrics, University of Tokyo*, **14**, 125-138.

Glottal Timing in German Voiceless Occlusives

P. Hoole, B. Pompino-Marschall and M. Dames
Munich, FRG

1. Introduction

The coordination of laryngeal and lingual gestures in voiceless sounds has proved a fruitful area in which to study the motor organisation of speech (e.g. Löfqvist, 1980). However, little information is yet available on German. Preliminary investigations with one speaker, using isolated nonsense words of the type 'CV, suggested that both place of articulation of the consonant and height of the following vowel have an influence on the length of the interval from the moment of peak glottal opening (PGO) to release of the oral occlusion (/p/ > /t/, /i/ > /e/ > /a/).

The purpose of the present study was threefold; (1) to reproduce the place of articulation and vowel height results using material more typical of German word structure; (2) to determine to what extent differences in stress result in articulatory reorganisation, (3) to compare fricatives and stops.

2. Method

Glottal activity was registered by means of the photo-electric glottograph (F-J Electronics).

Oral air-pressure was recorded with a Hansen Manometer via a catheter inserted through the nose into the pharynx. The air-pressure trace was principally used as a means of identifying the moment of oral closure. These two signals together with the audio signal were digitalized and stored on magnetic tape. After recording, a segmentation program was used to measure the durational parameters of interest; the most important being Implosion to Glottal Abduction (GA), PGO to Release, the oral occlusion, and aspiration.

3. Material and Subjects

Six pseudo-German words with stress on the first syllable (/pi:pə/, /ti:tə/, /pe:pə/, /te:tə/, /pa:pə/, /ta:tə/) were embedded in the sentence frame 'Ist der fette/feste-da?'. Each word was produced ten times in randomized order, i.e. five times each after 'fette' and 'feste', the latter two words providing the material for the plosive/fricative comparison. Three male native-speakers from different parts of Germany acted as subjects.

4. Results

The pre-stress plosives

To set the scene the average values for occlusion and aspiration for the three subjects are given in Table I. The values conformed to expectations. The subjects had very similar occlusions but differed somewhat in aspiration. Fig. 1 shows the results for PGO to Release. For all subjects this interval is longer for /p/ than for /t/, quite clearly for HGT and KD, slightly less so for BPM who has rather large standard deviations. However, the effect of the following vowel shows no clear pattern. In view of the results for the consonants we tested for a correlation between PGO to Release and Occlusion. HGT and KD had highly significant correlations but BPM did not. The values for Implosion to GA were fairly constant over the different vowels and consonants:

BPM: 20 ms, sd = 4 ms; HGT: 35 ms, sd = 10 ms; KD: 58 ms, sd = 13 ms.

Most striking here is the very high value for KD.

The Post-Stress Plosives

The results for these plosives showed the widest range of strategies among the subjects. All subjects had significantly shorter occlusions for the post- stress plosives and BPM and HGT also had much less aspiration but KD surprisingly did not.

Table I. Pre-stress plosives. Length of occlusion and aspiration in ms

		/pi/		/ti/		/pe/		/te/		/pa/		/ta/	
		mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
BPM	Occ	139	12	126	9	133	6	129	12	138	19	120	5
	Asp	62	10	74	7	44	5	60	13	45	8	53	12
HGT	Occ	137	14	129	21	134	17	114	12	122	14	98	9
	Asp	38	8	55	13	26	8	41	6	29	15	30	5
KD	Occ	144	20	115	15	144	17	112	13	127	14	104	10
	Asp	22	10	46	10	18	4	34	5	16	2	25	3

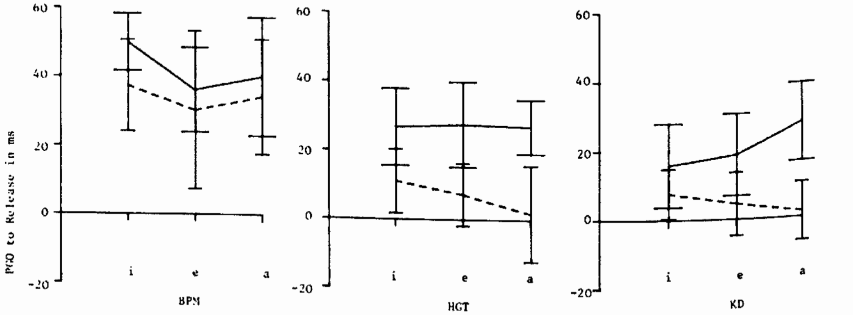


Fig. 1. Mean and sd for PGO to Release for the pre-stress stops; /p/ with unbroken line and with long horizontal bars delimiting the sd; /t/ with dashed line and with short bars.

The glottal parameters: for HGT no reliable measurments could be made as only slight glottal activity was observable. However, the air-pressure trace retained the typical form for voiceless plosives (quadrilateral) so we would guess that the subject actively slackened his vocal folds but adjusted the cartiliginous part of the glottis only very slightly. PGO to Release in the other two subjects showed the same consonantal dependencies as in the pre-stress case displayed in figure 1 but had overall lower values (a negative value means PGO comes after the release):

BPM /p/ mean = 13 ms, sd = 9 ms; /t/ mean = 0 ms, sd = 10 ms

KD /p/ mean = 12 ms, sd = 7 ms; /t/ mean = -8 ms, sd = 7 ms

This can probably be regarded as a fairly simple spin-off from the shorter occlusions. These results suggest that PGO amplitude must also have been less and a Wilcoxon test tended to confirm this. Implosion to GA revealed a striking example of articulatory reorganisation for KD. While BPM's values were much the same as in the pre-stress case (mean = 26 ms, sd = 10 ms) KD's values differed drastically: mean = 17 ms, sd = 8 ms.

This reorganisation was refected in the air-pressure curve; in the pre-stress case where tongue and glottal gestures are widely separated the air- pressure first rises quite steeply, followed by a shallower rise from the moment when the glottis begins to open up to the absolute pressure maximum. In the post-stress case where the gestures almost coincide there is a single, very steep rise.

'fette' vs. 'feste'

Other investigators (e.g. Yoshioka, Löfqvist and Hirose, 1980) have reliably found that fricatives are distinguished from plosives by a glottal opening gesture that is more vigorous and starts earlier with respect to oral closure. The present study confirmed this. Our specific interest here was to try and develop a simple motor plan for the vowel + consonant dyad (/et/ vs. /est/) by taking into account the above observation and the length of the vowel preceding the consonant (Fig. 2). The results for BPM and KD are given in Table II. BPM shows a very neat coincidence of the instants of GA for fricative and plosive when measured from the preceding vowel onset. KD has a similar overall trend but the coincidence is far from perfect. This can be attributed in part to his reorganisation of GA for post-stress plosives. HGT showed similar results for vowel length but once again the glottal signal for the post-stress plosive was too weak to be evaluated. This subject and KD also spoke an additional set of 12 sentences at a more deliberate pace (thus including 6 examples wach of 'feste' and 'fette'). The results, in Table III, for this very small sample showed some support for the hypothesis.

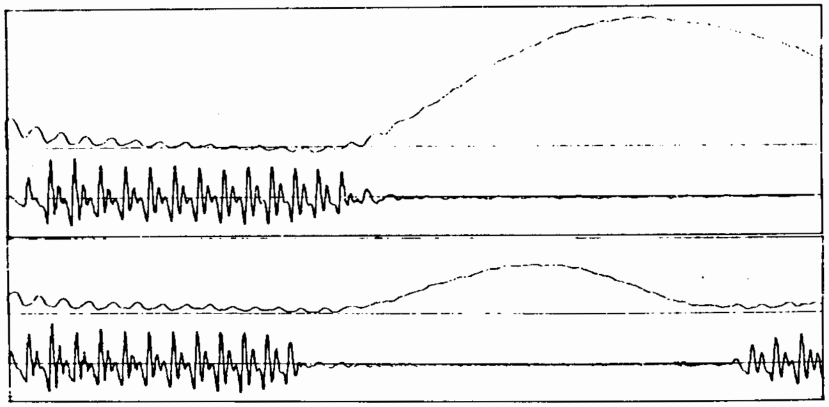


Fig. 2. Glottis signal and audio signal for /est/ (top) and /et/ (bottom), lined up at vowel onset to show the coincidence in GA.

Table II. 'feste' vs. 'fette'. Complete material

		Vowel duration		Implosion to GA		Vowel onset to GA	
		mean	sd	mean	sd	mean	sd
BPM	/est/	88 ms	5 ms	2 ms	3 ms	90 ms	6 ms
	/et/	69 ms	6 ms	19 ms	3 ms	88 ms	5 ms
KD	/est/	83 ms	4 ms	13 ms	4 ms	95 ms	5 ms
BPM	/et/	69 ms	5 ms	17 ms	4 ms	85 ms	6 ms

Table III. 'feste' vs. 'fette'. Additional material

		Vowel duration		Implosion to GA		Vowel onset to GA	
		mean	sd	mean	sd	mean	sd
BPM	/est/	196 ms	26 ms	6 ms	10 ms	202 ms	31 ms
	/et/	163 ms	26 ms	51 ms	13 ms	215 ms	28 ms
KD	/est/	106 ms	5 ms	14 ms	6 ms	121 ms	8 ms
	/et/	93 ms	12 ms	25 ms	7 ms	122 ms	9 ms

5. Discussion

Both the pre- and post-stressed plosives showed differences in the timing of PGO relative to release depending on place of articulation. This is a finding that theories of aspiration for German (e.g. Haag, 1979) must take into account. The difference in timing can be regarded as a largely passive effect of the difference in occlusion. A rather stereotyped glottal gesture is superimposed on a variety of oral articulations. A more active retiming would have had to be assumed if the following vowel had had an effect but no convincing evidence for this was found.

In a sense, the fricative/plosive comparison leads to the same conclusion. Within certain limits the articulatory mechanism seems to prefer to reorganize oral gestures rather than glottal ones. Though not perfect the coincidence in the timing of GA is worth investigating further and could lead to a simpler interpretation of the finding made for Japanese (Yoshioka, Lofqvist and Hirose, 1980) that the moment of maximum speed of glottal opening in fricatives and plosives coincide.

Of course, there is still a large fly in the ointment, namely KD's 'unnecessary' retiming of GA for post-stress plosives. This is a stable characteristic of his accent and manifests itself acoustically in an extremely abrupt decay in vowel amplitude before post-stress plosives.

References

Haag, W.K. (1979). An articulatory experiment on Voice Onset Time in German stop consonants. *Phonetica* **36**, 169-181.

Löfqvist, A. (1980). Interarticulator programming in stop production. *Journal of Phonetics* **8**, 475-490.

Yoshioka, H., Löfqvist, A. and Hirose, H. (1980). Laryngeal adjustments in Japanese voiceless sound production. *Haskins Laboratories SR-63/64:* 293-308.

Assimilation of Voice in Dutch

I.H. Slis
Nijmegen, the Netherlands

1. Introduction

In the Dutch literature on the assimilation of voice - which stems for the larger part from before 1960 and is therefore mainly based on non-instrumental investigations - a number of controversies can be observed. Yet, the following generalisations which are relevant for assimilation of voice can be made, be it with some reservation:

a. As a rule, syllable-final obstruents are voiceless in Dutch. Therefore, assimilation of voice can only occur if a voiceless syllable-final obstruent is followed by a voiced syllable-initial consonant.
b. If the second consonant of a two-consonant cluster (= C2) is a voiced fricative, progressive assimilation can nearly always be observed.
c. If C2 is a voiced plosive, regressive assimilation can be observed in the majority of cases.

Relatively speaking, the third rule is the most controversial one: also, progressive assimilation is reported frequently in clusters with a voiced plosive as C2. A number of causes are suggested for this devoicing, such as the emotional state and sex of the speaker, stress position, informational load of the following word and the phonological composition of the cluster.

Our major objective in the present study is to survey various influences on assimilation of voice. To this end, we make a distinction between:

a. non-linguistic influences such as organic conditions of the vocal cords, and
b. linguistic influences, such as the phonological or stress context.

1.1. Non-linguistic influences

At least some of these factors concern the way in which the vocal cords vibrate. In this paper we will pay attention to the influence of vocal pitch, voice quality and sex of the speaker.

Voice activity depends on conditions within the larynx. If, for example, the vocal cords are too far apart or too tense no voicing will occur. During the production of a consonant cluster the airflow through the oral passage can be obstructed to such an extent that the boundary conditions for vocal fold vibrations are approached. We presume that in a number of cases raising the pitch by increasing the tension of the vocal cords leads to a situation in which these boundary conditions are surpassed, and consequently to an increase in the number of interruptions of voicing. If these interruptions start within 50 ms after the preceding vowel and continue until the moment of oral disclosure (see operational definition below), progressive assimilation takes place. Another factor that can be expected to influence boundary conditions of vocal cord vibration is the organic state of the vocal apparatus. It is very likely that in speakers with a poor voice the organic state is less favorable for vocal cord vibration than in speakers with a good voice.

In addition we expect differences in assimilation between the speech production of men and women, since male and female larynges are known to be of different size.

Another factor which might influence the degree of assimilation is speech rate. Successive speech gestures will overlap to a greater extent with increasing speech rate, resulting in a higher proportion of assimilated clusters.

1.2. Linguistic influences

Less obvious is the relation between voicing and assimilation on the one hand and linguistic variables on the other. In this paper we will study the influences of place and manner of articulation, linguistic stress and the phonological length of the preceding vowel.

From a comparison of the voice onset times of dental and labial plosives it appears that there is a relationship between place of articulation and voicing. In almost all languages studied voice onsets in dentals are delayed compared with those in labials (Lisker and Abramson, 1964).

As far as manner of articulation is concerned, differences with respect to voicing are to be expected between clusters containing a fricative and a stop and clusters containing only stops. By means of transillumination of the glottis we observed a larger glottal aperture during the production of fricatives than during that of stops (Slis, 1970). This may have consequences for the assimilation of voice, because the larger glottal opening may lead to a higher proportion of interruptions of voice activity.

Since stress seems to be effected by a greater articulatory effort (Slis, 1971), we expect that the voice character of a consonant belonging to a stressed syllable dominates one belonging to an unstressed one. This implies that in pre-stress condition, in which the voiced C2 belongs to the stressed syllable, regressive assimilation is most probable. On the other hand we expect progressive assimilation to occur when the voiceless C1 belongs to the stressed syllable.

Furthermore, voiceless consonants tend to be longer and are pronounced with more effort after short than after long vowels. Consequently, we expect a greater prominence of the voiceless element in clusters after short than after long vowels which will lead to more progressive assimilation after short vowels. In view of the above relations we tentatively suggest that assimilation of voice is not a mentally programmed process, but rather the result of unintentional articulation.

2. Method

2.1. Independent variables

We investigated the following variables which are directly related to voice activity or coarticulation:

1. Pitch: normal intonation vs. low, medium and high monotonous speech.
2. Voice quality of the speakers: good vs. poor quality.
3. Sex of the speakers: male vs. female speech.
4. Speech rate: slow, normal and fast speech.

In addition, we studied some other variables whose relationship with voice activity is less obvious, but which may contribute as argued above:

5. Phonological composition of the clusters with respect to:
   a. manner of articulation: stop-stop vs. fricative-stop.
   b. place of articulation: /p,t,k,f,s,χ/ followed by /b/ or /d/.
6. Stress conditions: before or after stressed syllable or without stress on the adjacent syllables.
7. Phonological ength of the preceding vowel: /a:,o:,ø:,e:/ vs. /ɑ,ɔ,œ,I/.

2.2. Speech material

The production of words can be regarded as the performance of sets of coherent motor commands. In this light, assimilation within words may be regarded as engrained motor patterns. This is not the kind of assimilation we want to study in this paper. We will restrict ourselves to assimilation across word boundaries, where a 'new program' is needed since each combination is unique. Clusters in which C2 is a fricative are also excluded from this report. Since nearly 100% of these clusters show progressive assimilation (Slis, 1981) we do not expect that the few deviations from the rule will significantly contribute to our experiment.

For these reasons we only performed experiments on fricative-stop and stop-stop clusters across word boundaries, the second (stop) consonant being voiced. Seven different series of measurements were carried out on clusters in sentences that were read aloud. In each of the series between 20 and 40 sentences were recorded of between 20 and 45 subjects. In this way we obtained a total of 4661 two-consonant cluster-tokens.

2.3. Measurements

From preliminary observations we knew that assimilation, either regressive or progressive, need not always be complete: a large number of intermediate cases occurred. We therefore needed a strict definition of regressive and progressive assimilation. This definition should be based on the general definition which states that we speak of assimilation if one sound segment exerts influence upon the articulation of another, so that the two sounds become more alike or even identical (Crystal, 1980:35).

From U.V.-oscillograms (paper speed 50 or 100 mm/s) of the speech signal we derived the moments of closing and opening of the vocal tract from the moments of sudden changes in the amplitude. From electrolaryngograms we deduced the moments of opening and closing of the glottis; in nearly all cases these moments occurred at the end and the beginning of observable voice activity. In few cases the amplitude of the larynx signal was not completely reduced to zero. We assumed that the low-amplitude part of the electrolaryngogram depicted edge vibrations of the vocal cords with an open glottis. In these cases we made a 'best guess' based on the slope of the amplitude envelope.

From previous measurements on intervocalic voiceless plosives (Slis, 1970) we learned that voicing may continue about 20 ms after the moment of closing the oral tract. We called this continuation of voice a 'voice tail'. The moment of voice onset (VOT = voice onset time) occurred at or after the moment of opening of the vocal tract. We defined a syllable- final obstruent as voiceless when the voice tail was shorter than 50 ms, and an initial consonant as voiceless when the VOT took place at or after the moment of opening the vocal tract. Although these moments could not be determined with a high degree of precision (glottal period duration 5 to 10 ms) only a few cases proved to be ambiguous.

To sum up we arrived at the following definitions of assimilation categories:

1. No assimilation: The voice tail is less than 50 ms and VOT is negative.
2. Regressive assimilation: Voicing continues during the closed interval or the voice tail is longer than 50 ms. (interruption of voicing after so long a voice tail is attributed to aerodynamic causes and not to glottal opening).
3. Progressive assimilation: The voice tail is less than 50 ms and the VOT is zero or positive.

3. Results

3.1. Pitch

In line with the expectation mentioned above the results showed an increase of progressive assimilation with increasing pitch from 20% to 32% ($\chi^2 = 24$, df = 4, p < .001). Additionally, more progressive assimilation was found in low-pitched monotonous speech (20%) than in normally intonated speech (10%). We assume that this is brought about by a higher tension of the laryngeal musculature in monotonous speech in order to keep the larynx in a fixed position.

3.2. Good vs. poor voices

The results of our experiments, based on 1103 clusters, showed about twice as much progressive assimilation in poor (27%) as in good voices (13%, $\chi^2 =$ 44, df = 2, p < .001). This result confirms our expectation on this point.

3.3. Male vs. female voices

The differences in assimilation between the speech of men and women were studied in four experiments. In total 708 clusters of female speakers were obtained. These were compared with the pooled results of the male speakers in all experiments. In three of the four experiments, the results showed that women have significantly more progressive assimilation (45%) than men (25%) in all the conditions examined ($\chi^2 = 140$, df = 2, p < .001).

3.4. Speech rate

In one experiment (973 clusters) we measured the influence of speech rate in clusters in a pre-stress position comparing slow, normal and fast speech. With all cluster types studied we found a significant increase in degree of assimilation going from slow, via normal, to fast speech ($\chi^2 = 122$, df = 4. p < .001). This increase was from 51% to 92% when C2 was a stop and from 77% to 100% when C2 was a fricative. Normal speech took an intermediate position. No remarkable change in direction of assimilation could be observed.

3.5.a. Phonological composition with respect to manner of articulation

A comparison of stop-stop with fricative-stop clusters for the summed data of all experiments shows highly significant differences in assimilation ($\chi^2 =$ 162, df = 2, p < .001) in all the conditions investigated, i.e. before, after and without stress, in the speech of both men and women. More assimilation was found in stop-stop clusters (83%) than in fricative-stop clusters (77%). In fricative-stop clusters we observed more progressive assimilation (40%) than in stop-stop clusters (30%).

3.5.b. Phonological composition with respect to place of articulation

Clusters in which C2 was a /d/ showed more progressive assimilation (38%) than those with a /b/ (32%) in all the conditions investigated (in which manner of articulation of C1, the stress condition and the sex of the speakers, were variables). Five out of eleven differences were significant ($\chi^2 = 60$, df = 2, p < .05).

3.6. Stress condition

In all the conditions investigated (4236 clusters) we observed highly significant differences in assimilation ($\chi^2 = 276$, df = 2, p < .001): before stress we found twice as much regressive assimilation (54%, vs 24% progressive assimilation). Conversely, we found that after stress the frequency of progressive assimilation (50%) was twice that of regressive assimilation (19%). This confirms our expectation.

3.7. Phonological length of the preceding vowel

In one of the experiments we investigated the influence of the length of the preceding vowel on the assimilation of a cluster. Although we did not find significant differences, there was a tendency towards a higher frequency of assimilation after long stressed (65%) than after short stressed vowels (56%, $\chi^2 = 5.5$, df = 2, p<.10). This was mainly attributable to the higher frequency of regressive assimilation (26% after long and 13% after short vowels). This result is in line with our prediction mentioned above.

4. Discussion and conclusion

All the expectations put forward in the introduction are confirmed by the data. These expectations were based on the assumption that assimilation of voice is a result of unintentional coarticulation rather than a mentally programmed process. Alternative explanations for our results are possible. However, they are heterogeneous; for one aspect there is one alternative, for another we have to present another alternative. For instance, with respect to the influence of pitch, the only plausible explanation is the organic state of the vocal cords as put forward in the introduction. Differences in assimilation in the speech of men and women, however, may also be a sociolinguistic phenomenon. similar to other sex differences in language (e.g. Smith, 1979). Differences due to linguistic influences may also be the result of the intentional application of different rules depending on the composition of the cluster, or its phonological context.

Although no convincing proof can be given that all the effects which we observed are consequences of coarticulation, we feel attracted to an esplanation that covers all oppositions studied; coarticulation seems to be the only possibility.

References

Crystal, D. (1980). *A first dictionary of linguistics and phenetics.* London: André Deutsch.

Lisker, L. and Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word,* **20**, 384-422.

Slis, I.H. (1970). Articulatory measurements on voiced, voiceless and nasal consonants: A test of a model. *Phonetica,* **21**, 193-210.

Slis, I.H. (1971). Articulatory effort and its durational and electromyographic correlates. *Phonetica,* **23**, 171-188.

Slis, I.H. (1981). The effect of speaking rate on assimilation of voice. *Proceedings of the Institute of Phonetics Nijmegen,* **5**, 160-176.

Smith, P.M. (1979). Sex markers in speech. In: K.R. Scherer and E. Giles (eds.) *Social markers in speech.* Cambridge: Cambridge University Press, 109-146.

English Diphthongs, [ai, oi, ou]

I. Solomon and S.J. Sara
Washington, D.C. USA

A diphthong, by definition, implies a sequence of two different vocalic segments. Vocalic segments are characterized by the vocal tract configurations that are effected by the movements and positions of the tongue and lips. It is part of linguistic tradition to describe each vocalic segment in terms of the tongue raising or lowering, tongue advancement or retraction and lip positions.

The purpose of this presentation is not to discuss all the vowels, but rather to narrow the discussion to the three diphthongal sequences in American English: [ai, oi, au], as they occur in words like 'buy, boy, bough' respectively. Even within this narrow scope, the focus will be on the second element of these diphthongal sequences, i.e. [i,u].

There is no uniformity among phoneticians on how these sequences are perceived, and consequently these perceptions are reflected in the variety of phonetic transcriptions utilized, e.g. [ai, aι, ay; oi, oι, oy; au, aω, ay] by Jakobson et al (1952), IPA (1949), and Chomsky and Halle (1968), respectively. In adaption from Smalley (1964), the diphthongal sequences are traced on the vocalic chart below in Figure 1.

Since the variations in the phonetic transcriptions are due to perceptions of different tongue positions, palatography provides data that determine the positions of the tongue in the production of diphthongs and vowels. The present experiments used artificial palates to record tongue contacts with the palate, as described in Sara (1979). It included seven female and five male native speakers of American English to measure tongue height as indicated by channel width (measurement a), and tongue advancement as indicated by (measurement b) in Figure 2.

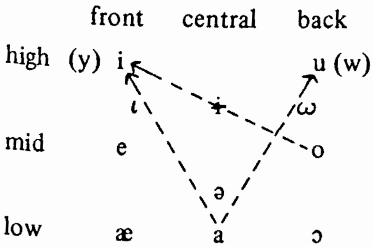


Figure 1. Diphthongal movement.

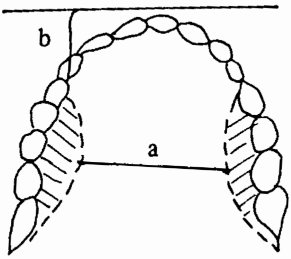


Figure 2. Channel width & Tongue advancement.

In the production of diphthongs and vowels, the higher the tongue the shorter measurement (a), and the more advanced the tongue, the shorter measurement (b) should be. The measurements from the thirteen subjects for the three diphthongs and the vowels [1, ι, e, ε, u, ω, o] as produced in the words: 'peep, pip, babe, pep, hoop, hoof, pope' respectively, are tabulated in Table I below. All measurements are to the nearest milimeter.

In correlating the measurements obtained for individual high and mid vowels with those for diphthongs as displayed in Table I. One can see how channel width (measurement a), and tongue advancement (measurement b) vary for each individual vowel in comparison with those for diphthongs. Table II tabulates the individual comparisons in terms of diphthongal measurements being 'wider' than, 'equal' to, or 'narrower' than those of individual vowels.

It is evident from the statistical comparisons in Table II that the tongue is lower and its advancement/retraction is less for diphthongs than for the high vowels. Based on the data of Table 2 , it is accurate to say that the tongue motions in the production of the diphthongs [ai, oi, ou] reaches the mid vowel position twice as often as it reaches the high vowel position. The diphthongal motions, then, are more appropriately represented by the solid line triangle than by the traditional phonetic transcriptions as represented by the broken lines of Figure 3.

Table I. a. Channel width; b. Tongue advancement

		i	ι	e	ε	ai	oi	u	ω	o	au
1	a	18	31	24	31	30	31	30	30	—	26
	b	19	25	14	23	24	26	24	21	—	22
2	a	15	33	27	29	15	22	33	32	35	37
	b	22	24	26	30	24	22	48	51	54	53
3	a	10	23	18	20	20	20	19	23	19	19
	b	9	20	15	22	16	23	43	48	49	41
4	a	19	26	19	25	22	19	—	—	—	—
	b	35	47	38	39	43	42	—	—	—	—
5	a	20	25	21	23	25	30	22	—	—	33
	b	14	19	15	25	17	27	40	46	—	23
6	a	10	25	18	24	24	31	17	25	24	34
	b	12	24	15	14	18	49	24	48	48	50
7	a	21	29	28	28	31	27	29	32	28	—
	b	20	19	26	33	25	28	39	39	41	—
8	a	13	26	25	27	29	28	25	—	—	—
	b	16	24	21	25	47	48	50	—	—	—
9	a	12	19	19	24	22	22	24	25	29	29
	b	10	16	19	16	25	25	30	33	46	43
10	a	18	24	23	30	20	24	44	44	—	43
	b	10	11	14	16	10	23	31	34	—	31
11	a	20	31	27	36	29	31	32	33	39	37
	b	17	19	19	31	22	32	41	38	41	37
12	a	20	27	25	33	32	30	31	32	32	35
	b	10	13	19	26	26	25	31	27	27	25
13	a	21	28	22	29	36	25	28	—	—	32
	b	14	21	22	23	27	28	27	—	—	25

Table II. Tabulated comparisons

Diph	Measure	Degree	Vowels				Total
			i	ι	e	ε	
[ai]	a	wider equal narrower	12 01 00	05 01 07	10 00 03	04 02 07	31 04 17
[ai]	b	wider equal narrower	12 01 00	06 01 06	10 00 03	06 01 06	34 03 15
[oi]	a	wider equal narrower	12 01 00	05 03 02	10 01 01	03 02 03	30 07 06
[oi]	b	wider equal narrower	12 01 00	11 00 02	12 00 01	10 00 03	45 01 06
			u	ω	o		
[au]	a	wider equal narrower	07 01 00	05 00 02	03 02 01	03	15 03 06
[au]	b	wider equal narrower	03 01 06	04 00 05	01 00 05		08 01 16

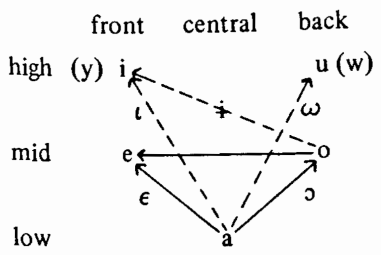


Figure 3. Diphthongal motions: --- predicted; ——— obtained.

References

Chomsky, Noam, and Morris Halle (1968). *The Sound Pattern of English*. New York: Harper and Row Publishers.

Jakobson, Roman, Morris Halle, and Gunnar Fant (1952). *Preliminaries to Speech Analysis*. Cambridge, Mass. The MIT Press.

The principles of the International Phonetic Association (1975). London: International Phonetic Association.

Sara, Solomon I. (1979). Vocalic Variability in Palatographic Impressions. In: Hollien and Hollien, eds. *Current Issues in Phonetic Sciences*. Amsterdam: John Benjamins. pp. 435-442.

Smalley, William A. (1964). *Manual of Articulatory Phonetics*. rev. ed. New York: Practical Anthrophology.

Section 7
Psychoacoustics of Speech

Differentiating Between Speaking and Singing Vocal Registers

H. Hollien
Gainesville, USA

1. Definitions

It is now reasonably well established that attributes appropriately identified as vocal registers exist in the human voice. Generalized (if not operational) definitions can be seen in the literature; one of the earliest being that of Garcia (1840) who indicated that a voice register is ... 'a series of succeeding sounds of equal quality, a scale from low to high produced by the application of the same mechanical principle, the nature of which differs basically from another series of succeeding sounds of equal quality produced by another mechanical principle.' In the years to follow, many vocal pedagogists, laryngologists and phoneticians offered like definitions; virtually all described voice registers in much the same terms as did Garcia (see for example Appelman, 1967; Fields, 1970; Hollien, 1974; Large, 1972; Preissler, 1939, Ruth, 1963; Vennard, 1962). It would be of little value to list any portion of these hundreds of definitions or labels (see Morner et al., 1964, for examples of terms) as they add little to the rather simplistic concept articulated by Garcia. Our use of the term simplistic is not intended to be judgmental or negative. Rather, it indicates that Garcia's definition, while a good one, does little but scratch the surface of the issue. That is, it must be asked: what are the scales to which Garcia refers? And.... what series, what qualities, what mechanisms are involved? Admittedly, Garcia implies that the fundamental frequency level of the sung tone is one of the controlling elements. However, until all of the questions are answered, his definition must necessarily be viewed as superficial, or at least, incomplete.

In over a century, the cited definition has been improved upon very much. Many scientists have tended to coin, or at least be sympathetic to, definitions of the type offered by Hollien and his associates (1974, 1976) who suggest that a register is a 'series or range of consecutively phonated frequencies which can be produced with nearly identical vocal quality... that there should be relatively little overlap between adjacent registers and that, to be a vocal register, the mechanism should be laryngal in nature'. Thus, in some ways, they extend Garcia by postulating an entire second set of 'vocal tract' based registers that are parallel to, or overlap, vocal registers of laryngeal origin. However, questions again can be asked – what frequencies, qualities, mechanisms? In response, both Hollien (1974) and Titze (1980) insist that definitions of this type add only little to the knowledge of registration. Indeed, as Titze points out, it is necessary to describe the entire production – and with respect to as many levels and dimensions as is possible, 'starting ...from the neuromuscular level to the biomechanical level, to the kinematic level,... to the aerodynamic and to the acoustic levels, and finally, to the perceptual level,' Titze warns against relating acoustics to acoustics. Finally, and in the same vein, it is suggested that, if registers are to be understood, they must be operationally defined with respect to their perceptual, acoustic, physiological, aerodynamic and neurological elements/bases.

On the other hand, there are a number of vocal pedagogists who have taken the position that voice registers do not exist within the *singing voice*; several references/arguments could be cited in this regard (see Fields, 1970, for example). However, Johnson (1982) articulates the argument succinctly when he suggests: 1) that it is only the *untrained* singer who distorts productions in such a manner that unrelated (register) sounds are produced, 2) that many great singers developed their voices without even being aware of the concept of registers, 3) that 'smoothness of scale and tone' more functionally relate to good development of voice rather than does training 'pieces' of the voice and 4) that the so-called registers could be the result of *illusions* based on singers feeling vibrations in their chests (or head) when they sing certain frequencies. There is little doubt but that successful singers – at least those trained in the classical 'western' opera or concert mode – are able to conceal register differences when they sing. Indeed, while Sundberg (1982) agrees with Johnson to some extent (as do we), he points out (as do others), that some forms of singing depend on a singer's ability to covary 'articulation, subglottal pressure and formant frequencies' with phonatory frequency. Specifically, Sundberg stresses that register usage can be an important part of technique in certain types of singing. That such instances exist is conceded by Johnson. However, she argues that 'distortions' of this type can be, and probably are, dangerous to the singer

The cited controversies led Hollien and his associates (1966, 1968, 1974, 1976, 1982) to attempt to provide new perspectives for the study of voice registers. First, a series of experiments on voice were carried out; nearly all made contributions to the issue of voice registers even though this purpose often was only a secondary one. Further Hollien notices that few phoneticians experienced any real difficulty in their conceptualization of voice registers. Indeed, except for a controversy which concerned the nature (pathology/nonpathology) of the vocal fry register (see Hollien et al. 1966) few phoneticians disagreed in any major way as to the nature, boundaries and/or functioning of vocal registers. Moreover, it was noted then – and it should be noted now – that many singers could produce register-related sung tones which were perceptually identifiable whenever they were requested to do so. Thus voice registers exist in singers – even though the concept is subject to much controversy - just as they do in speakers.

2. Register Types

The contrasts cited above have resulted in new insights relative to vocal registers. They can best be understood as follows. Recent attempts by the CoMeT (Collegium Medicorum Theatri) committee on vocal registers to develop appropriate models for their work have led them to articulate a position that appears to have been long overdue (it is reviewed in their reports, which are edited by Hollien, 1982, 1983). While this postulate is so simple that it seems not to be very profound, its absence has resulted in a substantial amount of confusion relative to vocal registers. Simply stated, the concept suggests that singing registers and speaking registers are different entities. Of course it must be conceded that they may overlap in function; that they may (in part anyway) have similar physiological roots. Nevertheless, it is recognized that while voice (laryngeal?) registers exist and are sometimes used in speech, no attempt is made to 'train' them out of the productive repertoire of the speaker. Moreover, a physiological register (vocal fry, pulse, creak) exists in speaking that is virtually nonexistent in singing. The most serious problem in this regard relates to confusions resulting from research reported in the literature. Specifically, it appears that vocal register studies carried out on non-singers cannot be extrapolated to singers on the basis of some simple mathematical relationship.

Before proceeding further, the questions can be asked: what are some of these singing and/or speaking voice registers; what are their boundaries and/or dimensions? As was cited above, a rather substantial number have been proposed and labeled (Morner et al., 1964); indeed, Vennard (1967) reports different scholars to have suggested that there are as few as one or as many as nine. Figure 1 should provide some insight as to the number, classification and extent of vocal registers; data here are drawn from the writings of four representative scholars. It should be noted, however, that the cited data were not necessarily obtained from a single reference. Rather, they are compilations of the 'best' information each of the authors provide in their writings. It should be noted also that dashed lines extend certain of the register ranges and that they suggest areas of uncertainty (either on the part of the author or on our part when interpreting his data). Moreover, the boundaries, as given, are not those of an individual or even the means of a subject group. Rather, they are the maximum extent of the register as portrayed by the most extreme individual within a sex. Further observation will reveal that Vennard, Garcia and Appelman all suggest the presence of *three* registers whereas Hollien suggests only two. This difference is easily resolved. Hollien's data are based upon the registers encountered in the *speaking* voice (two of the three he suggests have been established – out of a possible five) whereas the registers proposed by the other three authors relate specifically to singers and the singing voice.

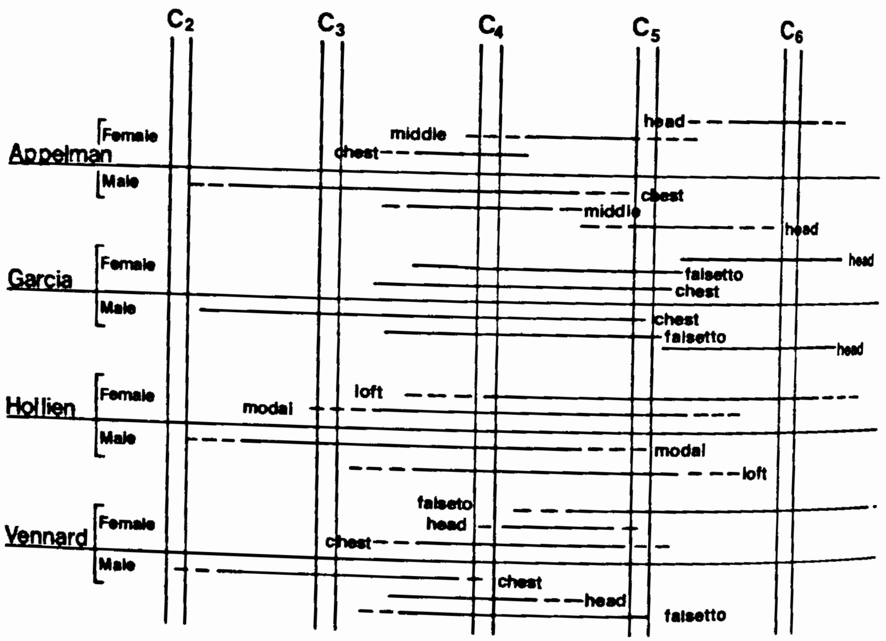


Fig. 1. Voice register types and extents as presented by four investigators.

3. Resultant Problems

But how do the two approaches do a disservice to each other? Consideration of Hollien's (1974) models (which predict/suggest the differences among voice registers), demonstrate how data on speakers may be misleading when applied to singers. For example, many of the contrasts he cited are not readily apparent when the phonatory productions of singers are reviewed. Relative to acoustics: PFR probably does not relate very well to singing range. Physiologically the two phenomena may not be so different but the efficiency and power of the singers are not very well portrayed by his models--nor are the aerodynamic contrasts. Briefly, data obtained on the registers of speakers probably do not predict singer's behavior very well at all. Research must be carried out on singers specifically – or on groups of singers and speakers – if these phenomena are to be well understood.

How can research on singers' registers mislead individuals interested in speakers – especially since very little research of this type has been carried out in the first place? Consider the following. Although the term 'modal' may be overtaking them, the two labels that appear more often than any other (in vocal music anyway) are 'chest' and 'head' – with *chest* referring to the lower register and *head* to the higher (if, indeed, there is only one higher register). As we all know, these terms are based upon singers sensations – i.e. on the mechanical response of the bodies of the singers to tones sung at or within certain frequency ranges. The generic connotation of these terms is such that they suggest certain relationships – specifically that the timbre for the lower register resides in the *chest*; that the quality/mechanism of the upper register results from activation of the vibratory properties of the sinuses and/or cavities in the head or 'mask'. Once these entities are considered physiologically, and/or mechanically, their use as definitions for vocal registers is shown to be illogical if not absurd. The lower register results from operation of the larynx – not from sympathetic vibrations of the chest to low frequency sung tones. The source of the upper register again is the *larynx* – *not* sympathetic vibrations (to higher sung frequencies) in the face. In short, while the sensations felt by singers, of course, are valid sensations (indeed, even the non-singer can experience them) *they have nothing to do with vocal registers.* It is only a chance relationship that brings the two into juxtaposition; that is, voice registers are frequency related and so are the sites of the sympathetically vibrating, sensation producing, structures of the torso and head. What a classic case of misdirection this is. For three hundred years, corrolary but indepent operations have been viewed as related – even causal – yet they were not and are not. Worst yet, this seeming relationship has led scholar after scholar astray – including many in the area of Phonetics.

4. Conclusions

The two examples cited above demonstrate how research concepts in one area can negatively effect those in the other. Accordingly, it is recommended that research on vocal registers be carried out (and interpreted) *independently* for singers and for speakers.

References

Appelman, Ralph (1967). *The Science of Vocal Pedagogy*, Bloomington, ID, The Indiana University Press.

Fields, V.A. (1970). Review of the Literature on Vocal Registers. *The NATS Bulletin*, February/March, 37-39; cover 3.

Garcia, Manuel (1840). *The Art of Singing*. Philadelphia, Oliver Ditson Co.

Hollien, H. (1974). On Vocal Registers. *Journal of Phonetics* **2** 125-143.

Hollien, H. (1982). *Report on Vocal Registers*. New York, Collegium Medicorum Theatri. (1983 Report in draft form).

Hollien H., Moore, G.P., Wendahl, R.W. and Michel, J.F. (1966). On the Nature of Vocal Fry. *Journal of Speech and Hearing Research* **9** 245-247.

Hollien, H. and Michel J.F. (1968). Vocal Fry as a Phonational Register. *Journal of Speech and Hearing Research* **11** 600-604.

Hollien, H., Gould, W.J. and Johnson, B. (1976). Two Level Concept of Vocal Registers. *Folia Phoniatrica* **26** 195-196.

Johnson, B. (1982). To Have or Have Not – That is the Question. In: *Report on Vocal Registers* (H. Hollien, Ed.). Collegium Medicorum Theatri, New York.

Large, J. (1972). Towards an Integrated Physiologic-acoustic Theory of Vocal Registers. *The NATS Bulletin*, February/March, 18-36.

Morner, M., Fransson, N. and Fant, G. (1964). Voice Register Terminology and Standard Pitch. *ASPR*, Speech Trans. Lab., KTH, Stockholm, **4** 17-23.

Preissler, W. (1939). Stimmumgange Gattungen der Menschlichen Stimme. *Arch. Sprach. Stimmheilh*, **3** 65-85.

Ruth, W. (1963). The Registers of the Singing Voice. *The NATS Bulletin*. May 2-5.

Sundberg, J. (1982). Comments on Voice Registers. In: *Report on Vocal Registers* (H. Hollien, Ed.). New York, Collegium Medicorum Theatri.

Titze, I. (1980). Have We Learned Anything About Registers (panel discussion). *Trans., Ninth Symp., Care Professional Voice*. New York, The Voice Foundation, 130.

Vennard, W. (1962). *Singing, The Mechanism and the Technic*. New York, Carl Fischer, Inc.

Motor Speech Characteristics in Diving

H. Hollien, J.W. Hicks, Jr. and P. Hollien
Gainesville, USA

1. Introduction

The importance of good voice communication among underwater workers has been accentuated as advances in saturated diving techniques have occurred. However, since the HeO_2 atmosphere in deep diving exhibits different sound transmission characteristics than does normal air and, since the resultant effects (coupled with those of high ambient pressure) severely degrade speech intelligibility, voice communication at depth is substantially reduced – to the detriment of efficient work capability and safety. For some years now we have been carrying out basic experiments in an attempt to identify the specific speech changes that occur as a function of HeO_2/P. In this regard, we have tended to focus our efforts on vowel formant shifts, changes in speaking fundamental frequency and related speech features. Our data argue that the observed reduction in speech intelligibility at depth cannot be accounted for solely by HeO_2/P relationships; hence, we suggest that radically new perspectives are needed. First, we have developed a new predictive model, adding acoustic radiation through vowel tract boundaries and the neurophysiological effects of HPNS (high pressure nervous syndrome) to the established variables. Second, we also have modeled the approach that we believe best meets these new perspectives. In this (second) model, we suggest a combination of four research thrusts – focused on 1) electronic processors, 2) articulatory modification by divers, 3) a special lexicon and 4) trained/specialized decoders – as relevant/necessary to mitigate the problem.

The first of the two models cited above includes three (major) contributing elements: change in density of the breathing gas, change in high ambient pressure and the neurological effects of the high pressure nervous syndrome (HPNS). As is well known, the life support atmosphere in saturation diving typically consists of mixtures composed predominantly of helium plys oxygen and, in some cases, small percentages of nitrogen. The reason for replacing nitrogen with helium is that, as ambient pressure increases, nitrogen becomes first narcotic and then toxic – also the danger of decompression sickness (the bends) becomes severe. The effects of helium in this regard are not nearly as great: therefore, it is used to replace nitrogen at depth. Indeed, even the overall percentage of oxygen is reduced due to its toxicity as a function of pressure. Taken as a whole, the available data indicate that the introduction of helium causes an upward shift in the formant frequencies of vowels, but *more importantly*, all speech sounds are effected (Beil, 1962; Fant and Lindquist, 1968; Rothman and Hollien, 1972; Sergeant, 1963; Tanaka, et al. 1974). However, taken *alone*, this upward shift does not appear to materially affect speech intelligibility (Sergeant, 1963) – even when the talker is totally within the HeO_2 environment (Hollien and Hicks, 1982).

Changes, (i.e. great increases) in ambient pressure also appear to degrade speech. Among the distortions noted is a non-linear shift in the lower formant (F_1) resulting from changes in vocal tract resonance. Moreover, one of the first distortions encountered in hyperbaric (helium) speech is a perceived 'nasality' and Fant theorizes that this perception may be due to 'a shunting effect between the cavity walls and the environment'. Therefore, at high pressures, there is a reduction in the impedance mismatch between the gas mixture and the cavity wall, and this latter effect can be a cause of distortion. In general, we agree with this position. Radiation of acoustic energy through the oral/facial wall undoubtedly does increase in parallel with ambient pressure and, while this factor probably does not constitute the 'main cause' of speech degradation at depth, it surely is of substantial importance.

Finally, the hyperbaric environment also produces physiological and psychological changes in the diver; changes which ultimately affect speech production. An example of such a potential disruption is the high pressure nervous syndrome (HPNS). While physiologists have understood the principal hazards involved in deep diving (i.e. oxygen toxicity, hypothermia, narcotic effects of gases and the effects of a dense atmosphere on respiratory function) at least since the early 1959's, the effects of the hyperbaric environment on the central nervous system were not identified until somewhat later. HPNS appears to be a complex and variable phenomenon; functioning as a consequence of: 1) absolute pressure, 2) rate of pressure change (compression rate); 3) gas mixture used, 4) individual susceptibility and 5) interaction among these factors. The overall result of HPNS appears to be disruption of normal neuromuscular activity as well as symptoms such as tremor, muscle jerks, convulsions and, in some cases, dysarthria (Vaernes et al., 1982). Accordingly we theorize that HPNS will have a measureable effect on the neuro-motor control required for speech production. In a sense, the relationship between speech production and HPNS can be thought to parallel the effects of Parkinson's Disease, at least on a temporary basis. In Parkinson's there is a general 'breakdown' in neuromuscular control which is associated with impairments to speech (Canter, 1963, 1965a, 1965b; Logemann and Fisher, 1981; Netsell et al., 1974). We predict that such is the effect of HPNS on divers' communicative ability.

2. Procedure

A consideration of the above review should suggest that speech intelligibility is not materially degraded by deep diving in HeO_2 mixtures. That is, taken alone none of the three effects appear to be severe enough to disrupt speech very much. But is this really the case? Four major thrusts have been carried out recently in an attempt to specify the exact extent (and nature) of this reduction in message intelligibility. A summarization of these data may be found in Figure 1. The data are from four of our projects; they were carried out at: 1) Sealab (DSSP) and EDU (Hollien et al., 1973), 2) the Duke and Westinghouse facilities (Hollien and Hicks, unpublished), 3) the Institute for Environmental Medicine, University of Pennsylvania (Rotman et al., 1980) and 4) at the Norwegian Underwater Technology Center, Bergen, Norway (Hollien and Hicks, 1981 and unpublished). As can be seen from examination of the figure, it can be generalized that speech intelligibility is degraded with increases in HeO_2/P. Of course, these projects exhibit rather substantial differences in research methodology (microphones and calibration, talkers

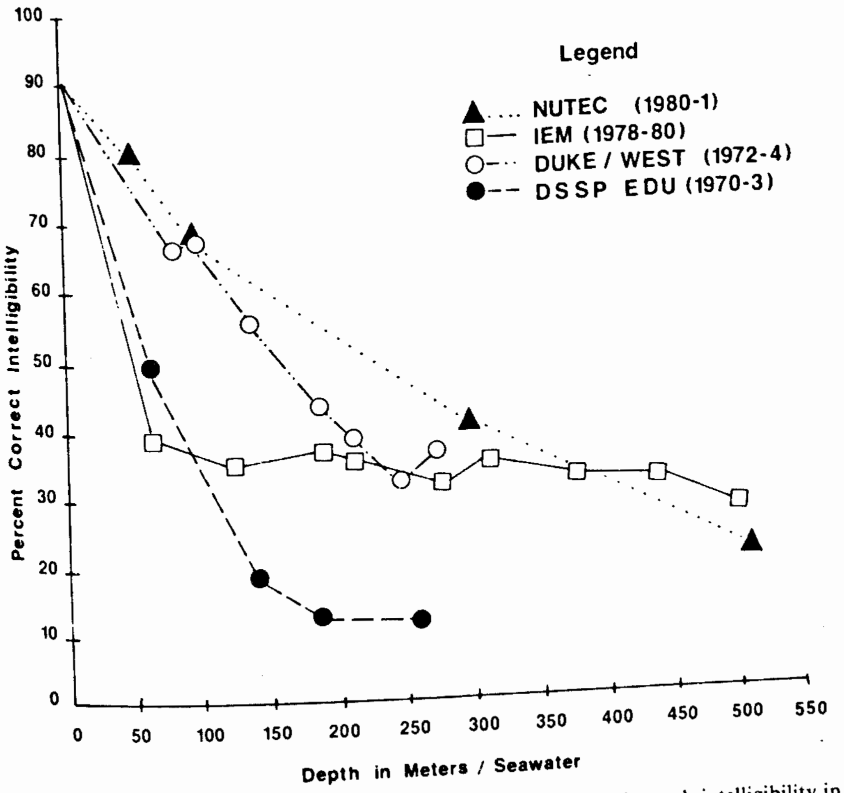


Figure 1. Summary figure of pooled data from several studies of speech intelligibility in the HeO_2/P environment.

differed in experience, noise levels varied and some speech was produced immediately after compression while other samples were obtained during decompression) and these factors tend to explain the differences among the curves. In any case, a good rule of thumb – relating speech intelligibility decrement to HeO_2/P level – appears to be that intelligibility level decreases approximately 10% per 50 meters for the first 100 meters, at a rate of about 10% per 100 meters for the next few hundred meters and finally at the rate of 3-5% per 100 m. So now it can be established that speech, indeed, is severely degraded at depth – and that it requires the addition of a third factor and the interactions among the three problem sources to account for the degradation. To be specific, what needs to be tested now are the physiological consequences of HPNS. One approach would be to assess the motor speech capabilities of saturated divers.

Research in this area was carried out at the Norwegian Underwater Technology Center's hyperbaric facility (Bergen, Norway). The dive was to 500 m and two of the three available aquanauts served as subjects.

Selection of a diadochokinetic test proved to be a problem. In our judgment, none of the available tests designed to assess the motor-speech function are better than marginally acceptable. Of course, development of such a test was not among the objectives of this research; hence we utilized the relatively informal tests proposed for, and already used in, research on motor-speech function. That is, we employed the Fletcher Time-by-Count Test of Diadochokinetic Syllable Rate (Fletcher, 1978) in this pilot study. Recording equipment included a calibrated B & K model 4166 condensor microphone (with a B & K model 2600 preamplifier) coupled to a calibrated laboratory quality tape recorder through a penetration tube in the wall of the chamber. In order to insure accuracy of timing, t-f-a spectrograms were made on a Voice Identification Model 700 unit and measurements were made with a 'time grid'.

3. Results

The results of this study are summarized in Table I, which provides summary data for the two subjects producing three single and one multiple phoneme as a function of depth. As can be seen there is a systematic reduction in the

Table I. Means of data for surface contrasted with means at depth. Values are number of productions as a function of time*

Condition	Single phonemes	Multiple phonemes
Surface (air)	5.65	1.92
Depth (HeO_2)	5.05	1.63
Difference	-0.53	-0.29

* Learning effects would bias toward improved scores at depth.

number of units the divers could produce as a function of increasing depth (i.e., increase in the proportion of helium in the environment, increases in ambient pressure and presumed decreases in motor coordination due to HPNS). When the data were converted to time required for a set number of repetitions, it was found that the two divers scored at the adolescent level at the surface and that their performance deteriorated 3-5 categories (to the norms for 9-year-old children) as a function of depth.

We should hasten to add that one variable affected our data. Prior to the project, we had no reason but to believe that the divers could carry out this task easily and that there would be no learning effect. Accordingly, we did not provide them with 'training' trials. Unfortunately, such was not the case and a very marked learning effect actually was observed – it was obvious both in subjects' hesitant behavior during the initial trials (at the surface) and in their comments about the task. Even with this variable operating to reduce the differences between surface and depth, a marked change was observed. We interpret these data to suggest that, indeed, HPNS was present when these divers spoke and it operated to degrade their speech. Further, the effects of the three sources of difficulty (HeO_2/P/HPNS) – even taken separately – begin to account for the substantial degradation of divers speech at depth. However, since these disruptions are mechanically induced, it may be possible to compensate for them, and increase speech intelligibility levels, by the application of training procedures suggested by relevant speech therapy. In any case, our theory that motor function is impaired at depth was supported – even though the neurological involvement did not appear to be as severe as with Parkinson's.

References

Beil, R.C. (1962). Frequency Analysis of Vowels Produced in a Helium-Rich Atmosphere. *J. Acoust. Soc. Amer.* **34**: 347-349.

Canter, G.J. (1963). Speech Characteristics of Patients with Parkinson's Diseases: I. Intensity, Pitch and Duration. *J. Speech Hear. Dis.* **28**: 221-229.

Canter, G.J. (1965*a*). Speech Characteristics of Patiens with Parkinson's Diseases: II. Physiological Support for Speech. *J. Speech Hear. Dis.* **30**: 44-49.

Canter, G.J. (1965*b*). Speech Characteristics of Patients with Parkinson's Diseases: III. Diadochokinesis and Overall Speech Adequacy. *J. Speech Hear. Dis.* **30**: 217-224.

Fant, G. and Lindquist, J. (1968). Pressure and Mixture Effects on Diver's Speech. *Quarterly Prog. Rept., Speech Trans. Lab.* Stockholm, 7-17.

Fletcher, S.G. (1978). *The Fletcher Time-By-Count Test of Diadochokinetic Syllable Rate*. Tigard, OR, C.C. Publications Inc.

Hollien, H. and Hicks, J.W., Je. (1981). Research on Hyperbaric Communication – A Progress Report. *IASCP/NUTEC-006/81*, 1-26 (Appendix A-C).

Hollien, H., Thompson, C. and Cannon, B. (1973). Speech Intelligibility as a Function of Ambient Pressure and HeO_2 Atmosphere. *Aerospace Med.* **44**: 249-253.

Logemann, J.A. and Fisher, H.B. (1981). Vocal Tract Control in Parkinson's Disease: Phonetic Feature Analysis of Misarticulations. *J. Speech Hear. Dis.* **46**: 348-352.

Netsell, R., Daniel, B. and Celesia, G.G. (1974). Acceleration and Weakness in Parkinson's Dysarthria. *J. Speech Hear. Dis.* **40**: 170-178.

Rothman, H.B., Gelfand, R., Hollien, H. and Lambertsen, C.J. (1980). Speech Intelligibility at High Helium-Oxygen Pressure. *Undersea Biomed. Res.* 7: 265-275.

Rothman, H.B. and Hollien, H. (1972). Phonetic Distortion in the HeO_2 Environment. *Proceed., Seventh Intern. Cong. Phonetic Sciences.* (A. Rigault and R. Charbonneau, Eds.), Mouton, The Hague, 589-598.

Sergeant, R.L. (1963). Speech During Respiration of a Mixture of Helium and Oxygen. *Aerospace Med.* **34**: 826-829.

Tanaka, R., Nakatsui, M. and Suzuki, J. (1974). Formant Frequency Shifts Under High Ambient Pressures. *J. Radio Res. Lab.* **2**: 261-267.

Vaernes, R., Bennett, P.B., Hammerborg, D., Ellertsen, B., Peterson, R.E. and Tonjum, S. (1982). Central Nervous System Reactions During Heliox and Trimix Dives to 31 ATA. *Undersea Biomed. Res.* **9**: 1-14.

An Acoustic Determinant of Perceived and Produced Anisochrony

P. Howell
London, United Kingdom

1. Introduction

Morton, Marcus and Frankish (1976) reported that digits presented with the same distance between onsets are not perceived as isochronous (equally timed). If subjects are allowed to position them so that they appear to be regular, the adjustments when measured between stimulus onsets, vowel onsets or peak amplitude value in the syllable are not equal. That is, the stimuli have to be physically anisochronous to appear perceptually isochronous. Marcus (1981) has examined what acoustic factors determine p-center location by editing speech to see what factors cause the p-center to vary. He found that varying the duration of the initial consonant of a syllable, lengthening its vowel and extending the period of closure before release of a syllable-final stop affected judgments about perceptual isochrony but that altering the amplitude of a final burst did not affect such judgments. Marcus interpreted his findings as showing that several acoustic factors determine p-center location.

A finding that may be related to those concerning perceptual isochrony is that when speakers are asked to produce isochronous lists, the timing corresponds to that which would be needed in order for the items to be perceived as isochronous (Fowler, 1979). Tuller and Fowler (1980) showed that certain syllables appear to be aligned with respect to orbicularis oris muscle activity. Fowler and her associates (Fowler, 1979; Tuller and Fowler, 1980) consider that the regularity in timed speech activity occurs in production with respect to activity in certain muscle groups and departures from physical isochrony in perception occur because perception is referred to production. Thus, perceptual adjustments do not align with respect to any acoustic referent because of the complex relationship between articulation and the sound produced. These results demonstrate that the acoustic onset of a syllable is not the same as the onset of the sound during production or perception. But before we accept that there are complex acoustic or productive determinants of p-center location, simple acoustic determinants should be ruled out. The criteria for a satisfactory factor that determine the location of the p-centers are, first, that it should vary in alignment across stimuli in the same way that the perceptual judgments do. Second, it should vary in location relative to stimulus onset in the same way that perceptual alignments vary when the acoustic properties of test stimuli are altered. Third, the factor should account for why the phenomenon occurs in perception and production.

To date, the acoustic factors that have been examined as candidates for determining p-center location have been acoustic reference points within a syllable, not acoustic factors associated with the syllable itself (e.g., reference points associated with the vowel). The principal acoustic factor that is associated with the syllable is the amplitude envelope (Mermelstein 1975). It is not directly related to any of the acoustic factors examined. So, for example, the parameters of the amplitude envelope are not fixed relative to acoustic factors associated with the vowel.

2. Experiments

2.1. Speech

The intention of the first experiment is to see whether variation in the amplitude envelope is a sufficient cue to cause variation in p-center location in perception with speech and non-speech sounds. The speech stimuli employed varied in the amplitude envelope alone. Thus, if p-center location varies it can only be attributable to this factor. Since variation in amplitude envelope can occur with non-speech sounds, there may be differences in the p-center of non-speech too.

To test this, different envelopes were introduced onto speech sounds by contouring the onset of a naturally spoken /ʃa/ to produce /tʃa/ (short rise) or /ʃa/ (long rise). The procedure only affects the envelope and, if the present account is correct, should be sufficient to cause variation in p-centers. To construct the stimuli, a recording was made of a /ʃa/ spoken by a male adult. Both the fricative and vowel were sustained so that they remained at the same amplitude for some time. The next step was to truncate the frication to 148.8 ms (measured back from vowel onset). It was then contoured by multiplying by a linear ramp of 40 ms and left at its original amplitude for the remainder of the frication. The rise of the stimulus with a value of 120 ms was constructed in the same way except that it was ramped over the first 120 ms. In each case the vowel was tapered by a linear ramp over 312 ms.

2.2. Nonspeech

The non-speech stimuli were constructed from a portion of white noise followed by a portion of sawtooth waveform. The noise and sawtooth were approximately the same peak-to-peak amplitude as the aperiodic and periodic portions had been in the recording of the original speech syllable. The contours of the speech stimuli with 40 and 120 ms rises were calculated from the rectified digital waveform low-pass filtered at 25 Hz (Fant, 1959).

These were used to contour the non-speech stimuli by multiplying the calculated contour by the non-speech stimulus just described. Subjects had to adjust the timing between one member of a pair of stimuli which were played repeatedly until they appeared to be perceptually regular. One of the intervals between the members of a pair was fixed at 750 ms and the duration of the other could be varied by means of a potentiometer. When the subject was satisfied that the items were regularly timed, he pressed a key and the duration of the adjusted interval was stored in the computer. The next pair was then selected and tested in like manner. On each trial the null position of the potentiometer was varied randomly and the subject was told this and informed that he would need to alter the position of the potentiometer. The stimulus pairings could both be the same (40/40 and 120/120) or differ (40/120 and 120/40) for both types of material. This gave four sorts of trial and there were three repetitions of each in a block of twelve adjustments. Altogether twelve subjects performed the experiment - six with the speech sounds and six with the non-speech sounds. Mean adjustments across subjects are presented in Table I separately for each type of adjusted interval. Inspection of the table shows that there is little difference in adjustments when stimulus pairs consisted of the same stimulus but with stimulus pairs with different envelopes at onset, longer adjustments were made when the interval between the stimulus with the slow rise at onset and that with the quick rise at onset was varied for both speech and non-speech sounds. The different adjustments were not significant by analysis of variance for stimulus pairs that were the same but were for both the speech and non-speech stimuli when the pairs had different envelopes at onset. ($F(1,5) = 39.1$, $p < 0.005$ and $F(1,5) = 15.5$, $p < 0.05$ respectively).

3. Discussion

Thus, altering the distribution of energy in the envelope is sufficient to cause variation in p-center location for both speech and non-speech stimuli. It is possible that amplitude envelope is just one other factor that determines p-center location (cf. Marcus, 1981). However, all Marcus's manipulations which were effective in altering p-center location are effective in altering the distribution of energy in the amplitude envelope. Another finding of note is

Table I. Mean duration of adjusted intervals (in ms) in Experiment 1

Speech			
All /tʃa/	All /ʃa/	Mixed lists	
		/tʃa/-/ʃa/ adjusted	/ʃa/-/tʃa/ adjusted
784	811	733	831
Non-speech			
All 40 ms	All 120 ms	Mixed lists	
		40 and 120 ms	120 and 40 ms
744	776	722	798

that variation in p-center location occurs for both speech and non-speech despite earlier claims to the contrary (Lehiste, 1973, Morton et al. 1976, though see Vos and Rasch, 1981 for another report of variation in p-center location of non-speech stimuli differing in amplitude envelope).

The correlation between the anisochronies in production and those in perception might occur because subjects judge the timing of their productions from the distribution of energy in the amplitude envelope of their own speech. If so, alteration to the distribution of energy in the amplitude envelope should affect their ability to produce isochronous sequences. Variations in the envelope during production should cause subjects to position syllables anisochronously.

This can be tested by requiring subjects to speak vowels that vary in length. A sustained vowel will have the energy in its envelope late relative to that of a short vowel. Thus speakers should advance the onset of the long vowel when it is spoken in alternation with a short vowel to make their onsets appear regular if subjects use the amplitude envelope to make this judgment. The ratio of the interval from the onset of a short vowel to the onset of a long vowel to the interval from the onset of a long vowel to the onset of the next short vowel in repeated vowel pairs should, then, be less than one, greater than one when the order of long and short vowels is reversed and equal to one when the vowels are of the same length. This prediction was tested in the same experiment. The vowels /i/ and /u/ were paired together or with themselves. With alternating vowel pairs the order of vowels was counterbalanced giving four sets of vowels in all. Each of the vowels could be of short, medium or long duration giving nine durations per vowel set (3x3). Eight subjects were told to repeat each of the vowel sets so that their onsets appeared regular. The mean duration for each set was calculated (omitting the initial and final vowel pair). The predicted ratios were obtained for all stimulus sets and there was no statistical difference between vowel sets.

4. Conclusion

It appears, then, that the amplitude envelope of speech and non-speech is an important factor in determining p-center location. Variations in this factor alone give differences in the location of p-centers in speech and non-speech. In addition, Marcus's (1981) manipulations affect the amplitude envelope and data have been presented showing that subjects may use the amplitude envelope to judge the p-centers of the productions.

References

Fant, G.M.C. (1959). Acoustic analysis and synthesis of speech with applications to Swedish. *Ericsson Technics*, **1**, 1-105.

Fowler, C.A. (1979). 'Perceptual centers' in speech production and perception. *Perception and Psychophysics* **25**, 375-388.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America*, **51**, 2018-2024.

Marcus, S.M. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics*, **30**, 247-256.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, **58**, 880-883.

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Rview*, **83**, 405-408.

Tuller, B., and Fowler, C.A. (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, **27**, 227-283.

Vos, J., and Rasch R. (1981). The perceptual onset of musical tones. *Perception and Psychophysics*, **29**, 323-335.

Pitch and the Perceptual Separation of Simultaneous Vowel Sounds

M.T.M. Scheffers
Eindhoven, the Netherlands

1. Introduction

Two experiments were carried out investigating identification of simultaneous vowel sounds by listeners. Our research is inspired by the intriguing question - first posed by Colin Cherry as the 'Cocktail Party Problem' (Cherry, 1953) - of how listeners are able to perceive the speech of a single speaker separately from a background of interfering voices. Cherry mentioned voice pitch as one of the factors possibly facilitating the separation. Much earlier, Stumpf (1890) had reported that the sounds of two musical instruments tended to fuse into a single percept when both instruments played exactly the same note, but were separately audible when different notes were played. More recently, Brokx and Nooteboom (1982) found that speech sounds presented in a background of speech from another or even the same speaker, could be identified considerably better when there was a difference of more than 1 semitone between the pitches of the two sounds. These observations prompted us to investigate the role of differences in pitch between simultaneous vowels in the perceptual separation process.

Identification of pairs of unvoiced vowels was investigated in the second experiment. This experiment was conducted in order to determine to what extent listeners could use information derived from the spectral envelope of the sound for identifying the vowels.

2. Experiment 1

The stimuli of the first experiment consisted of two different voiced vowels. The waveforms of the vowel sounds were computed using a software five-formant speech synthesizer. Eight vowels were used viz. the Dutch /i/, /y/ /I/, /ɛ/, /ə/, /ɑ/, /ɔ/, and /u/. Formant structures were taken from Govaerts' study of Dutch vowels (Govaerts, 1974). The duration of each vowel was 220 ms including cosine-shaped onset and offset ramps of 20 ms. The vowels were added with no temporal onset difference, starting in zero-phase. They had about equal subjective loudness. Six F_0 differences were used: 0, ¼, ½, 1, 2 and 4 semitones. The average F_0 was 150 Hz. For each pair of vowels with unequal F_0 two stimuli were made, one in which one vowel had the lower and one in which the other had the lower F_0. The waveforms of the 308 different combinations were digitally stored on disk.

Twenty subjects took part in Exp. 1. They had normal hearing and were familiar with synthesized speech sounds and psychoacoustic experiments. They were tested individually. The subjects were seated in a sound-treated booth and received the signals diotically through TDH 49-P headphones. The signals were band-pass filtered from 50 Hz to 5 kHz and presented at a level of about 60 dB SL.

A minicomputer controlled the presentation of the stimuli and recorded the responses. The subjects were instructed to respond to each stimulus by pushing two buttons on a panel of eight, each button representing one of the eight vowels used. All vowels were played to them before the experiment started. No feedback was given on their responses. The subjects attended the experiment in four sessions held on consecutive days. In every session, each of the 308 stimuli was presented once, in a random order that differed for each subject and for each session. A session lasted about half an hour.

3. Results

A synopsis of the results is presented in Fig. 1. The solid line in this figure gives the percentage correctly identified combinations (both vowels correct), averaged over the 28 combinations. The dashed line depicts the average percentage of individual vowels correct. No significant difference was found between the performance on the stimuli in which one vowel had the higher F_0

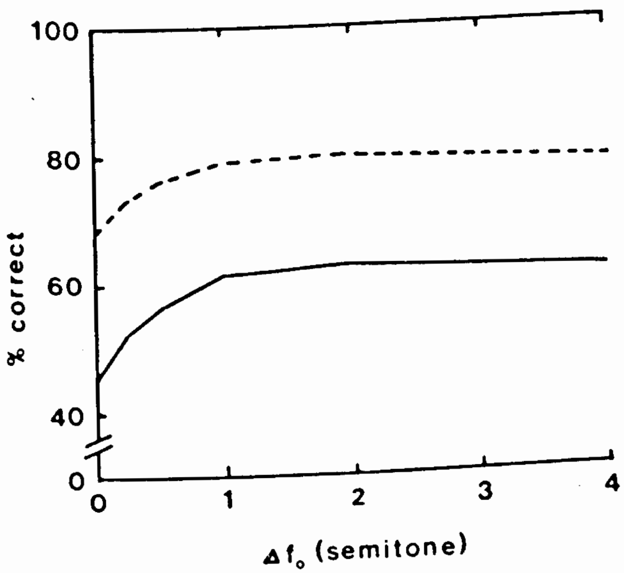


Figure 1: Percentage correct identification of two simultaneous voiced vowels as a function of the difference between the fundamental frequencies of the vowels. The solid line shows the average identification scores on the 28 combinations used (both vowels correctly identified) and the dashed line the average scores on individual vowels.

and on those in which it had the lower F_0. The results are therefore averaged over 'positive' and 'negative' F_0 differences.

The scores differed much for different combinations. They were in general lowest for combinations of similar vowels, such as two front vowels or two back vowels, and were highest for dissimilar combinations such as a front and a back vowel. The scores were down to around chance level (4%) for only a few combinations of vowels with equal F_0. It can be seen from Fig. 1 that the scores increased with increasing F_0 difference up to 1 or 2 semitones.

4. Experiment 2

When it was found that identification scores even on pairs of vowels with identical F_0's were generally well above chance level, a second experiment was devised. Stimuli in this experiment consisted of two different unvoiced vowels. They were constructed in the same way as the stimuli for Exp. 1. The unvoiced vowels had the same spectral envelopes as the voiced ones. The stimuli were D-A converted, band-pass filtered from 50 Hz to 5 kHz and recorded on magnetic tape with an ISI of 3 s. The tape contained every stimulus eight times in random order.

Eighteen subjects with normal hearing took part in this experiment. They were asked to identify the two vowels in a stimulus and to write down a phonemic transcription of both vowels on an answer form. The test method was the same as in Exp. 1 except for the use of a tape and written responses.

5. Results

Performance on the unvoiced vowels was significantly lower than for voiced vowels with equal fundamentals ($p < .01$). The identification score on combinations was 26% for the unvoiced stimuli and 45% for the voiced stimuli and the average score on individual vowels was 56% and 69% respectively. The same tendency for pairs of vowels with dissimilar formant structures to be better identified than pairs with similar structures was also found here.

6. Discussion and Conclusions

The most surprising result of the experiments is that identifiability of two simultaneous vowels was far above chance level even if both vowels had the same fundamental frequency or when they were unvoiced. The result that simultaneous unvoiced vowels were less well identifiable than simultaneous voiced vowels with equal F_0's cannot yet be explained. This was contradictory to what one would expect from the fact that formants are more sharply defined in unvoiced than in voiced vowels, although this is only true for the long-term spectrum. Identification scores on voiced pairs increased by about 18% on average when the F_0 difference between the two vowels was increased from 0 to 2 semitones. It is noteworthy that at least one vowel was correctly identified in 95% of the voiced stimuli and in 86% of the unvoiced ones.

Identification scores on combinations of vowels with strongly differing spectral envelopes like /i/ and /a/ were much higher than the scores on vowels with relatively similar spectral shapes like /i/ and /y/. This supports our theory of a 'profile' analysis (cf. Spiegel and Green, 1981) in the recognition process. A profile is considered to be a relatively simple image of the envelope of the spectral representation of the sound in the peripheral ear. Recognition is then a process of matching reference profiles to the one of the present spectrum and identifying the sounds on basis of the best fitting profiles. The profile is probably best defined around the first two formants of the vowel. The shape of the profile near the frequencies of these formants apparently weighs most in the matching (cf. Klatt, 1982; Scheffers, 1983). If there is a great difference between the profiles of the composing vowels, identifiability of the combination is relatively high and little influenced by F_0 differences. If the profiles are rather similar, however, F_0 differences can aid to separate the profile of the combination in parts belonging to one of the vowels and parts belonging to the other or maybe to both. Separation is supposed to be guided by the harmonic fine structure of the spectrum. This is only possible for relatively low frequencies because high harmonics are not separately detectable in the auditory system (e.g. Plomp, 1964). The theory is supported by the results of a pilot experiment in which it appeared that two pitches could be perceived when the F_0 difference was greater than 1 semitone, while for smaller differences only one (beating) pitch was heard. We may therefore expect little further improvement of the performance when the F_0 difference is increased beyond 1 semitone. A decrease in performance can even be expected for harmonic intervals between the two F_0's such as a major third (4 semitones) and especially for an octave because many harmonics of both vowels will then coincide. A clear decrease in performance for the 4-semitone difference was indeed found in the results for 8 combinations.

Acknowledgement

This research was supported by a grant from the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.), grant nr. 15.31.11.

References

Brokx, J.P.L. and Nooteboom, S.G. (1982). Intonation and the perceptual separation of simultaneous voices. *J. Phonetics* **10**, : 23-26.

Cherry; E.C. (1953). Some experiments on the recognition of speech with one and two ears. *J. Acoust. Soc. Am.* **25**, 975-979.

Govaerts, G. (1974). *Psychologische en fysische structuren van perceptueel geselecteerde klinkers, een onderzoek aan de hand van Zuidnederlandse klinkers* Doctoral thesis, University of Louvain.

Klatt, D.H. (1982). Predictions of perceived phonetic distance from critical-band spectra: a first step. *Proc. ICASSP* **82 (2)**, 1278-1281.

Plomp, R. (1964). The ear as a frequency analyzer. *J. Acoust. Soc. Am.* **36**, 1628-1636.

Scheffers, M.T.M. (1983). Identification of synthesized vowels in a noise background. In preparation.

Spiegel, M.F. and Green, D.M. (1981). The effects of duration on masker profile analysis. *J. Acoust. Soc. Am.* **70 (1)**, S86(A).

Stumpf, C. (1890). *Tonpsychologie*. Lizensausgabe des S. Hirzel Verlages, Leipzig. Republished in 1965 by Knef-Bonset, Hilversum-Amsterdam.

Perception of Speech in a Hyperbaric Helium-Oxygen Atmosphere

K. Slethei,
Bergen, Norway

This report presents results from an investigation of the relations between some linguistic factors and the intelligibility of speech produced in an atmosphere consisting of a mixture of helium and oxygen. The mixture in question is the one appropriate for saturation diving between 100 and 500 msw.

We have measured the intelligibility of unprocessed speech in helium-oxygen, using 20 respondents in multiple choice tests. The test design makes a strict distinction between initial and final ambiguity or confusion position, and the consonant phonemes have mainly been treated as either initial or final consonants. The findings are based on 24 000 individual judgments. The work is still in progress, and this report must therefore be considered as a preliminary one.

Intelligibility, as expressed by the mean percentage of correct identifications, can be broken down as a function of various test parameters. Table I gives the intelligibility as a function of depth.

Intelligibility has been broken down by consonant phonemes, as shown in table II. The ranking was done before rounding off.

The Spearman's Rank Order Correlation (r(s)) between the hierarchies for initial and final intelligibility is 0.83. Consequently, I consider the combined ranking in Table II as a general intelligibility hierarchy, derived from initial and final positions from depths between 100 and 500 msw.

The relations between the hierarchies for initial and final positions can be studied at individual depths.

Table I. Intelligibility as a function of depth

Depth (msw)	Intelligibility	S.D.	N (words)
0	97.5	9.2	300
100	67.3	28.0	150
200	62.4	29.0	150
300	59.7	27.8	350
400	56.4	27.0	200
500	60.0	22.5	50

Table II. Intelligibility and ranking for consonants initially, finally and combined. Depths from 100 to 500 msw pooled

	Initially		Finally		Both positions	
	Int.	Rank	Int.	Rank	Int.	Rank
/p/	81	3	63	6	69	5
/b/	52	11	38	12	49	9
/m/	50	12	33	13	42	12
/f/	67	8	72	4	69	6
/v/	75	6	63	5	69	4
/t/	68	7	61	7	64	7
/d/	49	13	39	11	45	11
/n/	65	9	57	8	59	8
/l/	83	2	74	2	78	2
/r/	83	1	75	1	78	1
/s/	63	10	74	3	70	3
/kj/	75	5	–	–	–	–
/j/	48	14	–	–	–	–
/k/	45	15	46	9	45	10
/g/	44	16	28	14	40	13
/ng/	–	–	41	10	–	–
/h/	76	4	–	–	–	–

Table III. Correlations (r(s)) between intelligibility hierarchies for initial and final ambiguity positions for 5 depths

Depth	r(s)	N (cons.)	p-level (dir.)	Sign.
100	0.49	13	$p < 0.05$	*
200	0.72	13	$p < 0.005$	***
300	0.76	13	$p < 0.005$	***
400	0.76	12	$p < 0.005$	***
500	0.30	8	$p < 0.05$	n.s.

At 500 msw the intelligibility hierarchies for initial and final position seem to be different from each other. At the other depths the hierarchies do not differ more than what may be expected by chance.

The intelligibility hierarchies for initial position have been studied in pairs of depths.

Correspondingly for final ambiguity position (Table V).

We see from Tables IV and V that the consistency of the intelligibility hierarchies does not depend on depths, but rather on the differences between depths.

Intelligibility can be related to structure type, to place of articulation or to phonation.

One of the purposes of this study is to provide adequate background knowledge for developing improved technical equipment for unscrambling

Table IV. Correlations (r(s)) for 10 pairs of depths with respect to intelligibility hierarchies for consonants in initial ambiguity position

Depths in comparison	Diff.	r(s)	p-level (dir.)	N (cons.)	Sign.
100 vs 200	100	0.64	$p < 0.005$	16	***
200 vs 300	100	0.85	$p < 0.005$	16	***
300 vs 400	100	0.90	$p < 0.005$	16	***
400 vs 500	100	0.69	$p < 0.005$	14	***
100 vs 300	200	0.54	$p < 0.025$	16	**
200 vs 400	200	0.76	$p < 0.005$	16	***
300 vs 500	200	0.75	$p < 0.005$	14	***
100 vs 400	300	0.48	$p < 0.05$	16	*
200 vs 500	300	0.41	$p > 0.05$	14	n.s.
100 vs 500	400	0.20	$p > 0.05$	14	n.s.

Table V. Correlations (r(s)) for 10 pairs of depths with respect to intelligibility hierarchies for consonants in final ambiguity position

Depths in comparison	Diff.	r(s)	p-level (dir.)	N (cons.)	Sign.
100 vs 200	100	0.98	$p < 0.005$	14	***
200 vs 300	100	0.91	$p < 0.005$	14	***
300 vs 400	100	0.69	$p < 0.025$	13	**
400 vs 500	100	0.90	$p < 0.005$	11	***
100 vs 300	200	0.86	$p < 0.005$	14	***
200 vs 400	200	0.48	$p > 0.05$	13	n.s.
300 vs 500	200	0.73	$p < 0.025$	11	**
100 vs 400	300	0.40	$p > 0.05$	13	n.s.
200 vs 500	300	0.61	$p < 0.05$	11	*
100 vs 500	400	0.53	$p > 0.05$	11	n.s.

Table VI. Intelligibility by depth by structure type. (a: Only finally. x: Reduced reliability – less than 100 judgments.)

	Depth (msw)				
	100	200	300	400	500
Plosive	62	54	54	51	60
Nasal	54	51	50	44	43
Lateral	85	81	77	70	67
Vibr./Tap.	64	82	83	79ax	75ax
Fricative	78	73	65	63ax	71ax

Table VII. Intelligibility by depth by place of articulation. (b: Only initially. x: Reduced reliability – less than 100 judgments.)

	Depth (msw)				
	100	200	300	400	500
Labial	61	53	61	60	63
Dent./Alv.	74	71	60	56	59
Palatal	78b	70b	53b	52b	58bx
Velar	47	42	42	45	45x
'/h/'	84b	76b	72b	73b	80bx

Table VIII. Intelligibility by depth by phonation

	Depth (msw)				
	100	200	300	400	500
Unvoiced	74	69	63	60	67
Voiced	62	57	57	53	53

speech distorted by helium. An experimental unscrambler is at present being developed in Bergen, with the Norwegian Underwater Technology Center (NUTEC) as main coordinator.

It is further believed that this will be useful knowledge for standardizing terminology and procedures in saturation diving.

Acknowledgement

This paper was supported by a grant from the Norwegian Underwater Technology Center - NUTECH.

Categorical Perception of Speaker Identity

H.G. Tillman, L. Schiefer and B. Pompino-Marschall
Munich, FRG

1. Introduction

In natural speech communication an important role is played not only by linguistically defined categories, which determine one part of the phonetic structure of verbal utterances, but also by other aspects, such as the expression of emotion or the characteristics of the perceived individual speaker. We do not yet know very much about how the latter types of information are encoded in the acoustic speech signal.

In the framework of a research project on non-linguistic categories of perceived natural utterances we have focused our interest on the acoustic parameters of speaker identity. To determine the relevant parameters (or combination of parameters) it seems obvious to work with naturally produced material. The consequence is a situation which is much more complex than in the case of synthetic speech, where single parameters can be controlled easily.

To extract relevant parameters of speaker identity the experimental paradigm of categorical perception should be an interesting instrument. Our first experiments described below were undertaken in order to see whether there actually exists the possibility of categorical perception in the domain of speaker identification. It is our aim to apply this instrument to (more or less) complex test material the parameters of which are manipulated in a more sophisticated way.

2. Method

The starting points for the production of our test stimuli were the two German sentences ‘Heute ist Donnerstag’ and ‘Aller Anfang ist schwer’ uttered by two male (HGT/GK) and two female (LS/GH) speakers, respectively. These pairs of utterances were digitally recorded and processed in order to generate new stimuli placed at exactly equal acoustic distances between the two original ones. The linear interpolation of the acoustic parameters was achieved in the following way. First, the two original digital speech signals were segmented into voiced, fricative-voiceless and silent parts. Bursts were handled as short fricative segments. The duration of silent intervals was interpolated directly. Fricative segments, after appropriate time-warping, were interpolated in the amplitude-time-domain. In order to interpolate the voiced parts four parameters had to be manipulated: intensity contour, pitch contour, spectral (harmonic) structure and the duration, measured in numbers of pitch periods, each of a defined length. The duration of the respective voice parts of the initial utterances was determined by counting the number of pitch periods and summing their durations. The time-warping of the stimuli in the continuum between the original ones results from interpolating between these values.

Each individual pitch period (of both original utterances) was separately transformed into the frequency domain by computing the Discrete Digital Fourier Transform. Interpolation of the harmonic spectrum and retransformation with the computed values of F_0 and intensity by computing the Inverse Fourier Transform yielded the pitch periods of the new signals (a more detailed description of our stimulus generating programs is given in Simon (1983)). It should be added that as soon as the two original signals have been segmented properly by our speech editing system the experimenter is free to choose the number of stimuli to be computed between the two original ones. Even extrapolation is possible. The computed stimuli sound quite natural. Listening to the continuum itself, one perceives the change from one speaker to the other in discrete steps.

For our experiments the male continuum Cm consisted of 10 stimuli (including the original ones at the ends of the continuum), the female continuum Cf had 7 stimuli. The tapes for running the identification tests contained each stimulus 10 times in randomized order. There was an interstimulus interval of 4 s and a pause of 10 s after each 10 stimuli. For the discrimination tests pairs of two-step-neighbours of the respective continua were chosen as well as identical pairs. Thus a set of 26 pairs resulted for the male Cm and a set of 17 pairs for the female Cf material. The Cm-tape contained each pair 5 times in randomized order, and to produce the Cf tape each stimulus-pair was repeated 10 times. Within a pair of stimuli the pause was 500 ms, between the pairs themselves 4 s. Blocks of 10 pairs were separated by 10 s again.

In the identification tests the original utterances of the two speakers (HGT/GK, LS/GH were demonstrated 5 times, and the subjects were instructed that the utterances of these speakers had been computer-manipulated to varying degrees, and they were then asked to identify the speakers. In the discrimination tests they were asked to decide whether the utterances were identical or not.

3. Experiments

In our first experiment (Exp. I) we presented the stimuli of continuum Cm to 11 members of our institute who are very familiar with the voices of HGT and GK. The results of both identification and discrimination tests are presented in Fig. 1.

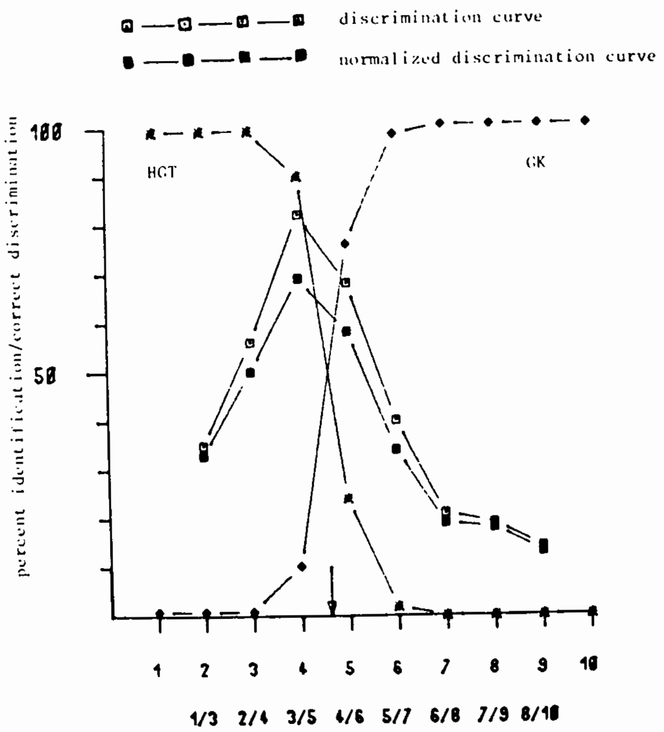


Figure 1. Results of Exp. I.

In Exp. II 15 subjects who had never heard the speakers of Cm before undertook first the discrimination test and then the identification test one week later (Cf. Fig.2)

In Exp. III the second continuum Cf was presented to the subjects of Exp. I. Again both speakers were known to them. (Cf. Fig. 3).

In Exp. IV the stimuli of Cf were presented to a group of 9 subjects to whom only one speaker, LS, was familiar. The discrimination test followed the identification test (Cf. Fig. 4)

In Exp. V the stimuli of Cf were presented to the subjects of Exp. II. This time the identification test was run first, and the discrimination test followed a week later. Again the speakers were not known to the subject. (Cf. Fig. 5)

4. Results and Discussion

The stimuli of continuum Cf and Cm were presented to three groups of listeners knowing either both, only one or none of the speakers, respectively. The results of the first group show clear categorical perception for both continua in the identification and discrimination tests (Exp. I and II, Fig. 1 and 3).

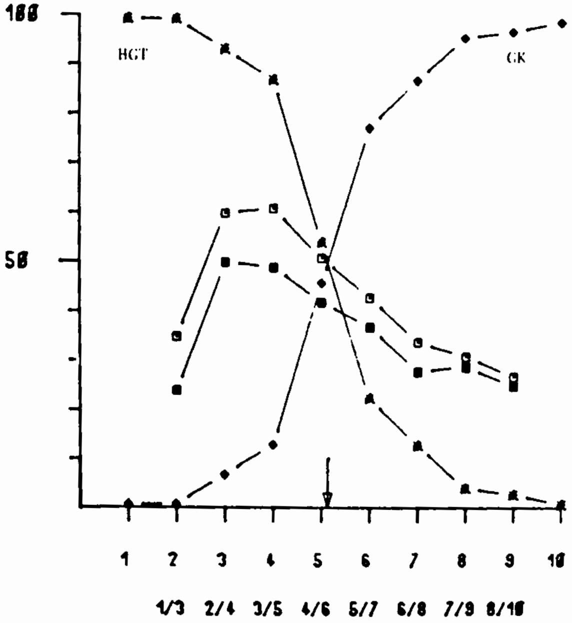


Figure 2. Results of Exp. II.

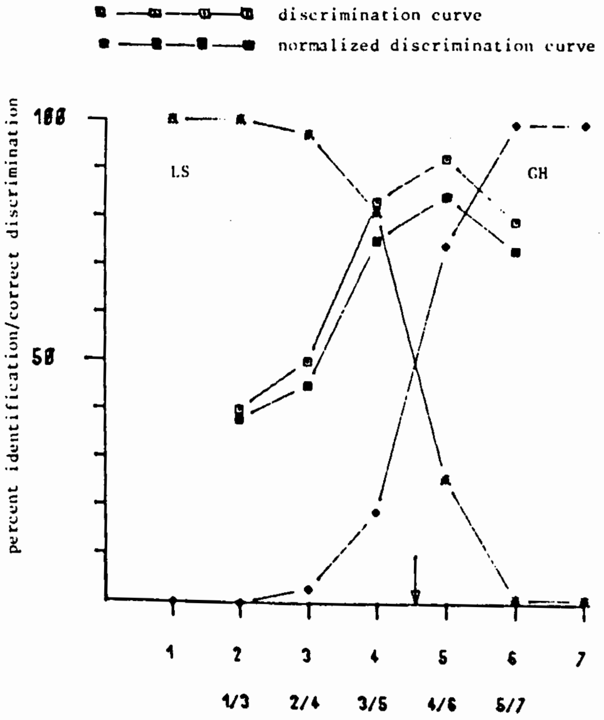


Figure 3. Results of Exp. III.

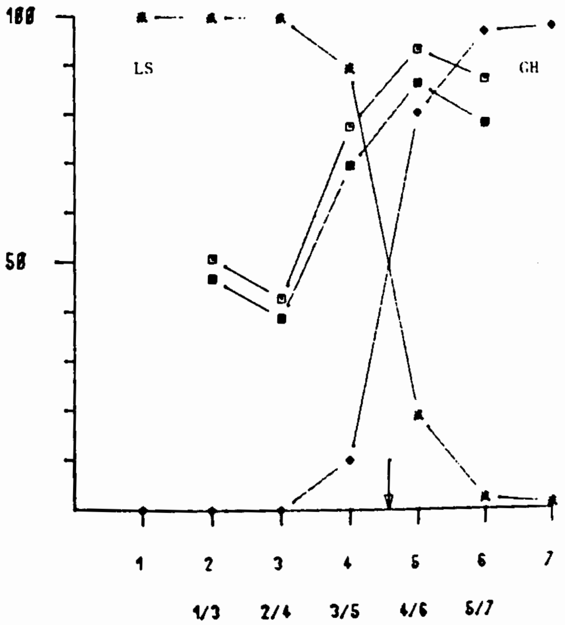


Figure 4. Results of Exp. IV.

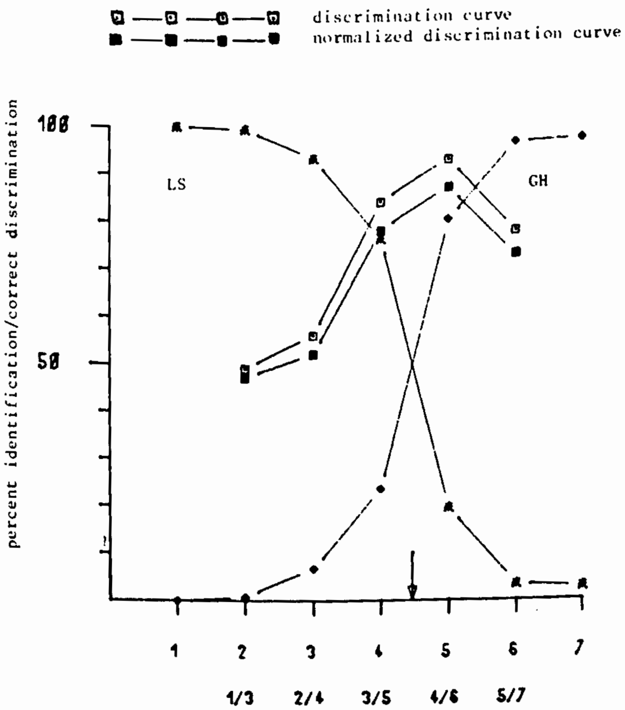


Figure 5. Results of Exp. V.

The second group obviously discriminated between the known and an unknown speaker and thus also produced categorical perception. (Exp. IV, Fig.4). No categorical perception was shown by the third group in Exp. II (Fig. 2) where the discrimination test was presented first. But this group reacted in a categorical manner when the discrimination test followed the identification test (Exp. V Fig. 5). Due to the ad-hoc complexity of the manipulated stimuli an interpretation of all details in the results cannot be given. Nevertheless some interesting facts should be mentioned. The discrimination curve (D-Curve) of group II for continuum Cf in Exp. IV (Fig.4) indicates that the difference between the original and the manipulated stimulus of the known speaker (i.e. stimulus 1/3, 3/1) leads to somewhat better discrimination. Another effect can be seen if one compares the original D-curves and those normalized according to the mean score for the respective identical pairs. Only in Exp. V (Fig. 5) do both D-curves have a nearly parallel form from the first until the last stimulus pair. In all other cases pairs of identical stimuli receive better 'same'- responses in the region of the identified speakers than in the region of the category boundary between the speakers. A third observation to mention is the dominance of speaker GK in Exp. I (Fig. 1), who wins 6:4 in the identification test, while on the other hand the D-curve shows better discrimination within the range of speaker HGT. This however correlates with the specific course of formants F_4 and F_5 in the utterance of speaker GK. In order to measure the influence of such different parameters of the phonetic form of utterances as F_0-contour, intensity contour, speech rate, different frequency regions of the spectrum etc., we are now preparing specific non-ad-hoc material which can be more easily manipulated in a systematic way. Finally it should be noted that also any 'artificial speaker' from the computed continua can be chosen as the starting point for the computation of a new continuum.

References

Simon, Th. (1983). Manipulation of natural speech signals according to the speech parameters of different speakers. *Forschungsberichte des Instituts fuer Phonetik und Sprachliche Kommunikation der Universitaet Muenchen (FIPKM)* **17** (in press).

Tillmann, H.G. (1974). *Das individuelle Subjekt und seine persoenliche Identitaet im phonetischen Kommunikationsprozess.* Hamburg: Buske.

Tillmann, H.G., Simon, Th. (1983). Kategoriale und nichtkategoriale Komponenten in der Wahrnehmung bekannter und unbekannter Sprecher. *Forschungsberichte des Instituts fuer Phonetik und Sprachliche Kommunikation der Universitaet Muenchen (FIPKM)* **17** (in press).

Section 8
Perception of Phonemes

The Role of Coarticulation in the Identification of Consonants

L.E.O. Klaasen-Don and L.C.W. Pols
Soesterberg and Amsterdam, the Netherlands

The study of the articulatory and acoustic properties of speech has failed to reveal basic segments at the phoneme level. On the contrary, every phonemic realisation reflects the influence of properties of preceding and/or following phonemic environments. Thus on the one hand there is the notion that 'pure', 'basic form' speech segments may exist (e.g. Daniloff and Hammarberg, 1973), whereas on the other hand it is found that information present in the speech sound is coarticulated (e.g. Bell-Berti and Harris, 1982). The concept of articulation assumes that, during (the process of) speech, basic forms, usually phonemes, are transformed in such a way that their articulatory target values become modified due to the interaction with properties of contiguous phonemes. Of course, some overlap in articulatory movements is inevitable, given that the speech organs are not capable of infinite acceleration. However, what makes coarticulatory influences interesting is that their explanations go beyond simple inertial factors, although inertia must play some role. Coarticulatory influences have been found that reflect planning in motoric programming. It seems that the speech apparatus can make preliminary adjustments for different phonemes. Obviously, the system of motor control has information about several phonemes at once. Perhaps the range of phonemes over which simultaneous information is available defines the organisational scheme of speech articulation.

Research concerning coarticulation has mostly been concentrated on the articulatory aspects of the phenomenon. Nowadays, several studies also deal with its perceptual aspects. In analogy with studies on the articulatory aspects the perceptual studies are particularly interested in the range of speech over which simultaneous information about different phonemes is available, related sometimes to perceptual units. Usually, these studies are confined to a few sounds, like plosives, and use synthetic speech or carefully pronounced utterances.

The present study investigates the relative contribution of vowel transitions to the identification of *all Dutch* consonants, both in utterances pronounced in isolation and in excerpts taken from conversational speech. By means of speech editing, parts of these utterances are deleted and listeners have to identify the (absent) consonant on the basis of information in the remaining part of the utterance. Figure 1 presents an example of the relevant CV part of the waveform of a CVt utterance /na:t/ pronounced in isolation,

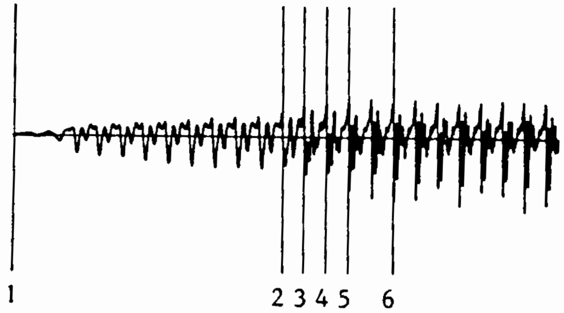


Figure 1. Example of the relevant CV part of the waveform of /na:t/. The vertical lines mark the segmentation points which define the 6 segments.

with the vertical lines marking 6 segmentation points by means of which 6 segments could be defined. Segmentation point 1 left the whole utterance intact and 5 other segments could be made from this CVt utterance by removing more and more of the vocalic transition.

Analogous to spectral analyses, the identification results show that presentation of the stationary vowel part alone does not contain enough information to identify adjacent consonants. However, the initial part of the vowel transition in CV syllables and the final part in VC syllables do contain some information about the adjacent consonant. The amount of information and its extension into the vowel transition differ widely for the various articulatorily defined groups of consonants. In utterances spoken in isolation voiced plosives, liquids and semi-vowels can be identified rather well on the basis of the vowel transition only, for nasals and unvoiced plosives this is more difficult, and for fricatives it is impossible. Thus it seems that Dutch initial voiced plosives on the one hand and fricatives on the other hand reflect extreme groups in this respect. Results of experiments (Don and Pols, 1983) using fast reaction times to detect and classify the consonant and vowel in CV syllables, reveal that this may be related to the fact that voiced fricatives are identified *earlier* than the following vowel, whereas voiced plosives are identified later than the following vowel. Moreover, fricatives are identified *earlier* than plosives, since in CV syllables containing a fricative, identification can start right from the physical beginning of the utterance, whereas in CV syllables containing a plosive identification is hindered by the vowel murmur preceding the burst.

A very important aspect in the study on coarticulatory effects is the direction of the perceptual influence. Is it more advantageous to have additional cues about a consonant before than after it is heard? The literature is not clear on this point. Most studies show that anticipatory influences are either dominant or at least equal to carryover influences. Results of the present study show that Dutch voiceless plosives are slightly better identified from VC than from CV vowel transitions. For voiced plosives and nasals the situation is reversed, with nasals being very much better identified from CV than from VC vowel transitions. Broadly speaking, we can say that some consonants are better identified on the basis of vowel transitions in CV syllables, whereas others are better identified from vowel transitions in VC syllables. No universal dominance can be found of anticipatory or carryover influences.

Information about the consonant in CV vowel transitions did not differ according to whether the segments were isolated from CVC or VCV utterances. However, some consonants are better identified from vowel transitions in VC segments excerpted from CVC utterances, whereas others are better identified from vowel transitions in VC segments excerpted from VCV utterances.

For excerpts from *running* speech the relative behaviour is similar, but the absolute scores are much lower.

There are various explanations for the observed perceptual interactions between vowels and consonants. One has to do with a more or less fixed time span over which the perceptual mechanism integrates. Another supposes a functional symbiosis between production and perception in such a way that whenever segments are coarticulated they are also perceived in a dependent way. The actual confusions made, which are often related to place of articulation, point in this direction.

For utterances pronounced in isolation, results of the present study show that perceptual information in the vowel transition about the contiguous consonant is in fact redundant since the consonant segment alone already causes almost perfect consonant identification.

In conversational speech consonant parts are often not well articulated or masked. Moreover, the durations of the steady-state vowel segments become shorter and the durations of the vowel transitions become relatively longer than in utterances pronounced in isolation (Schouten and Pols, 1979). For these reasons, we may expect the vowel transition to become relatively more important for consonant identification in utterances excerpted from running speech. The fact, however, that vowel transitions do not seem to contain much information about the adjacent consonants in conversational speech rejects this hypothesis and is an indication of reduced redundancy at the acoustic-phonetic level, which is probably compensated for by additional information at the lexical, syntactic, and semantic levels (Marslen-Wilson and Tyler, 1980).

Future research on 'coperception' should concentrate, more than has often been done in recent studies, on fundamental units of speech perception. If e.g. a specific form of a syllable (VCV,CV or VC) is found to be basic to speech perception, perceptual influences among speech segments can then be explained in terms of the perceptual organisation of the syllable cycle. Besides, studies on units of speech perception also allow one to study the relation between perceptual units and units of speech production.

Acknowledgement

This research was supported by grant 17-21-11 from the Foundation for Linguistic Research funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.).

References

Bell-Berti, F. and Harris, K.S. (1982). Temporal patterns of coarticulation. *J. Acoust. Soc. Amer.,* **71**, 449-454.

Daniloff, R.G. and Hammarberg, R.E. (1973). On defining coarticulation. *J. Phon.,* **1**, 239-248.

Klaassen-Don, L.E.O. and Pols, L.C.W. (1983). Dependencies of processing of consonants and vowels in CV syllables. Paper presented at the 11th International Congress on Acoustics, Toulouse.

Marslen-Wilson, W. and Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition,* **8**, 1-71.

Schouten, M.E.H. and Pols, L.C.W. (1979). Vowel segments in consonantal contexts: A spectral study of coarticulation - Part I. *J. Phon.,* **2**, 1-23.

Confusions between Dutch Consonants under Various Conditions of Noise and Reverberation

L.C.W. Pols and I. Stoop
Amsterdam and Leyden, the Netherlands

1. Introduction

Ever since the Miller and Nicely (1955) consonant identification experiment there has been a vivid interest in representing the perceptual differences and similarities between consonants. Idenfication under noisy or other disturbing conditions, paired or triadic comparison, scaling, and memory recall are some of the procedures used to achieve confusion or similarity data.

Miller and Nicely only had rather simple means available to process their confusion matrices and furthermore described their results in terms of predefined distinctive feature systems. We would prefer to use multidimensional scaling techniques and let the data more or less speak for themselves. Over the years this has been done several times with the original Miller and Nicely data, e.g. Shepard (1974), Wish and Carroll (1974), Soli and Arabie (1979), and Kroonenberg and De Leeuw (1980).

However, within the framework of a project about speech intelligibility and listening comfort in noisy and reverberant conditions, we had the opportunity to collect a large amount of new consonant identification data. We feel that there are various interesting aspects to these new data and to the way they have been processed:

- they concern a language different from English, namely Dutch;
- it is a large data set (17 consonants, 28 acoustically different conditions, 6 speakers, 5 listeners);
- initial, medial, and final consonants in CVCVC words were identified;
- not just different types of noise were used as acoustic disturbances but also reverberation, plus combinations of the two;
- advanced multidimensional scaling techniques were applied to the data;
- the present material is relevant with respect to noise legislation and listening comfort, although this point will not be discussed any further in this paper (see Pols, 1981a), nor will the links with the speech transmission index STI related to speech intelligibility (see Pols, 1981a and Houtgast et al., 1980).

2. Experimental Procedure

Three male and three female Dutch speakers read 20 different lists with 51

CVCVC nonsense words embedded in short carrier phrases, for instance /hɛt wort *limal* ɛində/. The 17 Dutch consonants /p,t,k,b,d,f,s,χ,v,z,h,m,n,l, r,w,j/ occurred three times in each list. Since in Dutch several voiced consonants do not occur in final position, there were 11 different final consonants /p,t,k,f,s,χ,m,n,ŋ,l,r/. The consonants were combined with 12 vowels and three diphthongs. Five subjects listened to these recorded word lists under various conditions of noise and reverberations, and had to identify the three consonants in each CVCVC word. The 28 different conditions were a mixture of four reverberation times (T = 0,05, 1 and 1,5 s), five signal-to-noise ratios (SNR > 50, + 15, + 8, + 1, and -6 dB), and five noise spectra, see Table I. The different noise spectra mainly differed in their high-frequency energy and were supposed to be representative of various types of ambient noise.

Table I. Specification of all 28 listening conditions in terms of type of noise spectrum used, signal-to-noise ratio, and reverberation time

T in S	SNR in dB					Type of noise spectrum
	50	+15	+8	+1	-6	
0	1	2	3	4	5	Speech noise
0.5	6	7	8	9	10	Speech noise
1	11	12	13	14	15	Speech noise
1.5	16	17	18	19	20	Speech noise
0.5			21	22		Traffic noise
0.5			23	24		Indoor traffic noise
0.5			25	26		Train/airplane noise
0.5			27	28		Industrial noise

After two days of training the listeners got, on each of the following days, all 28 conditions in random order for one speaker, plus some standard conditions for practice and reference. The identification experiment was computer controlled and all responses were stored on-line for subsequent data processing.

3. Results

For an evaluation of the various percentages correct scores under the various conditions we refer to Pols (1981a, 1981b); in this paper we will concentrate on the structure of the confusion matrices. The present data require a so-called three-mode model, namely stimuli x responses x conditions. These models are also referred to as individual difference models; these so-called individual differences are replaced in our data by differences caused by the acoustic conditions. For the time being we have neglected the real individual differences in our data, namely those caused by different speakers and different listeners, and used cumulative data. Appropriate multidimensional scaling programs like PARAFAC (Harshman, 1970), INDSCAL (Carroll and Chang, 1970), and ALSCAL (Takane et al., 1977) represent the stimuli and/or the responses as points in a multidimensional space with fixed axes, whereas the conditions are represented as factors weighing the overall configuration. Differences between these programs have to do with the structure of the input data the optimization algorithm, freedom of axes etc.

Kroonenberg and De Leeuw (1980) recently presented the method of principal component analysis by alternating least squares algorithms to solve the three-mode model in its most general form. Results from this program, called TUCKALS, will be presented here and we will limit ourselves to the initial-consonants data. A mean-squared loss function is used to minimize the difference between the low-dimensional model and the original data. A 3-dimensional stimulus configuration is presented in Fig. 1. The response configuration happens to be almost identical to this one. If one studies the configuration of Fig. 1 one will realize that clusters of consonants are positioned in the centre and at the four corner points of a tetrahedron. One can distinguish the following clusters: /l,r,w,j,h/, /z,s/, /v,f,χ/, /m,n/, and /p,t,k/ plus perhaps /b,d/. Although quite different from, for instance, the

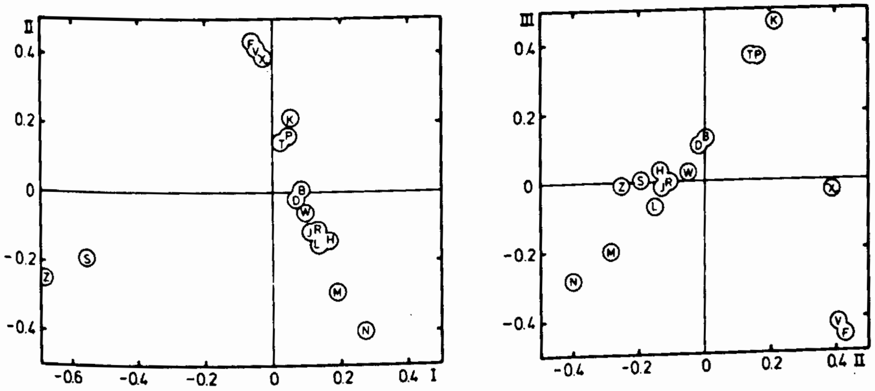


Figure 1. Stimulus configuration of initial consonants in the I-II and II-III plane of TUCKALS2 solution in three dimensions, based on double-centered confusion matrices for 28 different conditions.

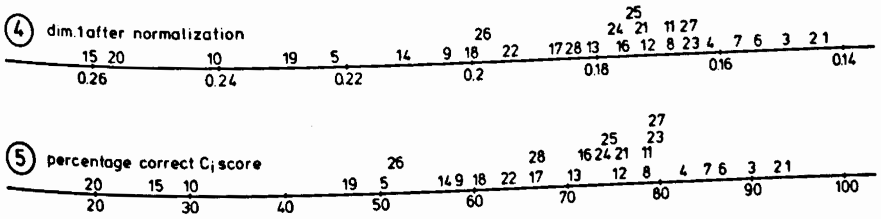


Figure 2. Representation of the first, and most important, dimension for the 28 conditions. The lower panel gives the percentage correct scores for the initial consonants for the same 28 conditions.

configuration found by Soli and Arabie (1979) by using INDSCAL on the Miller and Nicely data, this configuration seems to be quite attractive for Dutch consonants and certainly reflects the major structure in the confusion matrices. In those matrices one sees, for instance, almost no confusions between /f/ and /s/, or between /v/ and /z/, and more confusions between /p/, /t/, and /k/ than between /p/ and /b/, or between /t/ and /d/. This of course, is partly related to the types of disturbances used.

Despite the freedom given to the program to use three dimensions to represent the conditions, this representation of the 28 conditions turns out to be one-dimensional, see the upper panel of Fig. 2. The lower panel in this figure represents the percentage correct score for the initial consonants. Apart from a few deviations, e.g. conditions 26, 10, 15 and 20, there is a striking similarity. This could be an indication that both (speech) noise and reverberation have similar effects on consonant intelligibility and confusability behavior.

References

Carroll, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 35, 283-320.

Harshman, R.A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multimode factor analysis. *UCLA Working Papers in Phonetics No. 16.*

Houtgast T., Steeneken, H.J.M., and Plomp, R. (1980). Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. *Acustica* **46**, 60-72.

Kroonenberg, P.M. and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* **45**, 69-97.

Miller, G.A. and Nicely, P.E. (1955). An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* **27**, 338-352.

Pols, L.C.W. (1981*a*). Effect of reverberation and ambient noise on consonant intelligibility. *Report IZF 1981 C-5* (in Dutch).

Pols, L.C.W. (1981*b*). Consonant intelligibility in reverberant and/or ambient noise conditions. *Proc. 4th FASE Symposium on Acoustics and Speech, Venice, Italy*, 87-90

Shepard, R.N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika* 39, 373-421.

Soli, S.D. and Arabie, P. (1979). Auditory versus phonetic accounts of observed confusions between consonant phonemes. *J. Acoust. Soc. Am.* **66**, 46-59.

Takane, Y., Young, F.W., and De Leeuw, J. (1977). Non-metric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika* **42**, 7-67.

Wish, M. and Carroll, J.D. (1974). Applications of INDSCAL to studies of human perception and judgements. In: *Handbook of Perception, Vol. II*, Carterette, E.C. and Friedman, M.P. (eds.). New York: Academic Press, 449-491.

Amplitude Envelope and the Perception of Breathy Stops in Hindi

L. Schiefer and K. Kotten
Munchen, FRG

1. Introduction

In common with other Indo-Aryan languages of Northern Indian Hindi is characterized by a four-way contrast within the stop consonant series, the following phonological categories being present: voiceless, voiceless-aspirated, voiced, and breathy or murmured. While the first three categories are differentiated by Voice-Onset-Time (VOT), this does not hold for the description of the breathy stops. Breathy stops in Hindi are produced with prevoicing as the voiced stops are, but there is a significantly different release of the closure, i.e. the first part of the following vowel is breathy.

Fischer-Jørgensen (1967) and Dave (1977) who examined breathy and clear vowels in Gujarati did not find any difference in the formant structure of these vowels. The analysis of !Xõó and Gujarati carried out by Bickley (1982) showed that the relationship of the first and second harmonic is significantly different in breathy and clear vowels. The amplitude of the first harmonic is always higher than the amplitude of the second harmonic in breathy vowels.

The aim of the present study is 1) to check Bickley's results with Hindi material, and 2) to investigate the influence of the amplitude of the breathy part of the stop on the perception of breathy stops.

2. Material and Procedure

A list of single words containing all Hindi stop phonemes in initial position was produced by 4 native speakers of Hindi. The recordings were made in the sound-proofed room of the Institute. A digital copy of the material was prepared and the first syllable of all words containing a breathy stop in initial position was segmented in the following way. Four parts of the syllable were defined:

1. prevoicing,
2. burst+voiceless aspiration,
3. breathy part of the vowel,
4. steady part of the vowel.

In addition to these 4 parts all voiced portions were segmented into single periods to allow manipulation of these portions. Beside other values the following ones were calculated: the amplitude of the single pitch periods, average values for the amplitude of all harmonics for the single pitch periods, and the relation of the 1st and 2nd harmonic for the breathy and steady part of the vowel.

The relation between the 1st and the 2nd harmonic agreed with Bickley's results. The amplitude of the first harmonic was significantly greater in the breathy part than in the steady part of the words produced (from 3.46 dB to 20.02 dB). The comparison of the amplitude envelope of the breathy part with that of the steady part of the vowel also showed a significant difference. The amplitude of the breathy part was between 9.8 dB and 30.05 dB lower than that of the steady part. Taking these results as the point of departure test stimuli for two perception tests were produced by manipulation of naturally produced speech material.

3. Test Material

Previous investigations of the breathy stops in Hindi (cf. Schiefer 1983) showed that a breathy-to-voiced continuum can be generated by reducing the breathy part of the vowel in equal steps. The first stimulus of the continuum thus contains the whole breathy part of the vowel while the breathy part is completely eliminated for the last stimulus. Clear categorical perception between breathy and voiced is found.

The original stimulus chosen for the manipulation was taken from the word /dhobi/ containing the following values for the individual parts:

1. prevoicing 134 ms,
2. burst+aspiration 30 ms,
3. breathy part 105 ms,
4. steady part 154 ms.

The stimuli for TEST I were produced by reducing the breathy part in steps of 3 periods, i.e. in steps of about 15 ms, thus generating a continuum of 8 stimuli. The pitch periods used were taken from different parts of the breathy section to avoid distortion of the articulatory movements as much as possible. The newly defined stimuli were then constructed from the original stimulus with the help of a resynthesis program.

The stimulus used as point of departure for the manipulation had a burst of 30 ms with an amplitude of -45.7 dB and thus differed both in length and amplitude of the burst from normal realisations in breathy dental plosives. In order to study the influence of the amplitude as well as the duration of the burst the basic continuum was subjected to a further manipulation with respect to these two parameters. For the second continuum the amplitude of the burst was reduced by 30% in the third continuum the aspiration was eliminated and in the fourth continuum both these changes were made.

In order to investigate the influence of the amplitude envelope of the breathy part on the perception of breathy stops the whole breathy part of the vowel was eliminated in TEST II, generating a stimulus that consisted of prevoicing, burst+aspiration, and the steady part of the vowel; thus the breathy phonation characteristic of the category was removed. We then superimposed the amplitude contour of the breathy section on the first 21 pitch periods of the steady section by adjusting the amplitude of each period to the amplitude of the corresponding period. Pre-tests with phonetically trained staff and students of the Institute ensured that the manipulated stimulus was unambiguously assigned to the breathy category. The continua 2-4 of TEST II were generated in the same way as described for TEST I.

Two identification tests were prepared, each containing a set of 32 stimuli, which were repeated 5 times and were presented in randomized order with interstimulus intervals of 4 s and a pause of 10 s following a block of 10 stimuli. The tests were run at the Centre of German Studies at the Jawaharlal Nehru University in New Dehli. One group of listeners participated in the tests; all subjects were students of the Centre and native speakers of Hindi. The subjects were instructed that the presented stimuli were manipulated by computer. They were asked to decide to which phonemic stop category /t, th, d, dh/ the presented stimuli belong.

4. Results and Discussion

The results for TEST I show clear categorical perception in the identification test in the continuum from breathy to voiced (cf. Fig. 1). (For this preliminary investigation no discrimination tests were run. The tests were repeated in March 1983 using other stimuli as point of departure. This time clear categorical perception was found in both identification and discrimination tests.) The category boundary for all continua is located near the end of the continuum. The point of intersection for both breathy and voiced scores lies between the 6th and the 7th stimulus, to be precise at 6.54. A boundary shift was found in the other three continua, where the category boundary lies at 6.49, 6.3 and 5.97 (cf. Figs.2-4). This shift was found to be not significant in a correlated t-test.

The results for TEST II again show clear categorical perception from breathy to voiced; the category boundary lies at 6.99 (cf. Fig. 4). But it may be noticed that the identification rate is not as high as in TEST I. There is also a shift in the category boundary found for the continua 2-4 where the category boundary lies at 6.58, 6.26 and 5.93, which again is not significant (cf. Figs. 5-8). The values for the points of intersection for both tests are nearly the same.

The perception test run by Bickley (1982) for Gujarati showed that vowels were judged as breathy when the amplitude of the first harmonic was increased. Thus she assumed that 'the increased prominence of the fundamental with respect to the first harmonic is the acoustic correlate of breathiness'. The aim of our study was to examine the importance of the amplitude

Results of Test I

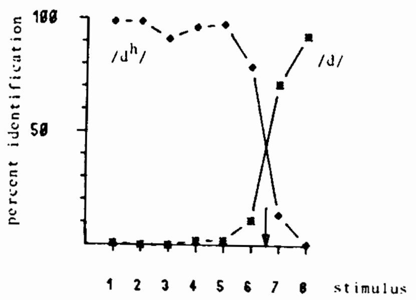


Figure 1. Continuum 1.

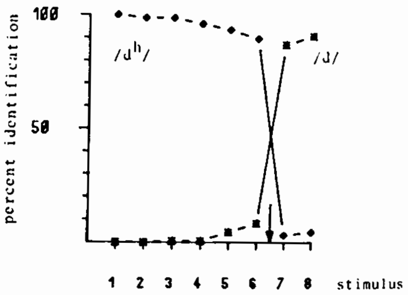


Figure 2. Continuum 2.

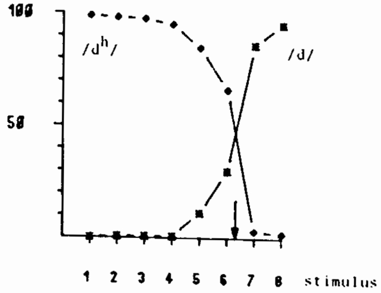


Figure 3. Continuum 3.

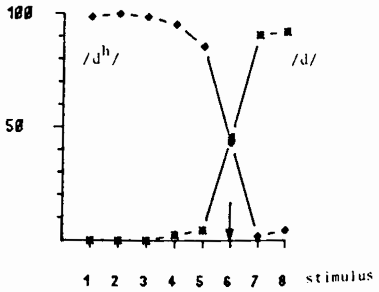


Figure 4. Continuum 4.

Results of Test II

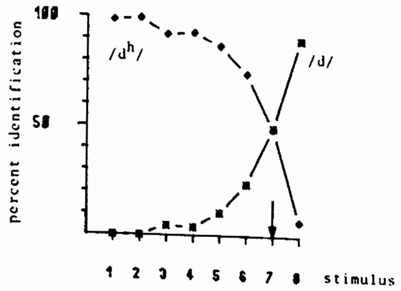


Figure 5. Continuum 1.

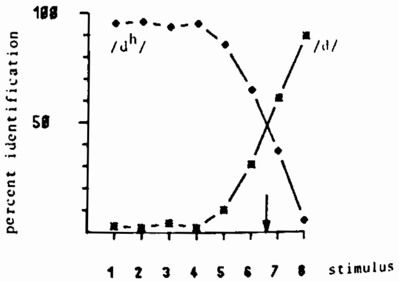


Figure 6. Continuum 2.

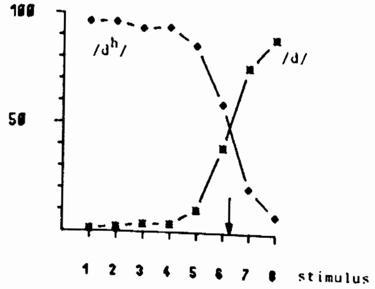


Figure 7. Continuum 3.

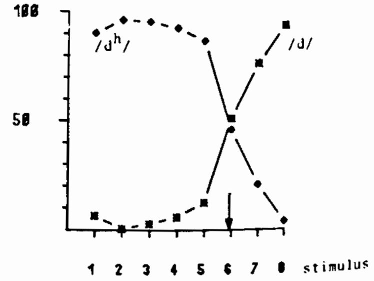


Figure 8. Continuum 4.

envelope of the breathy part of Hindi stops. Our results show that the amplitude envelope is also an important acoustic correlate for breathy stops. In addition we found that the manipulation of only 6 pitch periods is enough for the identification of breathy stops. Comparing the points of intersection of both breathy and voiced scores in both tests some interesting observations can be made. It is noticable -as found by Bickley too- that the breathiness of the productions does not contribute in a significant degree to the perception of breathy stops. The manipulation of the burst leads to nearly the same results in both tests. A shift in the category boundary was found in both tests, but it was obviously not significant.

This leads to the concluding observation, that rather than the 'acoustic content' of the signal the 'articulatory gesture' is judged by the listener (cf. Tillmann 1980). Further investigations should be undertaken to verify this hypothesis on the basis of a variety of natural stimuli differing in the duration as well as in the amplitude contour of the breathy part of the stop.

References

Bickley, C. (1982). Acoustic Analysis and Perception of Breathy Vowels. *Working Papers, Massachusetts Institute of Technology, Speech Communication Group*, Vol. **1**, 71-81.

Dave, R.V. (1977). *Studies in Gujarati phonology and phonetics*, Ph.D. thesis, Cornell University, Ithaca, New York.

Fischer-Jørgensen E. (1967). Phonetic analysis of breathy (murmured) vowels, *Indian Linguistics* **28**, 71-139.

Schiefer, L. (1983). Anwendungen von TISYS im Bereich der Analyse und Manipulation von Sprachsignalen - dargestellt am Beispiel der breathy Plosive des Hindi. To appear in: *Forschungsberichte des Instituts fuer Phonetik und Sprachliche Kommunikation der Universitaet Muenchen (FIPKM)* **17**.

Tillmann, H.G. and Mansell, Ph. (1980). *Phonetik. Lautsprachliche Zeichen, Sprachsignale lautsprachlicher Kommunikationsprozess*. Stuttgart.

Identification of Intervocalic Plosive Consonants: the Importance of Plosive Bursts vs. Vocalic Transitions

M.E.H. Schouten and L.C.W. Pols
Utrecht and Amsterdam, the Netherlands

1. Introduction

When plosive consonants occur between vowels, as they usually do in conversational speech, their information-bearing elements are silence, vocal murmur, the plosive burst, and the two vocalic transitions (VC and CV). In the literature there has been some controversy about the relative perceptual importance of burst, CV-transition, and VC-transition; for a short survey see Schouten and Pols (1983). It is clear that the plosive burst is the most invariant one of the three but it is also true that the vocalic transition is probably more resistant to masking, and may therefore be more important perceptually.

We have run a series of experiments in which subjects had to identify plosive consonants from signals in which various parts of those consonants had been deleted (replaced by silence). The aim was to find out what the relative contributions of the plosive burst and of the two kinds of vocalic transition are to the identification of plosive consonants. In this presentation we restrict ourselves to a subset of our data: we shall only discuss those conditions which all experiments had in common; a more comprehensive report will be published at a later date. There were four experiments:

1. Pairs of ambiguous sentences: each pair consisted of two nearly identical sentences, which differed only in one plosive consonant. Segments of this consonant were deleted in various ways, and the consonant had to be identified; subjects chose between the two possible alternatives.
2. In order to find out whether the sentence context, which, strictly speaking, was completely redundant, had any influence on the scores, we lifted the VCV sequences containing the relevant plosives from the sentences and presented them in isolation. Again, the two possible alternatives were available for subjects to choose from.
3. As a link to experiment 4, the stimuli of experiment 2 were presented to the same subjects, who this time were allowed to choose among the five plosive consonants Dutch has: /p, t, k, b, d/. This experiment was actually run before experiment 2.
4. In a re-run of the experiment described in Pols and Schouten (1982) and Schouten and Pols (1983), we asked subjects to identify the three plosive consonants in CVCVC nonsense sequences. There was, however, one difference: in the earlier experiment every transition was defined operationally as consisting of three pitch periods, but this time the transitions were defined as beginning (VC) and ending (CV) at the midpoints of the vowels. This made the segmentation procedure the same as the one employed in experiments 1, 2, and 3. As in experiment 3, subjects had five alternatives to choose from. In what follows, we shall only deal with the second of the three plosive consonants in the CVCVC utterance, since this was the only intervocalic one.

2. Method

In all four experiments the segment defined as the VC-transition preceding the burst started near the midpoint of the vowel and ended where the plosive burst (including vocal murmur or silent interval) began; everything was reversed for the CV-transition following the burst. The deleted segments were replaced by silence, and a 5 ms smoothing window was applied at the segmentation points. Four conditions are considered here: nothing deleted, only the burst left in, only VC left in, and only CV left in.

Experiment 1: ambiguous sentences

We constructed 50 pairs of sentences, each pair consisting of two sentences which were identical except for one plosive consonant (example: 'hij telde twee keer' = 'he counted twice' vs. 'hij belde twee keer' = 'he rang twice'). Each of the five Dutch plosive consonants occurred five times opposite each of the other four Dutch plosives in varying intervocalic contexts; since there were ten possible pairings, this resulted in 50 sentence pairs. The 100 sentences were read by one of the authors, who took great care to pronounce both members of each pair identically. The recordings were digitised (20 kHz, 12 bits) and regenerated in a random order of sentences and of segmentation conditions.

Subjects were 19 university students: half of them did this experiment before experiment 4 on the same day, and half of them did experiment 4 first. For each of the sentences they had to encircle one of two letters on a response sheet which gave the whole sentence. On the whole they enjoyed this task, which was experienced as quite a natural one.

Experiment 2: VCV segments from the sentences, 2 alternatives

The stimuli were the VCV segments containing the 'ambiguous' plosive consonants from experiment 1, from the middle of the preceding vowel to the middle of the following one. Either the whole segment was presented, or the burst, or one of the transitions.

Subjects were 24 university students: they took part in this experiment after experiment 3. They had to encircle the appropriate one of two letters on a response sheet; this sheet did not contain any information on the sentences. The task was strongly disliked.

Experiment 3: VCV segments from the sentences, 5 alternatives

This experiment, which always took place before experiment 2, and in which the same 24 subjects participated, was identical to experiment 2, except that subjects now had the five Dutch plosive consonants to choose from.

Experiment 4: CVCVC sequences

45 Meaningless CVCVC sequences were read out by the other author, digitised (12 kHz, 10 bits) and regenerated in a random order of sequences per condition. Whatever was deleted from one of the consonants was deleted from all three, so that subjects were presented with three bursts, or two CV-transitions, or two VC-transitions, or the whole utterance (as in all the experiments reported here, there were more conditions, but those do not concern us now). However, in the CV-condition the third consonant in the CVCVC-utterance was represented by its burst; the same applied to the initial consonant in the VC-condition. This was done in order to maintain the impression that each stimulus consisted of three consonants. Stimuli were embedded in carrier phrases.

Experiment 4 was alternated with experiment 1, and the same 19 subjects participated. Since this was the only one of the four experiments to be run on-line (tapes were used in the other three), in this experiment responses were given by typing in the appropriate letters. Subjects found it difficult to remember the correct order of the three plosives, but they expressed no strong dislikes.

3. Results and Discussion

All results that concern us here are summarised in Fig. 1, which shows the percentages of correctly identified plosives in the four experiments in the form of histograms. Unvoiced and voiced plosives are shown separately. The hatched bar in the middle of each histogram represents the condition in which only the plosive burst was heard; the bars to its left and to its right represent the conditions with only VC-transitions and only CV-transitions, respectively. The top horizontal line reflects the identification score for the uncut utterance.

We should like to highlight three features from Fig. 1.:

1. Overall, the burst elicits many more correct identifications than do either of the two transitions. This suggests that the vocalic transitions contribute relatively little to the perception of plosive consonants. Transitions help, presumably, but in many cases the plosive burst seems to contain enough information by itself. The results from experiment 4 contradict this conclusion to some extent. Since in experiment 4 the speaker was different from the one used in the other experiments, it could be that our main speaker just happens to be a good articulator of bursts. We would tend to support a different explanation, however: in experiment 4, subjects had to identify three plosives, which came in rapid succession. Hearing them as three vocalic transitions, one perceived a certain amount of continuity and tended to remember a (nonsense) word, whereas with three usually very brief plosive bursts the effect was one of three completely disjointed sounds suggesting a short burst of machine-gun fire. In such a situation low scores are almost inevitable, and the conclusion could be that vocalic transitions are indispensable under (normal) circumstances in which many consonants have to be identified in a short time: if transitions were absent, continuity could fall by the wayside.
2. Occasionally, the CV-transition rivals the burst in the percentage of correct identifications. This happens only in the case of voiced plosives, and then only in experiments 1 and 4. The same applies to the VC-transitions of unvoiced plosives in experiment 4. In experiment 4, these are probably the result of the low scores for the bursts, but in experiment 1 no such explanation is available. In fact, this is just one of the differences between the results from experiments 1 and 2. It should be remembered that stimuli and procedure were the same in experiments 1 and 2, except that in experiment 1 a supposedly completely redundant context surrounded the stimuli. Either the context was not as redundant as we believed it to be, or the 'naturalness' of the stimuli does have a strong influence on identification behaviour, turning CV-transitions into important cue-bearers in voiced plosives.

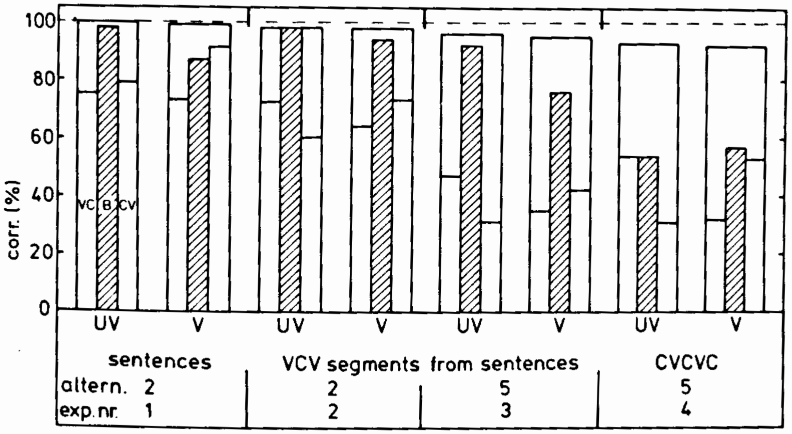


Figure 1: The results in terms of percentage of correct plosive identification responses in all four experiments. The hatched bar represents the score for the bursts without the vocalic transitions; the bars to its left and right represent the scores for the VC- and the CV-transition respectively. The top line in each histogram indicates the score for the whole (uncut) utterance. UV means unvoiced; V means voiced. Chance levels are 50% in experiments 1 and 2, and 20% in experiments 3 and 4.

3. There appears to be a fairly systematic difference between VC- and CV-transitions: the latter are more important in voiced plosives, whereas the former are more important in unvoiced plosives. Contrary to English, there is virtually no voice-onset-time between the burst and the vocalic parts of Dutch plosives, so our finding here may be a language-specific one. We predict, however, that voice-onset-time cuts off a part of the CV-transition and thus reduces the overall score for CV-transitions, but should not affect the relative scores. For the time being, therefore, we claim that our finding should also hold for a language such as English.

We realise that forcing people to listen to utterances from which bits have been deleted does not tell us everything about the role of the remaining bits in normal connected speech. By introducing silence into an utterance one does not simply eliminate cues - one also adds new ones. We have attempted to avoid that trap by conducting an experiment with sentences in which one plosive was mutilated, and which the subjects regarded as quite natural: they hardly noticed anything unusual about the sentences. However, we have not managed to avoid the trap: in all experiments, including the sentence experiment, the gaps in the stimuli provoked a strong tendency towards /p/- and /b/- responses, a tendency which seemed to overrule subjects' written preference for one sentence rather than the other of a pair. These things will, however, have to wait for future evaluation.

References

L.C.W. Pols and M.E.H. Schouten (1982). Perceptual relevance of coarticulation. In: R. Carlson and B. Granström (eds.). *The representation of speech in the peripheral auditory system*. Amsterdam, Elsevier Biomedical Press, 203-208.

M.E.H. Schouten and L.C.W. Pols (1983). Perception of plosive consonants. The relative contributions of bursts and vocalic transitions. In: M.P.R. van den Broecke, V.J.J.P. van Heuven and W. Zonneveld (eds.), *Sound Structures*, Dordrecht, Foris.

The Effects of Visually Presented Speech Movements on the Perception of Acoustically Encoded Speech Articulation as a Function of Acoustic Desynchronization

H.G.Tillmann, B. Pompino-Marschall, U. Porzig
Munich, Federal Republic of Germany

1. Introduction

Under certain conditions, visually presented speech movements have a strong influence on what is "auditorily" perceived when the acoustic signal contains the mapping of another speech movement. If, for instance, an acoustic [ga] is synchronized to a seen <ba> movement and not presented too clearly to the listener he perceives a heard "[da]". In a series of pretests we found that the different effects of VPSM (Visually Presented Speech Movements) on the perception of AESA (Acoustically Encoded Speech Articulation), which have been observed since the appearance of McGurk and Mac Donalds (1977), cannot be adequately accounted for by their division into fusions and combinations, e.g. in the respective cases of <ga> + [ba] = "[da]" and <ba> + [da] = "[bda]".

The 'winning eye' effects are to be judged quite differently depending on whether or not the subjects see that a labial articulation is taking place. For convenience we put the description of VPSM in angled brackets and if there is a difference between <VPSM> and [AESA] we enclose what is actually auditorily perceived in quotation marks. To express that it is true that the labial movement is visually present, we write <+L>, when it is not present <-L>.

Yet the nondominant auditory mode may also gain in influence depending on how clearly the true 'articulatory content' (cf. Tillmann 1980, 68ff, 244ff) of the given utterance is mapped onto the acoustic speech wave. Along these lines one finds an explanation for combinations such as "[bda]". Here we would like to make quite another observation. In a noisy computer room, looking at a not so clearly visible <ga>, which is presented with its original [ga] leads in most cases to a perceived "[da]".

We would like to distinguish three different VPSM/AESA effects. If the contradiction between <+L> and [+da] leads to the perception of "[bda]", we will speak of a phonetic combination. If in cases of <-L> the place of articulation of [+L] is moved from the lips into the mouth of the speaker (<ga> + [ba] = "[da]") we speak of the resulting "[-L]" as a phonetic fusion; two different phonetic categories fuse into a new category in between. But if in the case of <+L> there is also a transfer of labial manner of articulation, say "[p-]" or "[b-]", we would prefer to call the resulting combination a phonemic fusion, because this effect strongly resembles the effect of phonological fusion found in dichotic listening experiments (Cutting 1976). This is clearly the case with <+ba> + [la] = "[bla]".

In the experiments described below we are less interested in producing the different VPSM/AESA-effects but rather in destroying them by systematically desynchronizing the temporal coincidence of <VPSM> and [AESA].

2. Experiment I

In our first experiment we tested the phonetic fusion using the two German words "Gier" and "Bier", taking the VPSM from the first word and the AESA from the second one. We expected that <Gier> + [Bier] would result in "[dir]", which is also a German word. To prepare a test tape for the Sony-Umatic-recorder a female speaker was filmed uttering the two sentences "Ich habe - Gier - gesagt", "Ich habe - Bier - gesagt". The hyphens indicate a pause of nearly 1 s. As the lips were closed during the pause before "Gier" any preparatory tongue movements were masked and only the <g>-release could be seen. The plosive of the word "Bier" was produced with a short noticeable lip pressing.

The test tape had six blocks each consisting of eight copies of the "Gier"-sentence and of two copies of the "Bier"-sentence randomly inserted. The original speech signals from the test tape were digitally recorded and properly segmented. Based on this segmentation (for the details of this procedure cf. Tillmann, 1983) the second track of the test tape received the sentence frame via direct AD/DA-conversion from the original track whereas the desynchronized [Bier]-signals came from the computer memory. The desynchronizations of the acoustic signals had the values of 200, 250, 300, 350, 400 and 500 ms (deviation 0.05 ms) in positive and negative direction, respectively, and each of these items occurred four times (the zero condition was omitted because in pretests we had found that the effect of phonetic fusion is very resistant to small desynchronizations). The dummy "Bier"-copies received the original acoustic signal without desynchronization via direct AD/DA-conversion.

The tape was presented to 17 subjects who were instructed to watch the screen and to report whether they had heard "Bier" or "dir". It appeared that 10 of the subjects always heard the original acoustically given word "Bier", which means that they did not show any phonetic fusion. According to our pretests this was probably the case because no zero delayed <Gier> + [Bier] items were used in this test. The results of the seven fusioners are given in Fig. 1.

The same tape was presented a second time to the subjects who were now instructed to judge the quality of the synchronization as good or incorrect. The results of the seven fusioners are given in Fig. 2.

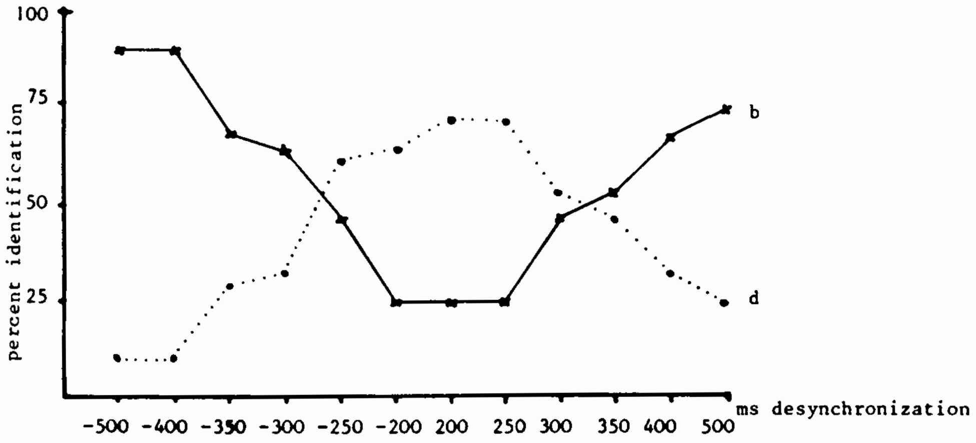


Figure 1. Identification results of experiment I ('undecided' responses omitted).

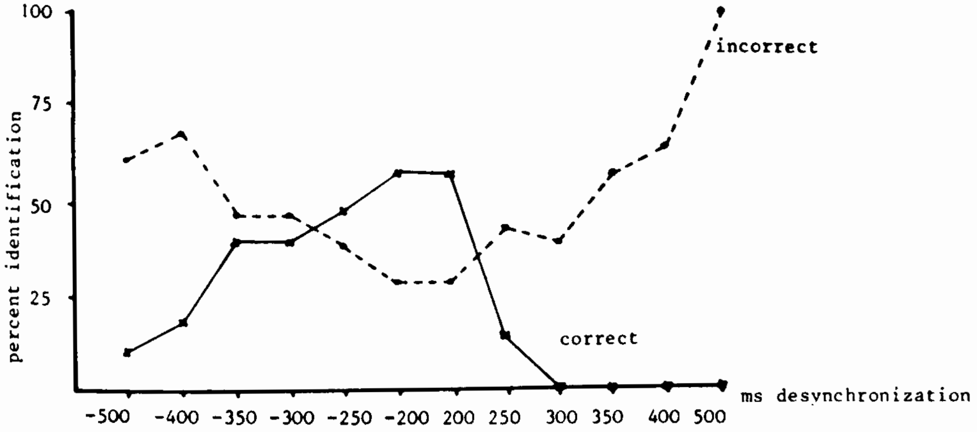


Figure 2. Results of quality judgement in experiment I ('undecided' responses omitted).

3. Experiment II

Phonemic fusion was tested in Exp. II where we used the German sentences "Ich habe - ba - gesagt", "Ich habe - la - gesagt". We expected that seen <ba> and acoustic [la] would give the fusion "[bla]". The tape was prepared as in Exp. I but we introduced only positive desynchronizations (acoustic delays) in this case: 0, 100, 200, 300, 400 and 500 ms. The tape was presented to 6 subjects who were instructed to report whether they had heard "bla" or "la". The results are shown in Fig. 3.

Analogously to Exp. I the second quality judging test was run. The results are given in Fig. 4.

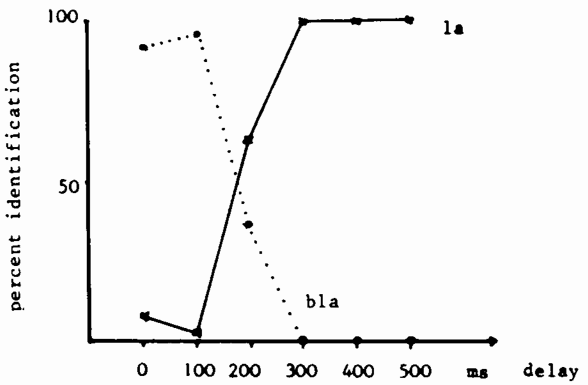


Figure 3. Identification of experiment II ('undecided' responses omitted).

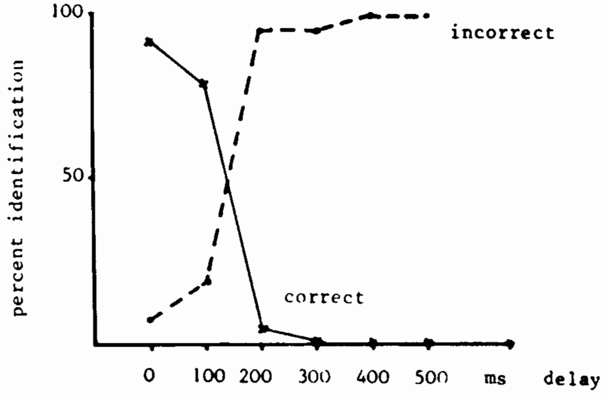


Figure 4. Results of quality judgement in experiment II ('undecided' responses omitted).

4. Discussion

The data of the fusioners in Exp. I (cf. Fig. 1) show that for one group of subjects there is a wide range of desynchronization where phonetic fusions stay predominant, the range going from -250 to +300 ms. We had already found this kind of asymmetry in our pretests. The data of the phonemic fusion experiment (cf. Fig. 3) show that in this case the timing relations are much more critical. The influence of the dominating eye breaks down as soon as the delay of the acoustic signal is more than 100 ms. Of interest is also the fact that exactly synchronized <ba> + [la] result in fewer "bla"-responses than in the case of the first desynchronized stimulus pair. This indicates that for phonemic fusion the timing relations of natural speech productions play a more critical role than in the case of phonetic fusion. A corresponding effect can also be seen in the judgements of the quality of the synchronization (cf. Fig. 4).

In general, the quality judgements are in agreement with the identification results. We find it very interesting that subjects react so much more critically to desynchronization in the case of phonemic fusion than in the case of phonetic fusion.

Further experiments are planned to determine more closely those articulatorily different conditions that influence the decreasing VPSM/AESA-effects as a function of desynchronization between VPSM and AESA.

The next experiment which has been prepared but not yet conducted concerns a situation where we have an amalgam of phonetic combination and phonemic fusion. As we have seen in the first pretests, <Bier> + [Gier] produces the phonetic combination "[bg]". We would like to see whether under the respective desynchronization conditions the phonemic fusions "[B'gier]" and "[G'bier]" result, since these could be understood by the subjects as allegroforms of the German words "Begier", "gebier".

Acknowledgement

We thank Prof. Müller, Münchner Hochschule für Fernsehen und Film, and his staff, especially Dr. Bamberg and Mrs. Schumann for helping us in cutting the original tapes.

References

Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening. *Psychol. Rev.* **83**, 114-140.

McGurk, H., Mac Donald, J. (1976). Hearing lips and seeing voices. *Nature* **264**, p. 746ff.

Tillmann, H.G. (1983, in press). Intra- und heteromodale Isochronie und temporale Koinzidenz von kategorialen Kontinua. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* **17**.

Section 9
Word Perception

Listening for Phonemes while Reading

M. Martin
Oxford, United Kingdom

1. Introduction

Investigations of the psychological processing of two or more concurrent linguistic signals have provided extensive evidence for general limits in central processing capacity (e.g.,Martin, 1978, 1980). A finding reported by Martin (1977), however, cannot be accounted for in this way. The auditory detection of phonemic targets (stop consonants) was assessed with or without concurrent reading and compared with that of syntactic targets (adjectives) and semantic targets (temporal reference words). It was found that relative to single-task performance (which was equated in the three conditions), dual-task performance was impaired to an approximately equal extent in syntactic and semantic conditions, and that the impairment was considerably greater in the phonemic condition.

Martin (1977) was able to discard a possible explanation of the phoneme-specific impairment in terms of the distinction between data-limited and resource-limited performance (Norman and Bobrow, 1975). If the three conditions differed in which of these two types of constraint were operative, then rate of presentation of stimuli should have exerted a differential effect. In practice however it did not. Thus an alternative explanation for the finding appears necessary.

The hypothesis to be examined here is the specific attentional impairment arising with stop consonants as a consequence of their very brief duration. It may be the case that in order to detect and discriminate stop consonants, subjects are forced to process in much shorter decision units, using a fine-timing mechanism. If this fine-timing mechanism were also involved in the control or allocation of attention when two tasks are carried out concurrently, it would account for the observed results.

In order to test this hypothesis a new experiment is reported in which detection of stop consonants is compared with that of several other phonemic targets. This allows comparison of (a) stop consonant detection with longer consonant detection, (b) short vowel detection with longer vowel detection, and (c) stop consonant detection with another short consonant detection. The stop consonant, other short consonant and short vowel used were /t/, /n/, and /ɪ/ while the long consonant and vowel were /s/ and /aʊ/. The durations in the short group, as in the long group, were approximately equal (Umeda, 1975; 1977). Thus the fine-timing hypothesis predicts significant effects within comparison (a) and (b), but not within comparison (c).

2. Method

Subjects

The participants were 20 right-handed males from the Oxford Subject Panel aged between 18 and 35 years.

Material

For the auditory monitoring task fifteen lists of 224 each were prepared. There were three lists for each of the five target phonemes /t/, /n/, /s/, /ɪ/, and /aʊ/. Each list contained examples of only one phoneme target type. There were twenty target words in each list embedded in 204 distractors which did not contain the target phoneme. Target words were arranged throughout the list, separated by an average of ten and a minimum of seven distractors, with on average the first five words and last nine words of each list also distractors. Target and distractor words were all polysyllabic. Target phonemes occurred once only in each target word, in any position excluding initial and final phoneme. The lists were spoken by a North American, as in the Umeda (1975; 1977) studies on vowel and consonant duration, at a rate of 1 word per second. Each list commenced 2 sec after a warning signal.

For the reading task seven passages of text were taken from the 'Discourse on Method' of Descartes (1968). This allowed the selection of passages that were two printed pages in length but contained no paragraph indentation.

Apparatus

The word lists were presented via headphones from a tape recorder. A response key was connected to a pure tone oscillator whose signal was fed through a mixer with that of the tape recorder into one channel of a second tape recorder. On the second channel of this tape recorder the subject's voice while reading was recorded.

Procedure

Each participant was tested individually. The experiment proper was preceded by practice trials on each of the types of detection tasks in isolation, on reading in isolation, and on both those concurrently. The order of occurrence of the different conditions in the experiment was balanced over subjects using balanced predecessor Latin squares.

3. Results

Analysis of variance showed that the speeds of reading in isolation and in the five target conditions differed significantly, $F(5,95) = 37.13, p < .001$, with mean speeds in isolation and with target detection of 223.4 and 158.0 words/minute, respectively. Duncan's multiple range test showed that while detecting consonants the reading rate (in words/minute) was slower for /t/ and /n/ (156.2 and 153.6, respectively) than for /s/ (172.3), and that while detecting vowels the rate was slower for /ɪ/ (145.9), than for /aʊ/ (162.1), $p <$ .05 in each case.

Mean target detection rates in isolation and while reading differed significantly at 84.6%, and 65.8%, respectively, $F(1,19) = 63.94, p < .001$. However, there was no significant interaction between this factor and the five types of detection, $F(4,76) = .52$, with dual-task detection at 79.5%, 80.2%, 79.6%, 74.8% and 74.2% of performance in isolation for /t/, /n/, /s/, /ɪ/, and /aʊ/, respectively.

4. Discussion

The results of this experiment provided evidence that the detection of a stop consonant is more injurious to dual-task performance than that of a longer fricative, while its effects do not differ significantly from those of another short, nasal consonant. Similarly, a short vowel is more injurious than a long one. The investigation of phoneme perception within a dual-task setting has thus been shown to provide significant evidence concerning the temporal characteristics of the human information-processing system, and in particular suggests that specific demands upon a posited fine timing mechanism may be placed both by phoneme detection and by the co-ordination of dual-task performance.

Acknowledgement

The author is grateful to the MRC and the SERC for their support of this research.

References

Descartes, R. (1968). *Discourse on Method* (trans. S.E. Sutcliffe). Harmondsworth, Middlesex: Penguin.

Martin, M. (1977). Reading while listening: A linear model of selective attention, *Journal of Verbal Learning and Verbal Behavior*, **16**, 453-463.

Martin, M. (1978). Retention of attended and unattended auditorily and visually presented material. *Quarterly Journal of Experimental Psychology*, **30**, 187-200.

Martin, M. (1980). Attention to words in different modalities: Four-channel presentation with physical and semantic selection. *Acta Psychologica*, **44**, 99-115.

Norman, D.A., and Bobrow, D.G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, **7** 44-64.

Umeda, N. (1975). Vowel duration in American English. *Journal of the Acoustical Society of America* **58**, 434-445.

Umeda, N. (1977). Consonant duration in American English. *Journal of the Acoustical Society of America,* **61**, 846-858.

Speech Quality and the Gating Paradigm

S.G. Nooteboom and G.J.N. Doodeman
Eindhoven, the Netherlands

1. Introduction

The correct recognition of a word from speech can often take place when only part of the spoken word form has been heard (e.g. Marslen-Wilson, 1980). In such a case the remainder of the word is, in some sense, redundant information. Of course, redundant does not mean superfluous. Redundancy serves to make speech communication less vulnerable to all kinds of degradation of the 'ideal' speech signal, caused for example by sloppy articulation, external distortion, or a hearing deficit in the listener.

In the present experiment we have set out to measure the effect of differences in speech quality, caused by differences in degree of data reduction in LPC vocoder speech, on the relative number of speech sounds needed for correct recognition of polysyllabic words. For this purpose we used an adaptation of the 'gating paradigm' introduced by Grosjean (1980).

Our aim was twofold. We wanted to find out whether we could obtain a reliable and relatively easily applied measure of speech quality. We also wanted to see whether the course of the probability of correct recognition, as controlled by the successively added speech sounds, has any diagnostic value with respect to the type of degradation of the speech signal.

2. Method

A set of 40 Dutch polysyllabic words was selected, with frequencies of usage of 10 or more per 720.000 words in the Uit den Boogaart word frequency count (Uit den Boogaart, 1975). Optimal spoken realisations of these words by a speaker of standard Dutch were recorded and stored on disk in digital form (pcm, 12 bit per sample, 10 kHz sampling frequency). From each word token an initial fragment was isolated, corresponding to the beginning of the word, and containing several speech sounds. This fragment was chosen such that it was long enough to successfully apply LPC analysis and resynthesis, and short enough to ensure a low probability of correct recognition. Further versions of the same word token were produced by adding segments of speech corresponding to successive speech sounds to the initial fragment. This was done under visual and auditory control. An example of a phonetic transcription of consecutive fragmentary word tokens of one word, in this case the word AUTORITEIT (Engl. AUTHORITY) is:

1. [oto], 2. [otor], 3. [otori], 4. [otorit] 5. [otoritɛi], 6. [otoritɛit].

All 40 sets of word fragments were prepared in four speech qualities:

1. the original digital recording, using 120,000 bits per second;
2. vocoder speech, obtained with an LPC-to-formant analysis-resynthesis system, using 16,000 bits per second (Cf. 't Hart, Nooteboom, Vogten and Willems, (1982);
3. idem, with further data reduction by parameter quantisation to 4,000 bits per second;
4. idem, with still further reduction to 1,000 bits per second.

From these 40 sets of word fragments in the four speech qualities, four stimulus tapes were prepared. Each tape contained four groups of ten words, each group in a different speech quality. Each group of ten words appeared in a different speech quality on each of the tapes. The order of speech qualities on each tape varied randomly from one word to the next.

Each tape was played over headphones to a different group of five listeners, who were tested individually. After the presentation of each fragment listeners were encouraged to guess and say aloud the word from which the current fragment was taken. If not able to guess, they were asked to repeat aloud the fragment heard. After each correct guess the experimenter switched to the next set of word fragments. Stimuli and responses were recorded on two separate tracks of a magnetic tape for later analysis.

3. Results

The results presented here will be limited to probabilities of correct recognition as a function of the number and kind of added segments. Probability of correct recognition as a function of the number of speech segments added to the initial word fragment, for the four speech qualities separately and averaged over all words and all subjects, is given in Fig. 1.

The difference between each pair of curves is significant ($p < 0.05$) on a sign test applied to estimated means for individual words in different conditions. As expected, the number of audible segments necessary for correct recognition systematically increases with decreasing speech quality.

In search for diagnostic indications in our data, we have calculated the relative contribution of consonant and vowel segments to correct recognitions. The proportion of the total number of correct recognitions occurring immediately after adding a vowel segment, and the proportion occurring immediately after a consonant segment, in the four speech qualities, is plotted in Fig. 2. We see that with decreasing speech quality the relative contribution of vowel segments increases at the cost of consonant segments.

We also investigated the relative contribution of stressed and unstressed syllables to recognition. For this purpose we focused on those 27 of the 40 words in which the initial word fragment did not contain the lexically stressed syllable. For each of those words we numbered the added segments,

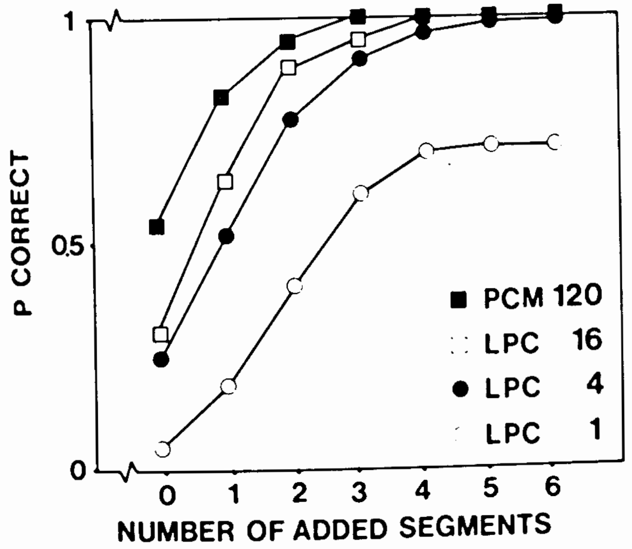


Figure 1. Probability ot correct word recognition as a function of the number of sound segments added to the initial word fragment, for four speech qualities.

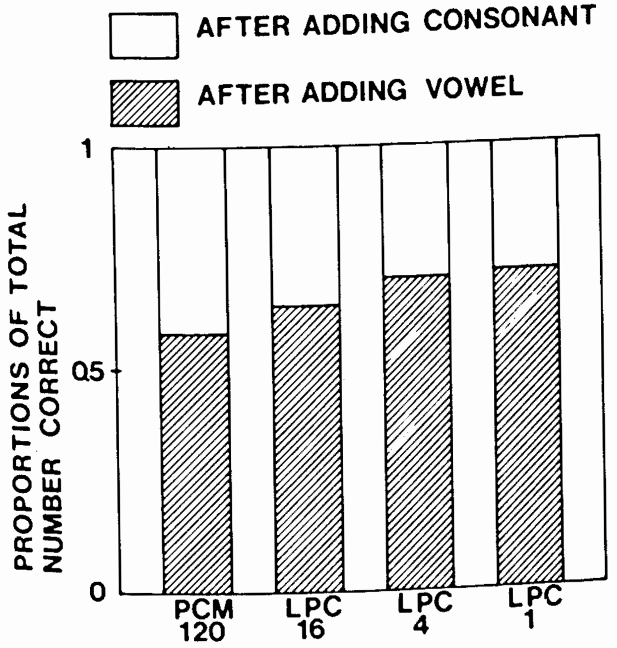


Figure 2. Proportions of total number correctly recognized words after adding a vowel or a consonant segment, for four speech qualities.

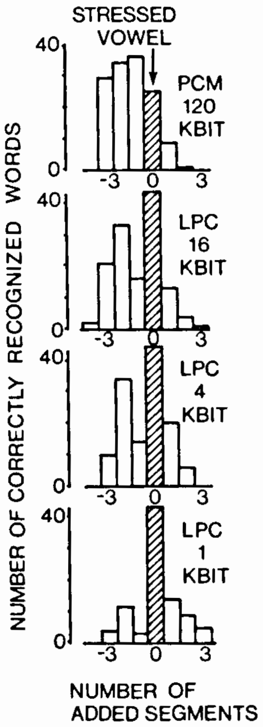


Figure 3. Frequencies of correct word recognition as a function of the position of the added segment. This position is taken relative to the position of the lexically stressed vowel.

starting with 0 for the vowel of the stressed syllable, negatively towards earlier and positively towards later segments. We then made frequency distributions of correct recognitions over the numbers of added segments. These are presented in Fig. 3. Obviously, as speech quality decreases correct recognition becomes more and more dependent on the availability of the vowel of the lexically stressed syllable.

4. Discussion and Conclusion

The results of this experiment show that the 'gating paradigm' can fruitfully be applied to the problem of measuring differences in speech quality. It proved possible to find significant differences between the four speech qualities used, with 40 words and only a few listeners per word, suggesting that measurement of relative speech qualities can be fairly easy and fast, given the availability of prepared sets of word fragments. The discriminative power of the test compares favourably with an adaptation of the Nakatani and Dukes (1973) test, as applied to approximately the same speech qualities by Vogten (1980). As exemplified in the results section, a simple analysis of the data distribution may give useful indications which parts of the speech signal are most seriously damaged in each speech quality.

References

Grosjean, F. (1980). Spoken word recognition and the gating paradigm. *Perception and Psychophysics,* **28**, 267-283.

Hart, J. 't, Nooteboom S.G., Vogten, L.L.M. and Willems, L.F. (1982). SPARX: manipulation of speech sound. *Philips Technical Review,* **40**, 134-145.

Marslen-Wilson, W.D. (1980). Speech understanding as a psychological process. In: J.D. Simon (Ed.) *Spoken Language Generation and Recognition.* Dordrecht: Reidel.

Nakatani, L.H. and Dukes, K.D. (1973). A sensitive test of speech communication quality. *Journal of the Acoustical Society of America,* **53**, 1083-1092.

Uit den Boogaart, P.C. (1975). *Woordfrequenties in geschreven en gesproken Nederlands.* Utrecht: Oosthoek, Scheltema en Holkema.

Vogten, L.L.M. (1980). Evaluation of LPC formant-coded speech with a speech interference test. *IPO Annual Progress Report,* **15**, 33-41.

The Detection of Mispronunciations and the Influence of Context

I.B. Ottevanger
Utrecht, the Netherlands

1. Introduction

This report reviews a series of experiments investigating word recognition and the influence of context. The series was set up to test the cohort theory of word recognition (Marslen-Wilson and Welsh, 1978), a model which makes precise predictions about the way in which recognition takes place. In short it claims that in the perception of speech 'the word' is the level at which data-driven and knowledge-driven processing strategies are optimally co-operative. The model assumes that on the basis of acoustic information a word-initial 'cohort' is activated, which contains all words in a language that begin with the same two or three phonemes as the input word. Next, word candidates are removed from the cohort as soon as their acoustic characteristics are no longer compatible with the acoustics of the flow of new input; the same happens when word candidates are in conflict with contextual specifications. When one word candidate is left, word recognition has been achieved. Going from left to right in the word, the phoneme that distinguishes that word from all others in the cohort is called the recognition point.

2. Method

2.1. Stimuli

Recognition points of twelve Dutch polysyllabic words were determined with the aid of a standard Dutch dictionary (Kruyskamp, 1976). Each word was mispronounced by changing one phoneme into another at four or five successive points, the 3rd point being the phoneme that functioned as the recognition point. Care was taken that the initial two phonemes of the stimulus words and the final one were not mispronounced, so that word boundaries were kept intact. Other requirements were that all mispronunciations were phonotactically legal and that the initial part of the words up to and including the misplaced phoneme was not identical with the beginning of any other Dutch word.

The stimulus words were spoken in isolation and in a final position in short auditory[1] context sentences. These sentences were alternative versions of the phrase 'The next word is ....' Out of the auditory context sentences the words

were spliced onto structural context sentences, again sentence-finally. The latter set consisted of five sentences which were ambiguously constraining, i.e., syntactically and semantically they led both to the target word and at the same time to another word that shared its first two phonemes with the stimulus word. The extent to which targets and alternatives turned out to be predictable on the basis of preceding context combined with the acoustics of the first phonemes was $\bar{X}$ 52% and 55%, respectively. The remaining seven sentences were uniquely constraining: syntactically and semantically they led to the target words only. On the basis of prior context and acoustic characteristics of the first phonemes their mean predictability was 92%.

2.2. Procedure

For each of the three conditions, isolation, auditory context and structural context (ambiguous and unique), five groups of eight to ten subjects were instructed to listen for mispronunciations (cf. Cole, 1973) and to press a response key as soon as an error was heard. The four or five different mispronunciations of a target word were presented to the different groups of subjects. Reaction times (RTs) were measured from the onset of the mispronounced phoneme.

2.3. Predictions

On the assumption that word recognition is prior to error detection, the cohort model predicts long RTs to mispronounced phonemes preceding the recognition point or coinciding with it, and short RTs to errors following the recognition point.

Because in the isolated condition recognition is based on the interaction of acoustic input and lexical knowledge only, the cohort theory predicts long RTs to mispronounced 1st, 2nd and 3rd points (the 3rd point being the recognition point), and short RTs to 4th and 5th points. The same prediction applies to the auditory context condition: since auditory context has no power to remove word candidates from the cohort, words are predicted to be recognized at the same point as when presented in isolation.

For the ambiguous structural context condition the model predicts long RTs to mispronounced 1st and 2nd points, since on the basis of acoustic input, lexical knowledge and contextual constraints two word candidates (the stimulus word and the alternative word) are left in the cohort, and therefore, word recognition has not yet taken place. Short RTs are predicted to 3rd, 4th and 5th points, because one of the two candidates has now been removed on account of its incompatibility with the acoustic input.

[1] The term is taken from Pollack and Pickett (1964); auditory context implies that the sentence-final word is not constrained by prior context in a syntactic and/or semantic manner, in the case of structural context prior context does have such constraints on the sentence-final word.

In the case of the seven uniquely constraining structural context sentences the cohort theory claims that word recognition has occurred on the basis of context well in advance of the earliest mispronunciation point and all RTs should be short.

3. Results

The results of the detection experiments are displayed in Table I and graphically represented in figure 1.

A one-way analysis of variance showed that for the isolated condition RTs to 4th and 5th points were significantly shorter than to 1st, 2nd and 3rd points as predicted (F(3,370)=2.90, $p < .05$); for the auditory and the ambiguous structural context condition there was no significance. For the unique structural context condition there was a highly significant difference between RTs to 1st and 2nd points on the one hand and 3rd, 4th and 5th points on the other (F(4,351)=4.69, $p < .01$); this was not in accordance with the prediction that RTs to the successive points would be equally short.

4. Discussion and conclusions

The recognition of words spoken and presented in isolation is adequately accounted for by the cohort model. The same is not true when words are presented in auditory and structural context. In auditory context the pattern which reflects the crucial role of the recognition point in word recognition, namely the large difference between the 3rd and the 4th mispronunciation point, has disappeared. The ambiguous structural context sentences have not achieved that recognition occurs at an earlier point in the word. The unique structural context results show that recognition has taken place at an earlier point, but not so early as the interaction of acoustic analysis, lexical knowledge and syntactic/semantic constraints permits.

For a more elaborate discussion of these results and for a presentation of the complete stimulus set of which these stimuli were a subset, the reader is referred to Ottevanger (1982; 1984).

Table I. Mean RT and standard error in ms per mispronunciation point for the three conditions

	Isolated	Auditory context	Structural context	
			ambiguous	unique
1st point	833 (25)	769 (31)		753 (75)
2nd	831 (18)	717 (24)	659 (50)	613 (28)
3rd	790 (23)	642 (26)	644 (58)	455 (28)
4th	601 (17)	556 (21)	660 (97)	454 (24)
5th	576 (18)	495 (18)	567 (59)	458 (19)

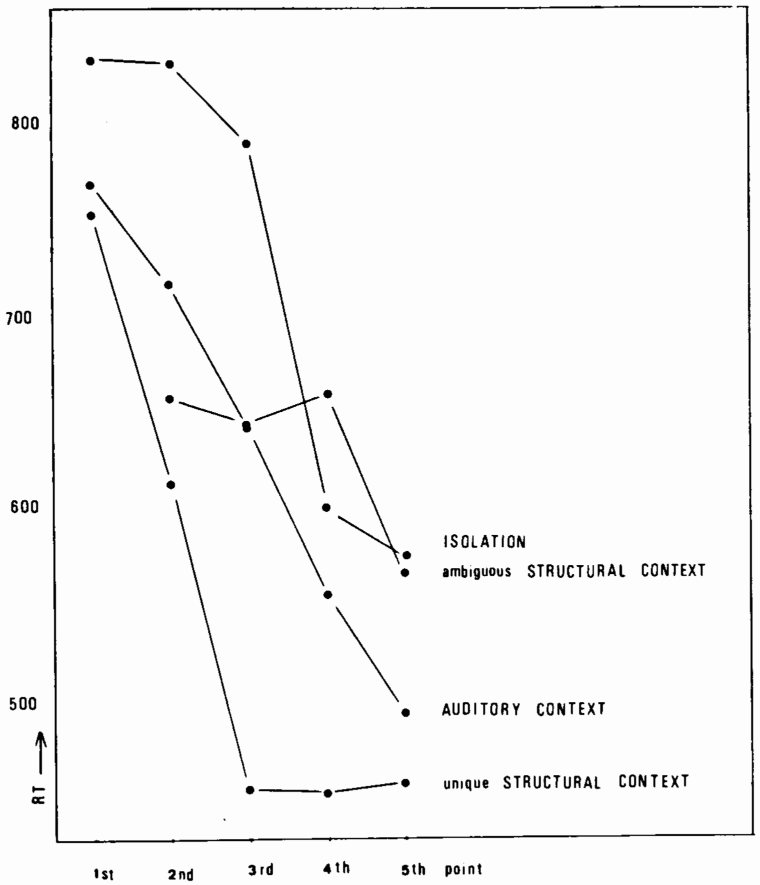


Figure 1. Mean RTs to the successive mispronunciation points in the target words as found in the isolated condition, in auditory context and in structural context (ambiguous and unique).

It can be concluded that, although reactions to mispronunciations in words presented in context are *faster*, context has no accelerating effect on word recognition in the sense that words are recognized *earlier*.

The finding that RTs are shorter to mispronunciations in words presented in auditory context compared to isolated words, fits in well with the results of Pollack and Pickett's (1964) experiment, in which they found that additional context contributed to the intelligibility of excerpts even though the contents were known to their subjects beforehand.

Finally, it is striking to see that, however long RTs to mispronunciations in isolation are, standard errors are small, indicating much conformity between subjects. The same degree of conformity is not found in the other conditions; the extremely high standard errors for ambiguous structural context indicate that subjects were very much hampered by the presence of alternative word candidates.

Acknowledgement

This research was made possible by a grant of the researchpool of Utrecht University.

References

Cole, R.A. (1973). Listening for mispronunciations: a measure of what we hear during speech. *Perception and Psychophysics* **11**: 153-156.

Kruyskamp, C. (1976). *Van Dale's groot woordenboek der Nederlandse taal*, 10th edition, Martinus Nijhoff, 's-Gravenhage.

Marslen-Wilson, W.D. and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology* **10**: 29-63.

Ottevanger, I.B. (1982). Word recognition in non-constraining context in comparison with recognition of isolated words. *Progress Report of the Institute of Phonetics* (PRIPU) **7**,(2): 41-56.

Ottevanger, I.B. (1984). Word recognition in syntactically and semantically constraining context. To appear in *Progress Report of the Institute of Phonetics* (PRIPU) **9**,(1).

Pollack, I. and Pickett, J.M. (1964). Intelligibility of excerpts from fluent speech: auditory vs. structural context. *Journal of Verbal Learning and Verbal Behavior* **3**: 79-84.

Effects of Anticipatory Coarticulation on Vowel Detection in Meaningful Words

V. J. van Heuven and M. Ch. Dupuis
Leyden, the Netherlands

1. Introduction

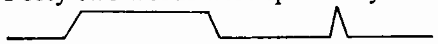
Acoustic and perceptual consequences of coarticulation have been extensively investigated for CV and VC structures in meaningless syllables, using either human or synthetic (pattern playback) speech. Generally, in VC syllables the place of articulation of C is clearly perceptible in the preceding V, and such characteristics as lip rounding, vowel height and backness are audible in the initial C of a CV syllable. Moreover, perceptual effects may not be confined to neighbouring sounds: the formant frequencies and transitions of V_1 in a V_1CV_2 sequence are systematically affected not only by C but also by V_2 (Öhman, 1966; Lehiste and Shockey, 1972). In spite of several attempts, however, it has proven very difficult to ascertain whether such anticipatory coarticulation between such non-adjacent sounds enables the listener to recover the identity of V_2 (or some feature of it) from an earlier portion of the utterance if both V_2 *and* the preceding C are (electronically) removed from the stimulus (cf. Lehiste and Shockey, 1972; Benguerel aud Adelman, 1975). This would seem to suggest that useful coarticulation information does not extend beyond adjacent sounds.

Recently, Martin and Bunnell (1981) were able to show that vowel detection latencies for V_2 were slightly (9 ms) faster with a properly coarticulated V_1 than when V_1 contained acoustically conflicting information after cross-splicing. However their stimuli were meaningless, and phonologically illegal structures of the type /kæ'zi:, kæ'za:, ku:'zi:, ku:'za:/, with a non-reduced first vowel, and stress on the second. It would be of interest to see if more convincing effects can be found across word boundaries rather than word internally using meaningful and phonologically legal words.

The present experiment assumes that the final portion of a word contains information that enables the listener to predict some or all properties of the beginning of the next word, i.e. the word initial consonant(s) and possibly even the following vowel. We wished to test the hypothesis that such feed-forward information facilitates the identification of the following vowel (V_2) in continuous utterances. Our approach was to compare two types of stimuli: one with the crucial word in its original coarticulated environment, and another in which the vowel immediately preceding the crucial word (V_1) had been replaced by a noise burst.

2. Method

Forty-two words were spoken by a male speaker of Dutch in the fixed carrier



1. ət fəlγəndə wor:t #ɪz də γəwərdə

(Het volgende woord is de ... geworden, 'The next word has become the ...'), i.e. preceded and followed by unstressed syllables containing schwa. The crucial words were disyllabic and with initial stress, which was consistently realised with the same accent-lending pitch movement (see stylized pitch trace in (1)). Six stressed vowels (i:, y:, u:, e:, o: a:) were combined with 6 initial consonants p, t, k, b, d, or no C). Finally, 6 three-syllable words beginning with V were added to the stimulus set.

The 42 stimulus sentences were recorded per target vowel in blocks of 7, interspersed in a quasi-random fashion with an equal number of foils (drawn from a pool of 42 two or three-syllable words containing any of the Dutch non-target full vowels or diphthongs). The schwas immediately preceding the crucial words were excised from the utterances and replaced by pink noise with an amplitude equal to that of the removed vowel (10 ms rise/decay time). In the control condition the schwas were left intact, but now the vowel /I/ in *is* was replaced by noise. Stimuli were sampled from both conditions (coarticulation removed *vs.* coarticulation intact) in equal numbers. Two test tapes were prepared such that when coarticulation intact) in equal numbers. Two test tapes were prepared such that when coarticulation was removed on tape I, it was left intact on tape II, and *vice versa*. As a result each tape contained each word only once.

Forty-one Dutch listeners were instructed to press a button as soon as they heard a particular target vowel. After each block of 14 stimuli they were informed of the change in target.

3. Results

Subjects scoring more than 4 errors (misses or false alarms) were eliminated from further analysis, leaving 2 balanced groups of 14. The results showed that one group (tape I) had significantly faster detection times than the other (tape II) (310 *vs.* 493 ms). To redress subject variability, the data were Z-normalised for individual means and standard deviations. Figure 1 plots normalised vowel detection latency as a function of the initial C (panel A), and as a function of the stressed vowel (panel B). A second abscissa provides a rough translation of Z-scores to milliseconds, on the basis of a grand mean of 401 ms and an average standard deviation (= Z) of 171 ms per subject.

Detection latencies do not differ for any of the 6 target vowels when coarticulation is preserved, F(5,571)=2.1 (ins.). Removal of coarticulation has the overall effect of slowing detection down by .16 Z (or 27 ms),

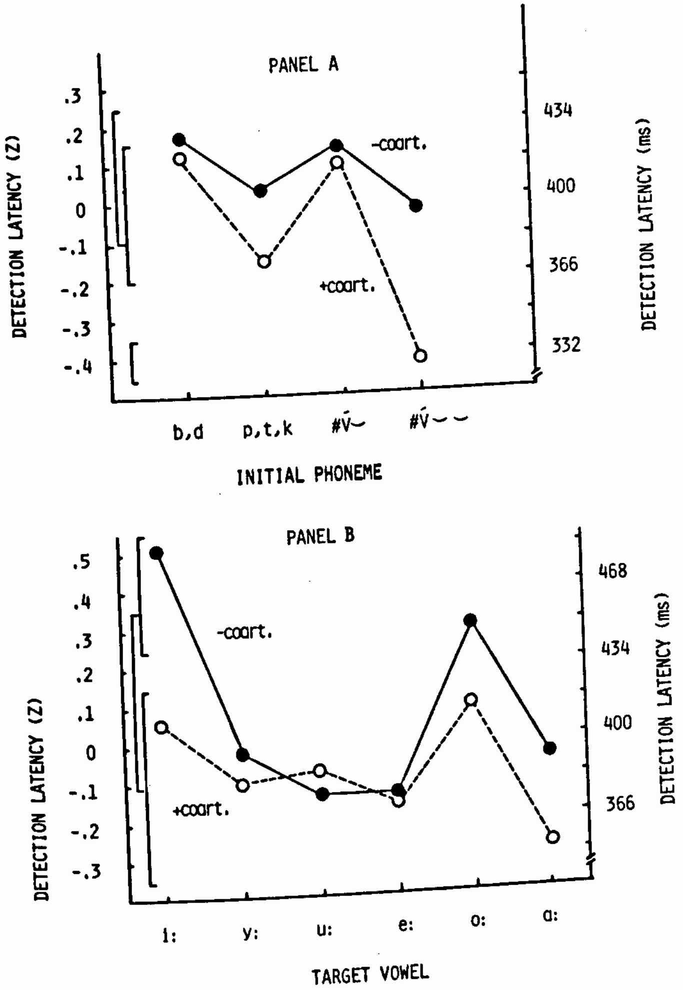


Figure 1. Z-transformed detection latency (normalised per subject for individual mean and standard deviation) as a function of the target vowel (panel B) and type of initial phoneme (panel A), separated out for removed vs. intact coarticulation with the preceding word. Z-scores have been reconverted to milliseconds along the right-hand axis. The ordinate does not represent continuous variables. Braces enclose means that do not differ significantly from each other by a Newman-Keuls test with p<.05.

F(1,1155)=7.9 (p=.005), but not all the vowels are affected to the same extent. Specifically, /i:/ and /o:/ are now significantly slower than the other vowels in the same condition, F(5,574)=6.8 (p<.001), and in fact slower than any of the other vowels in either coarticulation condition, F(11,1145)=4.8 (p<.001). For /a:/, /y:/, and /e:/ the effect of coarticulation is in the predicted direction, but it is too small to reach significance. Finally, for /u:/ the effect is in the wrong direction, but again insignificantly so.

It is apparent from panel A that vowels are detected sooner after voiceless plosives or initially in three-syllable words than after voiced plosives or initially in two-syllable words, F(3,1153)=6.5 (p<.001).

4. Conclusions and Discussion

First and foremost, our data support the hypothesis that target vowels are detected earlier when anticipatory coarticulation is provided in the preceding syllable (word), even across an intervening consonant.

One might object, of course, that the longer latencies in the coarticulation removed condition are simply due to the disturbing influence of the noise burst immediately preceding the target word. Had this been the case, however. the delay should have been the same for each of the 6 target vowels, which clearly it was not. Similarly, its effect should have been more disruptive for targets in initial position (i.e. abutting the noise) than for targets separated from the noise by a consonant. Again this effect does not obtain (cf. panel B).

As a final point we want to speculate on the origin of the difference in detection latency for initial vowels in two as opposed to three-syllable words. Why would targets in the longer words be detected sooner than in the shorter words? Explanations on the basis of word-frequency differences can be ruled out, as both types were selected from the low frequency brackets in the Dutch lexicon. Neither do the two types differ with respect to the point at which the word can be uniquely distinguished from all other words in the lexicon: both two- and three-syllable words have their theoretical recognition points (cf. Marslen-Wilson and Welsh, 1978) after 4½ phonemes, on average. Therefore, we reason that the effect must have a phonetic origin. We know that in Dutch the duration of a long vowel in an initial stressed syllable is 20 to 30 ms shorter in a three-syllable word than in a two syllable word (Nooteboom, 1972). It might thus be reasonable to assume that our subjects delay their decision until they have heard the end of the target vowel, which comes earlier in the three-syllable words (all else being equal). We have re-analysed our data, using target offset rather than onset as the reference point for the latency measurements, and again normalising the new results for individual means and standard deviations. Generally, having eliminated an uncontrolled source of error, residual variance in the data should now decrease, the statistical significance of all effects should go up, except for those that are contingent on vowel duration. Error variance does indeed go down, viz. from 1079 to 1035, and the differences between the target word onset conditions are substantially reduced, but fall short of statistical significance, F(3,1153)=2.4 (ins.). Counter to the prediction, however, the differences between the 6 target vowels increase, especially between the phonetically long vowels /e:/, /o:/, /a:/ on the one hand, and the phonetically shorter vowels /i:/, /y:/ and /u:/. Possibly, therefore, identification of the shorter vowels is delayed until the end of the vowel, but may take place at some earlier point for the longer vowels.

Acknowledgement

Research supported by a grant from the Foundation of Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (ZWO) under project nr. 17-21-22. The experiment was carried out by D. Lemoine and P. van den Mortel.

References

Benguerel, A.P., Adelman S. (1975). Coarticulation of lip rounding and its perception. In: A. Cohen, S. Nooteboom (eds.), *Structure and process in Speech perception*, Berlin: Springer, 283-293.

Lehiste, I., Shockey, L. (1972). On the perception of coarticulation effects in English VCV syllables. *Journal of Speech and Hearing Research*, **15**, 500-506.

Marlsen-Wilson, W.D., Welsh, A. (1978). Processing interactions and lexical access during words recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Nooteboom, S.G. (1972). Production and perception of vowel duration: a study of durational properties of vowels in Dutch. Doct. Diss, University of Utrecht.

Martin, J.G., Bunnell, H.T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, **69**, 559-567.

Öhman, S.E.G. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, **39**, 151-168.

Section 10
Stress and Accent

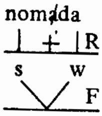
Extrametricality and Italian Stress

E.A. den Os
Utrecht, the Netherlands

In standard generative phonology stress was regarded as a property of single vowels. Liberman and Prince (1977) proposed another way to look at stress. According to them stress must be seen, not as an absolute property, but as a relative prominence between syllables, represented in a hierarchical, prosodic structure. It is the result of a) the way in which elements are grouped in binary branching trees and b) the relation strong-weak which is attached to pairs of sister-nodes.

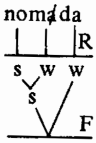
This new 'metric' theory, which was based on English, induced Harris (1980) to reinvestigate the Spanish stress system. In his description of the stress system of Spanish nouns, adjectives and adverbs, Harris used the theoretical notion of 'Extrametricality'. Extrametrical elements are elements which are temporarily skipped over when the prosodic trees are formed. Harris gives as an example the stress assignment in *nomada* 'wandering'; according to Harris the first 'a' is extrametrical and the following structure is formed, where '/' marks extrametricality on the segmental level and '-' extrametricality on the rhyme level:

(1)



The extrametrical element must be incorporated in the prosodic structure. This is accomplished by 'Stray Rhyme Adjunction' (SRA) which adjoins a skipped element as a weak node (w) to the adjacent foot:

(2)

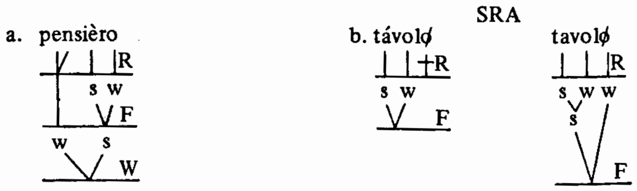


I have used Harris' paper as a starting point for the analysis of Italian stress. Describing stress assignment in Italian nouns, adjectives and adverbs I start from the following principles:

a. All suffixes are attached to words and not to derivational stems, e.g. *gatto* + *ino* and not *gatt* + *ino*. A phonological rule, V → ø/____ + V, takes care of vowel deletion: *gatto* + *ino* → *gattino* 'kitten'. Besides this phonological rule a morphological rule is needed for the deletion of the 'e' in e.g. *formale* + *mente* → *formalmente* 'formally'. This is a morphological rule since it is a deletion before the specific suffix, *mente*.
b. Rules for the segmental structure (in this case vowel deletion) work before the stress rules.
c. Stress is assigned cyclically. This is necessary given e.g. the words *ìmpossìbilitá* 'impossibility' (from *impossíbile*) and *ìrrazionalitá* 'irrazionality' (from *irrazionále*). In these words the relative prominence relations of the inner cycle are maintained.
d. In the lexicon segments can idiosyncratically be marked as extrametrical. These extrametrical segments must be peripheral in the word. Hayes (1981) remarks that it is a universal characteristic of extrametrical elements that they are peripheral in the stress domain, which in Italian is the word. Thus, the Italian word *tavolo* 'table' has an extrametrical 'o' which is peripheral in the stress domain.
e. Primary stress is assigned from right to left and is determined by language specific rules for foot and wordtree structure. In Italian these rules are 1) Foottrees are left branching, labelled s w; Non-branching nodes are labelled at word level, because in metric theory there must be at least two elements, since stress is expressed as 'relative' prominence. 2) The maximal footform is s w w and the rightmost foot with this form is only possible when there is an extrametrical final segment. 3) The Branching Condition, which says that footnotes labelled w may not branch, is applicable. 4) Word trees are right branching, labelled w s.

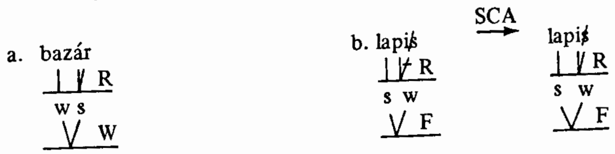
In the following words, the first one has no extrametrical final segment, the second one has:

(3)



In *pensiero* 'thought' the Branching Condition is not violated, since the branching rhyme is labelled at word level and not at foot level. In *tavolo* 'table' Stray Rhyme Adjunction (SRA) takes care that the last 'o' is attached to the preceding foot as a weak node; SRA is structure preserving: the branching direction and the labelling of the foot in Italian are not violated. Words which end with a consonant can also have an extrametrical final segment. In the following words, the first one has no extrametrical segment, the second one has:

(4)

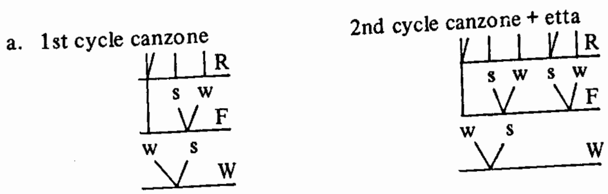


In *bazar* the last syllable must get the label s(trong), since it has a branching rhyme. In *lapis* 'pencil' the last syllable can have the label w without violating the Branching Condition: the final segment is temporarily invisible. This segment is incorporated in the structure by Stray *Coda* Adjunction (SCA), since it is not a whole rhyme that is involved here.

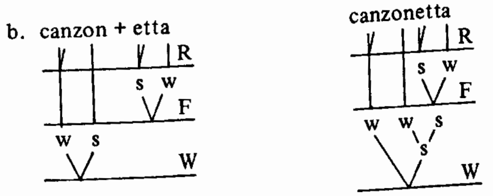
There are a few Italian words whose stress pattern cannot be explained by this theory. These words have branching rhymes in the penultimate or final syllable and primary stress on the antepenultimate syllable, e.g. *mandorla* 'almond', *festival*. Since so few words are involved, I do not want to consider these words as counterexamples.

Italian has three wellformedness conditions in relation to stress: 1) There are no stress clashes 2) Words begin with a stressed syllable and 3) There are no successions of more than two unstressed syllables (see Vogel and Scalise, 1982). When we regard the forming of *canzonetta* 'little song' (from *canzone* 'song'), we see that the first wellformedness condition is automatically fullfilled by the rules for foot and wordtree structure. However, the second wellformedness condition is violated:

(5)



First, vowel deletion takes place, because this rule works before the stress rules. After vowel deletion a word tree is formed:



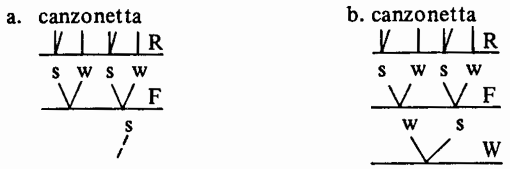
The stress clash is removed; a rule is needed for stress on the first syllable:

(6) Initial Stress Rule

$_W[\,_w X_\alpha \rightarrow \,_W[\,_F[\,s\,w\ldots]_F$ $\alpha \neq$ primary stress

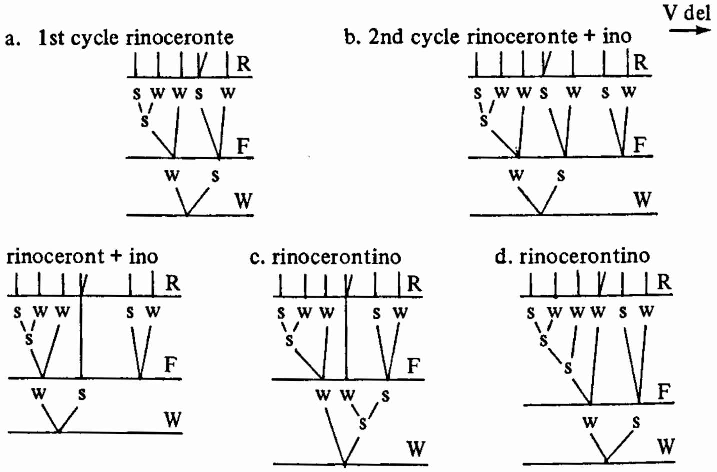
This rule says that the first w at word level is formed together with at least one syllable following the foot s w. The Initial Stress Rule gives in the case of *canzonetta* structure (7a); now the right word tree can be formed (7b):

(7)



The prosodic structure of *rìnocèrontíno* 'little rhinoceros' looks like (8d) after primary stress assignment in both cycles, after vowel deletion, after word-treeforming and after the 'elegance' principle which says that a w at word level, which is not labelled at foot level, is attached to the preceding foot as a weak element of that foot:

(8)



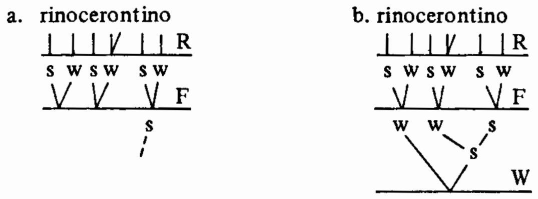
Now a foot with the form s w w w is formed. This is not allowed because the wellformedness condition which says that there are not more than two successive unstressed syllables in Italian, is violated. We need a rule which transforms this big foot into two feet with rhe form s w:

(9) Maximal Foot Rule

$$_F[\text{s w w w}] \rightarrow {}_F[\text{s w}]\,{}_F[\text{s w}]$$

In the case of rìnocèrontíno this rule gives the form (10a); the new right wordtree is formed (10b):

(10)



We see that the Branching Condition is violated: 'on' is a branching rhyme with the label w. I want to conclude from this that it is more important in Italian that stress clashes which arise when two morphemes are put together, are avoided than that the Branching Condition is maintained.

Acknowledgement

Irene Vogel's help is gratefully acknowledged.

References

Harris, J.W. (1980). *Extrametricality and Spanish Stress.* MIT.

Hayes B.P. (1981). *A Metrical Theory of Stress Rules.* Ph. D. Diss. Yale University.

Liberman, M. and Prince A. (1977). On Stress and linguistic Rhythm. *Linguistic Inquiry* **8**, 249-336.

Vogel, I. and Scalise, S. (1982). Secondary stress in Italian. *Lingua* **58**, 213-242.

The Stress Pattern and Its Acoustic Correlates in Beijing Mandarin

M.C. Lin, J.Zh. Yan, G.H. Sun
Beijing, the People's Republic of China

1. Introduction

Chinese is a tone language, and it also has the feature of stress in syllable-groups (including words and phrases) and sentences. Phonemically, only three different degrees of stress are found: weak stress (i.e. neutral tone), normal stress and contrastive stress.

2. The perceptual result of normal stress

Actually, the syllables in syllable-groups that have neither neutral tone nor contrastive stress do not have the same degree of phonetic stress. The stress in such syllable-groups is defined as normal stress.

In our experiments, 103 two-syllable groups were pronounced with normal stress by m1 and f1 and 154 three-syllable groups were spoken by m2 and f2. The normal stress in the speech sounds of the two-syllable groups and of the three-syllable groups was judged by 8 listeners and by 7 listeners (all phoneticians) respectively.

Figure 1 shows the probability distribution histogram on the perceptual result of normal stress in the two-syllable groups judged by 8 listeners. From Figure 1 we can see that in 103 two-syllable groups, there were 95 groups pronounced by m1 and 92 groups done by f1 in which the second syllable was judged as having the normal stress by the great majority of 8 listeners.

8 students of linguistics were asked to pronounce the same two-syllable groups and judge normal stress of his or her own speech sounds. The perceptual results of normal stress are represented in Figure 2. Figures 1 and 2 identically demonstrate that the second syllable was judged by the great majority of the listeners as having normal stress.

Some scholars claimed that the contrast between the second syllable and the first syllable with normal stress does exist in such two-syllable groups like 工事 'fortification' vs. 攻势 'offensive', 报到 'register' vs. 报道 'report', 公鸡 'cock' vs. 攻击 'attack', 散步 'take a walk' vs. 散布 'spread', 生气 (v.) 'get angry' vs. 生气 (n.) 'vitality'. In order to verify this claim, we put these two-syllable groups into sentences. They were then pronounced by m1 and f1. The results show that the second syllable was often judged as having normal stress by our informants. We conclude that in two-syllable groups, normal

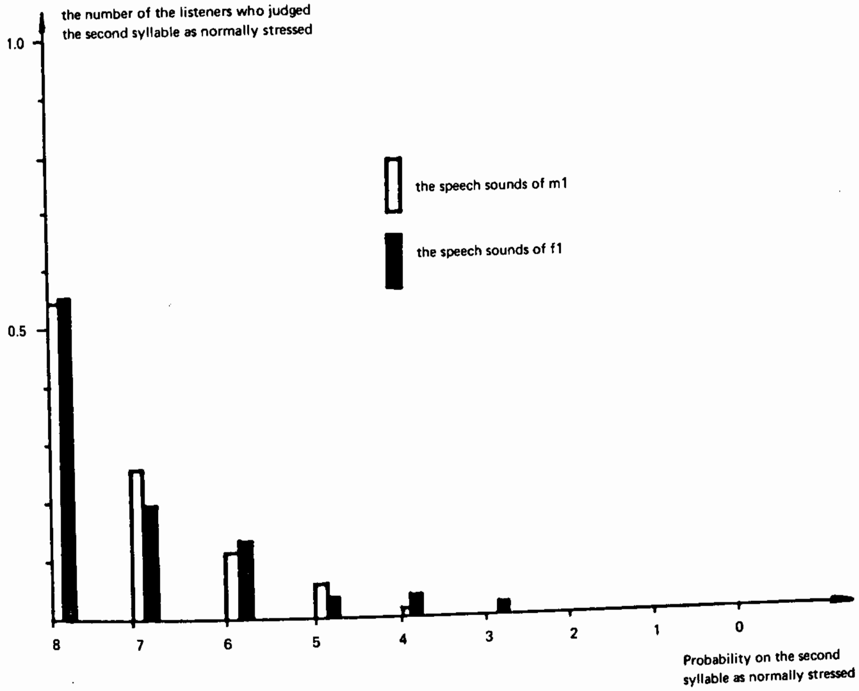


Figure 1. Probability histogram on the perceptual data by 8 phoneticians.

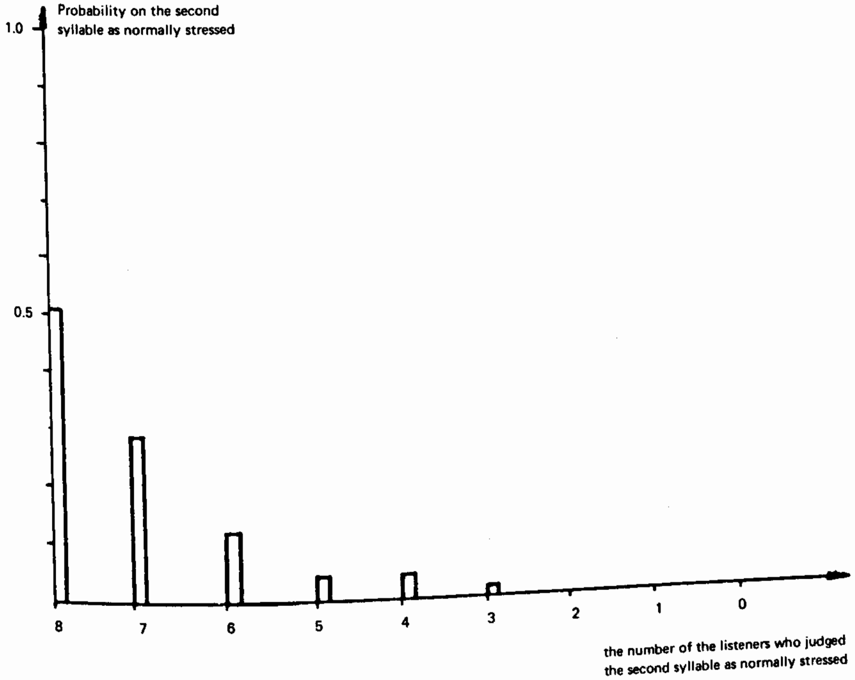


Figure 2. Probability histogram on the perceptual data by 8 students.

stress usually occurs on the second syllable. It is also the case that there is no such a two-syllable group in which the first syllable carries normal stress.

In three-syllable groups, normal stress is usually on the last syllable. Which is more stressed, the first syllable or the second one? The judgement is not consistent.

3. The acoustic data on normal stress

Figure 3 indicates the relative distribution of syllable duration in the two-syllable groups. There were 71 groups pronounced by m1 and 84 groups by f1 in which the duration of the second syllable was longer than that of the first one. Correlation coefficients of .82 for m1 and .80 for f1 were found between the normal stress and syllable duration.

Figure 4 shows the relative distribution of syllable duration in the three-

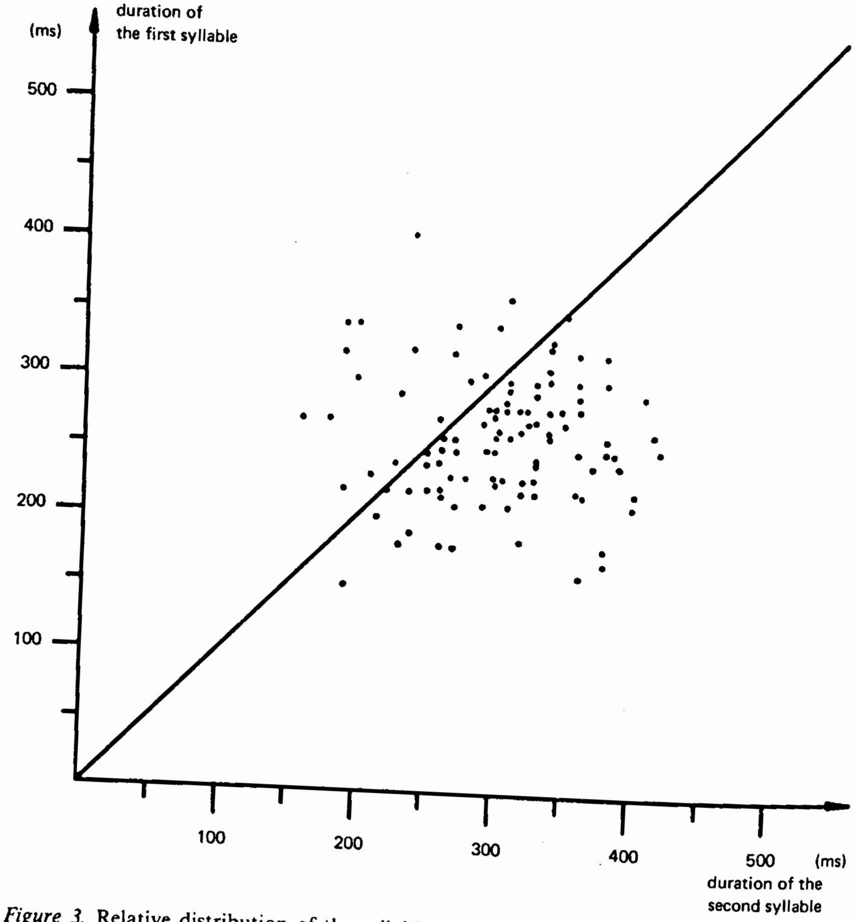


Figure 3. Relative distribution of the syllable duration in two-syllable groups by m1, f1's is similar to m1's.

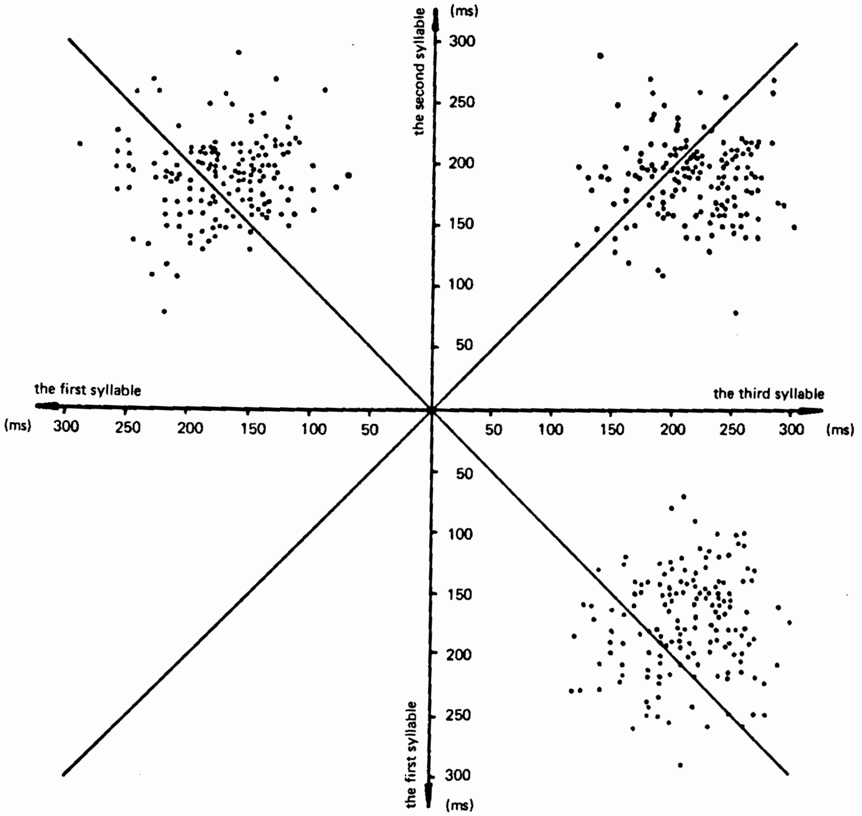


Figure 4. The relative distribution of the duration in three-syllable groups by m2, f2's is similar to m2's.

syllable groups. The duration in the great majority of the third syllables is longer than that of either the first one or the second one, not only for m2 but also for f2. It must be pointed out that the duration in the majority of the second syllables is longer than that of the first one.

The peak intensity in the second syllable in the great majority of the two-syllable groups or in the third syllable in the great majority of the three-syllable groups is not higher than that of the first one or the preceding ones. This can be seen in Figures 5 and 6.

The pitch contour of the last syllable in the two-syllable groups or in the three-syllable groups frequently approximates the tone pattern of the syllable in isolation. But there is a variation between the pitch contour of the first syllable in the two-syllable groups and that of the first two syllables in the three-syllable groups and their standard tone pattern. These facts are illustrated in Tables 1 and 2, in which the average pitch for each syllable is given.

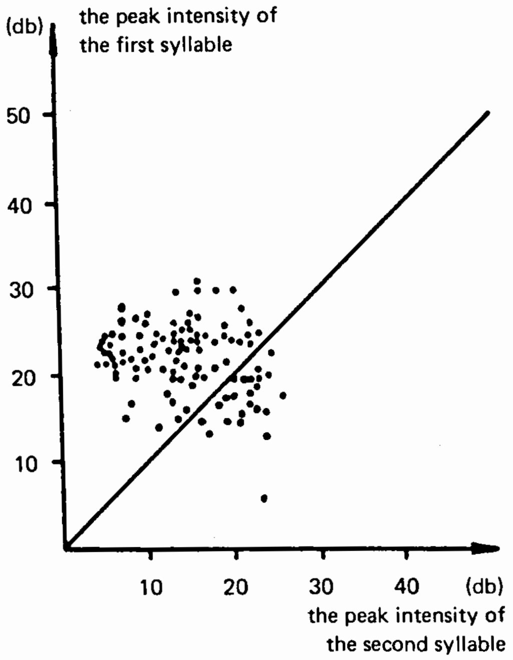


Figure 5. The relative distribution of the peak intensity in two-syllable groups by m1, f1's is similar to m1's.

From the above analysis, it is concluded that in a syllable group the last syllable is the syllable with the normal stress. The acoustic correlates of normal stress are given below: the pitch contour of the syllable with normal stress approximates its tone pattern in isolation; its duration is longer; as for peak intensity, it takes little part in normal stress.

4. The acoustic properties of neutral tone

The acoustic characteristics of neutral tone (cf. Lin and Yan, 1980) will be presented here briefly. A syllable with neutral tone loses the original tone pattern of the syllable, and the duration is shortened by an average of 50%, compared with the stressed syllable. When a syllable is pronounced with neutral tone, the tongue position of the main vowel more or less shifts toward that of the central vowel. But its peak intensity is not always decreased. These results come from the acoustic data of 29 minimal stress pairs, for example, 东 西 'east and west' vs. 东 ·西 'thing', 兄 弟 'brothers' vs. 兄 ·弟 'younger brother', 莲 子 'lotus seeds' vs. 帘 ·子 'curtain', 火 烧 'to burn' vs. 火 ·烧 'baked wheaten', 老 子 'the philosopher Lao-zi' vs 老 ·子 'a father' and 大 意 'main paints' vs. 大 ·意 'careless', etc. In each of these pairs, the three constituents (the initial, the final and the tone) of the first syllables are the same, but the second syllables, with identical initial and final constituents, can be

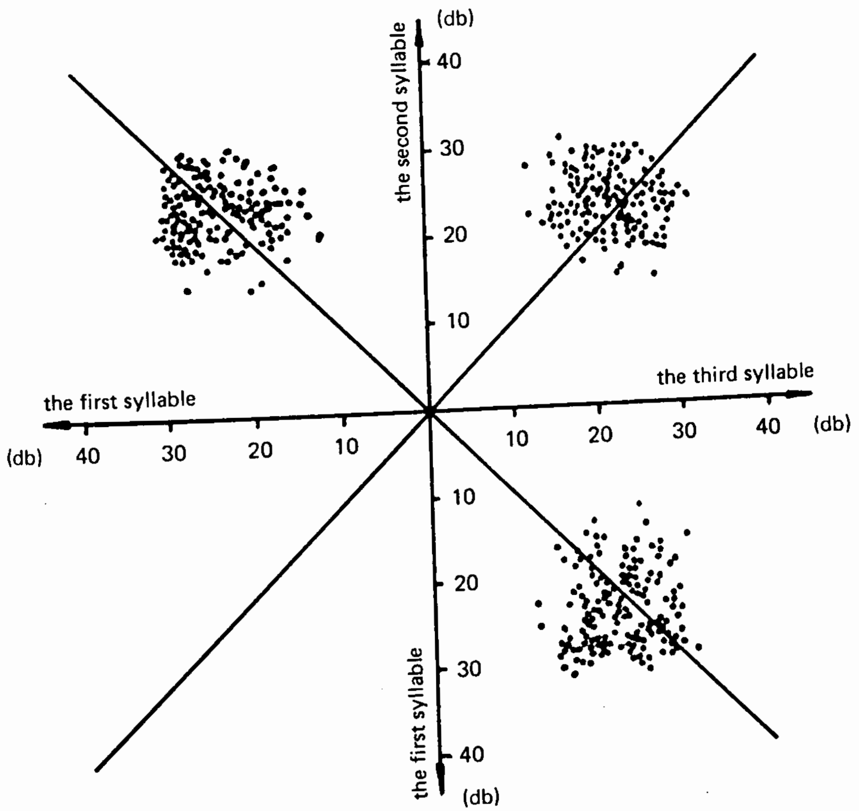


Figure 6. The relative distribution of the peak intensity in three-syllable groups by m2, f2's is similar to m2's.

pronounced with normal stress or neutral tone. To mark a neutral tone, a dot has been placed before the Chinese character.

As for the contrastive stress, it implies that an emphasis is put on some syllable or syllable-group.

Table 1. The average pitch* and its tone in two-syllable groups

m speaker		tone arrangement	f speaker	
the average pitch and its tone			the average pitch and its tone	
the first syllable	the second syllable		the first syllable	the second syllable
178-179* 55**	174-170 55	tone 1 + tone 1	221-221 55	224-222 55
194-195 55	122-165 35	tone 1 + tone 2	225-226 55	170-219 35
188-189 55	128-83-115 312	tone 1 + tone 3	224-225 55	168-92-147 312
188-187 55	192-87 51	tone 1 + tone 4	228-229 55	252-131 51
138-187 35	179-179 55	tone 2 + tone 1	176-249 35	243-241 55
134-192 35	118-158 24	tone 2 + tone 2	174-240 35	171-224 35
135-199 35	125-90-124 313	tone 2 + tone 3	169-243 35	127-92-139 112
114-179 25	190-90 51	tone 2 + tone 4	171-240 35	262-126 51
141-112 42	158-163 44	tone 3 + tone 1	178-147 32	224-225 55
144-109 42	112-167 25	tone 3 + tone 2	178-147 32	155-228 25
118-190 25	118-86-121 213	tone 3 + tone 3	167-247 35	149-93-135 211
131-99 31	181-87 51	tone 3 + tone 4	187-149 42	242-134 51
189-109 52	157-157 44	tone 4 + tone 1	231-149 52	220-226 55
196-110 52	100-154 14	tone 4 + tone 2	231-155 52	158-212 35
196-110 52	104-82-114 212	tone 4 + tone 3	242-152 52	143-82-156 212
210-113 52	179-90 51	tone 4 + tone 4	250-165 53	234-130 51

* Pitch = fundamental frequency (Hz).

** The value on the five-point scale.

Table 2.1. The average pitch and its tone for tone 3 + tone 3 in three-syllable groups

m speaker			tone arrangement	f speaker		
the average pitch and its tone				the average pitch and its tone		
the first syllable	the second syllable	the third syllable		the first syllable	the second syllable	the third syllable
166-166 55	159-163 55	110-70-103 313	tone 1 + tone 3 + tone 3	205-200 55	180-191 55	139-112-142 313
123-162 35	154-156 55	111-77-107 312	tone 2 + tone 3 + tone 3	142-196 35	185-200 55	138-107-138 313
123-158 35	124-104 32	158-154 55	tone 3 + tone 3 + tone 1	154-193 45	138-131 32	181-181 55
137-161 45	134-109 43	113-139 34	tone 3 + tone 3 + tone 2	150-200 35	154-111 41	119-159 24
132-159 45	159-161 55	103-73-102 212	tone 3 + tone 3 + tone 3	150-185 35 135-132 32	185-189 55 132-200 25	138-102-138 313
120-158 35	133-96-118 413	144-77 51	tone 3 + tone 3 + tone 4	138-185 35	150-107-123 312	199-107 51
118-108 52	123-154 35	116-72-92 311 103-103 22	tone 4 + tone 3 + tone 3	208-138 53	154-193 45	138-92-131 312

Table 2.2. The average pitch and its tone for tone 2 on the second syllable with tone 1 or tone 2 on the first syllable in three-syllable groups

m speaker			tone arrangement	f speaker		
the average pitch and its tone				the average pitch and its tone		
the first syllable	the second syllable	the third syllable		the first syllable	the second syllable	the third syllable
160-155 55	134-157 45	153-151 55	tone 1 + tone 2 + tone 1	188-186 55	176-183 45	179-179 55
158-154 35	144-152 55	116-147 35	tone 1 + tone 2 + tone 2	207-185 55	191-198 55	140-188 35
156-156 55	146-158 55	108-86-99 212	tone 1 + tone 2 + tone 3	200-200 55	174-190 45	138-87-126 312
158-158 55	133-157 45	166-79 51	tone 1 + tone 2 + tone 4	200-200 55	163-198 45	208-100 51
123-169 35	147-157 55	159-159 55	tone 2 + tone 2 + tone 1	146-200 35	173-204 45	208-212 55
124-159 35	159-164 55	124-159 35	tone 2 + tone 2 + tone 2	149-207 35	189-191 55	133-180 35
123-162 35	144-154 55	110-85-118 313	tone 2 + tone 2 + tone 3	152-208 35	197-213 55	108-92-138 313
122-154 35	138-154 45	159-77 51	tone 2 + tone 2 + tone 4	46-196 35	62-200 45	215-85 51

Table 2.3. The average pitch and its tone (excluding tone 3 + tone 3) in three-syllable groups

	tone	tone 1	tone 2	tone 3	tone 4
	speaker	the average pitch and its tone			
the first syllable	m	159-158 55	124-161 35	120-113 33	161-105 52
	f	201-198 55	146-196 35	151-134 33	206-132 52
the second syllable	m	159-158 55	133-156 45	115-86-106 312 125-101 32	157-101 52
	f	198-200 55	161-194 45	156-111-124 312 147-117 32	201-121 52
the third syllable	m	157-154 55	117-150 35	109-80-111 213 74-89 12	163-81 51
	f	187-187 55	133-180 35	135-98-133 315	203-104 51

References

Lin, M.C. and Yan, J.Zh. (1980). Acoustic Characteristics of Neutral Tone in Beijing Mandarin, *Dialect,* **3**, August 1980 (in Chinese).

On the Distinctive Features of Byelorussian Utterance Accents

Anna A. Metlyuk
Minsk, USSR

1. Introduction

The present paper reports on an experimental investigation of linguistically relevant prosodic units and their distinctive features in Byelorussian. Utterance accentemes actualized as accents are the smallest elements of the accentual subsystem of utterance prosody. Paradigmatic types of accents are determined on the basis of their functioning in larger phonological complex-accentual structures, which stand out as models of syntagmatic organization or patterns of combination of the smaller elements.

It is widely accepted that in the accentual structure of an utterances accents are peaks of prominence, hierarchically ordered due to the differences in their functions.

In Byelorussian, like in Russian and English, and, presumably, in a number of other languages, functional differentiation of accents in an utterance, i.e. their paradigmatic classification, is based on the discrimination of the following types: nuclear accent, non-nuclear (full and partial), and zero accent.

The nuclear accent differs from the non-nuclear in its specific semantic and organizing role: it singles out semantically the most important word in an utterance and is its obligatory element. Non-nuclear accents give prominence to semantically less important words and function as optional elements. So, the relations between the nuclear and non-nuclear accents within the structure are those of subordinative syntagmatic hierarchy.

The basic characteristics of the paradigmatic relations of these accents are their qualitative distinctions. They manifest themselves in the predominant co-occurrence of the nuclear accent with the kinetic tone and the non-nuclear accent with the static tone. In addition, the nuclear and non-nuclear accents can be opposed to each other by the degree of pitch contrast - strong/weak.

Analogous to the above differentiation is the discrimination of non-nuclear accents into full and partial, which also reflects the semantic inequality of the words.

All the types of accents are represented in concrete speech realizations by definite phonetic variants, conditioned by positional and combinative factors. For example, the first and the second full pre-nuclear accents belonging to one functional type differ as positional variants by their pitch characteristics.

2. Method

In analysing the acoustic correlates of utterance accents it seems reasonable to consider each type of accent separately. In our investigation the nuclear, and non-nuclear full and partial accents were analysed in 114 statements and in 128 general questions, which were read by 5 native speakers of Byelorussian. The full non-nuclear accent was analysed in its two variants - the first and the second positions. Acoustic prominence was determined by relative values of fundamental frequency, intensity and duration of the accented syllables as compared to those of adjacent unaccented syllables.

3. Results

The data obtained demonstrate the greatest prominence of accented syllables in duration features: accented syllables in Byelorussian utterances are much longer than unaccented ones. Intensity contrast between accented and unaccented syllables appears to be weaker than duration contrast, whereas fundamental frequency contrast is the weakest.

The highest peak of duration prominence is with the nuclear syllable (in statements duration contrast is sharper than in questions), the second peak is with the first accented syllable. Intensity prominence increases from the first accented syllable to the nuclear one. F_0-prominence is the greatest with the nuclear syllable; the second peak of F_0-prominence is with the first accented syllable. Thus the nuclear accent is effected by the strongest contrasts in all the parameters.

The types of accents also differ in the combination of acoustic contrasts, i.e. in their acoustic structures. The nuclear syllable is most frequently effected by the structure for all the 3 parameters, whereas the structure of non-nuclear accent is most frequently constituted by 2 parameters (Metlyuk, 1976).

In creating the effect of utterance accent irrespective of their type the leading parameter, as to the frequency of occurrence of its contrast, is fundamental frequency (84% of the cases on the average). In the acoustic structure of the nuclear accent F_0-prominence occurs in 100% of the experimental utterances (both statements and questions). In the structure of the full non-nuclear accent it occurs in 76% of the cases (90% in the first accent and 62% in the second one). There is practically no F_0-prominence in the acoustic structure of the partial accent.

I-prominence appears to be more frequent (82% of the cases on the average) than t-prominence (68% of the cases). It occurs in 88% of the cases in the structure of the nuclear accent, and in 83% in the structure of the full-nuclear one, whereas t-prominence is observed in the structure of the nuclear accent in 72% of the cases, and in the structure of the full non-nuclear accent - in 63% of the cases.

In investigating the acoustic nature of utterance accents we are not equally interested in all the existing differences between accented and unaccented syllables but mainly in those which are significant on the perception level. Therefore the data obtained as a result of the acoustic analysis were checked against the results of the auditory test of synthetic Byelorussian utterances - (statements and questions) which testified the validity of the above conclusions.

4. Discussion

From the linguistic point of view the results of our investigation can be interpreted in the following way.

The distinctive features of the above accent types are of a compound nature (cf. compound distinctive features of phonemes, Dzhaparidze, 1979). The acoustic structures of the compound distinctive features are essential for the perception of these accents. The components of the distinctive feature of the Byelorussian nuclear accent are fundamental frequency prominence and duration prominence, F_0- prominence playing the leading role. At the same time duration prominence can function as the only distinction in cases when there is no F_0-prominence (the so-called level tone).

The non-nuclear full and partial accents are differentiated by the prevalence of intensity or duration prominence in their acoustic structures, intensity prominence being more essential for the full accent and duration for the partial accent.

Qualitative distinctions of the accents are closely related to their quantitative distinctions: the degree of prominence increases from the zero to the nuclear accent. The quantitative characteristics of Byelorussian accents are less conspicuous than in Russian and even less so when compared to utterance accents in English.

References

Metlyuk, A.A. (1976). Ab akustichnikh strukturakh frazavikh natsiskau u belaruskai move. In: *Belaruskaya Lingvistika.*. Minsk.

Dzhaparidze, Z. (1979). O komponentakh raslichitelnikh priznakov fonem. In: *Fonologiya, fonetika, intonologiya.* Moskva.

Stress in Standard Swahili

C.W. Temu
Dar Es Salaam, Tanzania

1. Introduction

The word 'stress' generally means the greater force exerted in the articulation of one part of an utterance compared with another, thus accentuating that certain part of the utterance, giving it more prominence. In Swahili, when a word or phrase is pronounced in isolation, stress is realised on the penultimate syllable.
Examples:

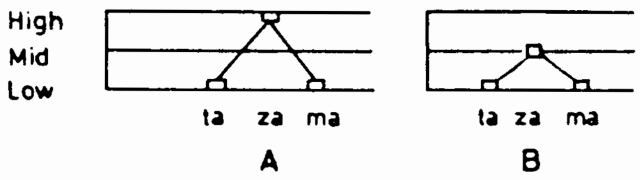
piga (hit) [ˈpi-ga]	imba (sing) [ˈi-mba]	mtu (person) [ˈm̩-tu]
tazama (look at) [ta-ˈza-ma]	waite (call them) [wa-ˈi-te]	
anacheza (he is playing) [a-na-ˈtʃe-za]	walimpa (they gave him) [wa-li-ˈm̩-pa]	

It is now generally agreed that stress is manifested by pitch (fundamental frequency), force (intensity) and duration (time), but how these parameters combine to produce stress in Swahili still needs to be examined.

2. Stress parameters

2.1. Pitch

This term is used to describe a high or a low frequency of vibration of the vocal cords. The syllable -ba- can be pronounced in many pitch levels but in actual normal speech the syllables of Swahili can be said to belong to three general pitch levels i.e. high, mid (normal) and low, while other syllables have a rising pitch or a falling pitch. Stress can be manifested by a higher pitch of a given syllable in relation to the others appearing together in the same utterance.

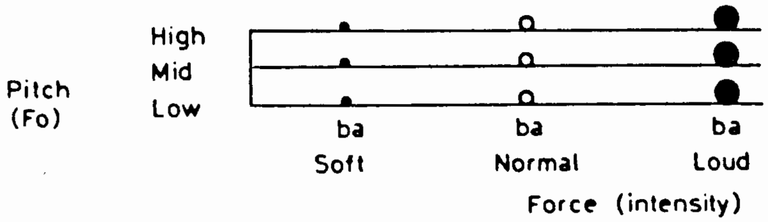


In the above diagram the second syllable in each group can be taken to be stressed as it has a higher pitch in relation to the others. A subsidiary question can now be asked, Is pitch the only manifestation of stress in Swahili? Before an answer is sought, the other parameters must be examined.

2.2. Force

A syllable can be produced loudly (i.e. with greater intensity), normally or softly (i.e. with less intensity) in relation to surrounding syllables. A syllable produced with more force in relation to the others around it is louder and hence is heard more clearly. The question is whether a stressed syllable in Swahili is associated with greater force. From the definition of stress one would say that naturally if a syllable is stressed then it must be produced with greater force. However, if a combination of force and pitch is made one could also argue that it is possible to produce a syllable with normal force but with a high pitch and still produce the effect of 'stress'.

Let a dot be selected to represent a softly pronounced syllable and a larger circle to represent a loudly pronounced syllable. Let us also assume there are three pitch levels in Swahili. One could then come up with a picture such as the one below showing the possibilities of pronouncing the syllable /ba/ with force.



2.3. Duration

Any sound takes time to be produced and in the case of syllables one can measure their duration in (milli) seconds. If a syllable takes longer to be produced, one could say that it is longer than the other. In Bantu languages it is customary to talk of long and short syllables, so in this case we shall assume that the syllables in Swahili are either short or long. The question then is whether a stressed syllable is longer than its unstressed counterpart.

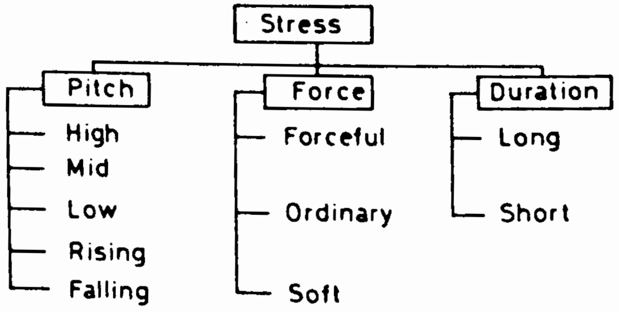
Ashton (1947:5) stated that 'stress accompanied by length and a falling tone lies normally on the penultimate syllable of a word or phrase'. The author thinks that by 'stress', Ashton meant both 'pitch' and 'force'. The author thinks that length is part and parcel of 'stress' in Swahili.

3. Discussion

If we now synthesize the three parameters the picture becomes a bit more complicated and the third question to be posed in respect of Swahili stress becomes relevant, and that is: To what extent does syllable length (duration) play a part in the manifestation of stress in Swahili?

At this stage one can speculate and say that it is most likely that a stressed syllable has a higher pitch, is more forcefully produced and has a longer duration. However, as stated earlier, this statement needs proof based on results of actual measurements using phonetic instruments such as the sound spectrograph, the intensity meter, the fundamental frequency meter and even a computer. The second main question of practical importance is: What signs or symbols should be used in order to explain and to teach Swahili stress, rhythm and intonation? When such symbols are used one must take into consideration the simplicity of writing and printing and the probability of acceptability of those signs for the speakers and learners of the language.

4.1 Summary of Stress Parameters:



4. Marking

Pitch can be shown by the placement of the mark between two lines so that the mark is on the upper level for a high pitch and on a lower level for a low pitch. In the case of force, a dot can signify a soft syllable while a heavy bigger circle shows a forceful syllable, leaving an ordinary unshaded circle to signify a syllable produced with normal force. The problem lies with the choice of a symbol for showing syllable length, rising or falling tone. Would the following symbols be acceptable?

Eˋ A high falling forceful syllable
E. A low rising soft syllable
E‾ A high long forceful syllable
E• A high short forceful syllable
E A high short soft syllable etc.

References

Ashton, E.O. (1947). *Swahili Grammar Including Intonation.* London: Longmans.

Malmberg, B. (ed.) (1968). *Manual of Phonetics.* Amsterdam; North Holland Publ. Co..

Nyqwist, A. (1961). Stress, Intonation, Accent, Prominence in Disyllable Doublestress compounds in Educated Southern English. In: *Proceedings of the Fourth International Congress of Phonetic Sciences, Helsinki.*

Onishi, M. (1961). Prominence and its Phonetic Nature. In: *Proceedings of the Fourth International Congress of Phonetic Sciences, Helsinki.*

Welmers, W.E. (1973). *African Language Structures.* Los Angeles: UCP.

Phonetic Reality of Linguistic Structures: the Case of (Secondary) Stress in French

S. Paul Verluyten
Antwerp, Belgium

1. Introduction

A great deal of the investigation of stress phenomena in French has been carried out by scholars who belong to what might be called the French school of prosody and intonation (cf. Ladd, 1982: 161), whose approach is 'characterized by an emphasis on instrumental phonetics, a tendency to translate instrumental results into fairly 'concrete' or 'surface-y' phonological analyses, and an interest in the prosodic correlates of grammatical organization' (Ladd, ibid.). These scholars work under the assumption that there exist one or more phonetic correlates of the phonological category of *stress*, and that these correlates can and must be identified before we can further our understanding of stress-related linguistic phenomena. In this paper, I wish to argue that (i) it is by no means either necessary or obvious that stress may be correlated with phonetic parameters, and phonetic data up to now tend to show that the correlation, if it exists at all, is at the very least not a straightforward one, and (ii) the necessity of positing a *phonological* category of stress can be demonstrated without references to the phonetic parameters this category is usually correlated with. Stress phenomena are one of the surface manifestations of the *underlying* prosodic ('metrical') structure of the language. We claim that the underlying prosodic structure of French (monomorphemic) words is characterized by an *alternating rhythm* of the type S W S (W). Most of the data adduced in this short paper concern secondary stress in French, but the same argument could be equally well upheld for primary stress in French, and possibly in most other languages too.

2. Perception of stress

The 'phonetician's assumption' that there must be phonetic correlates to a phonological category such as stress is very probably based on the (implicit) idea that something cannot be perceived which is not, in some form or another, physically present in the signal. Yet psychological literature abounds of examples where this is not the case. Just as there are many well-known instances of so-called 'optical illusions', there also exist cases where the acoustical signal and the perceived categories do not correspond. To take an example which concerns us more directly, it has been established

that the human being may perceive (and normally does perceive) rhythmic groups in stimuli which are, in fact, not grouped together in any fashion at the physical level:

> Si nous écoutons tomber dans le silence des gouttes d'eau provenant d'un robinet mal fermé, nous les percevons groupées par deux ou par trois, plus rarement par quatre, même si la cadence de leur chute est parfaitement régulière. On parle alors de rythmisation *subjective* parce qu'aucun facteur lié à la suite objective des stimulations ne détermine le groupement (Fraisse 1974: 74).

Thus, there are in fact no *a priori* reasons to assume a correlation between phonetic parameters and any given phonological category. The existence itself of such a correlation is an empirical question, and the absence of any correlation at all cannot be rejected without prior examination.

3. Absence of secondary stress

In reality, phonetic investigations as they have been carried out since the beginning of the century, if one examines them without bias, tend to indicate that a correlation between the phonological category of stress and phonetic parameters such as intensity, pitch or duration is either absent, or else much less simple and straightforward than has been assumed. Initially, the major correlate of stress in French was thought to be intensity (Grammont, 1914: 105); since Delattre (1938), it is often taken to be vowel duration; for some, the major parameter is pitch (cf. Wunderli, 1978: 71-74, 99-100); recently, Crompton (1980) has adduced evidence that stress is not correlated to duration, but that 'it seems that contrary to the views expressed by certain other investigators (...), intensity is a significant correlate of accent in French' (Crompton, 1980: 230-1). Thus, we are back at the beginning... Clearly, there exists no *stable* phonetic correlate of stress; typically, the investigator's conclusions vary depending on the data he or she obtained, i.e., ultimately, on the test and on the subjects tested. There may be regional or individual variation as to the major parameter; or else, as we argued above, it may be the case that stress is not correlated to any phonetic parameter at all (this view is also expressed by Crompton, 1980: 211).

Rigault (1970) has set up a phonetic experiment by means of which he wishes to show that there exists no secondary stress in French. In this experiment, three subjects were asked to read a number of French trisyllabic and quadrisyllabic words; in these words, Rigault searched for a possible correlation between pitch, intensity and duration and the systematic occurrence of secondary stress (as defined by those phonetic parameters) on any but the final syllable. It comes as no surprise that the three parameters themselves exhibit a different behavior:

> Pour les trisyllabes, on constate que les schémas des trois facteurs de la proéminence se répartissent de façon différente: les schémas de hauteur et d'intensité sont en majorité décroissants ou croissants-décroissants (...). Pour les quadrisyllabes, la situation est moins nette. Remarquons que dans l'ensemble la hauteur et l'intensité suivent des schémas inverses de celui de la durée (Rigault 1970: 286-7).

Although the author concedes that we do not know the relative weight of each of the phonetic parameters, he does not consider the possibility, however clearly it seems to us to be indicated by the data, that there is a total absence of correlation between secondary stress and the phonetic parameters that are examined. Rather, Rigault's conclusion, hardly surprisingly on the basis of his assumptions, will be that there exists no secondary stress in French. In fact, the experiment is irrelevant to the question of the existence of a *phonological* category of secondary stress in that language.

4. Phonology of stress

The relevance of the phonological category of secondary (as well as primary) stress in French can be demonstrated without reference to phonetic parameters (except in the trivial sense that, for instance, vowel deletion is of course also reflected in the phonetic output). The prosodic model we adopt is inspired by Schane's (1979*a*, 1979*b*) proposals for English (which are themselves a rather strongly modified version of Liberman and Prince's (1977) metrical theory). Within the word, only syllables are assigned a binary relative prominence marker (S or W). According to Schane, the basic rhythm pattern of English words is *dactylic*: SWW ($\overset{\text{s w w}}{\textit{Canada}}$, $\overset{\text{w s w w}}{\textit{America}}$); our hypothesis is that the underlying rhythm of French words is iambic, or *alternating*, of the type WSWS(W) (in general the final syllable is W if it contains schwa, S otherwise). Thus the penultimate syllable of French words (lexemes) is relatively less prominent (W) than the antepenultimate (S): this captures the more traditional idea of antepenultimate secondary stress (which most French phonologists reject anyway). If we accept this prosodic structure of French words, various phonological phenomena can be explained in a simple and general way. Some of these phenomena have not been observed as such before, because there was no theoretical framework which made their observation feasible.

Firstly, we can explain why schwa is deleted more easily in penultimate position, whereas it tends to be maintained or restructured towards [e] or [ɛ] in antepenultimate position. Thus, the initial schwa is deletable in *seconde* (where it is W), and tends to become [e] in *secundo* (where it is S); it is deletable in $\overset{\text{w s w s}}{\textit{demi-saison}}$, but not in $\overset{\text{s w s w}}{\textit{demi-heure}}$. In words such as $\overset{\text{s w s}}{\textit{derechef}}$, $\overset{\text{w s w s}}{\textit{briqueterie}}$, $\overset{\text{w s w s}}{\textit{buffleterie}}$, the schwa which is in W position is deletable; the one in S position is not, and tends to become [ɛ]. Many more data from standard French can be adduced (most data are from Martinet and Walter 1973).

Secondly, it appears that the phenomenon of high-vowel syncope in Quebec Freuch occurs preferentially in W position: in the corpus we examined (from Hammond, 1980), there were 45 cases of syncope of [i], [y] or [u] in penultimate or ante-antepenultimate (i.e. W) position, against 12 in antepenultimate (S) position: *versification*, *inscription*, *capucine*, *rajoute*, etc.

Thirdly, positing an alternating rhythm (and hence secondary stress) on French words allows us to explain some observations made in the realm of French verse. It is known that in French verse, some of the metrical positions (such as the 6th and the 12th in the *alexandrin*) have to be filled with a (prosodically) stressed syllable. Whereas this syllable had to bear primary stress (hence, in French, had to be word-final) in the classical and postclassical period (17-18th century), this constraint has been weakened in the course of the 19th century. In particular, the French poet Paul Verlaine (1844-1896) gradually starts writing lines where the 6th metrical syllable is not word-final. Given that all of Verlaine's violations of the classical metrical constraints on the French alexandrin are clearly becoming more numerous and more extreme over the years, it is interesting to observe that lines where the 6th metrical position is filled with a syllable that is prosodically S in our model (mostly antepenultimates) appear much earlier in his works than lines where this is not the case (i.e., where the 6th metrical position is filled by a W syllable such as a penultimate). The first type of line also remains much more frequent than the second type. Clearly, it is felt the first type of line is less unmetrical; this observation receives a straightforward explanation in our model of French prosodic structure, where antepenultimate syllables carry secondary stress because they are underlyingly S in virtue of the alternating rhythmic pattern of French words.

5. Conclusion

In summary, there appears to be a wide array of facts which tend to establish the necessity of positing secondary stress in French. Some phonologists may dispute these facts, or try to explain them differently. Either way, however, the question of the existence of phonetic parameters of secondary stress is at best an interesting side-problem, at worst completely irrelevant to the issue phonologists are discussing. It is neither necessary nor advisable to assume phonetic reality for all linguistic structures.

References

Crompton, A. (1980). Timing patterns in French. *Phonetica* **37**, 205-234.

Delattre, P. (1938). L'accent final en français: accent d'intensité, accent de hauteur accent de durée. *The French Review* **12**, 141-45.

Fraisse, P. (1974). *Psychologie du rythme*. Paris: P.U.F.

Grammont, M. (1914). *Traité pratique de prononciation française.* Paris: Delagrave.

Hammond, M.-A. (1980). *La chute des voyelles hautes en français québécois.* Unpublished M.A. thesis, Université de Montréal.

Liberman, M. and Prince A. (1977). On stress and linguistic rhythm. Linguistic Inquiry **8**, 249-336.

Martinet, A. and Walter, H. (1973). *Dictionnaire de la prononciation française dans son usage réel.* Paris: France-Expansion.

Rigault, A. (1970). L'accent secondaire de mot en français: mythe ou réalité? *Actele celui de-al XII-lea congres international de lingvisticà s filologie romanicà* **I**, 285-90.

Schane, S. (1979*a*). Rhythm, accent and stress in English words. *Linguistic Inquiry* **10**, 483-502.

Schane, S. (1979*b*). The rhythmic nature of English word accentuation. *Language* **55**, 559-602.

Wunderli, P. et al (1978). *Französische Intonationsforschung. Kritische Bilanz und Versuch einer Synthese.* Tübingen: Gunter Narr.

Section 11
Temporal Organisation of Speech

On the Perception of Juncture in English

W.J. Barry
Kiel, Federal Republic of Germany

1. Introduction

In English, where relatively unconstrained syllabification and free stress patterning exist, word-boundary disambiguation is more critical than in languages where fixed word stress or more restricted syllable structure reduce potential ambiguities to a minimum. Word identity is, of course, guaranteed to a large extent by situational information and by on-going processing of syntactic and semantic structure (Marslen-Wilson, 1975; 1980). However, it may be presumed in this as in other areas of speech perception that top-down processing must be supported by a systematic signal structure.

In the two major experimental studies of juncture production and perception (Lehiste, 1960; Gårding, 1965), systematic differences in the speechsignal structure of juncturally ambiguous strings of segments were found. With few exceptions (e.g. the style-dependent use of glottalized vocalic onset in initial vowels) the differences can be subsumed under the parameters duration and intensity, both of which contribute to the perceptual prominence of a segment. Word initial segments generally have greater prominence than cognate word-final segments. In a later study by the present author (Barry, 1981), the main trends of the previous analyses were confirmed, though there was a tendency, to a differing degree from speaker to speaker, to neutralize word-initial and word-final parameter values in running texts. Also, uncertainty in junctural identification was found to coincide with one or more atypical parameter values.

Neither the studies mentioned above nor the exclusively perception oriented studies by O'Connor and Tooley (1964) and by O'Connor-Dukes and Nakatani (1979) have carried out perception tests with systematic variation of the signal properties which have been found to differ.

The present paper reports on a perception experiment which aimed to examine the contribution to junctural identity of various parameters specified in the three studies mentioned above.

The junctural pairs: See Mabel - seem able; why choose - white shoes; keep sticking - keeps ticking were spoken several times by the author and one pair selected on the basis of an auditory judgement as to junctural distinctiveness and prosodic identity. Care was taken not to produce glottal constriction during 'seem able', nor to glottalize the /t/ in 'white shoes'. These pairs were selected from the nine pairs used in the previous study (Barry, 1981) for reasons of differing trans-junctural complexity. The durational characteristics of the individual utterances are given in Table I.

Apart from the modified properties, no changes were made in the stimuli. Table II gives the parameter values used for modifying the stimuli.

Signal manipulation was carried out with the Speech Signal Processor (Schäfer, 1982) at the Kiel Institute of Phonetics. To avoid undue concentration on the acoustic structure of each juncture type, all three pairs were offered together in a single test.

Five repetitions of the 32 generated stimuli from the 3 junctural pairs (6 + 8 + 18) were offered with a 3 s. decision pause in randomized order to 20 native speakers of English for identification. Subjects were required to mark the member of the pair identified.

Table I. Durational values for signal segments of junctural pairs (ms)

a.	/s/	/i:/	/m/	eɩ/	/b/	/l/	Σ
See Mable	153	149	76	187	75	157	797 ms
Seem able	144	161	61	187	82	135	770 ms
b.	/waɩ/	/t/	/ʃ	/u:z/	Σ		
Why choose	223	57	110	438	828 ms		
White shoes	218	54	174	404	850 ms		
c.	/ki:/	/p/	/s/	[t]	[-h]	/ɩkɩŋ/	Σ
Keep sticking	139	102	95	28	16	335	715 ms
Keeps ticking	124	80	81	56	39	334	714 ms

Table II. Parameter values for modification of junctural-pair members

a.					
See Mabel	59	67,6	76		
vs.					[m] duration (ms)
seem able	61	69,7	78,4		
step	1	2	3		
b.					
why choose					
vs.	174	152	131	110	[ʃ] duration (ms)
white shoes					
step	1	2	3	4	
c.					
keep sticking	161	179	197		[ps] duration (ms)
vs.					
keeps ticking	44	69	95		[t^h] duration (ms)
step	1	2	3		

2. Results

Group average scores are displayed graphically in Fig. 1a - c. It is immediately apparent that the degree of certainty with which the junctural distinction is made varies considerably. Although the number of 'Mabel'-judgements increases systematically with increasing [m]-duration (all rank-sum diffs. > 12, n = 20, k = 3; $p < 0.05$ after Wilcoxon-Wilcox, one-sided), even the unmodified, original 'Mabel' does not score better than 60%. The 'choose' judgements, on the other hand, range from 24% - 93%, and 'sticking' from 10% - 88%.

These differences can be explained partly by the degree of acoustic change imposed on the original stimuli (cp. Table II). Whereas the difference in [m]-duration between 'See Mabel' and 'seem able' is only 17 ms, the fricative element in 'white shoes' and 'why choose' differs by 64 ms. However, the even greater difference between 'keep sticking' and 'keeps ticking' (36 ms + 51 ms) does not result in the expected unanimity of judgement for the original stimuli.

With the exception of the 'sticking-ticking' pair the parameters modified in this experiment appear to explain the junctural distinction completely. In neither of the other two pairs is there any difference in the judgements as a function of the original stimulus. The stimuli derived from original 'ticking', however, receive significantly more 'ticking' judgements than those derived from 'sticking' (T = 32,5 < 52, n = 20, $p < 0.05$ after Wilcoxon). This indicates that remnant junctural information is contained in the stimuli apart

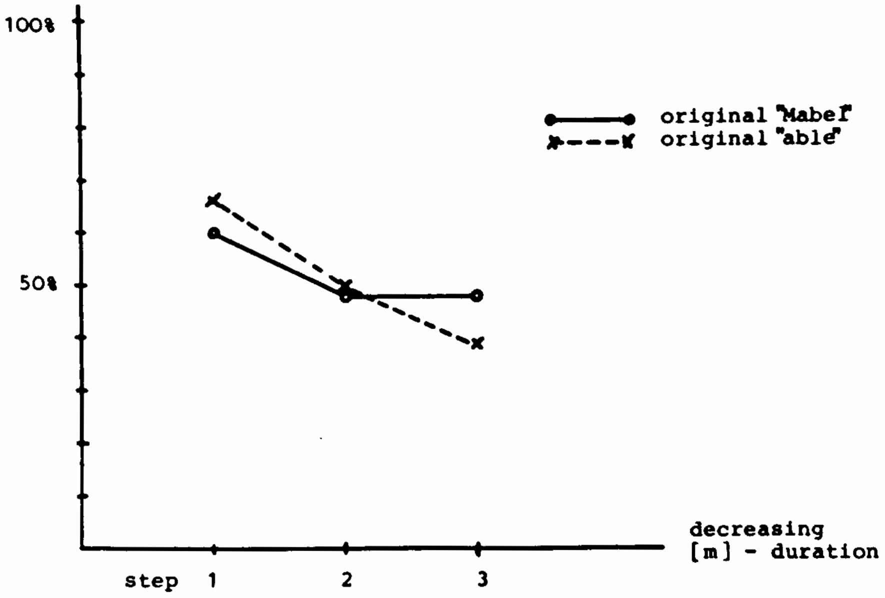


Figure 1. 'Mabel' judgments (English subjects).

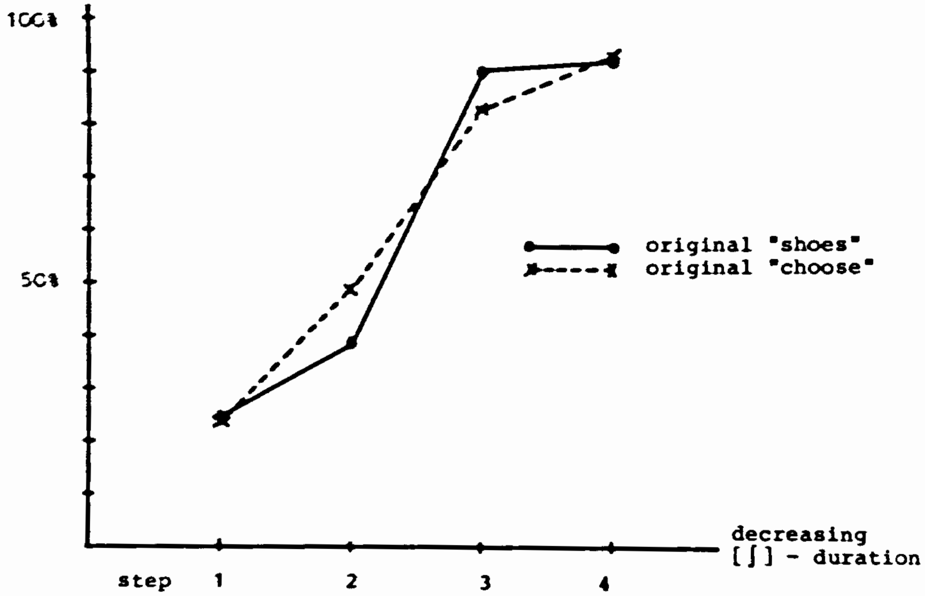


Figure 1b. 'choose' judments (English subjects).

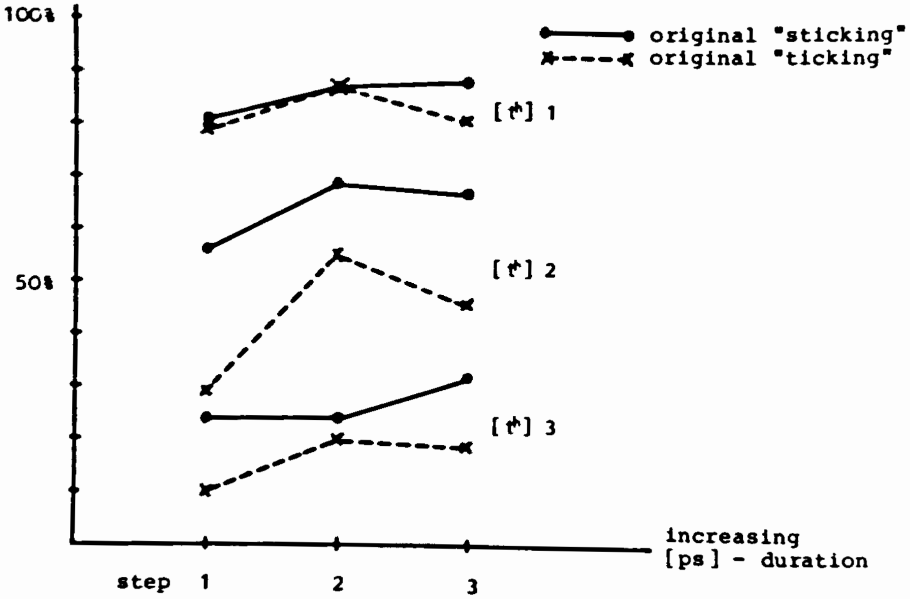


Figure 1c. 'Sticking' judgments (English subjects).

from the durational information in the [pst^h] segments that were manipulated. The likeliest candidate is the greater release burst intensity of [t^h] since it remained unmodified by the durational manipulation applied to the section following the burst spike. This assumption is plausible in the light of the dominant influence which the aspiration has on junctural identity: changes in /ps/ duration have only a slight effect on judgement distribution (see Fig. 1c) though the influence again appears to be systematic between the extreme parameter values (diff. > 12, $n = 20$, $k = 3$; $p < 0.05$ after Wilcoxon-Wilcox).

In a test with 15 German subjects the regularities in perception were much less clearly defined (Fig. 2a - c), though for the 'choose-shoes' and the 'ticking-sticking' distinction there was a significant perceptual effect of [ʃ] and [t^h]-duration, respectively (diffs. > 15 for 'shoes-choose', $k = 4$; diffs. $>$ 11 for 'sticking-ticking', $k = 3$; $p < 0.05$). The German subjects were unable to distinguish the 'Mabel-able' pair, understandably in view of the predominance of glottalized vowel onsets in German, and they failed to react to the [ps]-duration in 'keeps ticking' vs. 'keep sticking'.

3. Discussion

The following tentative conclusions may be drawn from the results:

1. Junctural distinctions vary considerably in their perceptibility, but under circumstances of mixed presentation do not achieve scores suggesting clear junctural categories even in their original realisation. This would suggest that the *phonetic* decoding of word boundaries is secondary to the

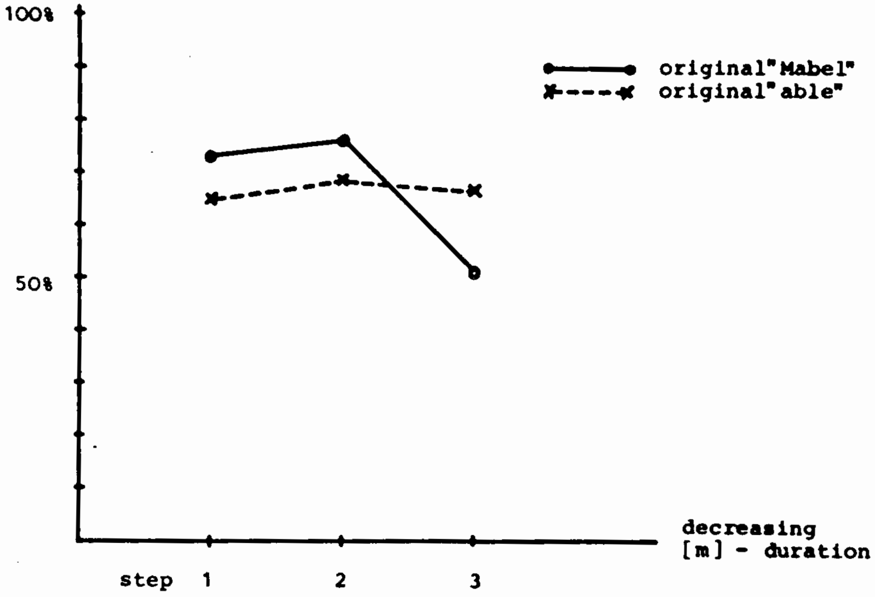


Figure 2a. 'Mabel' judgments (German subjects).

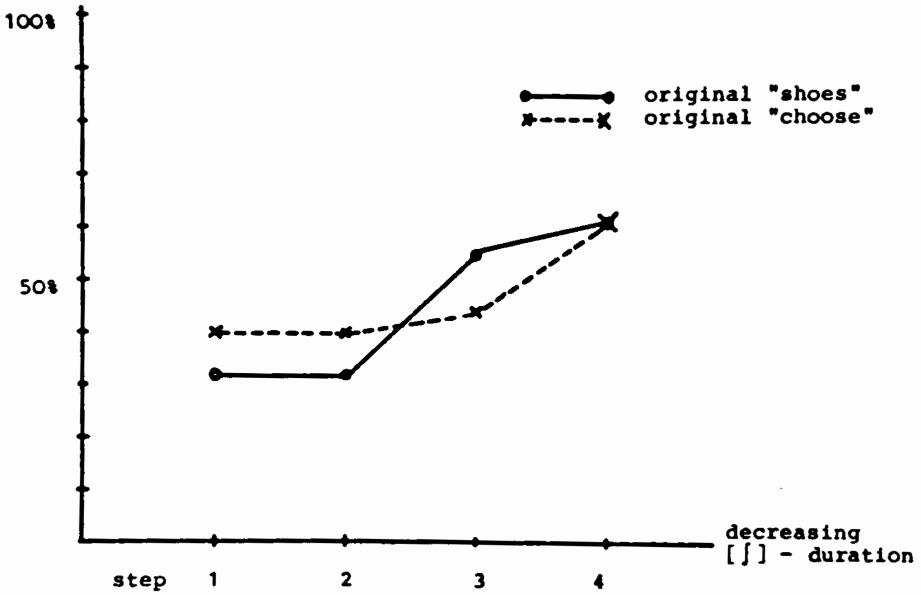


Figure 2b. 'choose' judgments (German subjects).

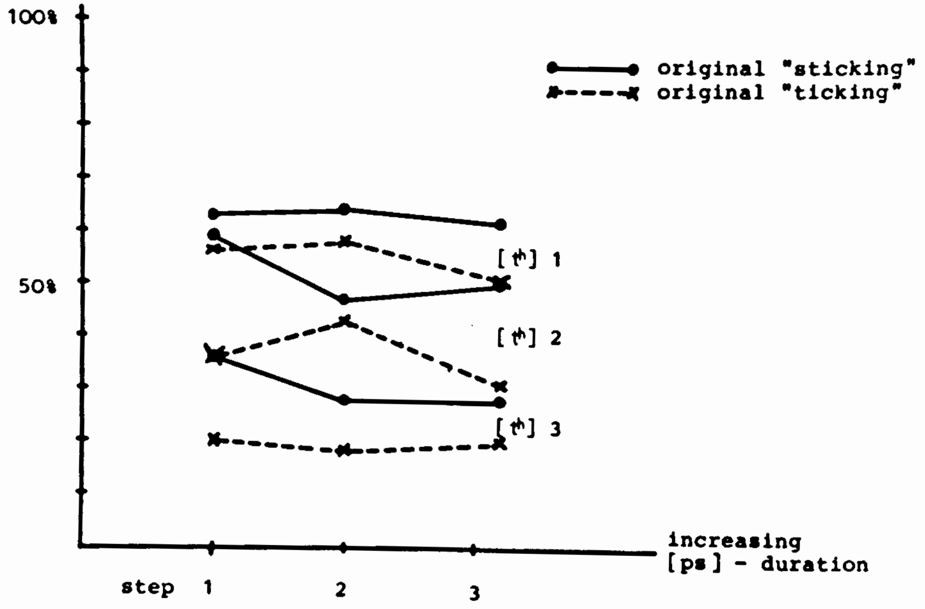


Figure 2c. 'sticking' judgments (German subjects).

clearly categorical function of phonemic identification, and that top-down information has a more prominent place in the perceptual strategy. This is in accord with the tendency to neutralize junctural distinctions in running text *production* (Gårding, 1965; Barry, 1981).

2. Not all systematic signal differences contribute to the perception of juncture. There is some support for Gårding's (1965, p. 15) hypothesis that *post*-junctural information is more important then *pre*-junctural information. The degree of post-junctural aspiration was dominant in the 'ticking-sticking' decisions, and although the 'choose-shoes' pair was only varied in a single (post-junctural) parameter it was distinguished quite as well as 'sticking-ticking'. Also it may be assumed that the missing [ʔ] in 'able' contributed to the extreme uncertainty of decisions on the 'Mable-able' pair. However, a global post- vs. pre-junctural dichotomy is obviously too undifferentiated; the longer post-junctural /s/ in 'sticking' had no effect on decisions.
3. With mixed presentation (i.e., under slightly less 'clinical' conditions than often exist in perception tests), junctural identification scores could be explained completely by a single parameter. This suggests that subjects' extreme sensitivity to a multiplicity of signal differences, which has been used as an argument against feature-based perception (Bailey and Summerfield, 1978), might be less a component of normal speech perception than a sign of human perceptual flexibility and learning potential under extreme condition (cp. Barry, 1980, p. 116-7).
4. Despite the 'production-universal' character of the junctural differences, the perceptual exploitation of them appears, at least partly, to be language specific. The German subjects, accustomed to a language with strong distributional constraints in the phonemic system (final devoicing, glottal onset in initial vowels) showed generally weaker reactions to the potential junctural signals.

References

Bailey, P.J. and Summerfield, A.Q. (1978). Some observations on the perception of [s]+stop clusters. *Haskins Labs. Status Report on Speech Research*, SR-**53**.2, 25-60.

Barry, W.J. (1980). *Die Verarbeitung akustischer Information in der lautsprachlichen Wahrnehmung.* (= Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK), **13**).

Barry, W.J. (1981). Internal juncture and speech communication. In: *Beitrage zur experimentellen und angewandten Phonetik* (eds. W.J. Barry and K.J. Kohler) (= Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK), **16**), 229-289.

Gårding E. (1965). *Internal Juncture in Swedish.* Lund: Gleerup (= Travaux de l'Institut de Phonétique de Lund **VI**).

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica* **5**, Supplement 5-54.

Marslen-Wilson, W. (1975). Sentence perception as an interactive parallel process. *Science* **189**, 226-228.

Marslen-Wilson, W. (1980). Speech understanding as a psychological process. In: *Spoken Language Generation and Understanding* (ed. J.C. Simon), Dordrecht, Boston, London: Reidel Publ. Co., 39-67.

O'Connor, J.D. and Tooley, O.M. (1964). The perceptibility of certain word boundaries. In: *In Honour of Daniel Jones* (eds. Abercrombie, Fry, MacCarthy, Scott and Trim), London: Longmans, 171-176.

O'Connor-Dukes, K.D. and Nakatani, L.H. (1979). The role of morphology in a word-boundary decision task. *J. Acoust. Soc. Am.* **66**, Suppl. 1, 550 (A).

Schäfer, K. (1982). Concepts in the SSP programme. In: *Phonetic Data Processing at Kiel University* (eds. W.J. Barry and K.J. Kohler), (=Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK), **18**, 110-121.

'Koartikulation' and 'Steuerung' as Factors Influencing the Perception of 'Momentary Tempo'

B. Pompino-Marschall, H.G. Piroth, P. Hoole, H.G. Tillmann
Munich, Federal Republic of Germany

1. Introduction

In earlier experiments we were able to replicate the results of Ventsov's (1981) investigations into the influence of open and closed syllables on the perception of 'momentary tempo' but the results of our modified experiments led us to the conclusion that it is rather the sequence of the articulatorily defined point of 'Koartikulation' - in accordance with the results of p-centre experiments (cf. Tuller/Fowler 1980) - than the sequence of closed syllables as proposed by Ventsov, which determines the perception of 'momentary tempo'. The tests described below were designed firstly to examine the effect of the direction of vowel manipulation and secondly to investigate sequences with consonant clusters instead of just one consonant.

2. Procedure

The subjects were asked to decide wheter a [mamamamamam] sequence had the same rhythmic structure as an [a a a a a] sequence (all [a] sounds being separated by pauses), whereby the repetition rate of the open [ma] syllable was different from the repetition rate of the closed [am] syllable. A natural utterance of [mamamamamam] was segmented into single pitch periods normalised with regard to F_0 as described in Pompino-Marschall et al. (1982). In the following the pitch period will be used as the unit of time. For Experiment 1 the periods from this inventory were combined according to (1):

(1) (4m+[26a)+(8m]+[22a)+(12m]+[18a)+(16m]+[14a)+
(20m]+[10a)+24m]

In this way a test sequence was produced in which the [ma] syllable is always 30 periods long, and the [am] syllable always 34 periods long (round brackets vs square brackets).

Five sequences for comparison were constructed, each consisting of five vowels separated by pauses. The vowels were constructed from the above inventory of [a] periods, and were equal in length with those of the test sequence. The pauses were so chosen as to give five different sequences with vowel repetition rates corresponding to a time equal to 28, 30, 32, 34 and 36 periods respectively. The five sequences for comparison with the test sequence are analogously described in (2), where 'nP' stands for a pause of length n periods:

(2.1) (26a + 2P) + (22a + 6P) + (18a + 10P) + (14a + 14P) + 10a
(2.2) (26a + 4p) + (22a + 8P) + (18a + 12P) + (14a + 16P) + 10a
(2.3) (26a + 6P) + (22a + 10P) + (18a + 14P) + (14a + 18P) + 10a
(2.4) (26a + 8P) + (22a + 12P) + (18a + 16P) + (14a + 20P) + 10a
(2.5) (26a + 10P) + (22a + 14P) + (18a + 18P) + (14a + 22P) + 10a

Stimuli for an AX test were formed by combining sequence (1) with each of the sequences (2.1)-(2.5). The five stimuli, which occurred ten times each, were presented to the subjects over loudspeakers. The interval between the end of the first sequence and the beginning of the second sequence within each stimulus was 250 ms. The stimuli were separated by a pause of 3 s. Between each block of 10 stimuli there was a longer pause of 5 s. The subjects were required to decide whether the two sequences in a given stimulus had the same rhythmic structure. 12 students and members of the institute took part in the experiments.

3. Results

The results of Experiment 1 largely confirmed the findings of Ventsov (1981) (see fig. 1). As in his Experiment 2 the stimulus with a vowel repetition rate of 34 periods in the second sequence was the stimulus most often judged to have the same rhythmic structure in both sequences. In contrast to Ventsov however we found an asymmetry between the number of 'same' responses to stimuli (1 - 2.3) and (1 - 2.5). In other words, the sequence with the slowest vowel repetition rate (36 periods) was more sharply discriminated from the test sequence than was the sequence with a vowel repetition rate of 32 periods. To test our hypothesis that this asymmetry is caused by a perception of increasing tempo within the test sequence, for Experiment 2 we constructed a complementary test sequence with lengthening [a] according to (1.1):

(1.1) (24m+[10a)+(20m]+[14a)+ (16m]+[18a)+(12m]+[22a)+
(8m]+[26a)+4m].

Analogously to Experiment 1 the vowels of the sequences for comparison were chosen so as to reflect the vowel structure of the test sequence. Procedure and subjects remained the same.

The results (cf. fig. 1) are clearly in agreement with our hypothesis (the closed syllable is here 30 periods long) since there is a similar asymmetry but in the opposite direction.

To test whether these results are interpretable along p-centre lines, for the next pair of experiments we complicated the consonantal part of the sequence by introducing the variable-length fricative [ʃ] before the constant [m]:

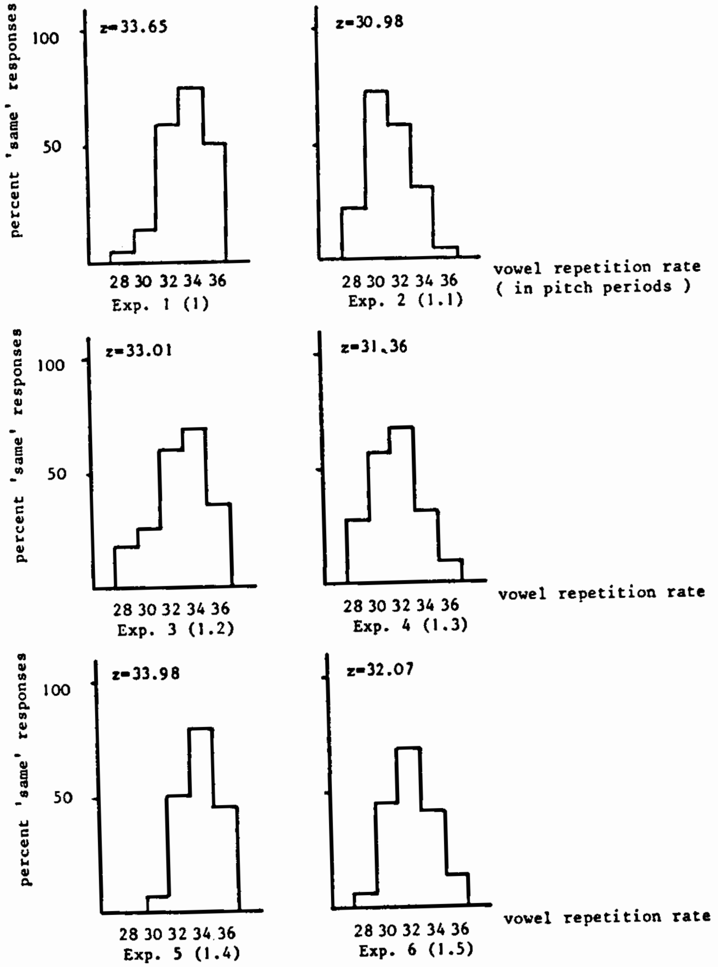


Figure 1. Results of experiments 1-6 (numbers in brackets referring to the formula in the text).

(1.2) (4ʃ+[4m+22a)+(8ʃ]+ [4m+8a)+(12ʃ]+[4m+14a) +(16ʃ]+[4m +10a)+(20ʃ] +[4m+6a)+24ʃ].

In analogy to Experiment 2 the complementary sequence (cf. 1.3) was tested in Experiment 4:

(1.3) (24ʃ+[4m+6a)+(20ʃ]+ [4m+10a)+(16ʃ]+[4m+14a)
+(12ʃ]+[4m+18a)+(8ʃ] +[4m+22a)+4ʃ].

The results of this pair of experiments clearly show the same reversal in the asymmetry of responses (see fig. 1); it is also noticeable that the median for the 'same' responses now lies slightly further away from vowel onset than in the simple [mam] sequences. The results are all in agreement with those of the p-centre experiments: lengthening of prevocalic consonant clusters shifts the point of psychological syllable onset away from vowel onset and the complementary sequence shows the mathematically opposite effect.

In the last pair of experiments we tested our hypothesis that these results as well as the results of the p-centre experiments are interpretable along the lines of the theory on 'Koartikulation' and 'Steuerung' proposed by Menzerath/de Lazerda (1933), i.e. that the psychological moment of syllable onset is the point of 'Koartikulation', defined by the simultaneous production of the prevocalic consonant and the vowel.

Here we kept the vowel constant and varied the relationship of [m] to [ʃ] in a consonant cluster also of constant length, sequence (1.4) starting with long [ʃ], short [m] and sequence (1.5) starting with short [ʃ], long [m]:

(1.4) (24ʃ+[4m+6a)+(20ʃ]+ [8m+6a)+(16ʃ]+[12m+6a)+
(12ʃ]+[16m+6a)+(8ʃ]+ [20m+6a)+4ʃ]

(1.5) (4ʃ+[20m+6a)+(8ʃ]+[16m+6a)+(12ʃ]+[12m+6a) +(16ʃ]+[8m
+6a)+(20ʃ] +[4m+6a)+24ʃ]

The experiment with sequence (1.4) could also be easily interpreted along the lines of 'Koartikulation' (see fig. 1). However, the complementary sequence (1.5) was perceived as beginning in [ʃma] and ending in [maʃ], and gave results showing that the subjects perceived the tempo as lying half-way between [ʃma] and [maʃ]. We would interpret this as indicating that the effect of 'Steuerung' becomes more pronounced in the course of the sequence i.e. as syllable-final [ʃ] becomes longer.

References

Menzerath, P., Lacerda, A. de (1933). Koartikulation, Steuerung und Lautabgrenzung. Eine experimentelle Untersuchung. *Phonetische Studien* **1**. Berlin and Bonn: Dummlers.

Pompino-Marschall, B., Piroth, H.G., Hoole, P., Tilk, K., Tillmann, H.G. (1982). Does the closed syllable determine the perception of 'momentary tempo'? *Phonetica* **39**.

Tuller, B., Fowler, C.A. (1980). Some articulatory correlates of perceptual isochrony. *Perc. and Psychophys.* **27**, 277-283.

Ventsov, A.V. (1981). Temporal information processing in speech perception. *Phonetica* **38**, 193-203.

Patterns of Temporal Compression in Spoken Italian

M. Vayra, C. Avesani and C.A. Fowler
Pisa, Italy and New Haven, USA

1. Introduction

According to many linguists (e.g. Classe, 1939; Pike, 1945; Abercrombie, 1965), languages can be classified by their rhythmic structures as stress timed (Germanic languages) or syllabic timed (Romance languages). A strong version of a stress-timing hypothesis for Germanic languages - that intervals between stressed-syllable onsets are isochronous - has been disconfirmed (e.g. Lehiste, 1973; Lea, 1974 among many others). This disconfirmation, coupled with evidence that listeners are insensitive to acoustic departures from isochrony in speech (e.g. Lehiste, 1973; Donovan and Darwin, 1979), has suggested that stress timing is, at least in part, a perceptual illusion.

However, some 'stress-timed' languages exhibit temporal structures that may support linguists' and other listeners' intuitions. Lindblom and Rapp (1973) report that stressed syllables in a word shorten substantially in the context of following unstressed syllables and weakly in the context of preceding unstressed syllables. This pattern of temporal compression, found also in Dutch (Nooteboom, 1973), English (Fowler, 1981; Huggins, 1978) and other languages, suggests a metrical foot structure in which unstressed syllables cohere with preceding stressed syllables. This durational compression is perceptible (Nooteboom, 1973), and may underlie the perceptual reports of stress timing. (See Dauer, 1980, for another explanation).

A strong version of a syllable-timing hypothesis for Romance languages is also disconfirmed (e.g. Navarro-Tomas, 1946, and see our Figure 1 below). Our research is designed to ask whether the linguists' reports that Romance languages, in particular Italian (cf. Bertinetto, 1981), are syllable-timed can be explained by patterns of temporal compression. Compression suggestive of syllable timing would include shortening of a vowel in the context of preceding and following intrasyllabic consonants. Our research tests predict that strong syllable-level compression, but, at most, very weak foot-level compression will occur in Italian.

2. Experiment

To test the first hypothesis (syllable-timing) we measured the duration of [a] in $C_0^2VC_0^2$ syllables. To test the second hypothesis (stress-timing), we used 14 stress patterns corresponding to real words or to noun-verb combinations of words. For each stress pattern, [a] was produced in reiterant versions ([da da...]) of the target words. The stress patterns realized four different conditions in which different numbers of unstressed vowels preceded and followed the target stressed [a], and in which the position of word boundary intervened between stressed and unstressed vowels or did not intervene. All utterances were spoken in a sentence frame.

For the syllables, the interval of periodicity for the vowel's voicing was measured from a waveform display. For the reiterant productions, vowel duration was measured as the interval between burst release and the first evidence of closure for the following consonant.

One subject, a native speaker of Italian who was naive to the purposes of the study, repeated each of the syllables three times. In addition, he repeated each stress pattern both in its real-speech and reiterant forms three times each.

3. Results

Figure 1 shows a weak and inconsistent effect of syllable structure on duration of the vowel. This outcome was not expected, based on the idea that temporal compression effects at the level of the syllable could explain reports that Italian is syllable-timed.

Figure 2 shows the effect of preceding and following unstressed vowels on a stressed vowel within a pair of word boundaries. The shortening effects are regular and asymmetrical like those occurring in Swedish and English. Figures 3 and 4 show anticipatory and backward shortening effects, respectively, when a word boundary intervenes between the stressed vowel and one or more of the unstressed vowels. Shortening is present but is weaker and less consistent than that shown in Figure 2.

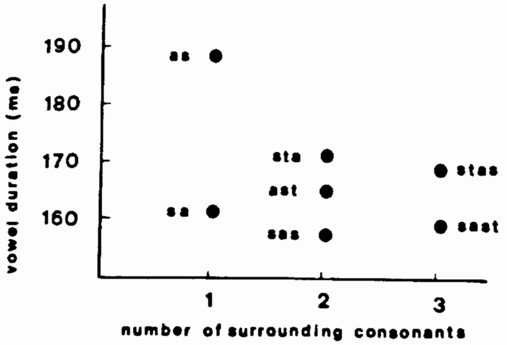


Figure 1. Syllable level. Effect of syllable structure on vowels duration.

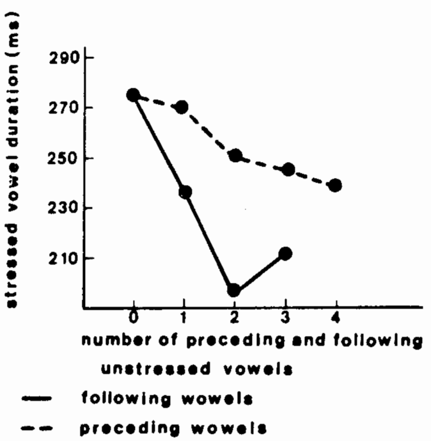


Figure 2. Inter stress interval. Within word boundary. Average duration of stressed vowels as a function of the number of preceding and following unstressed vowels with in the word.

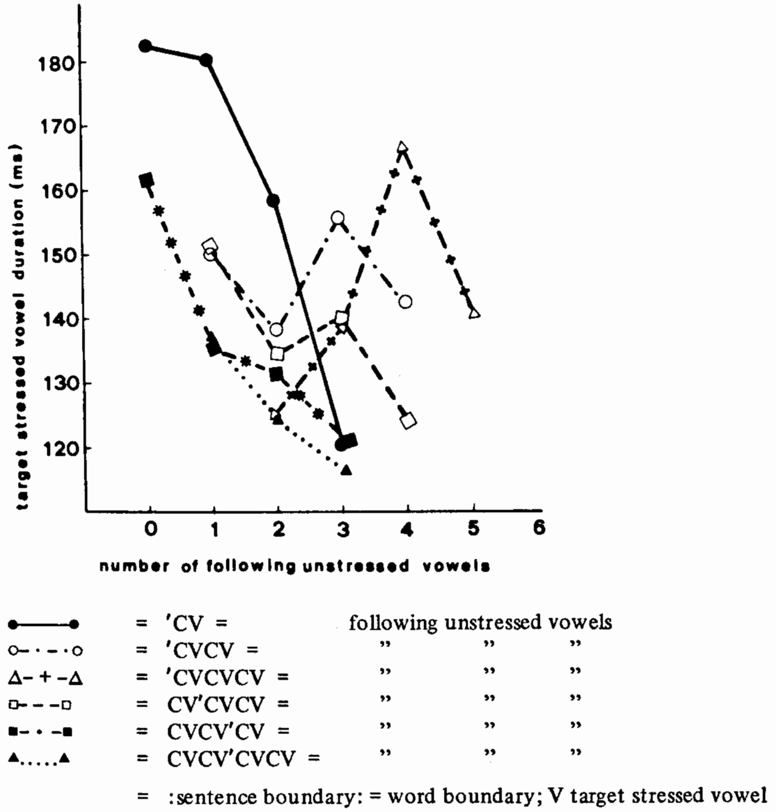


Figure 3. Inter stress interval. Across word boundary. Average duration of stressed vowels as a function of the number of following unstressed vowels across word boundary. Like in Fig. 4 six compression curves are plotted, one for each of the prosodic contexts on the left side of the word boundary (#).

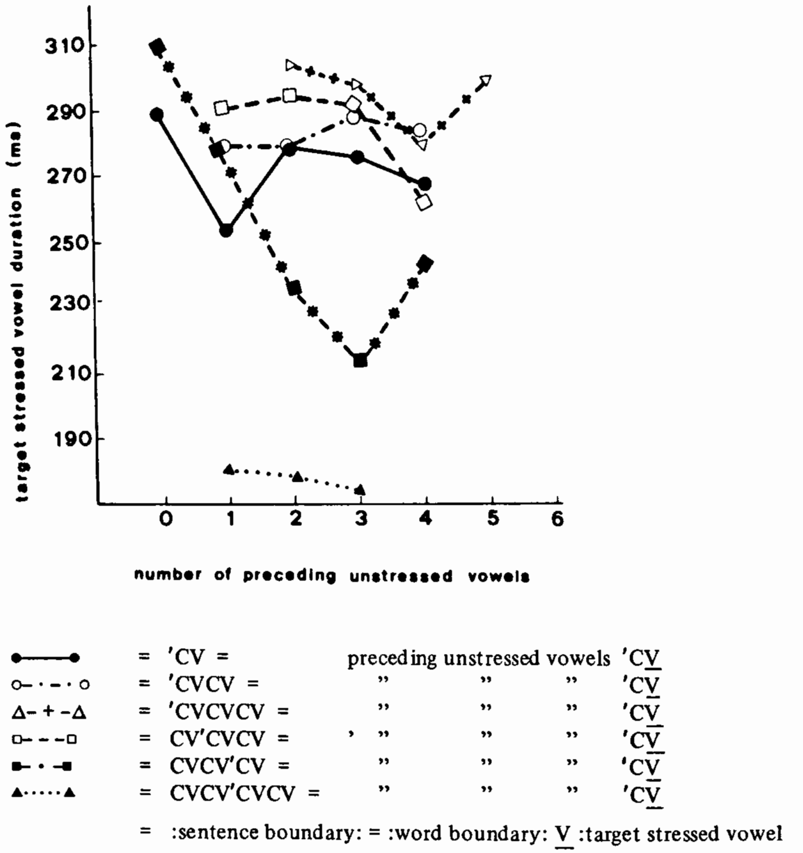


Figure 4. Inter stress vowel. Across word boundary. Average duration of stressed vowels as a function of the number of preceding unstressed vowels across word boundary.

4. Discussion

To compare the strength of compression effects in Italian with those reported for speakers of stress-timed languages, we submitted our foot- level data to modelling procedures similar to those described by Lindblom and Rapp (1973). The syllable-level data were not modelled because the vowels did not show regular shortening influences of surrounding consonants.

Lindblom and Rapp (1973) modelled patterns of temporal compression using the following formula: $D_0 = D_i/(1 + a)^{\alpha} (1 + b)^{\beta}$. Applied to stress vowels surrounded by various numbers of unstressed syllables, D_i is the duration of a stressed vowel in a monosyllable. In a polysyllable, this underlying duration is compressed by: 'a' the number of following unstressed syllables in a word; and 'b' the number of preceding syllables. α and β are parameters of anticipatory and backward compression. For Lindblom and Rapp's data, the best-fit values of α and β were .48 and .12, the asymmetry reflecting the foot structure of Swedish.

Fowler (1981) modified the model slightly, adding a parameter of final lengthening and counting unstressed vowels across a word boundary in the value of 'a' and 'b'. In her data, $\alpha = .6$ and $\beta = .07$ (average error, 9.3 msec per stressed vowel) for utterances produced to a metronome. For the same utterances without the metronome (Fowler, unpublished data) $\alpha = .5, .4$ and $\beta = .05, .09$, for two speakers (average error: 3 msec).

For our Italian speaker, we modelled two sets of utterances: those in which unstressed vowels are within the same word as the stressed vowel ($\alpha = .08$, $\beta = .07$, average error: 10 msec), and those in which some unstressed vowels are not ($\alpha = .17$, $\beta = .03$, average error: 11 msec). In accordance with our expectations, the speaker of Italian showed much weaker foot-level shortening (that is, the values of α are small) than speakers of two-stress-timed languages.

5. Conclusion

Consistent with our predictions, we found evidence for relatively weak compression at the level of the foot in Italian. Unexpectedly, however, syllable-level compression also was weak. Indeed, overall, stressed vowels in Italian showed rather weak shortening effects.

We hypothesize that Italian is reported to be syllable-timed for two reasons: first, unstressed vowels cause little foot-level shortening and second, reportedly (Bertinetto, 1981), unstressed vowels reduce less in Italian than in stress-timed languages

Acknowledgements

This research was supported by the following grants to Haskins Laboratories: NIH-NICHD Grant HD-16591 and NSF Grant BNS-8111470. We are grateful to the Laboratories for the friendly and inspirational help which made this research possible for us.

References

Abercrombie, D. (1964). Syllable quantity and enclitics in English. In D. Abercrombie, D. Fry, P. MacCarthy, N.S. Scott and J. Trim (eds.), *In honour of Daniel Jones.* London: Longman.

Bertinetto, P.M. (1981). *Strutture prosodiche dell'italiano.* Firenze: Accademia della Crusca.

Classe A. (1939). *The rhythm of English prose.* Oxford: Blackwell.

Donovan, A. and Darwin, C. (1979). The perceived rhythm of speech. *Proceedings of the Ninth International Congress of Phonetic Sciences,* **2**, 268-274.

Dauer, R. (1980). Stress timing, syllable timing and isochrony. *Linguistic Society of America Annual Meeting.*

Fowler, C.A. (1981). A relationship between coarticulation and compensatory shortening. *Phonetica,* **38**, 35-40.

Huggins, A.W.F. (1978). Speech timing and intelligibility. In: J. Requin (ed.), *Attention and performance VII.* Hillsdale, N.J.: Lawrence Erlbaum Assoc., 279-298.

Lea, W.A. (1974). Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding. *UNIVAC Rep. No. PX10791.* St. Paul, Minn.: Sperry Univac, DSD.

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *JASA,* **54**, 1228-1234.

Lindblom, B. and Rapp, K. (1973). Some temporal regularities of spoken Swedish. *PILUS,* **21**, 1-59.

Navarro-Tomas, T. (1946). *Estudios de Fonologia Española*. New York: Las Americas.

Nooteboom, S.G. (1973). Perceptual reality of some prosodic durations. *J. of Phonetics,* **1**, 25-45.

Section 12
Communicative Functions of Prosody

Acceptability of Accenting and De-accenting 'NEW' and 'GIVEN' in Dutch

J.G. Kruyt and S.G. Nooteboom
Leyden, the Netherlands

1. Introduction

Thanks to the work of 't Hart and Cohen (1973) and 't Hart and Collier (1975) we have at our disposal a 'grammar of Dutch intonation', a rule system that generates stylized pitch contours that are melodically equivalent to naturally occurring Dutch pitch contours. These rules can easily be implemented in a system for speech synthesis-by-rule, and then supply the synthetic speech with acceptably sounding pitch contours. Figure 1 shows some possible pitch contours for the sentence 'Het Concergebouworkest speelt Mahler's negende symphonie' (The Concertgebouworchestra plays Mahler's ninth symphony). The contours differ - among other things - in number and location of the pitch accents (a pitch accent is an accent lending pitch movement on the lexically stressed syllable).

For example, in (a) the words *Concertgebouworkest, Mahler's, negende* and *symfonie* are accented, whereas in (b) the words *Concertgebouworkest* and *Mahler's* remain unaccented (or are de-accented). The grammar of Dutch intonation can generate these and other possible pitch contours, but at present it is still necessary to indicate by hand which words have to be marked with a pitch accent. In other words, rules that automatically and correctly determine pitch accent locations are still lacking.

Part of the research in our laboratory is concerned with the question which

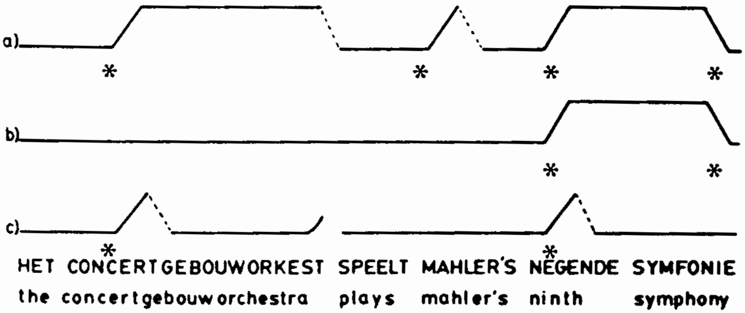

a)
b)
c)
*
HET CONCERTGEBOUWORKEST SPEELT MAHLER'S NEGENDE SYMFONIE
the concertgebouworchestra plays mahler's ninth symphony

Figure 1. Some examples of stylized pitch contours. Accent lending pitch movements are marked with '*'.

factors systematically affect the accenting of words in read aloud texts. Central questions are (a) which factors influence the speaker's choice of the words to be marked with a pitch accent, and (b) how do accents affect the listener's perception and/or comprehension of an utterance. In this paper we are concerned with only one of the factors related to accentuation, viz. the 'newness' vs. 'giveness' of the information referred to (Chafe 1976; Firbas 1979; Halliday 1967). In an earlier production experiment, in which subjects had to read aloud newsbulletin-like items, we defined the terms NEW and GIVEN operationally as 'not mentioned' and 'mentioned in the immediately preceding utterance', respectively. A syntactic constituent could be repeated in the same form or in the form of a synonym. In accordance with the literature we found that NEW information was generally accented. However, rather unexpectedly, GIVEN information was very often accented, too. Only a few speakers de-accented the GIVEN information (Kruyt 1982). This raised the question whether or not context has much effect on accenting in this kind of newsbulletin-like language material. Or, more specifically, whether or not accenting of just mentioned information is acceptable to the listener, and if so, whether or not listeners have a preference for accenting or de-accenting. This question was addressed in the following perception experiment.

2. Experiment

In the present experiment several instances of plus and minus accent on particular words were judged as to their relative acceptability, as a function of the 'newness' vs. 'giveness' of the referents of these words. NEW and GIVEN were again defined as 'not mentioned' and 'mentioned in the immediately preceding utterance', respectively. For this reason pairs of sentences were used, in which the grammatical subject of the second sentence was or was not mentioned in the first. Four 'first sentences of a news item' and two 'second sentences' were constructed. These are shown in Figure 2. The grammatical subject of the 'second sentence' is *Gouda* or *gemeente*. *Gouda* is NEW in the sentence pairs Aa and Ba, and GIVEN - in the form of a repetition of the name *Gouda* - in Ca and Da. *Gemeente* can only be GIVEN, viz. in the sentence pairs Cb and Db, in the form of a synonym of *Gouda* (Ab and Bb do not make sense). All these sentences were spoken by a speaker of standard Dutch, and, with the aid of an LPC-analysis-resynthesis system ('t Hart, Nooteboom, Vogten and Willems 1982), provided with pitch contours as shown in Figure 2. In the second sentences the pitch contour on the words *Gouda* and *gemeente* was systematically varied: 1, 2 and 3 are accent lending movements, 4 and 5 are not. The predicate always had a 'hat pattern', consisting of an accent lending rise, a stretch of high declination line, followed by an accent lending fall. All these re-synthesized utterances were recorded onto Bell and Howell Language Master cards, so as to enable easy recombination of utterance pairs during experimental sessions.

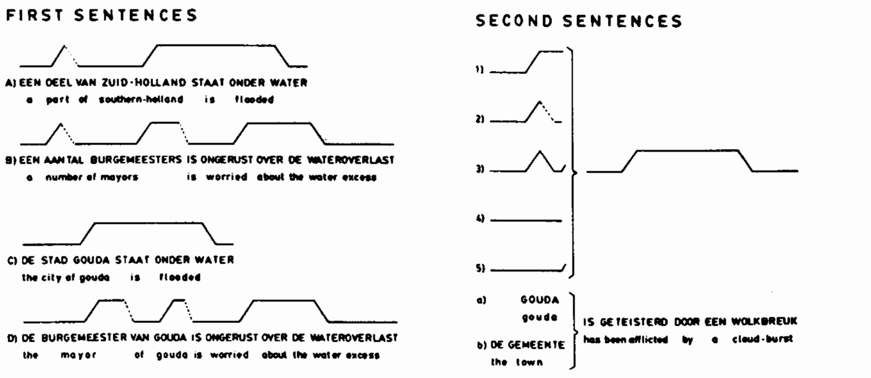

FIRST SENTENCES
A) EEN DEEL VAN ZUID-HOLLAND STAAT ONDER WATER
a part of southern-holland is flooded
B) EEN AANTAL BURGEMEESTERS IS ONGERUST OVER DE WATEROVERLAST
a number of mayors is worried about the water excess
C) DE STAD GOUDA STAAT ONDER WATER
the city of gouda is flooded
D) DE BURGEMEESTER VAN GOUDA IS ONGERUST OVER DE WATEROVERLAST
the mayor of gouda is worried about the water excess
SECOND SENTENCES
1)
2)
3)
4)
5)
a) GOUDA
gouda
b) DE GEMEENTE
the town
IS GETEISTERD DOOR EEN WOLKBREUK
has been afflicted by a cloud-burst

Figure 2. Survey of stimulus utterances and their stylized pitch contours, used in the experiment.

Sixteen listeners had to fulfill two tasks. In a forced choice task they had to select the most appropriate first sentence (A, B, C or D) for each intonational variant of the utterance 'Gouda is geteisterd door een wolkbreuk' (Gouda has been afflicted by a cloud-burst) (a1-5). In a scaling task they had to listen to each possible utterance pair (A, B, C and D combined with a1-5, C and D combined with b1-5) and indicate on a ten-point scale (1-10) how well the utterance pair sounded. The ten-point scale was chosen because of the familiarity of our subjects with this scale, as it is generally used in the Dutch educational system (10 means very good, 1 means very bad). Tasks and order of presentation of the utterance pairs were balanced over the listeners.

3. Results

Table I shows the results of the forced choice test. Rows show the two categories [+accent] (a1-3) and [-accent] (a4,5) on *Gouda*. Columns show how many times subjects responded with [*Gouda* NEW] (A or B selected) and [*Gouda* GIVEN] (C or D selected) (differences within the categories [+accent] and [-accent] were small). [-accent] is clearly associated with [*Gouda* GIVEN] (30 out of 32). [+accent], however, has been assigned to both categories, albeit with preference for [*Gouda* NEW] (30 out of 48).

In Figure 3 the results of the scaling test are presented. Panel A shows

Table I. Distribution of utterances intonationally marked with and without accent on *Gouda*, judged most suitable to introducing a new referent (NEW) or to repeating a previous mention (GIVEN). The effect of plus vs. minus accent is significant ($\chi^2 = 28$, df = 1, $p < .001$)

Type of pitch movement	GOUDA in 2nd sentence		Total
	NEW	GIVEN	
+accent	32	16	48
-accent	2	30	32

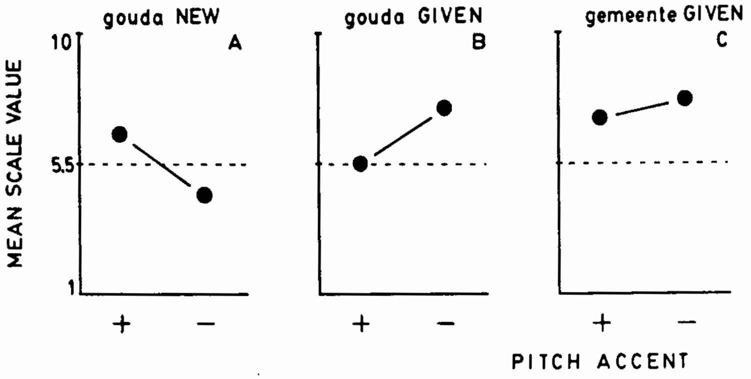


Figure 3. Mean acceptability of accenting or de-accenting a subject noun introducing a new referent (NEW) or repeating a previous mention (GIVEN), either by repetition (B) or by a synonym (C). All differences are significant (A: t(158) = 8.1, p < .001; B: t(158) = -8.0, p < .001; C: t(158) = -3.5, p = .001).

mean acceptability for [+accent] and [-accent] pitch movements on *Gouda* when introducing a NEW referent. [+accent] is preferred over [-accent] (6.5 vs. 4.4). In panels B and C mean acceptability for [+accent] and [-accent] are shown for *Gouda* and *gemeente*, respectively, both repeating a previously introduced referent either by repetition (B) or by a synonym (C). Now [-accent] is preferred over [+accent], both in the case of *Gouda* (7.4 vs. 5.5) and in that of *gemeente* (7.7 vs. 7.0). However, in the case of *gemeente* the difference (although significant) is so small that we may conclude that both [+accent] and [-accent] are acceptable.

4. Discussion

From these observations we can conclude that listeners prefer NEW referents to be mentioned in accented verbal expressions, as one would predict from the literature and from the earlier mentioned production experiment (Kruyt 1982). It would have been in line with the literature if GIVEN referents could not acceptably be referred to with accented verbal expressions. Here, however, - and this agrees with the production experiment - the data, although suggesting a preference for [- accent], also indicate that [+ accent] is not entirely unacceptable. If we draw the line between acceptable and unacceptable in the middle of our ten-point scale, at 5.5, we see that accenting a constituent referring to what is GIVEN is never clearly unacceptable, and is even clearly acceptable if the constituent is a synonym of the referring expression used in the previous utterance.

If these results will be confirmed in other experiments, then this could be taken to mean that 'newness' and 'givenness', as defined here, are not necessarily important determinants in accent location rules for this kind of text.

Acknowledgements

We thank J. 't Hart for re-synthesizing the pitch contours of the utterances used in the experiment. The investigations were supported by a grant from the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research Z.W.O., under project number 17-21-14.

References

Chafe, W.L. (1976). Givenness, contrastiveness, definiteness subjects, topics and point of view. C.N. Li (ed.), *Subject and Topic*. New York: Academic Press. pp. 25-55.

Firbas, J. (1979). A functional view of 'ordo naturalis'. *Brno Studies in English* **13**, 29-59.

Halliday, M.A.K. (1967). Notes on transitivity and theme in English II. *Journal of Linguistics* **3**, 199-244.

Hart, J. 't and Cohen, A. (1973). Intonation by rule: a perceptual quest. *Journal of Phonetics* **1**, 309-327.

Hart, J. 't and Collier, R. (1975). Integrating different levels of intonation analysis. *Journal of Phonetics* **3**, 235-255.

Hart, J. 't, Nooteboom S.G., Vogten, L.L.M. and Willems, L.F. (1982). Manipulation of speech sounds. *Philips tech. Rev.* **40**, 134-145.

Kruyt, J.G. (1982). The effect of 'given' and 'new' information on (de-)accentuation in Dutch. S. Daalder and M. Gerritsen (eds.) *Linguistics in the Netherlands 1982*. Amsterdam: North-Holland. pp. 167-171.

Syntactic Dispositions as a Factor in Discourse Segmentation

Z. Palková
Prague, Czechoslovakia

1. Introduction

In the past we have voiced the hypothesis that the decisive unit in the sound-form of a continuous (read) text is, in Czech, what we called the discourse segment (Palková, 1963). It is a unit of the same order as the colon in classical metrics, and has been explicitly formulated for Czech in the works of Daneš (1957). Of the units used in the modern descriptions of other languages it probably comes closest to Trager and Smith's (1951) 'phonemic clause' as applied in Dittmann and Llewellyn's study (1967) or the 'tone unit' of Křeckel's works (1981). Our subsequent investigations have shown that the degree to which the grammatical and content properties of a text support or inhibit its division into discourse segments is what determines the ease with which a speaker reads or a hearer can perceive the text. We have described texts which conspicuously support segmentation as rhythmical. The experiments we have conducted enable us to assert that, for Czech, the major factors through which a text contributes to its own segmentation are its syntactic structure and the semantic dispositions for locating sentence stress (Palková, 1974).

Dispositions in favour of segmentation arising out of syntactic structure have so far been investigated from the point of view of the admissibility or otherwise of an intersegmental divide between individual syntagmas, under the assumption that the dimension of the excerpt favours segmentation (in Czech a length of five or six stress groups is already enough for the tendency towards segmentation to arise). Accordingly we looked at those tendencies which are linked to the 'quality' of the syntagmas verifying experimentally the types of close-knit word combinations that are unlikely to be split by an intersegmental divide (e.g. attributive adjective in agreement with noun - the *velký dům* (=big house) type).

2. Subject of Analysis

In this paper we shall be describing another way in which a text's syntactic structure affects segmentation. It stems from the speaker's need properly to maintain the hierarchy of segmental divides within a closed syntactic unit (clause or sentence).

This hierarchy comes about as a consequence of the confrontation between the linear progression of the spoken rendering and the not always linear construction of the sentence as a structured unit. Individual adjacent parts of the sentence need not always be equally strongly connected, so for example:

1. Z mohutného kopce/1 porostlého nízkými keři/2 je dobře vidět/3 i za špatného počasí/4 na blízké městečko (=lit.: From an impressive hill/ overgrown with low shrubs/is easy to see/ even in bad weather/ to the nearby township).

The potential segmentation in accordance with the basic tendencies towards phrasing in Czech is indicated by slashes. However, the ties between adjacent segments at 1 and 3 are stronger than those at 2 and 4. 1 and 3 separate segments between which there is a direct syntactic relation, so that if the divide happens to be omitted a 'meaningful' formation is left. The segments divided by 2 and 4 have no such direct syntactic connection, and taken together they give no meaningful interpretation.

This mutual imbalance between consecutive potential segmental divides often works as a conditioning factor in the actual phonic realisation of certain divides. In the example, it is mostly up to the speaker to segment at 1, but if he does so he must also segment at 2, otherwise the hearer will sense something wrong in the spoken rendering. The same holds for 3 and 4. From the phonic-stylistic point of view, with all four divides being made, this hierarchy ought ideally to be preserved; i.e. between loosely joined segments the divides should be more conspicuous than between more closely connected ones. In this way a competent speaker can facilitate his hearer's orientation in the text.

3. Discussion

This effect of sentence structure on segmentation influences a large part of a text. It is most striking that wherever the structure of the text permits two interpretations, segmentation in the spoken rendering resolves the ambiguity in one or other way. If the speaker does not appreciate this consequence of phrasing, the resultant mis-phrasing can give the wrong interpretation (a common case in complex texts in radio and television journalism).

The classical instance of this is the different possible grouping of a compound sentence, as in:

2. Znovu došlo k neshodě/1 mezi mým otcem/2 a jeho přítelem Janem/3 a mou nejmladší tetou (= lit.: again there was disagreement/ between my father/ and his friend John/ and my youngest aunt).

Only the relative proportions of divides 2 and 3 determine who disagreed with whom.

In Czech, textual ambiguity may also come from different interpretations of syntagmas, which is often supported by case homonymy:

3. Informovali jsme už/1 o výsledcích nových pokusů/2 našeho dlouholetého spolupracovníka X.Y (=lit.:We have already informed/ on the results of the recent experiments/ of our long-standing colleague X.Y.).

In Czech, the ambiguity arises from the two possible evaluations of NP (*our long-standing colleague X.Y.*) in the last segment:
a. it is a genitive and the syntagma is: the experiments of our colleague;
b. it is an accusative and the syntagma is: we have informed our colleague.

The hearer's selection of one or other variant depends on the strength of divide 2: if it is weaker than 1, we have the a) variant.

4. Syntactic framework

It is not easy to describe systematically what we have called the 'hierarchy of segmental divides' because of the sheer variety of real sentence structures and because syntactic structure does not operate in isolation. Nevertheless analysis of large amounts of material (from television news programmes) lead us to believe that for Czech the description of inter-segmental relations as endo- or exocentric may offer at least a partial syntactic motivation for this phenomenon (Lyons 1969); this characterisation would follow from the relation of the distribution of a given pair to their distribution separately.

The point of departure here is the a priori (hypothetical) segmentation of a text in accordance with the basic established phrasing tendencies in Czech. The resultant (hypothetical) segments usually represent noun or verb phrases, most often expanded, and are often long enough to constitute segments in a genuine rendering.

Determination of the endo- and exocentric relations between these (for us elementary) units enables us to assign to the subsequence of units within the sentence a structure which is part of the overall syntactic structure of the sentence, e.g.:

4. A(Informovali jsme už)B(o výsledćich nových pokusů)/C(našeho dlouholetého spolupracovníka)

which can be interpreted either as

5. $A \times (B \leftarrow C)$

or

6. $(A \leftarrow B) \times C$

where x and → denote exo- and endocentric relations respectively. Then in the phonic rendering the relationship between endo- and exocentric relations is represented by two degrees of divides, the exocentric being characterised by a stronger one than the endocentric. So in a real phonic rendering the ratio

7. Endo: Exo $= T_i : T_j,\ i < j$

ought to be maintained, where Endo, Exo represent endo- and exocentric relations between pairs of phrases, T is the strength of the respective intersegmental divide, and i,j denote the degree of strength; in principle i and j acquire the values $i=1$, $j=2$, or $i=0$, $j=1$. In the extreme case the difference may be eliminated ($i=j$), if the grammatical structure of the text does not admit two interpretations. The reverse ratio ($i>j$) should not arise or the hearer will sense that the rendering is wrong.

5. Conclusion

So far we have assumed a two-degree hierarchy of intersegmental divides, which is what is most commonly found in standard phonic texts, and we believe that this matches the aptitude of the average professional speaker and the perception of the average listener. How far this assumption of just a two-degree hierarchy is a generally valid constraint is the subject of further investigation.

References

Daneš, F. (1957). *Sentence Intonation in Present-Day Standard Czech.* Prague: Academia (in Czech).

Dittmann, A.T. and Llewellyn, L.G. (1967). The Phonemic Clause as a Unit of Speech Decoding. In: *Journal of Personality and Social Psychology,* **6**, No. 3, 341-349.

Lyons, J. (1969). *Introduction to Theoretical Linguistics.* CUP.

Kreckel, M. (1981). *Communicative Acts and Shared Knowledge in Natural Discourse.* London: Academic Press.

Palková, Z. (1963). The Question of Rhythm in Czech Prose. In: *Slovo a slovesnost* **24**, 221-230 (in Czech).

Palková, Z. (1971). On the Relation of Phonic and Syntactic Segmentation. In: *AUC-Philologica, Phonetica Pragensia II,* 19-29 (in Czech).

Palková, Z. (1974). *The Rhythmical Potential of Prose.* Prague: Academia (in Czech).

Trager, G.L. and Smith, H.L. (1951). *An Outline of English Structure.* Norman, Okla.: Battenberg Press.

The Effect of Accentuation on Comprehension: An Experiment

J.M.B. Terken
Eindhoven, the Netherlands

1. Introduction

Psycholinguistic investigations (Cutler, 1976; Cutler and Foss, 1977) have proposed that accents direct the listener's attention to the words carrying the accents. This would make sense if, as linguistic investigations have proposed, accents mark words which convey important information in a given context, and deaccentuation (that is, not accenting a word that may be accented for syntactic reasons) marks information that is recoverable from the context (Halliday, 1967; Chafe, 1974). Accentuation and deaccentuation might signal to the listener how to distribute processing efforts in an efficient way; for instance, deaccentuation might tell him not to spend much effort on doing a precise analysis of the speech signal, since the interpretation is already available on the basis of the preceding context (Nooteboom and Terken, 1982; Terken, in preparation).

If the distinction between presence and absence of accent is a functional one, we should be able to measure the extent to which the presence or absence of accents affects the listener's comprehension of the incoming information. More specifically, we may ask what the effect of appropriate application of accentuation rules is, that is, whether the listener's comprehension is seriously disrupted if a word which should be accented is not accented or vice versa. For the clearly structured material used in the present experiment we will use the following straightforward rule: if an expression has been mentioned in the preceding utterance in the same syntactic function, it need not be accented. Deviation from this rule defines 'inappropriate' accentuation.

2. Method

To investigate these questions we have chosen a verification task: a listener is watching a screen displaying a simple configuration of alphabetic characters. Subsequently, he hears a description of some aspect of the configuration, and he is asked to decide as quickly as possible whether the description is right or wrong. Decision time is taken as an index of comprehension time.

To give an example: we present the letter configuration shown in figure 1. After a warning signal the *p* moves to the top of the *k*. Subsequently, we present the description 'the p moves to the top of the k'. The listener should decide whether the description is right or wrong. Next the *p* moves to the left of the *q*. We present the description 'the p moves to the left of the q'. In the second description, the 'p' should be deaccented because it has been mentioned in the preceding utterance. If it were accented, it would be labelled 'inappropriately accented'. On the other hand, the 'q' has not been mentioned in the preceding utterance and should therefore be accented. If it had not been accented, it would be labelled 'inappropriately deaccented'. The successive descriptions are interrelated, so that we can define the appropriateness of accentuation with reference to the preceding context.

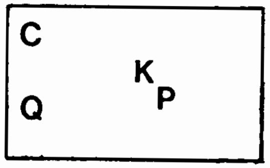


Figure 1. Example of initial configuration of alphabetic characters.

By manipulating the presence or absence of an accent on a specific word, and by measuring the decision time from the word onwards, we get an indication of the effect of accentuation and appropriateness of accentuation on comprehension time.

The presence or absence of accents and 'appropriateness' is manipulated independently for subject nouns and predicate nouns. For each condition, ten sentences are presented to eleven subjects, and comprehension times measured.

3. Results and Discussion

For subject nouns we find the following data (we do not consider the data for 'true' subject targets, since if the subject is described correctly, the listener has to wait for the remainder of the sentence before he can make a decision, while we are interested in the immediate effects of accentuation. So the data presented are only for 'false' subject targets. Means are in ms. Arrows between means in the matrices indicate that the means differ significantly in a sign test on comparisons of individual measurements. Each mean is based on at most 110 measurements).

$$\begin{tikzcd} & -\text{Accent} & +\text{Accent} \\ \text{Appropriate} & 606 & 584 \arrow[dl, <->] \arrow[d, dash] \\ \text{Inappropriate} & 624 & 613 \end{tikzcd}$$

We see that the presence of an accent leads to faster decisions, and that both for plus and minus accent the appropriate application of the rule leads to faster decisions than the inappropriate rule application.

For predicate nouns we have two sets of data: for true descriptions and for false descriptions. Let us first consider the data for false descriptions.

	–Accent	+Accent
Appropriate	435	389
Inappropriate	453	427

Again, we see that the presence of an accent leads to faster decisions than the absence of an accent, and that appropriate rule application leads to faster decisions than the inappropriate application of the rule. The data for 'true' predicate targets are the following.

p = .06

	–Accent	+Accent
Appropriate	266	294
Inappropriate	407	323

Again, we see that inappropriate rule application leads to longer decision times. However, for true descriptions the presence of an accent leads to faster decisions only when the rule is applied incorrectly. When the rule is applied correctly the absence of an accent leads to faster decisions.

4. Conclusion

We see that appropriate application of the rule in all cases leads to faster decisions, and that the presence of an accent leads to faster decisions except when the description is true and the word may be appropriately deaccented, that is, when there is no conflict between the description and the word described, and there is also no conflict between the lexical information and the accentual information. Thus, the present results suggest that the 'appropriateness' of deaccentuation should not only be defined with respect to the preceding linguistic context, but also with respect to the domain of reference: we have found that accentuation speeds up decision time when there is a conflict between the information conveyed and the actual state of affairs. This would make sense if accentuation signals to the listener that he cannot easily supply the intended interpretation by himself, so he should give proper attention to the analysis of the speech signal, and if deaccentuation conversely signals that the listener has the intended interpretation already available on the basis of the preceding context, so that he need not spend too much effort on analysing the speech signal. This new definition of 'appropriateness' is being tested in current experiments.

5. Applications

In the design of rule systems for automatic accent assignment the present results suggest the following. Words which may be accented for syntactic reasons (e.g. nouns) should be marked by an accent, except in cases where there is no doubt concerning two matters:

1. deaccentuation is appropriate with respect to the preceding context;
2. what is said is a true description of the state of affairs.

This strategy will probably lead to the presence of an accent in a number of cases where speakers would apply deaccentuation. We have no evidence yet about the possible cumulation of negative effects of such a conservative strategy in coherent discourse.

Acknowledgements

This research was funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.).

References

Chafe, W.L. (1974). Language and consciousness. *Language* **50**, 111-133.

Cutler, A. (1976). Phoneme monitoring reaction time as a function of preceding intonation contour. *Perception and Psychophysics* **20**, 55-60.

Cutler, A. and Foss, D.J. (1977). On the role of sentence stress in sentence processing. *Language and Speech* **20**, 1-10.

Halliday, M.A.K. (1967). Notes on contrastivity and theme II. *Journ. of Linguistics* **3**, 199-244.

Nooteboom, S.G. and Terken, J.M.B. (1982). What makes speakers omit pitch accents? An experiment. *Phonetica* **39**, 317-336.

Terken, J.M.B. The distribution of accents in coherent monologues. In preparation.

Section 13
Pitch and Intonation

Perception and Interpretation of Non-Native Intonation Patterns

M. Cruz-Ferreira
Lisboa, Portugal

1. Introduction

Most of the few studies on non-native use of intonation are *production* studies (Anderson, 1979, MacNaught, 1978), analyzing overt 'foreign accents' but failing to reveal the less obvious though no less impairing interference of defective *comprehension* of intonation (Berkovits, 1980; Scuffil, 1982).

This paper outlines an experiment designed to assess non-native comprehension of intonation in (European) Portuguese and (British) English, using as informants thirty native speakers of each language who could speak the other (Cruz-Ferreira, 1983). Material was, for each language, a set of thirty minimally paired sentences, differing in intonation only, or presenting the same intonation twice ('filler' pairs) spoken on tape by a native speaker of each language (the intonational framework used is based on Halliday, 1967 and O'Connor and Arnold, 1973). Each pair conveys meaning distinctions which may or may not be rendered by the same intonational means in L2 (or by intonation at all). The task was a closed-choice binary decision on two levels: i) *perception*: decide whether the sentences in each pair have the 'same' or 'different' meaning, and ii) *interpretation*: match one of the two meaning glosses provided for each pair with one or both of the sentences, according to the previous decisions. This layout enabled conclusions regarding non-native ability to *discriminate* the patterns (by perceiving them as the same or as different); to *interpret* the patterns (through the meaning assigned to each or to both); and to *identify* the patterns (i.e. to recognize them as meaningful in the filler pairs).

Each native group was asked to provide answers for their native tongues too, and the results were used as control values for statistical treatment of the non-native data ($p \leq .05$). Some very general characteristics of Portuguese intonation are (as relevant in explaining the results presented below): there are no falling-rising intonations in Portuguese, but there is a very low starting falling tone, the 'low low-fall'; the nucleus has a fixed final position within the intonation group.

2. Analysis of Some Results

Some typical non-native replies.

The non-natives had no difficulties

1. *where the same meaning contrast is conveyed by the same intonational device in L_2 and in L_1.*

1. *Same meaning conveyed by same intonational device in* L_1 *and* L_2 *object v apposed subject*

Ptg	_ela co'meu a galinha	*int. dev. tonality*
	ela comeu/a galinha	*n–n reply.* 1 IG: NP_2 = object 2 IG: NP_2 = apposed subject
Eng.	they've'left the children	
	they've left/the children	

The interpretation of F--RF as conveying 'neutral' vs. 'impressed' statements in both languages was also correct for both groups of listeners.

2. *where the meaning contrast in L_2 may be associated with a typical use of high versus low nuclear pitch in L_1.*

2. *Meaning contrast in* L_2 *may be associated with a typical use of 'high' versus 'low' nuclear pitch in* L_1
command v warning

Ptg.	não ↘bebas dessa água	*int. dev. tone* Low LF~LF
	não 'bebas dessa agua	*n–n reply.* Low LF = command LF = warning
Eng.	|don't leave the `house	*int. dev. tone* F~FR
	↘don't leave the ᵛhouse	*n–n reply.* F = command FR = warning

English F-~FR, conveying 'neutral' vs. 'reserved', statements, was also correctly interpreted by the Portuguese listeners. Even where one of the contrasted patterns does not exist in L_1, the non-natives seem to be sensitive to differences in the gross phonetic shape of the patterns, and to associate overall higher pitch with a more 'open' meaning, and lower pitch with a more 'final' overtone (see Cruttenden 1981).

The non-natives had difficulties

3. *with idiosyncratic uses of intonation in L_2.*
These fall into three classes:

3.1. *where the meaning conveyed by intonation in L_2 has no intonational equivalent in L_1.*

3. *Idiosyncratic uses of intonation in each language*

3.1. *Meaning conveyed by intonation in* L_2 *has no intonational equivalent in* L_1
'not... `any' 'not...ᵛany'

Eng.	they ´|don't admit `any students	*int. dev. tone* F = 'none at all'
	they ↘don't admit ᵛany students	FR = 'only some' *n–n reply. perception*: random *interpretation*: random

The distinction between *ànу* and *ˇany* is made through lexical means in Portuguese. In a filler pair presenting the *ˇany* pattern twice, the Portuguese perceived the patterns as the 'same' but assigned to both the 'none at all' meaning of a straightforward negative.

On the other hand, in a Portuguese pair presenting constant polarity tags with LR and LF, which convey the same 'patronizing' overtone, the English perceived the tones as 'different' and gave random replies to their interpretation.

3.2. *where the meaning contrast is conveyed by intonation in both languages, but the device used in L_2 is non-existent in L_1.*

3.2. *Meaning is conveyed by intonation in both languages, but* L_2 *device is non-existent in* L_1
positive vs negative presupposition

Eng.	I ᵛthought she was married	*int. dev. tonicity* V = pos
	I 'thought she was `married	obj = neg *n–n reply. perception*: random *interpretation: random*

3.3. *where the intonational devices exist in both languages, but are put to different uses.*

3. *Idiosyncratic uses of intonation in each language*

3.3. *Intonational devices exist in both languages, but are put to different uses positive v negative presupposition*

Ptg	¯eu sempre achei que ele vinha	*int. dev. tone* LF = pos
	¯eu sempre achei que ele `vinha	HF = neg *n–n reply. perception*: 'different' *interpretation*: random

In these three cases, the contrastive meanings dependent on various aspects of intonation are idiosyncratic to L_2. The use of intonation in L_2 cannot therefore be deduced from uses of intonation in L_1.

4. *with items or syntactic structures usually associated with one particular meãning.*

4. *Syntactic structure or item usually associated with one meaning*
question vs exclamation

Eng. ˈdidn't John enˋjoy it — *int. dev. tone* F = exclamation
ˈdidn't John enˏjoy it — LR = question
n-n reply. perception: random
interpretation: random

In English, inversion structures are commonly associated with an interrogative meaning. Random results were also obtained by the English listeners to a Portuguese pair presenting two alternative interpretations of a word (*'como'*), usually associated with an interrogative meaning too.

3. Conclusions

On the whole, the hesitation of the non-native in assigning meaning to the intonation patterns of L_2 was clear. Most patterns were either misunderstood through the assigment of a wrong meaning to them –, or not understood at all – as proved by the large number of random replies. These results argue for the arbitrariness of certain intonational meanings. But the results also show that the non-natives do not always hesitate when the foreign pattern is non-existent or has no equivalent use in their mother tongue. They apparently do not have the exact native patterns in mind and, in some cases, seem to make use of some 'internalized' abstract functions of intonation. These results show that some features of intonation and associated intonational meanings seem to lie beyond particular uses in each language. It seems possible to set up some interpretive *strategies* that non-natives use in listening for intonational meaning in L_2:

- *the transfer strategy*: where the same intonational difference operates on a familiar structure in L_1, the meaning conveyed by intonation in L_1 is generalized to L_2 (see 1);
- *the pitch height strategy*: where the meaning contrast conveyed by intonation in L_2 can be associated with broadly similar uses of pitch contours or pitch heigth in L_1, 'abstract' generalizations regarding meaning seem to be made, correlating higher overall pitch with 'openness' and lower pitch with 'finality' (see 2);
- *the lexico-syntactic strategy*: where certain lexical items or grammatical patterns of L_2 are most commonly associated with one particular meaning, this straigthforward interpretation suggested by the words will tend to override alternative interpretations which are intonation-dependent (see 4); if the patterns are perceived as the 'same' (viz. in the filler pairs), this strategy results in the assignment to both patterns of the meaning favouring a more straightforward interpretation of the words (see comment for the pair presenting the ˇ*any* pattern twice, in 3.1).

Acknowledgment

I am indebted to the Gulbenkian foundation in Portugal, for financially supporting my attendance to this Congress.

References

Anderson, K.O. (1970). *Some aspects of English language interference in learning German intonation.* Unpublished PhD thesis, University of Colorado.

Berkovits, R. (1980). Perception of intonation in native and non-native speakers of English. *Language and Speech*, **23**(3): 271-280.

Cruttenden, A. (1981). Falls and rises: meanings and universals. *Journal of Linguistics* **17**(1): 77-91.

Cruz-Ferreira, M. (1983). *Non-native comprehension of intonation patterns in Portuguese and in English.* PhD Thesis, University of Manchester.

Halliday, M.A.K. (1967). *Intonation and grammar in British English.* The Hague: Mouton.

McNaught, J. (1978). *The prosodic competence of a sample of French speakers of English.* Unpublished M.A. Diss., University of Manchester.

O'Connor, J.D. and Arnold, G.F. (1973). *Intonation of colloquial English.* London: Longman.

Scuffil, M. (1982). *Experiments in comparative intonation. A case study of English and German.* Tübingen: Niemeyer.

Experiments on the Stylization of British English Intonation

J. 't Hart and J.R. de Pijper
Eindhoven, the Netherlands

1. Introduction

On earlier occasions we have reported on a method applied in intonation analysis which we call the stylization method. Straight line segments are used to construct approximations to F_0 curves in such a way that these are simplified as much as possible, but nevertheless give a melodic impression similar to that of the originals. After stylization it is far less difficult to make an inventory of the melodic possibilities of the intonation system of a language, as was demonstrated for Dutch ('t Hart and Cohen, 1973; 't Hart and Collier, 1975). Since this approach boils down to analysis by synthesis, an additional advantage is that it yields reliable recipes for the control of F_0 in speech synthesis, by virtue of its continuous testing of the perceptual adequacy in the process of the analysis.

This study examines whether the stylization method is also applicable to the intonation of British English (henceforth BE).

2. Close-copy stylizations

The first question to be answered for BE intonation is whether such a straight line approach is feasible at all. Instead of aiming at fully standardized contours in one step, we introduced an intermediate step of constructing close-copy stylizations. These are defined as composed of the smallest number of straight line segments with which perceptual equality can be obtained.

Close-copy stylizations are made following a trial and error method with the experimenter as sole judge. The alleged perceptual equality is, of course, tested afterwards in an appropriate listening experiment.

In such an experiment, 64 native English subjects were presented with 60 pairs of sentences, varying in duration between one and three seconds. Each sentence was processed by means of an LPC analysis-resynthesis system in order to obtain three versions: a resynthesized original, with the original F_0; a close-copy stylization; an 'alternative' contour, made in such a way that the differences with the originals would be easily audible. Sentence pairs came in three categories: (A) a resynthesized original and a close-copy stylization; (B) two identical versions; (C) a resynthesized original or a close-copy stylization and an alternative contour. Subjects were asked to indicate for each pair

whether they thought its members to be exactly equal or not. In the instruction they were acquainted with the notion of close-copy stylizations, not only verbally, but with demonstrations on tape as well. Table I summarizes the results of the test.

Category A shows a very high number of responses 'equal'. One reason for this could have been that, since the pairs of category C differed so much, subjects' criteria might have been biased in favour of judging the pairs of category A to be equal. However, a post hoc analysis of the data did not substantiate this possibility. Moreover, the scores show that the number of 'equal' is lower with A than with B. This can be attributed to a number of cases in which some listeners were able to hear differences so small that they had not been observed by the experimenter. The listeners would therefore certainly have scored a higher number 'different' for category A if the objective differences would have been only slightly bigger than they actually were. The conclusion is that it is possible to stylize BE intonation on by means of a restricted number of straight line segments, and nevertheless to maintain perceptual equality.

3. Standard stylizations

The ultimate aim of the stylization method is to replace the movements in close-copy stylizations by pitch movements with standard specifications of slope, duration and position in the syllable. The purpose of such a standardization is to facilitate the finding of correspondences between otherwise different contours, and thus reveal the melodic structure of the various patterns of the intonation system of the language. Such standard recipes have been developed for a limited variety of different intonation patterns. As examples of these patterns, two representative items were taken from each of the seven Primary Tones as recorded on the tape which goes with Halliday's course (1970). These are fully acceptable samples of BE intonation, as has been confirmed in the experiment to be discussed below.

Standard stylizations are generally not indistinguishable from their corresponding originals. But if we want to claim that standard stylizations are still representative of normal BE intonation, they should at least sound as acceptable as resynthesized originals.

Table I. Total number of responses, and number of cases in which members of a pair were judged equal or different. Objectively, the members in a pair were different in categories A and C, and equal in category B

Category	No. of responses	No. 'equal'	No. 'different'
A	1280	1111	169
B	1280	1222	58
C	1280	75	1205

This claim was tested in an acceptability experiment. Of each of the fourteen chosen utterances, five versions were made, only differing in intonation. The five versions were: resynthesized originals (ORIG), fully standardized stylizations (FS), partly standardized stylizations (PS), DUTCH-versions and WITTEN-versions.

In PS-versions not all parameters that characterize pitch movements are fixed, but two of them are allowed to vary freely, viz. overall position of F_0 and overall size of the excursions. DUTCH-versions were constructed following the rules developed for Dutch intonation. Care was taken to select Dutch patterns that resembled the given BE patterns as much as possible. WITTEN-versions were constructed following the recipes for the synthesis of Halliday's seven Primary Tones, as published by Witten (1977).

DUTCH-versions were incorporated with the expectation that, when recipes for another language are used to build the contours, the outcome will be less acceptable. Witten based himself on Halliday's impressionistic descriptions of the Tones. In view of the notorious unreliability of the unaided ear in intonational matters, these versions were also expected to be less acceptable than the ones based on instrumental analysis. Sixty-six native speakers of BE were asked to judge the acceptability of each of the test items on a five point scale. By means of the method of successive intervals (Edwards, 1957) a psychological continuum was derived from the raw data, and each of the stimuli assigned a scale value which refers to that continuum. See Table II.

Unlike the raw data, the scale values lend themselves to a statistical analysis. Such an analysis revealed that ORIG, FS and PS do not differ significantly in acceptability, whereas DUTCH and WITTEN do differ from each of these, and from each other. We may therefore conclude that the FS- and PS-versions sufficiently agree with the internal representations the listeners have of the intonation patterns of their language. But at the same time, the outcome for DUTCH- and WITTEN-versions shows that the recipes used for the standardized versions are necessary to agree with the listeners' internal representations: as soon as other recipes are applied, the resulting contours sound less acceptable. Thus, the attentive subjects are sensitive to any violation of the rules of their intonation system. However, their critical behaviour did not keep them from approving of the FS- (and PS-) versions. These versions can therefore be considered as fully representative of some patterns of the BE intonation system.

Table II. Scale values, averaged over the 14 stimuli, and standard deviations, for each of the five conditions used in the acceptability test

	ORIG	FS	PS	DUTCH	WITTEN
Mean	2.676	2.712	2.476	1.839	.654
s.d.	.438	.384	.541	.625	.454

Acknowledgements

This research was funded by the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.) under grant no. 17-21-07.

References

Edward, A.L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century Crofts.

Halliday, M.A.K. (1970). *A course in spoken English: intonation*. Oxford: Oxford University Press.

't Hart, J. and Cohen, A. (1973). Intonation by rule: a perceptual quest. *J. of Phonetics* **1**, 309-327.

't Hart, J. and Collier, R. (1975). Integrating different levels of intonation analysis. *J. of Phonetics* **3**, 235-255.

Witten, I.H. (1977). A flexible scheme for assigning timing and pitch to synthetic speech. *Language and Speech* **20**, 240-260.

Gradations in Pitch Accents?

T. Rietveld
Nijmegen, the Netherlands

1. Introduction

Much of the work on Dutch intonation is based on the framework developed by Cohen and 't Hart (1967), and Collier and 't Hart (1978). This framework of concepts and terms is presented in a diagram which was taken from a course in Dutch intonation by Collier and 't Hart (1978), see Figure 1.

The diagram presented in Figure 1 distinguishes only one type of pitch accent. This pitch accent is regarded as conspicuous whereas all other accents, realized by duration, are regarded as less conspicuous. If one wants to use the terms 'primary stress' and 'secondary stress' for these types of sentence stress, the latter has to be assigned to the duration accent.

In our work on the automatic detection of prominence in Dutch (Rietveld and Boves, 1979), we have come across several samples of pitch movements with characteristics typical for prominence-lending movements, which did not lead subjects to score 'sentence stress'. These pitch movements very often occurred on that syllable in polysyllabic words which carries word stress. The task the subjects had to perform was an (0,1)-task; this means they only had to mark the stressed syllables in the sentence '1'. The prosodic phenomenon in question cannot be classified as micro intonation since the observed F_0-movements show fairly large excursions and since their occurrence is not determined by phonetic factors but by a grammatical one, i.e. word stress in polysyllabic words.

The questions we want to discuss should be seen against the background of the construction of a prominence (= sentence stress) detector. This detector has to cope with F_0-movements which - apart from the range - are quite similar to the movements which give rise to sentence stress, but are not prominence-lending themselves.

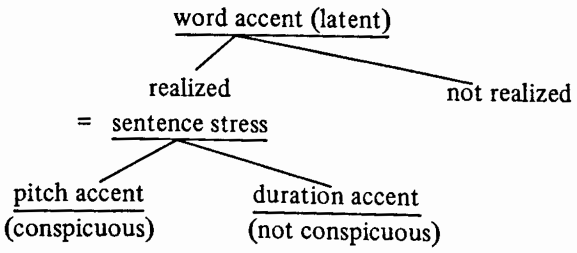


Figure 1. Framework of concepts from Collier and 't Hart (1978).

The relevant questions are:

a. To what extent can F_0-movements which are similar to but not identical with the prominence-lending movements, be expected on syllables carrying word stress but not sentence stress?
b. In how far does the F_0-range of these movements overlap with that of the prominence-lending movements?
c. Are the syllables mentioned under a) perceived as carrying some kind of stress?

The first two questions are relevant for the construction of the detector, the last one pertains to its evaluation: 'false alarms' on syllables which carry some kind of 'secondary stress' are less serious than false alarms on other syllables.

Experiment 1 and 2 are carried out in order to answer the first two questions. In the first experiment we gave explicit instructions to the speakers to stress only specific syllables; in the second experiment speakers were free to realize the speech material in their own way. This time, however, we asked them to give stress judgments of their own speech, thus we could say with greater certainty which syllables were meant to take sentence stress and which not. The third and fourth experiments aim at answering the question whether an extra response category ('Half stressed') is a useful and reliable one and can be related to the occurrence of pitch movements which do occur on syllables with word stress.

2. Experiment 1

Procedure

In this experiment three male speakers had to realize one carrier sentence with eighteen different polysyllabic words (target words).

The carrier sentence was:

'Je moet *niet* over de .... gaan *klagen*, maar er *wel* over *na*denken'.
(You should not complain about...., but think about it)

The speakers were asked to stress the italic syllables only. The blank (....) had to be completed with a three- or moresyllabic compound ('target word').

Six of these had the stress pattern 1 3 2:
'fruittelersbond, suikerbietenoogst, autobusdienst,
schoonmaakploeg, vuurwapenwet, veevoederprijs'

six had the stress pattern 1 2 3:
'zakwoordenboek, kinderspeelplaats, scheepstimmerlui,
legerwoordvoerder, tandheelkunde, filmtijdschrift'

Six had the stress pattern 2 1 3:
'schooladviesdienst, boerenleenbank, stadsschouwburg,
staatsbosbeheer, rijkswaterstaat, burgemeestershuis.'

2.1.Analysis

Pitch and intensity were measured with the IFN-pitchmeter (Van Rossum and Boves, 1978) and the IFN-intensitymeter, respectively. The analog output was fed into the computer with a sample frequency of 100 Hz; subsequently the pitch curves were plotted in semitones together with the lower declination line, according to a procedure borrowed from 't Hart (1979), but adapted by us.

The measurement procedure was as follows: (a) If a target word contained an F_0-rise which started early in the syllable or a fall which started late and if the F_0-range exceeded 1 semitone, the corresponding pitch jump was measured. We are quite aware of the fact that only an average of more than 1 semitone can be seen as a meaningful result. (b) Pitch jumps were also measured on the syllables which subjects had been instructed to stress.

2.2. Results

We observed pitch movements to or from the lower declination line on all target words. In most cases these pitch movements occurred on the syllable carrying the main word stress.

The following pitch jumps were observed:

Target words; mean F_0-excursion: 4.04 semitones, s.d.: 1.65 semitones ($n = 54$)

Stressed syllables in other words; mean F_0-excursion: 6.34 semitones, s.d.: 1.78 semitones ($n = 208$).

The difference in pitch jumps between the two types of syllables was significant : $t = 8.56$, $p < 0.01$, $df = 260$.

Thus, we found evidence that word stress which is not intended to be sentence stress is accompanied by pitch movements in many cases. The fact that the excursion of these movements is smaller than the one for sentence stress, can be regarded as an indication that 'secondary pitch accent' may be a useful category in Dutch prosody.

3. Experiment 2

This experiment was carried out to answer the same questions as in experiment 1. This time, however, the speakers were not instructed to stress any particular syllable. Furthermore, the speakers had to judge their own material twice: once immediately after the realization of the complete set of utterances, and once in a normal listening test which took place a week later. In this way we were able to determine which syllables were meant as stressed by the speakers themselves. In addition a panel of ten listeners performed the listening test too. In all cases the test was an (0.1)- task: all syllables had to be categorized as either 'stressed' or 'unstressed'.

3.1. Speech material

Nine sentences read out by five speakers. In this material three samples from each of the three categories of polysyllabic words mentioned before ('target words') were included (see experiment 1).

The speakers were asked to imagine a situation for each utterance so that they could be used as 'naturally as possible'.

3.2. Analysis

The pitch jumps were determined in the same way as described in experiment 1.

We determined:

a. the jumps on the target words
b. the jumps on the syllables which were labeled 'stressed' by the majority of the listeners.

3.3. Results

1. None of the target words were scored as stressed by the speakers when they had to judge their utterances straight away.
2. In the second task (the listening task) four of the five speakers did not change their mind; only one speaker now scored eight of the nine target words as 'stressed'.
3. Fourteen of the 45 target words (= 5 × 9) were judged stressed by the panel of ten listeners.
4. In 42 of the 45 target words clear pitch movements were found on the syllable carrying word stress.
5. The pitch jumps on the target words (mean: 4.87 semitones, s.d.: 2.04 semitones) were significantly smaller than those on the syllables in non-target words which had been judged stressed by the majority of the panel (mean: 8,45 semitones, s.d.: 1.98 st): $t = 1.72$, $p < 0.05$, $df = 156$.
6. A significant, though not a strong correlation was found between the excursion of the pitch jumps on the target words and the number of stress scores given by the panel: Pearson's r: .65, $p < 0.01$ ($n = 44$).

4. Experiment 3

In the third experiment we presented the same speech material as in experiment 2. This time, however, listeners had the possibility to use three response-categories in their stress-judgments: 'full-stress' 'half-stress' and 'no stress': a (0,1/2,1)-task.

In experiment 1 we observed noticeable F_0-movements on the syllable carrying word stress in polysyllabic words. These syllables were not judged 'stressed' in an experiment with two response-categories: no stress and stress. In the present experiment with three response-categories we expected the majority of the syllables involved to be judged as 'half-stressed'.

4.1. Results

Thirty out of 45 syllables carrying word stress (= 67%) were judged as 'half-stressed' by the majority of the listeners. This result implies that the dichotomy of pitch accents (= sentence accents) and other accents as introduced by Cohen and 't Hart is not confirmed by our data. The presence of pitch movements on syllables which do not carry sentence stress is related to a stress impression of a certain kind: 'half-stressed'. Experiments with resynthesized speech and manipulated F_0-movements may show to what extent this relationship is a causal one.

5. Experiment 4

A further indication of the reality of a 'secondary accent' may be deduced from the interrater reliability of scores involving this response category.

To this effect we presented a text consisting of 10 sentences with 221 syllables to two panels of 12 listeners each.

One panel was given a (0,1)-task, the other a (0,1/2,1)-task. We did not use all the scores gathered in this experiment since the method we applied to determine the reliability of the scores is based on the assumption of independence of scores: Ebel's reliability coefficient by analysis of variance. Stress scores, however, are not independent; for that reason we formed subsets of all available scores by quasi-random sampling.

The reliabilities we obtained in both tasks for 'full stress' were high: 0.92; the reliability for 'half-stress' was lower, but still satisfactory: 0.80.

From the latter value we can deduce that the 'secondary-accent' is indeed a real perceptual category in Dutch.

6. Conclusion

In the measurements reported here, clear pitch movements were found on the syllables carrying word stress in words of three or more syllables. The great majority of these syllables were not scored as 'stressed' by the speakers and/or listeners. The range of these pitch movements was smaller than the range on other syllables which had to be pronounced as 'stressed' (exp. 1) or were judged 'stressed' (exp. 2). In another experiment (exp. 3) we found that the majority of the syllables carrying word stress in words of three or more syllables were judged 'half-stressed' in a task which included three possible responses: 'unstressed', 'half-' and 'full stressed'. F_0-measurements (exp. 2) showed considerable pitch movements on the syllables carrying only word stress.

Our data do not confirm, therefore, the suggestion of the diagram by Collier and 't Hart that accents which are not 'conspicuous' do not carry noticeable pitch movements and are mainly durations accents.

References

Cohen, A. and Hart, J. 't. (1967). On the anatomy of intonation. *Lingua,* **19**, 177-192.

Collier, R. and Hart, J. 't. (1978). Cursus Nederlandse intonatie. *Brochurereeks Dona.* Diepenbeek Wetenschappelijk Onderwijs Limburg.

Hart, J. 't. (1979). Explorations in automatic stylization of F_0 curves. *IPO Annual Progress Report,* **14**, 61-65.

Rietveld, A.C.M. and Boves, L. (1979). Automatic detection of prominence in the Dutch language. *Proceedings Institute of Phonetics Nijmegen,* **3**, 72-78.

Rossum, N. v. and Boves, L. (1978). An analog pitch-period extractor. *Proceedings Institute of Phonetics Nijmegen,* **2**, 1-17.

An Analysis of Regional Variation in English Intonation

T.I. Shevchenko
Moscow, USSR

1. Introduction

English dialectology which is traditionally deeply rooted in the history of the English language has accumulated ample material on sound change in regional varieties of English. The specific regional features which manifest themselves at the suprasegmental level, however, on account of their complexity have been left, as a rule, beyond the systematic research of both phoneticians and dialectologists. The fact that mention was made of the phenomena in question by Ida Ward (1948) Roger Kingdon (1958) and Kenneth Pike (1965) is but a proof of their objective reality.

In the context of what has been said above, the papers on regional intonation, in which is it viewed either in the information aspect (Brown et al. 1980) or sociolinguistically (Trudghill, ed. 1978) are of great interest. After David Crystal who emphasized the necessity to correlate social factors with prosodic and paralinguistic features in speech analysis (Crystal 1975), John Pellowe and Val Jones (Trudgill, ed. 1978: 101-121) as well as Gerry Knowles (Ibid: 80-90) have made succesful attempts at establishing such correlation in their study of Tyneside speech and Scouse, respectively. The results highlighted a few aspects relevant for our comparative intonation analysis of different regional varieties of English.

Our experiment is concerned with intonation peculiarities of Midland, Northern, Scottish, Irish and Welsh types of English as compared with the intonation of Southern standard English (RP). The material, 12 readings of an identical text containing 318 words was in part recorded at the Phonetic Laboratories of Edinborough University and the the Experimental Phonetics Laboratory of Moscow Foreign Languages Institute by English students, aged 18 to 22, male and female, with various regional backgrounds, two speakers for each type. The corpus was subjected to auditory and acoustic analyses.

2. Results

The data obtained in the course of the analysis give evidence to the following:

1. Besides the nuclear tones which are generally considered socially and

regionally significant, configurations of the melodic contour and its pitch range characteristics, the inventory of heads, their variants and relative frequency prove to be relevant too. Thus, for instance, narrowing the pitch range is a typical regional modification of the contour, when compared with a similar realization in RP; another common feature is the descending character of the head which prevails in all regional types of English. Certain specifications, however, ought to be made here: In Scottish English, for instance, the descending head is mainly realized in the medial part of the range and is shaped wavy-like; in contrast with it high level and low level heads are more frequent in Midland English.

The importance of the inventory of heads employed by the speaker can thus be demonstrated: the informants from London and Edinburgh make use of 4-5 types of heads, while the Irish speakers limit the amount to three; the Hebrides Islands speaker makes it one.

2. Nuclear tones, which have always been the focus of attention in English intonology, reveal their regionally distinctive relevance in the inventory, distribution, choice of the structural variant and functional load. E.g. most of the regional speakers tend to use more level tones (compared with RP) in non-final tonegroups of statements. Hence, the level tones are equivalent to the rising tone in expressing non-finality. Alongside with this general observation the use of the level tone in special questions (the Irish speaker) may be treated as a specific feature.

It is common knowledge that falling tones are most frequent in the English language. But their configuration may be modified, levelled out and not reaching the lowest pitch level at the end. Besides that, an initial rise is commonly registered at the beginning of the tone in Scottish and Northern English. This variant is identified by phoneticians who have a long experience in listening to RP, as a risefalling tone. Functionally, however, it is quite devoid of the connotations described for the RP rise-fall (O'Connor and Arnold, 1961). It is deemed necessary, therefore, to qualify the tone in question as a regionally marked structural variant of the falling tone. Thus we can conclude that phonetic variability, however great, may not be functionally supported: the process of communication, with the exception of the connotations mentioned above, appears to be in no other way impaired by these modifications.

3. Common to all the readings which represent different regional varieties of English are similarity of accentuation, the division into tonegroups and the location of pauses. The length of pauses and the rate of delivery, it may be assumed, is determined, to a great extent, by sex and individual variability. One cannot, however, make generalizations about the data from the limited amount of the material available.

4. The fact that in a number of national varieties of English, such as General

American and General Australian, for instance, similar tendencies have been observed, among them a narrow range of pitch, a mid-level wavy head and a rising- falling-to-mid terminal tone (Shakhbagova 1982) gives us grounds to believe that the latter constitute inherent English features, part of the prosodic system of English.

References

Brown, G., Currie, K.L. and Kentworthy, J. (1980). *Questions of Intonation.* Croom Helm.

Crystal, D. (1975). Prosodic and Paralinguistic Correlates of Social Categories. In: Ardener E. (ed.), *Social Anthropology and Language.* London: Tavistock Publications, 185-205.

Kingdon, R. (1958). *The Groundwork of English Intonation.*. London.

Knowles, G. (1978). *The Nature of Phonological Variables in Scouse.* In: Trudgill, P., ed. (1978), 80-90.

O'Connor, J.D. and Arnold, G.F. (1961). *Intonation of Colloquial English.* London.

Pellowe, J. and Jones, V. (1978). *On Intonational Variability in Tyneside speech.* In: Trudgill, P., ed. (1978). 101-121.

Pike, K. (1965). *The Intonation of American English.* Ann Arbor.

Shakhbagova, D.A. (1982). *Varieties of English Pronunciation.* Moscow.

Trudgill, P. (1978). *Sociolinguistic Patterns in British English.* London: Arnold.

Ward, I. (1948). *The Phonetics of English.* Heffer.

Surface Tones in Chaga: Towards a Tonetic Classification

C.W. Temu
Dar es Salaam, Tanzania

1. Introduction

The Vunjo dialect of the Chaga (Bantu) language is spoken by approximately two hundred thousand Wachaga people on the slopes of Mount Kilimanjaro in Tanzania. A prevalent or standard form does not exist in the Chaga language and it is not actively written except for a few religious hymn books. Occasionally friends write to one another in the dialects but tone is not marked so that the dialects exist side by side influencing one another and being much influenced by Swahili vocabulary and syntax.

1.1. Stage of Research on the Chaga language

The first published work (Raum, 1964) described the Moshi dialect but contains very little on tones.

1.2. Corpus

The material analysed for this paper is not based on a strictly delimited corpus since the author speaks the dialect as his mother tongue.

1.3. Terminology

The terminology used in this study is freely drawn from that of modern Bantuists and that of general phonetic and linguistic description. The symbols are those of the IPA.

2. Main Structural Features of Vunjo

Since tonetics and tonology are intricately interrelated to segmental phonology, morphology and syntax it is proposed here that only the main features will be presented in outline.

2.1. Vowels

Vunjo has five vowel phonemes represented as /i/, /e/, /a/, /o/ and /u/.

Phonetic realisations are [i], [e], [a], [o], [u]. Minimal pairs involving long and short vowels are not found in Vunjo vowels. All vowels are phonemically short. Phonetically long vowels occur as a result of morphophonemic processes (see 2.4). There are no phonemic diphthongs. What appear to be diphthongs are quick transitions from one vowel to another.

Vowel Distribution

All vowels can occur alone as independent syllable-carrying tones of the structure -V. Moreover all vowels can occur alone initially before a consonant or another vowel and finally after a vowel as most of Vunjo syllables are open hence /#-C, /#-V, /C-C, /CV-#. In all these cases they carry tone.

2.3 Consonants

Phonologically Vunjo has twenty-eight consonant phonemes which are realised as twenty-nine consonant sounds phonetically.

2.3.1.Phonetic Chart of Vunjo Consonants and their realisations

	Bilabial	Labio Dental	Dental	Alveolar	Palatal	Velar	Glottal
Stops	p b		t d		c	k	
Prenasalized	mb		nd			ŋg	
Fricatives		f	s	ɹ	ʃ		h
Prenasalized		mv	nz				
Africates		pf	ts		tʃ		
Prenasalized					ndʒ		
Nasals	m			n	ɲ	ŋ	
Laterals				l	λ		
Flapped				ɾ			
Rolled				r			
Semi Vowels	w				j		

The phonetic realisation of the consonants is as follows:
Stops: *p*ata MM (field) *mbo*ɹa HM (blessing u*r*ifo MLM (foot) *nd*ifo MM (footsteps), *d*uka MM (shop), i*k*oru MML (snail), *c*ija ML (yam), ŋgofi HH (blows); Fricatives: i*ʃ*ila MML (weakling), *mv*uo HH (rain), *i*sembo HHL (fool), *nz*ie HL (locust), ɹema HH (farm), *m* ʃiki MML (sister) *h*andu HL (place); Affricates: *t* ʃoma MM (mucus), n*dʒ*ui HH (hair), *pf*umu HH (spear), ki*ts*i ML (waist); Nasals: *m*anake MMM (boy), *n*atsi ML (dry grass), ŋama HL (meat), ruŋu ML (leopard); Laterals: *l*oɹika HHL (stool) maλa ML (edibles); Flaps: koɹa HM (moss); Rolls: ŋge*r*e*r*o HML (beard); Semivowels: *w*oλi ML (wedding), *j*ewa HL (shoulder).

2.3.2. Consonant Distribution
All consonants can occur in initial and intervocalic and after syllabic nasals without noticeable phonetic modification.

2.4. Morphonotactics (Morphophonology)

The most common morphophonemic changes observed in Vunjo (Chaga) are also very common in most other Bantu Languages when sounds are in contact. These are contraction, assimilation, elision, devocalization and palatalization.

2.4.1. Contraction
a + a > a: /wa+ana/ > [wana] (HH) (children).

2.4.2. Reciprocal Assimilation without Contraction
a + i > ee: /wa+iwi/ > weewi HLH (thieves).

2.4.4. Elision
i. V+CV+o > V+C+o: /i+tʃu+o/ [itʃo] HH (that one Cl 1)
ii. CV1+V2>CV2V2: /wa+ou>[woou] HLH (fearful ones Cl 2)

2.4.5. Devocalization
i+V > j: /i+a+ko/ > [jako] ML (mine).

2.4.6. Consonant Assimilation
ɾ>d, w>b: /n+rusu/ > [ndusu] HH (strings) /n+weri/ > [mberi] MM (feathers).

2.4.7. Consonant Loss
n>ø /-p, t, k, l v, s, f, (optional): /n+fana/ > [fana] MM (birth mark) /n+teri/ > [teri] HL (soil).

2.4.8. Palatalization
This affects the sound /k/ mostly so that it is always palatalized before front vowels: /ki+ndo/ [cindo] HL (thing).

3. Tonetic Classifation of Nouns

3.1.

A simple noun in Vunjo (Chaga) consists of at least one independent prefix and a noun root: mndu</mu+ndu/ (person) pl (wandu). On the phonetic level some of the prefixes are not realised, e.g. [pfumu]<i+pfumu/ (spear) pl [mapfumu]. Compound nouns consist of two independent prefixes and at least two stems: [mndumka]</m+ndu+m +ka/ (woman). Complex nouns

consist of at least three prefixes and at least two stems: e.g. [mawanduwasoro] < /ma +wa+ndu+wa+soro/ (big men).

3.2. Phonetic Realisations: Syllabic Structure of Nouns.

A noun always occurs with its prefix in speech, except where the prefix has been dropped by specific morpheme rules. It is within this framework that the tonetic classification of nouns is given.

Most of the nouns in Vunjo are of the syllable structure CVCV and CVCVCV. Monosyllabic, quadrisyllabic and pentasyllabic nouns occur very infrequently. Of a total 484 nouns examined only 6 (1.2%) were monosyllabic, 197 (39.4%) disyllabic, 223 (46%) trisyllabic, 32 (6.4%) quadrisyllabic, and 6 (1.2%) pentasyllabic. Disyllabic and trisyllabic tonetic realisations accounted for 86% of the total.

3.3. Tonetic Patterns of Nouns

The nouns were tonetically classified into different patterns depending on their tonal realisations in isolation and also in a very limited context. A question such as 'what is the word for 'stranger' in Vunjo' the answer, 'mjenu' HML. This was then put in the context, /ni mjenu tupu/ 'it is a stranger only'. On the basis of this criteria the following patterns were found. The monosyllabics have one pattern: H ɹ (sleep). The disyllabics have five patterns: HH mana (child), HM kip fi (wasp), MM pata (meadow), HL kindo (thing), ML soka (axe). Trisyllabics: MMM it ʃeri (tick), HML kikut ʃu (bag), HHL loɹika (stool), MML mkuma (wind), HLH kikapu (basket), MLM mkoku (baby), HLL kileje (bird). The quadrisyllabics were found to have eleven patterns HLHL kilemeko (bend), MLML lavvutaŋi (brother-in-law), HLLM uaŋgari (thirst), HLLH kilodaŋa (hare), MLHH mfiɹiwa (cattle-pen), HHLH ndaɹakana (spark), HHMM warisinda (sister in law), HHHH manamae (cousin), HMMM ikuŋia (sack) HLLL kipopo ru (rodent), HHHL ndondokoɹo (ant). The pentasyllabics have three patterns: HLHLH (mneengeɹi), MLLHL it ʃoŋgololo (millipede) HLHHL (ikuɹuɹuma) (thunder).

References

Raum, J. (1964). *Versuch einer Grammatik der Dschaggasprache (Moschi Dialekt).* Ridgewood: Gregg.

Temu, C.W. (1975). *Nominals in Vunjo (Chaga). A Linguistics Description of the Class System.* (Unpublished M.A. Thesis). Dar es Salaam University.

Section 14
Speech Acquisition

Pitch and Timing Cues in Speech Intelligibility: the Case of Child Language

N. Bacri
Paris, France

1. Introduction

These last years, several investigators have presented evidence for a direct link between timing and pitch cues in sentence perception. It has been shown that prosodic parameters allow listeners to disambiguate sentences (Lehiste et al., 1976), whereas deviant durations and fundamental frequency contours decrease intelligibility (Huggins,1978, 1979). Moreover, data bearing on the relative effect of temporal cues and pitch movements point to the greater efficiency of the former. In experiments where the three suprasegmental variables for one meaning of an algebraic expression are interchanged for the other meaning of the expression, Streeter (1978) demonstrated that the magnitude of the pitch effect was about 14-20%, depending on the speaker. When conflicting cues are introduced, lengthening alone gave a correct score of 90% (Nooteboom et al. 1978), but F_0 overrode the cue value of pauses as boundary markers (Larkey 1981). However, if local variations of F_0, except in the latter case, are not a crucial factor for sentence comprehension, the contribution of prosody to speech perception appears to be more important than suggested by these data. Duration and pitch are analysed as connected cues and form an integrated percept. They supply different information, and specifically, prosodic continuity seems to be essential to the listener's ability to perceive speech at all (Nooteboom et al., 1978, Carlson et al., 1979).

In reply to the question as to what kinds of perceptual cues are available to allow a listener to understand speech with a low degree of intelligibility, I have put forward the hypothesis that compatibility between timing and pitch cues support intelligibility whereas conflicting cues result in misunderstanding. In this paper, I try to answer this question by comparing temporal organization and pitch patterns of two types of utterances, judged intelligible or not.

2. Subjects, test material and procedure

Three adults (two females and one male), native speakers of Parisian French, were asked to sort out 1450 utterances of spontaneous speech, produced by a child, into three categories: intelligible (IU), unintelligible (UU), undecidable. Utterances were classified as follows: 463 IU (32%) and 859 UU (59%);

128 utterances (9%) were judged undecidable by at least one judge. The child, aged 1;10 at the beginning of the study and 2;2 at the end, was visited six times at home, every three weeks. Each session lasted 45 minutes and was recorded. All utterances were transcribed in IPA by a phonetician who then calculated the number of syllables in each sequence. An independent rater likewise computed the number of syllables by sequence. A level of agreement of 0.96 was attained between the two ratings (de Bardies et al., 1981).

The present study does not bear on phonemic distribution. However, it is worth noticing that is is significantly different from adult French distribution.

For each utterance, a pitch detector plotted variations of F_0 and amplitude of the speech signal as a function of time. The following variables were measured: total duration of utterances, initial and final syllabic durations, overall speaking rate (duration of utterances / number of syllables / sec.), articulation rate or time spent articulating (duration of utterances minus duration of pauses / number of syllables / sec.), duration and frequency of pauses, mean F_0 value for the first F_0 p'ot, amounts of rises and falls, average number of fluctuations per second and per syllable in the F_0 patterns. For the last two variables, a fluctuation was defined as a point corresponding to a local maximum or minimum on the F_0 curve. In order to be sure of the perceptual relevance of F_0 movements, a F_0 change was labelled fluctuation if it exceeded 35 Hz during at least 50 msec.

3. Timing analysis: durational cues

Analysis of variance was carried out on 360 utterances from 2 to 5 syllables, equally distributed over the sessions. Total duration of utterances had no significant effect on intelligibility, and neither had the number of syllables. The main effects were due to initial and final syllabic durations. Initial syllables in IU were shorter than in UU ($F(1,356) = 6.30$, $p < .02$). The mean durations were 139 msec. and 156 msec., respectively. Final syllables in IU were longer than in UU ($F(1,356) = 5.84$, $p < .05$). The mean durations were 185 msec. and 166 msec. respectively. Moreover: final syllables in IU increased in length with the increased length of utterances. The interaction between final syllabic duration and number of syllables was significant ($F(1, 178) = 4.98$, $p < .05$). However, the degree of variation for syllabic durations was considerable, as previously noted by Oller and Smith (1977) for child language. So an analysis of syllabic durations was performed session by session, and on the whole corpus, data being partitioned in short and long utterances according to the number of syllables (2+3 vs 4+5 syllables). *t*-test showed significant differences between IU and UU, but only for initial syllabic durations of long utterances. *t*-test for paired observations performed on IU and UU respectively showed systematic differences between initial and final syllabic durations for IU (short IU: $t(5)=4.25$, $p < .01$; long IU: $t(5)=5.69$, $p < .01$), but not for UU (Fig. 1).

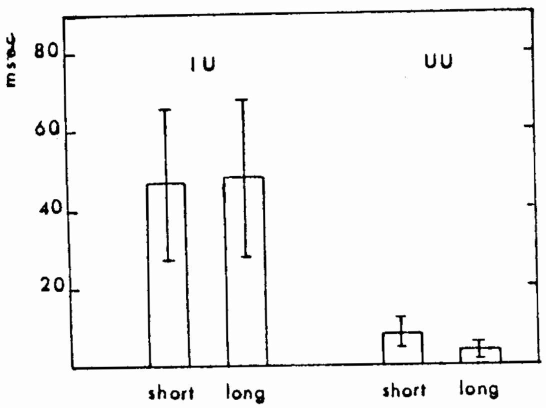


Figure 1. Mean differences and confidence intervals (in ms.) between initial and final syllabic durations, within short and long sequences, for intelligible (IU) and unintelligible (UU) utterances.

Moreover, changes in duration varied from 21% to 30% in IU and so, according to Klatt (1976), could serve as perceptual cues, whereas they varied from 1% to 8% in UU and it was not possible to discriminate them.

The main difference between IU and UU comes from the contrast between initial and final syllabic durations in IU, a contrast that does not appear in UU. IU alone present a regular temporal organization, and therefore a rhythmical structure, which can be used by the listener as a cue to the linguistic value of the utterances, mainly to their boundaries. These local variations are probably directly linked to changes in the spectral properties of speech (Miller 1981), and can facilitate the judgment of intelligibility. But it is also well known that rate of speech, and particularly articulation rate, influence temporal characteristics and the way in which intelligibility is achieved. So it is worth studying whether these local variations are linked to variations in speech rate, to some kind of extrinsic timing.

4. Timing analysis: pauses, speaking rate, articulation rate

For the two types of utterances, pause length was about 20% to 27% of the total mean duration. IU were characterized by one pause in median position, just after the first syllable (short utterances) or after the second or third one (long utterances). In all cases, pause was preceded by syllable lengthening. The mean prepausal syllabic duration was 168 msec. On the other hand, no preferential position for pauses and no lengthening were found for UU. Speaking rate and articulation rate were similar for the two types of utterances. Speaking rate was 4.23 syll./sec. for IU, 4.40 for UU; articulation rate was 5.75 syll./sec. for IU, 5.99 for UU;

Our data suggest the perceptual relevance of relative timing in the intelligibility of child language. But the role of speech rate (overall speaking rate and articulation rate) has not been confirmed. It seems likely that intrinsic timing, i.e. variations in initial and final syllabic durations, is not related to changes in the rate of speech. It is not possible to conclude from the data that listeners adjust to speech rate, in order to judge utterances. It can be assumed that it is intrinsic timing alone which serves as a cue because it supports the rhythmical structure of intelligible utterances. It can be thought that the contrast between initial and final syllabic durations in IU is a cue both to the phonetic identity and to the identification of utterances boundaries. This last hypothesis concerning the role of timing cues in intelligibility is confirmed by the close correspondence between temporal and speech patterns.

5. Fundamental frequency analysis

Analysis of F_0 initial mean value has showed its steadiness through the whole corpus for every kind of contour, rising, flat or falling (336 to 294 Hz for IU; 340 to 322 Hz for UU). Amounts of rises and falls had similar ranges of values, about 160 Hz for the rises, 100 Hz for the falls. No difference was found between types of utterances. However, it is worth noticing that F_0 changes exceeding 100 Hz could occur on arbitrary positions for UU, whereas they tended to occur on the last syllable for IU, or on a syllable preceding a pause. In this last case, the utterance was identified as a sequence of two constituents.

Analysis of the two 'dynamic' variables indicated that the average number of fluctuations per syllable and per second was quasi constant on the whole corpus, whereas for each session the two types of utterances differed significantly. The number of fluctuations per syllable was two to three times smaller for IU than for UU. Values of *t* (6) varied from 3.6, $p < .02$ at 1;10 year old to 9.80, $p < .001$ at 2;2 year old. Average number of fluctuations per syllable was .666 for IU, 1.26 for UU (Fig. 2).

These results suggest that one fluctuation per syllable impairs intelligibility. Moreover, small fluctuations occurred in all syllabic positions for UU, whereas they occur for IU on the last one or the last two syllables, final rise or fall being preceded by one or two small F_0 changes. In long IU, a third fluctuation occurred only before a 200 to 300 msec. pause. Simultaneity of occurrence of durational and pitch boundary markers seems to facilitate intelligibility judgment. This explanation is confirmed by analysis of the average number of fluctuations per second; 2.78 for IU, 4.95 for UU (Fig. 2). As a consequence, IU F_0 slopes were smoother than UU ones.

The close relationship between intelligibility judgment and the rate of F_0 change suggests that abrupt slopes and frequent jumps lower intelligibility, particularly when large F_0 changes do not precede a pause and when a lot of small fluctuations disturb the establishment of pitch continuity.

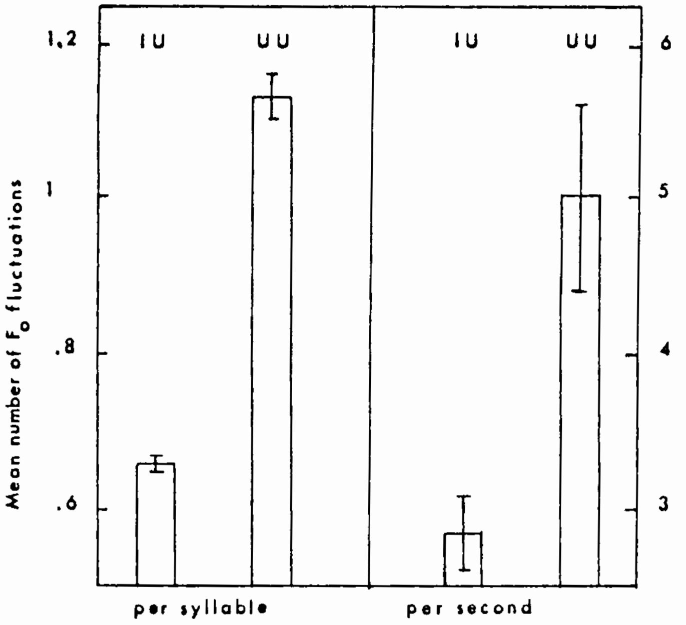


Figure 2. Mean number of F_0 fluctuations per syllable and per second for intelligible (IU) and unintelligible (UU) utterances, with confidence intervals.

6. Discussion

It is well known that the French language is characterized by a final accent marked by a final syllabic lengthening as well as by changes in F_0 (Delattre, 1966, Crompton, 1980; Rossi et al., 1981), but there is neither a rapid jump, nor a break in F_0 slope exceeding 40% or 50% of the former value (Delgutte, 1978; Vaissière, 1980). In this paper, I have attempted to show that prosodic patterns differ systematically between utterances perceived by adults as babbling or as first language. However, neither differences in overall duration, speaking or articulation rates, nor differences in F_0 initial values, amounts of rises and falls, can account for intelligibility judgments. Relative lengthening and pitch continuity seem to be the main perceptual cues.

In spite of the great variability in prosodic realizations, close correspondence between syllabic lengthening and F_0 movements preceding boundaries in IU, as well as the smoothness of F_0 slopes, confirm the importance of prosody in speech perception. Conflicting cues providing misleading information significantly impair intelligibility. I cannot specify here which is the main determinant of listeners' judgment: duration or pitch. However, the results suggest that pitch continuity contributes to intelligibility by establishing a process of backward perceptual normalization, related to the intrinsic timing of utterances, that facilitates syllabic identification and the detection of boundary markers.

Though the data show that boundary detection contributes to the listener's construction of an integrated percept, this does not exclude the role of segmental features, which I have not studied here. It is also worth noticing that this study bears on the language of a French child, and that it needs to be confirmed with other languages.

References

de Boysson-Bardies, B., Bacri, N., Sagart, L., and Poizat, M. (1981). Timing in late babbling. *J. Child Lang.* **8**, 525-539.

Carlson, R., Granström, B. and Klatt, D.H. (1979). Some notes on the perception of temporal patterns in speech. *Ninth Inter. Congress of Phonetic Sciences*, II 260-267, Copenhagen.

Crompton, A. (1980). Timing patterns in French. *Phonetica* **37**, 205-234.

Delattre, P. (1966). Les dix intonations de base du français. *French Review*, **40**(1), 1-14.

Delgutte, B. (1978). Technique for the perceptual investigation of F_0 contours with application to French. *J. Acoust. Soc. Am.* **64**(5), 1319-1332.

Huggins, A.W.F. (1978). Speech Timing and Intelligibility. J. Requin (ed.). *Attention and Performance* VII, Hillsdale, Erlbaum Ass.

Huggins, A.W.F. (1979). Some effects on intelligibility of inappropriate temporal relations within speech units. *Ninth Intern. Congress of Phonetic Sciences*, II 283-289, Copenhagen.

Klatt, D.H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am.*, **59**(5), 1208-1221.

Larkey, L.S. (1981). *The role of prosodic information in speech perception.* Doct. Dissertation, Univ. of Minnesota.

Lehiste, I., Olive, J.P. and Streeter, L.A. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *J. Acoust. Soc. Am.*, **60**(5), 1199-1202.

Miller, J.L. (1981). Effects of Speaking Rate on Segmental Distinctions. P.D. Eimas and J.L. Miller (eds.) *Perspectives on the Study of Speech.* Hillsdale, Erlbaum Ass.

Nooteboom, S.G., Brokx, P.L. and de Rooij, J.J. (1978). Contributions of prosody to Speech Perception. W. Levelt and G. Flores d'Arcais (eds.) *Studies in the Perception of Language.* N.Y.: Wiley.

Oller, D.K. and Smith, B.L. (1977). Effect of final-syllable position on vowel duration in infant babbling. *J. Acoust. Soc. Am.*, **62**(4), 994-997.

Rossi, M., Di Cristo, A., Hirst, D., Martin, Ph. and Nishinuma, Y. (1981). *L'Intonation: de l'acoustique à la sémantique*. Paris: Klincksieck.

Streeter, L.A. (1978). Acoustic determinants of phrase boundary perception. *J. Acoust. Soc. Am.*, **64**(6). 1582-1592.

Vaissière, J. (1980). Structuration prosodique de la phrase française. *Les Annales de l'Ecole Normale Supérieure de Pise*, Pise.

Audio-Vocal Self Control Functions: Pitch Fluctuations and Audio-Vocal Pitch Matching

S. Eguchi and M. Suto
Tsukuba, Japan

1. Introduction

Several sensory feedback loops – auditory, tactile and kinesthetic sensations – are thought to exist for controlling voice and speech. It is well known that the auditory feedback self-control functions are the most effective among them for speech motor behavior. We can realize the importance of the auditory feedback self-control function for speech motor behavior through the fact that the voice and speech of the deaf persons have very distorted acoustical features.

This paper involves the pitch fluctuations in the sustained voices of normal-hearing and hearing-impaired children and the audio-vocal pitch matching abilities of normal-hearing adults to show the auditory feedback self-control influences upon voice production functions.

2. Experimental procedure

The pitch fluctuations of sustained voices of five Japanese vowels by normal-hearing and hearing-impaired children were measured. The subjects were 21 hearing-impaired children and 60 normal-hearing children of 7 to 12 years old. Voice samples for 3-5 seconds were recorded. Each 200 ms of the onset and offset periods of voices was eliminated because voice pitch of these portions was not stable. Fundamental frequencies were measured at 100 ms measurement points in the stable portions of the analog output with a pitch meter and the results were shown in the Figures 1, 2 and Table I.

The relative pitch fluctuations to the mean fundamental frequencies (SD/F_0) of hearing- impaired children's voices (0.93%-3.10%) showed clearly higher values than those of normally hearing children (0.49% – 0.89%). The correlation of age with relative pitch fluctuations of hearing-impaired children was not pronounced but that of hearing threshold at the lower frequencies (250Hz and 500 Hz) with relative pitch fluctuations was manifest.

The audio-vocal pitch matching by normally hearing adults was examined.

The subjects were 5 normally hearing male adults without special musical training.

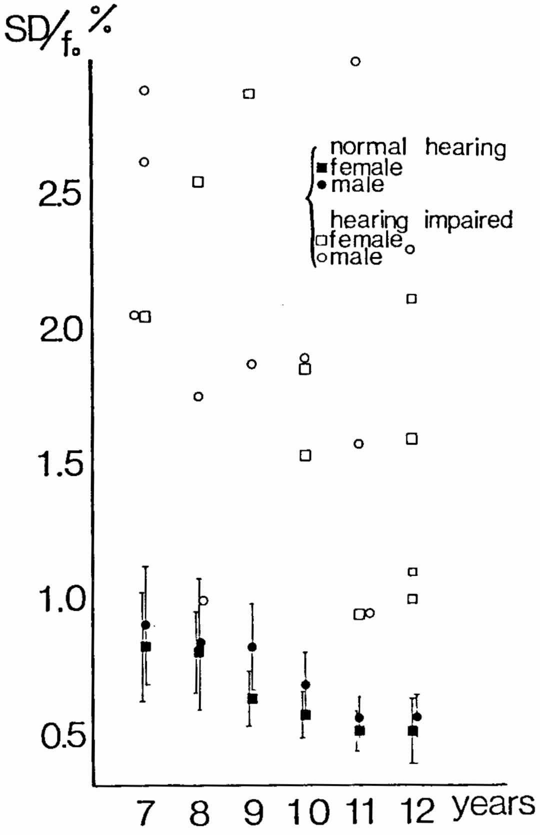


Figure 1. Relative Pitch Fluctuation as a Function of age.

130 Hz and 145 Hz tones were used as the base voice pitch for each subject to produce a natural sustained voice. The test target tone generated by an oscillator was given to the right ear of subjects at 45 dB HL through a headphone (TDH-49) and subjects produced the Japanese vowel /a/ at a pitch as near as possible to that of the target tone. When the sustained voice of a subject had been continued for about 2 seconds, the pitch of the target tone was changed higher or lower by 10 Hz or 20 Hz. Subjects perceived the pitch changes of target tones and then followed the changed pitch of target tones matching their own voice pitch with the pitch of target tones. The speech samples and target tones were recorded on a 2 channel tape recorder at the same time and their pitch was extracted by a 2 channel pitch meter and recorded on an analog display.

To measure the subjects' audio-vocal pitch matching abilities, three temporal time lags – from the beginning of the pitch change of a target tone to the beginning of the pitch change of the subjects' voice, from the beginning of the pitch change of a target tone to the end of the pitch change of the subject's voice, and from the beginning of the pitch change of a target tone to stabilizing of the subject's sustained voice – were obtained and the results are shown in Figure 3 and Table II.

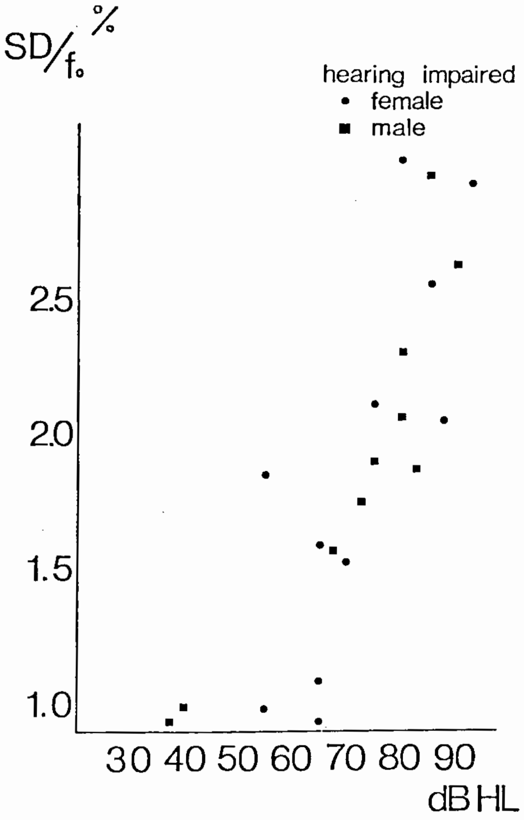


Figure 2. Relative Pitch Fluctuation as a Function of Hearing Threshold.

3. Results

The time lag from the beginning of the pitch change of a target tone to the beginning of the pitch change of the subject's voice ranged from 328 ms to 405 ms. The mean value of this time lag was 351 ms, the between-subject standard deviation in the distribution of this time lag was 82 ms, and the intra-subject standard deviation was 75 ms. There was no significant difference in the data of the time lags among the four conditions in which pitch changes of target tones were different.

The time lag from the beginning of the pitch change of a target tone to the end of the pitch change of the subject's voice ranged from 674 ms to 830 ms. The mean value of this time lag was 756 ms, the between-subject standard deviation in the distribution of time lag was 181 ms and the intra-subject

Table I.

Normal hearing (N = 60)

Age	Male	SD/F (%)	Female	SD/F (%)
7		0.89		0.81
8		0.82		0.79
9		0.81		0.63
10		0.67		0.58
11		0.56		0.50
12		0.56		0.49

Hearing-impaired (N = 21) HL: (250 + 500 Hz)/2

Age	Male	HL (dB)	SD/F (%)	Age	Female	HL (dB)	SD/F (%)
7		80	2.05	7		87.5	2.04
7		85	2.95	8		85	2.54
7		90	2.61	9		92.5	2.91
8		40	0.99	10		55	1.84
8		72.5	1.74	10		70	1.84
9		82.5	1.86	11		65	0.93
10		75	1.88	12		65	1.58
11		80	3.10	12		55	0.98
11		67.5	1.56	12		75	2.10
11		37.5	0.93	12		65	1.08
12		80	2.29				

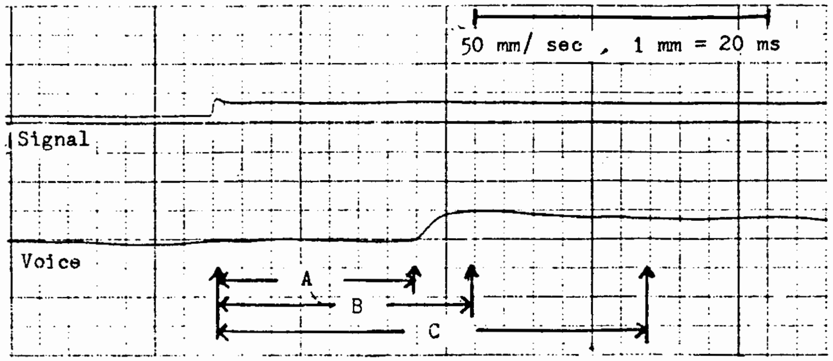
Table II. Latencies for the signal change

	Onset	End	Stable stage
Mean	351.5	756.9	1326.2 msec.
B.S.S.D.	82.9	181.3	325.6
I.S.S.D.	75.9	115.9	245.6

B.S.S.D.: Between subject standard deviation; I.S.S.D.: Intra subject standard deviation.

standard deviation was 115 ms. When the pitch of a target tone was changed to lower pitch, the time lag was shortened significantly in comparison with the condition in which the pitch of the a target tone was changed to a higher pitch.

The time lag from the beginning of the pitch change of a target tone to the stabilizing of the subject's sustained voice ranged from 1226 to 1409 ms. The mean value of this time lag was 1326 ms, the between-subject standard deviation in the distribution of this time lag was 329 ms and the intra-subject standard deviation was 245 ms. There was no significant difference in the data of the time lags among the four conditions in which pitch changes of target tones were different.



A: Latency to the onset of voice inflection
B: Latency to the end of voice inflection
C: Latency to the stable stage of sustained voice

Figure 3. A Sample of Actual Recording.

4. Conclusion

From our present experimental results concerned with the influence of the auditory self control functions on voice and speech behavior which are very important for the development of speech sounds in early childhood, it was shown that the fluctuations of sustained voice were not correlated with age but with hearing thresholds and that there were several time lags between auditory target tone and the achievement of voice responses in the audio-vocal pitch matching ability.

References

Chase, R.A. and Guilforyle, G. (1962). The effect of simultaneous delayed and undelayed auditory feedback on speech. *Journ. of Speech and Hearing Research,* **5**, 144-151.

Eguchi, S. and Hirsh, I.J. (1969). Development of speech sound in children. *Acta Otolaryng, Suppl.* 257.

Monsen, R.B. (1979). Acoustic qualities of phonation in hearing impaired children. *J. Acous. Soc. Amer.* **22**, 270-288.

Interactive Speech Synthesis in the Study of Normal Perceptual Development

V. Hazan and A.J. Fourcin
London, United Kingdom

1. Introduction

If normal speech perceptual development is largely dependent on learning processes in which auditory analysis plays an important role, three aspects of speech contrastive development in the normal child are likely to be of special consequence.

First, the earlier stages of speech perceptual development will be characterised by the influence of auditorily dominant speech pattern elements. Second, development will tend to proceed from the simple to the complex in auditory - rather than articulatory - terms. Third, individuals may differ markedly in their ability to use different speech pattern elements in the identification of phonetic contrasts.

The stages of development of a child's acquisition of perceptual ability in the processing of speech patterns have yet to be fully investigated. A normative study of the ages at which the child is likely to start making contrastive use of, for example, F_1, F_2 and F_3 transition cues, VOT information, aspiration cues is imperative in order to assess and facilitate the development of these abilities in hearing impaired children.

The present experiments make use of synthetic speech stimuli defining meaningful phonetic contrasts. These high quality stimuli are modelled on the utterances of a particular woman in order to provide a coherent pattern set and to minimise normalisation problems for the child.

The end-point stimuli are interpolated to provide a six step continuum. In our first work in this area, pre-recorded sequences of these stimuli were used. These were too long for young children. Now, an interactive test system is used, in which the syntheses are made on-line at a level of difficulty which is a function of the subject's response. The ongoing assessment of the subject's performance ensures that the most efficient presentation of stimuli is given, concentrated in the 75-25% labelling area, with fewer stimuli presented in those parts of the stimulus continuum where the subject is labelling with confidence. The test-retest reliability was found to be good.

The response situation has been improved by the use of touch sensitive pictures which relay the child's choice directly to the microprocessor controlling the running of the test. The whole apparatus used is relatively compact and portable so that it can be used in classroom and clinic. The tests available range in increasing pattern complexity from simple fundamental frequency contours; a vowel contrast between two phonetically contiguous forms; a plosive consonant voicing contrast; and finally a voiced plosive alveolar-velar contrast.

2. Main features investigated

1. Fundamental frequency

English intonation is a highly complex system mainly cued by the fundamental frequency contour pattern. It cues not only the difference between question and statement but also the placement of nuclear stress. An assessment of the child's ability to contrast Fx patterns is made using a very simple 'Oh? - Oh' test in which the Fx contour is systematically varied between an extreme rise and an extreme fall.

2. Vowels

The child's ability to contrast two vowels can be assessed by systematically varying F_1 and F_2 in between two extremes corresponding to two phonetically contiguous vowels. The relative importance of F_1 and F_2 in establishing a contrast can be assessed by presenting each formant separately. This proves very fruitful when trying to assess whether the hearing impaired child is able to make effective use of F_1 information.

3. Consonant place contrasts

Important acoustic features of place contrasts in initial consonants are provided by the F_2 transition into the vowel and the frequency of the noise burst. A 'date/gate' test is used in which these two cues can be varied systematically, either separately or together.

4. Voicing contrast

The voicing contrast in initial cognate plosives is cued mainly by a rising F_1 transition in the voiced consonant and the voice onset time following the noise burst. A 'coat-goat' or 'pea-bee' opposition are used to assess the perception of these contrasts.

3. Results

Results were first obtained for normally hearing adults in order to assess how the stimuli would be labelled by subjects who can potentially make use of a complete set of speech pattern features.

Since large listener to listener variations are typically found, results are analysed individually. The labelling curves below are plotted from the percentage of responses of one label versus step number and illustrate adults' responses to the place contrast in the 'date-gate' opposition and the voicing contrast in the 'coat-goat' opposition, see fig. 1.

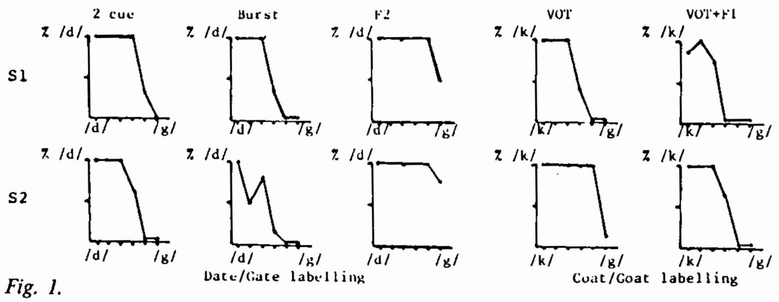


Fig. 1.

In the 'date-gate' test, results obtained for both subjects to the stimuli containing both the F_2 transition and burst frequency cues are characterised by sharp labelling. This is also the case for the stimuli in which the contrast is cued only by a change in the burst frequency value. However, both subjects had difficulties in making the contrast on the basis of F_2 transition alone. It therefore seems that they are primarily relying on the burst frequency as a main cue to this alveolar-velar contrast.

In the 'coat-goat' contrast cued by the change in VOT alone, the first subject does not seem perturbed by the absence of F_1 information and seems therefore to be making the contrast primarily on the basis of the VOT value. The second subject perceives most stimuli as voiceless in the absence of F_1 information. The labelling curve becomes more balanced when this second cue is added.

Results are needed from large numbers of children from age groups ranging from 4 to 14 in order to make a reliable estimate of the ages at which normal children are most likely to make consistent use of the various speech pattern cues. This will necessarily only be a general framework as great individual differences in strategies used by the children are to be expected.

The results given below illustrate responses obtained to the consonant contrast stimuli by two children of 6 (S1) and 9 (S2), see fig. 2.

Both children gave sharp labelling to the 'date-gate' stimuli containing the combined F_2 transition and burst frequency cues. They seemed able to make use of the burst frequency cue alone although their labelling is less sharp. There is uncertainty in their responses to the stimuli in which the contrast is cued only by the F_2 transition. The 'coat-goat' contrast seems to be more easily perceived, with evidence of sharp labelling of stimuli containing the combined VOT and F_1 transition cues. Both children also respond well to the one-cue stimuli but give a higher percentage of voiceless responses at all steps of the continuum.

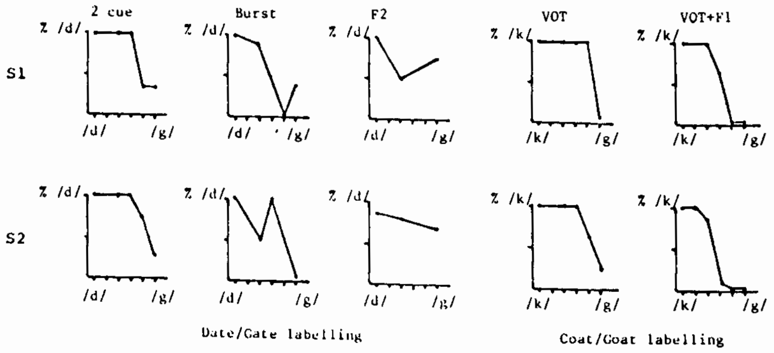


Fig. 2.

4. Conclusion

Interactive speech tests may have widespread future applications for the assessment of speech perception in hearing impaired and language retarded children. While providing a rough quantitative estimate of a child's perception of speech, classical speech audiometry gives no indication as to what features the child is making use of in contrasting two sounds. Interactive speech tests, however, carried out at regular intervals, may provide a real insight into the development of perceptual abilities. Similarly, such tests may be used to assess whether the perceptual development of the language retarded child is following a 'normal' course, albeit delayed, or whether his phonological development is deviant.

The Change of Voice during Puberty in Choir Singers Measured with Phonetograms and Compared to Androgen Status together with Other Phenomena of Puberty

M.F. Pedersen, E. Munk, P. Bennet, S. Møller
Copenhagen, Denmark

1. Introduction

The purpose of our work is to get better laryngophysiological understanding and possibilities to predict the time when choir boys lose the high frequencies of voice, so that they no longer can be used as sopranoes or altoes in a boys' choir. Two questions arise:

1. Which parameters will enlighten the changes in the best way.
2. How are the normal limits defined in relation to extreme variants and pathological divergences.

These relations have been discussed by Weiss (1950). Already Flatau and Gutzmann (1908) made a study on the voice of schoolchildren. Hartlieb (1957) concentrated on the tone range of youngsters. Sturzeberger, Wagner, Becker, Rauhut and Seidner (1982) examined in an up to date study singing formants and phonetograms in adults, but we have found no study where our earlier work on comparing the strictly biological endocrine factors to voice phenomena could be elucidated (Pedersen, Kitzing, Krabbe and Heramb, 1982). Krabbe, Christiansen, Robro and Transbol (1979) have compared pubertal hormone changes to bone growth and mineralisation. The steroid hormones were examined in elder Danish children by Pedersen, Bennett, Nielsen and Svenstrup (1983). A survey of endocrinology in puberty was given by Tanner (1981).

Our study is here concentrated on understanding the connection between phonetograms and androgen status. Many other parameters have been measured and those are included that elucidate the phenomena mentioned.

2. Material and Method

48 boys from 8-19 years in the Copenhagen Boys' Choir were included in the research programme. The selection was randomized with an equal number of subjects in each class. The children are taken into the music school on the third level with the help of the criteria musicality and good voice qualities.

Blood examination for androgen status together with somatic examination were carried out on the same day. The phonetograms were made after Seidner and Schutte (1981) and areas were estimated. Some other parameters

are shown, e.g. testis volume, fundamental frequency with electroglottograph in a reading situation after Kitzing (1979) and the lowest frequency in the phonetogram. The measuring of androgens was carried out at the hormone Dpt. of Statens Seruminstitut. The logaritmic transformations of the observations of all parameters were required to obtain a normal distribution. Data were investigated by one-way analysis of variance and the correlation coefficients were calculated comparing all parameters. The age distribution is described in a set-up alike to the earlier mentioned study of Danish children. The yearly changes in percent of the values were calculated by linear regression on the logarithmically transformed data.

3. Results

The characteristic phonetograms at different ages are shown in Figs. 1-4. The pubertal changes for an individual certainly are not linear as the mean values suggest. The computed scattergrams for serum testosterone illustrated our problem. The regression line is straight even if the material is divided in two averaged groups there is no big difference in the gradient. In Table I we therefore described the results in groups, but we have calculated the change per year in percent of each factor for the androgenstatus as well as the testis volume and voice function on the basis of all results. The testis volume has a growth rate of 36% per year compared to a fall of fundamental frequency of 11% per year. Many nuances have to be explored, but we found that the parameters are correlated significantly to height, weight, and pubic hair ($p < 0.01$).

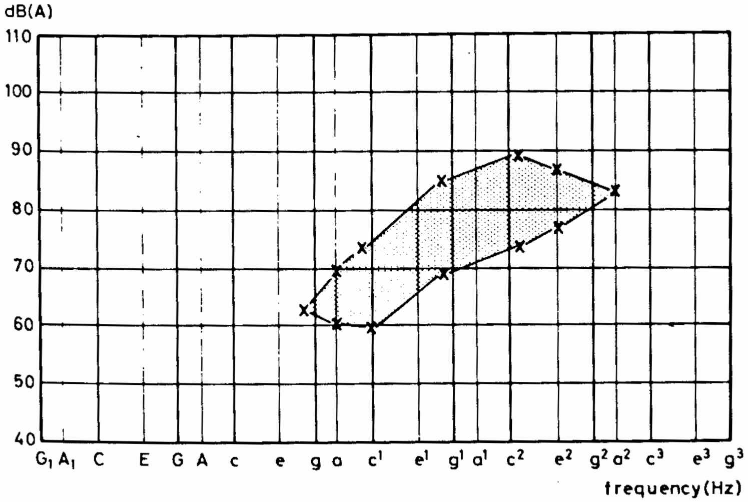


Figure 1. The characteristic phonetograms at 9.25 years when the boys start at the third school class level.

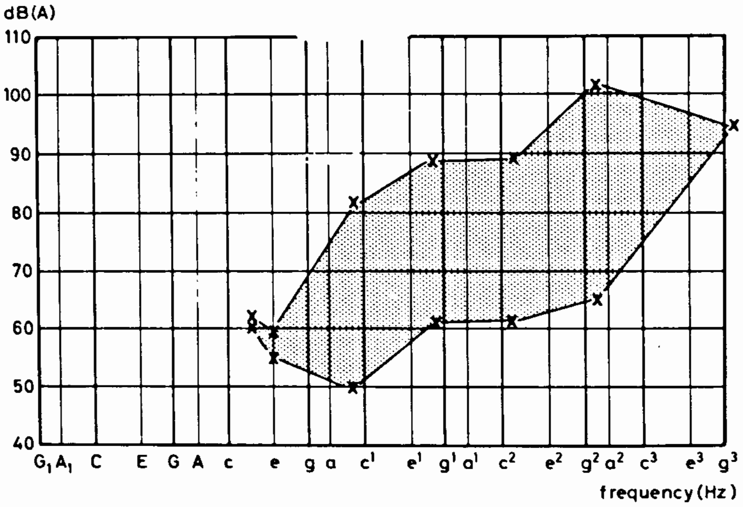


Figure 2. The characteristic phonetograms at 11.9 years when the singing quality is at its height.

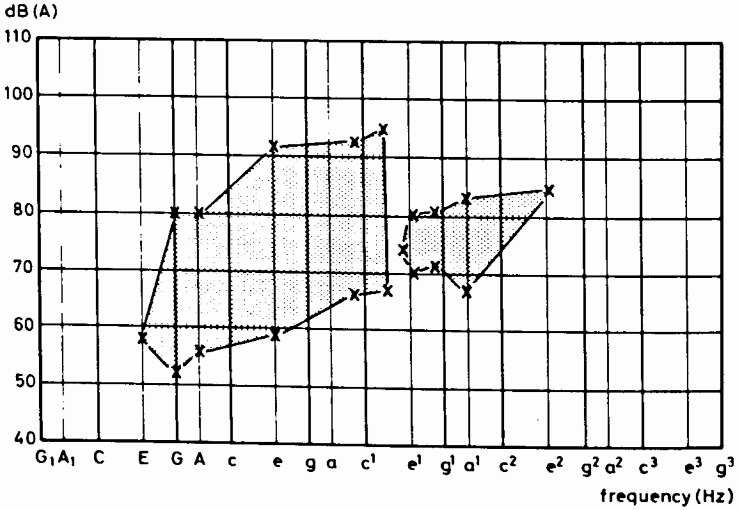


Figure 3. The characteristic phonetograms at 14.2 years when the boys can no longer sing properly.

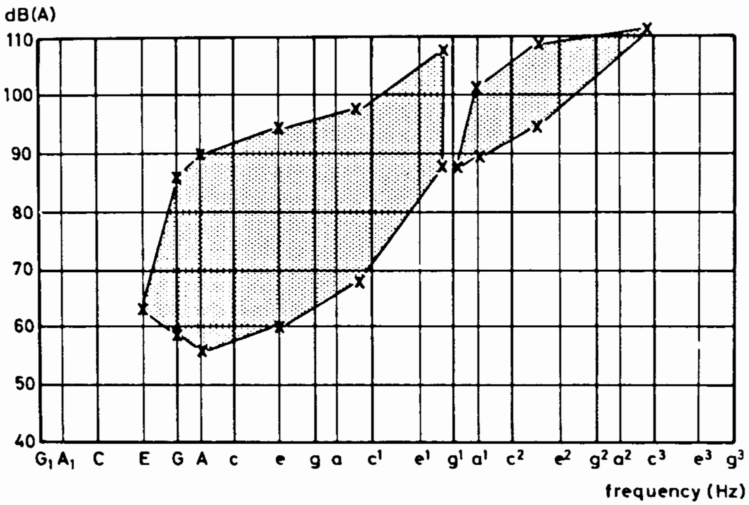


Figure 4. The characteristic phonetograms at 18.2 years when the young man sings again with a good voice.

Table I. Geometric means of androgen status, testis function and voice

Age	(years)	8.7-12.9	13.0-15.9	16.0-19.5	pr yr
No of boys		19	15	14	% change
Serum testosterone	(nmol/l)	0.54	10.5	18.9	68
Dihydrotestosterone	(nmol/l)	0.18	1.21	1.57	37
Free testosterone	(nmol/l)	0.007	0.14	0.33	77
Sexual hormone					
Binding globulin	(nmol/l)	134	66	45	-16
Delta 4 androstene					
Dione	(nmol/l)	0.59	1.7	2.5	24
Dehydro epi andro					
Sterone sulfate	(nmol/l)	1400	4100	5900	25
Testis volume	(ml)	2.3	13	20	36
Fundamental frequency	(Hz)	273	184	125	-11
Voice range	(semitones)	3.7	4.8	5.0	3.9
Phonetogram area	(cm^2)	19	28	34	9.2
Lowest biological tone	(Hz)	158	104	72	-12

4. Conclusion

A significant correlation between androgens, fundamental frequency voice range phonetogram area and lower border frequency in the phonetogram was found. We can therefore state that the hormone changes are related to the fundamental frequency and probably at least in our material to the lower border of the biological tone range and the phonetogram area with the same significance. We therefore have to continue our studies to encircle prognostic data for losing height of voice in puberty. Problems with differentiation of falsetto in the upper part of the phonetogram must be discussed. The limits can possibly be defined with singing formants which were found in some children (Seidner and Pedersen, 1983). In the material given, our results on the musical talents of the children are not discussed in detail. With a perception test we have tried to find the limits of the tone range usable in a choir situation and mark them in the phonetograms.

References

Flatau, Th.S., Gutzmann, H. (1908). Die Singstimme des Schulkindes. *Arch. Laryng. Rhinol. Berlin* **20**, 427.

Hartlieb, K. v. (1957). Der Umfang der Jugendstimme. *Folia Phoniatr.* **9**, 225.

Kitzing, P. (1979). Glottografisk Frekvensindikering. Thesis. Malmø: Lund Universitet.

Krabbe, S., Christiansen, C., Robro, P. and Transbol, I. (1979). Effect of puberty on rates of bone growth and mineralisation: with observations in male delayed puberty. *Arch. of Disease in Childhood* **54**, 950.

Pedersen M.F., Kitzing, P.; Krabbe, S. and Heramb, S. (1982). The change of voice during puberty in 11 to 16 years old choir singers measured with electroglottographic fundamental frequency analysis and compared to other phenomena of puberty. *Acta otolaryngol.*, supp. **386**, 189.

Pedersen, P.S.,Bennett, P., Nielsen, J. and Svenstrup, B. (1983). Undersøgelse af steroidhormoner hos større danske børn. *Ugeskr. laeger* **145**, 2010.

Seidner, W. and Schutte, H.K. (1981). Standardisierungsvorschlag Stimmfeld messung/ Phonetographie. *Congress Proceed. IXth Congress UEP*, 88.

Seidner, W. and Pedersen, M.F. (1983). Personal communication.

Sturzebecher, E., Wagner H., Becker R., Rauhut, A. and Seidner, W. (1982). Einrichtung zur simultanen Registrierung von Stimmfeld und hohen Sangerformant. *HNO Praxis, Leipzig* **7**, 223.

Tanner, J.M. (1981). Endocrinology of puberty. In: Brook, Ch.G.D. *Clinical Paediatric Endocrinology.* Oxford: Blackwell, 207.

Weiss, D.A. (1950). The pubertal change of the human voice. *Folia phoniatr.* **2**, 127.

Section 15
Sociophonetics

Sociophonetic Restrictions on Subphonemic Elements in Pirahã

D.L. Everett
São Paulo, Brasil

1. Introduction

The Pirahã language is spoken by approximately one hundred and ten individuals along the Maici river in the state of Amazonas, Brazil. The Pirahã are monolingual having had ony sporadic contact with outsiders (traders, laborers, etc.) until the last ten years or so. Further, most of these outsiders ridicule the Pirahã language commonly referring to it as *fala de galinhas* 'chicken talk'. An especially frequent source of such ridicule is the visual effects produced by the two phonetic (allophonic) segments [í̦] and [b̃] (to be described below).

In this paper, we want to examine the rather interesting sociophonetic restrictions on [ǐ̦] and [b̃] which result from this ridicule and to investigate a few of the implications of these restrictions for phonological theory. To begin, let us review briefly the notions 'phonemic' and 'subphonemic'.

2. Phonemic and Subphonemic

Traditionally, theories which attach psychological significance to the notion of phonemic or phonological segments have maintained something like the concept developed by Sapir. To Sapir, the phoneme was a basic segment of sound perceived by the native speaker as a discrete element rather than merely a point on a continuum of a particular articulatory or acoustic feature. In his terminology (Sapir, 1949), phonemes are as distinct from one another as 'poles' and 'clubs'. There is no halfway point at which a particular feature of 'clubness' disappears and a club becomes a pole. This is partially due to the fact that variations in the forms of phonemes in this theory, and to some degree generative phonology, are seen to be *sub*phonemic, i.e. without psychological status. As Sapir said, '... what the naive speaker hears is not phonetic elements but phonemes.... It is exceedingly difficult, if not impossible, to teach a native to take account of purely mechanical phonetic variations which have no phonemic reality for him.' (ibid:23).

We want to establish here that, somewhat contrary to Sapir's statements, an entire segment of the population of the Pirahã has 'taken account' of certain phonetic variations due to contact with outsiders and that the social context may act as a 'filter' to eliminate socially unacceptable, sub-phonemic features.

In order to better understand the restriction within the phonological system of Pirahã, the major processes and features of this system are given below.

3. Major Phonological - Phonetic Features of Pirahã

Phonemes

Pirahã has ten segmental phonemes: /p/, /t/, /ʔ/, /b/, /g/, /s/, /h/, /i/, /a/, /o/ and two (register) tonemes, /´/ 'high tone' and /`/ 'low tone'.

Palatalization

/t/ and /s/ are realized as [t ʃ] and [ʃ] respectively, when preceding /i/.

Nasalization

/b/ and /g/ are (optionally) realized as [m] and [n] respectively, following pause.

Reduction of constriction

/b/ and /g/ are (optionally) realized as the vibrants [b̃] and [ľ̦]. [b̃] varies with [b] in the environment, $\left\{\begin{matrix}/\text{i}/\\ /\text{a}/\end{matrix}\right\}$ –/o/ and [g] varies with [ľ̦] in /o/ – /i/.

4. Consonantal Tenseness (length)

In Grimes (1981), it is shown that voiceless consonants are longer than voiced consonants. In fact, a hierarchy of length exists in which voiceless stops are longest, followed by voiceless fricatives and, finally, by the voiced stops.

5. Male-Female Speech Distinctions

Phonetically, women's speech is marked by what might be described informally as a type of 'guttural posture' in which the walls of the pharynx are slightly constricted and occlusives are retracted in relation to their points of articulation in men's speech. Socially, women do not speak with outsiders, whereas men value acceptance by foreigners highly, even eliminating the 'offensive' phonetic segments [ľ̦] and [b̃] in their presence.

6. Description and Distribution of [ľ̦] and [b̃]

Let us examine more closely this elimination, or 'filtering out', of [ľ̦] and [b̃] in the presence of foreigners, beginning with an informal description of these segments.

[ľ̦] is produced with the tongue tip tapping the alveolar ridge continuing until it extends out of the mouth, with its sublaminal portion resting on the lower lip. In Everett (1982) I call this a (voiced) egressive apico-alveolar/sublamino-labial lateral flap. With regard to the present discussion, we should note that the visual impression resulting from the protrusion of the tongue tip from the mouth is quite strong.

[b̃] is a bilabial multiple vibrant (trill). The visual impression of this segment is also very unusual and quite obvious.

We have already described the phonetic distribution of [ľ̦] and [b̃] under 'reduction of constriction' above. As to their social distribution, as mentioned earlier, they do not appear in men's speech in the presence of foreigners. It is only as I have learned the Pirahã language and have been accepted by the Pirahã (they refer to me by the kinship terms *xàhàigí* 'brother' reserved exclusively for Pirahã) that I have observed these segments in men's speech. However, from the beginning of my fieldwork, I have observed these elements (indirectly) in women's speech. Pirahã men have subsequently explained to me that they only pronounce words 'in other ways' (i.e. use the variant forms [ľ̦] and [b̃] with me because I am 'one of them',

7. Questions

At first glance, at least four questions are raised by this pragmatic filtering of phonetic features. First, what is the relevant feature or conjunction of features which defines the class [ľ̦], [b̃] ? Is this purely phonetic or should nonphonetic features be allowed? It seems that in men's speech a purely phonetic feature, e.g. [± vibrant], is not sufficient to account for the restrictions mentioned since no such restrictions exist on the same elements in women's speech. What must be recognized is that it is precisely the reactions produced in non Pirahã by these segments which causes Pirahã men to eliminate them. That is, their elimination is based on their 'strangeness' in relation to Portuguese. Using a familiar sociolinguistic term, we migh label [ľ̦] and [b̃] [-superstrate] where 'superstrate' refers to segments possessed by the dominant (Brazilian) culture.

Such a decision raises another question. What then would be the nature of the relationship which obtains between the class of [-superstrate] elements and its environment? We can answer this simply by stating that in a [-familiar] environment (in which [-familiar] is a contextual feature), i.e. where foreigners are present, [-superstrate] elements are prohibited.

But granted this relationship between contextual features such as [± familiar] and sociophonetic classes described by the feature pair [± superstrate] the question remains of how to characterize the 'filter' or 'rule' involved. A likely answer is that something similar to the 'variable rules' of sociolinguistics is needed to represent the fact that once the speakers of Pirahã have been made aware of these subphonemic elements of their speech, they are able to *systematically* (cf. Everett, to appear) omit these features in the presence of outsiders. Such control demonstrates the importance of ethnographic factors at all levels, even the supposedly psychologically 'unreal' phonetic level (to reverse Sapir's terminology).

The implications of these phenomena raise and partially answer a fourth question, namely, what might such sociolinguistic filters have to tell us about the nature of phonological change? Part of the answer is clear. Change at the phonological level cannot be considered exclusively as a function of random idiolectical variations but also stems from cross cultural context in which an entire segment of the population can in fact cede to social pressure and regulate or modify its language at any level accordingly. I am aware of no published example as clear on this as Pirahã.

References

Everett, Daniel L. (1982). Phonetic Rarities in Pirahã. *Journal of the International Phonetics Association.* December.

Everett, Daniel L. (to appear). Dialogue and the Selection of Data for a Grammar. In: Marcelo Dascal (ed.) *Dialogue: an Interdisciplinary Approach.* Amsterdam: John Benjamins Publishers.

Grimes, Joseph (1981). Synthesis and Feedback in Field Linguistics. 101st meeting: Acoustical Society of America.

Sapir, Edward (1949). The Psychological Reality of Phonemes. In: Valerie B. Makkai (ed.) 1972. *Phonological Theory.* Lake Bluff: Jupiter Press.

Perceptual Description of Long-term Speaker-Characterizing Voice Features by means of Semantic Differential Ratings of Running Speech

W.P.F. Fagel and L.W.A. van Herpt
Amsterdam, the Netherlands

1. Introduction

For several reasons it is interesting to know which acoustic characteristics define our subjective impressions of running speech, so far as these impressions concern long-term speaker-characterizing voice features, or 'extra-linguistic voice features' according to the definition given by Laver and Trudgill (1979). These reasons vary from the need of more objective methods for the diagnosis of speech disturbances to the desirability of a reliable procedure for a phonetic description of dialects and sociolects.

If we want to analyse the relation between acoustic and perceptual features in an adequate way, we need an efficient and reliable instrument for collecting subjective ratings on speech in the first place. Our research is concerned with the development of such an instrument by means of the semantic differential technique (Osgood, Suci and Tannenbaum, 1957).

2. Rating Experiments

In Dutch, as in other languages, there are hundreds of adjectives that can be used to describe long-term speaker-characterizing voice features. These adjectives range from very general, like 'pleasant', 'slow' and 'powerful', to highly specialized like 'monophthongized', 'hyperfunctional' and 'breathy'. All these adjectives can be considered to represent potential criteria for judging a person's voice ('voice' being used here as including phonation *and* articulation).

However, such criteria differ strongly in reliability (consistency among judges when used for rating a speaker) and relevance (discriminability among speakers, proportion of variance in the speaker population that is accounted for). Apart from this, the criteria cannot a priori be considered to be independent of each other. On the contrary, it is clear that many criteria are strongly related to each other. This means that any arbitrary sample of descriptive adjectives selected to describe different speakers will yield more or less redundant information.

Therefore we conducted a number of experiments, the purpose of which was to construct a standardized procedure for perceptually describing a speaker's voice. We wanted this procedure to be based on the potential judgement criteria mentioned above, covering all relevant perceptual parameters in an optimally reliable and economical way.

For this reason, some 800 terms referring to special attributes of speech were collected first by our colleagues of the Institute. Various bipolar 7-point rating scales were obtained by pairing contrasting items from this collection. After several preliminary experiments a rating form consisting of 35 bipolar rating scales was put into use.

Ten speakers were subsequently judged on these rating scales by 235 listeners. The speakers were all native, normal speakers of Dutch. (One speaker had a rather husky voice.) We had our speakers read uniform material from typewritten texts, thus eliminating differences in vocabulary and grammatical accuracy.

The subjects who served as judges did not know or see the speakers but merely listened to their tape-recorded voices. Most listeners (211) were students from training courses of Speech Therapists (one group from Antwerp, Belgium and 6 groups from various regions in the Netherlands). The listeners varied as to degree of training (first to third year of training). The remaining 24 listeners were students of Dutch from the University of Amsterdam.

A more detailed discussion of this experiment and its results will be available elsewhere (Fagel, van Herpt and Boves, 1983).

3. Results

First, the rating scores were subjected to a scaling analysis based on Thurstone's Law of Categorical Judgement (Torgerson, 1958). This analysis was used to obtain information about the linearity of the scales, their discriminative power, the reliability of the scores and the extent to which the psychological continuum underlying the scales can be considered as unidimensional.

Subsequent factor analysis of the correlations between the 35 rating scales yielded 5 orthogonal factors. The first factor was interpreted as 'melodiousness' or 'variety' the second factor as 'articulation quality' and the third as 'static voice quality', strongly associated with perceived clarity or brightness as well as with subjective strength of voice. The last two factors were clearly associated with pitch and tempo.

At the University of Nymegen the same 35-scale rating instrument was used in an experiment where 6 male speakers were judged twice by 117 listeners. Factor analysis of the scores yielded virtually the same factor structure as we found in our perceptual data.

A subset of 12 carefully selected scales proved to be sufficient to generate this factor structure for both groups of speakers. The selection was based on the results of the foregoing scaling analysis and on such criteria as factorial purity and interrater reliability.

The factor structure we found appeared to be very stable over different groups of raters, suggesting strong validity of the perceptual dimensions involved. There is some evidence, though, that the third factor actually incorporates two different criteria, one 'clarity' criterion represented by items like 'husky--not husky' and 'dull--clear', and one 'potency' criterion, represented by 'powerful--weak' and 'loud--soft' for example. (The scale labels mentioned are, of course, only translations of the Dutch adjectives actually used in the experiments).

Since we want to check the stability of the factorial structure over different groups of speakers, a short rating form was designed on the basis of the results of our perceptual analysis. This rating form contains 14 bipolar rating scales, the 12 scales selected before plus two. Each of the 5 factors found in our study is represented by 2 rating scales, the third factor by four rating scales (two items for each criterion presumably confined in this factor). Although global evaluative ratings on the scales 'pleasant--unpleasant' and 'beautiful--ugly' are associated most strongly with the first factor, these ratings clearly cannot be predicted from speakers' scores on the first factor exclusively.

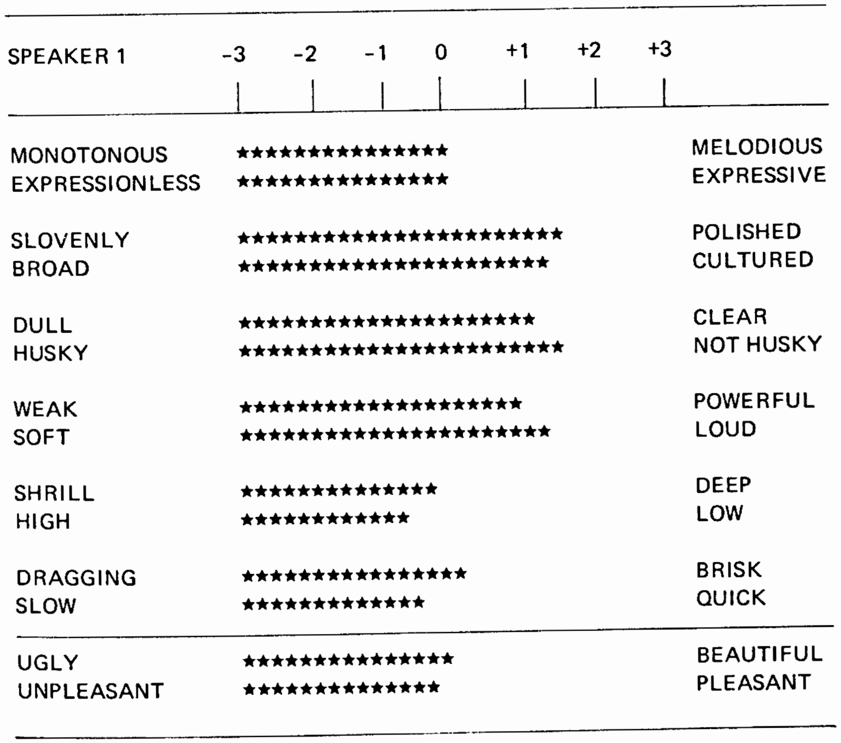


Figure 1. Speech profile for speaker 1 (male).

Therefore these two scales were adopted in our rating form as a separate component to keep the possibility of checking on the relations between global evaluative reactions and the judgements on the perceptual criteria we isolated. This short-hand rating form will soon be used in new rating experiments with larger groups of speakers.

Speech profiles can be composed from ratings on the 14 scales involved. Figure 1 shows such a profile for one of our speakers.

At least for the 10 speakers used in our experiments these speech profiles proved to be very reliable if based on the scores of 25 listener-judges or more. Most reliable were scores on the scales 'monotonous--melodious' and 'expressionless--expressive'. Ratings on the tempo scales 'dragging--brisk' and 'slow--quick' turned out to be least reliable, that is to say, listeners disagreed most on these scales. The new rating experiments mentioned above will have to confirm this reliability information over a larger set of speakers.

Acknowledgements

This research was made possible through a grant from the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (Z.W.O.), no. ST 17-21-13.

Further we would like to thank Loe Boves from the Institute of Phonetics at Nymegen with whom we have collaborated very closely in this research project.

References

Fagel W.P.F., van Herpt, L.W.A. and Boves, L. (1983). Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. Accepted for publication in *Speech Communication.*

Laver, J. and Trudgill P. (1979). Phonetic and linguistic markers in speech. In :K.R. Scherer and H. Giles (eds.), *Social Markers in speech.* University Press: Cambridge, 1-32.

Osgood, C.E., Suci, G.T. and Tannenbaum, P.H. (1957). *The Measurement of Meaning.* University of Illinois Press: Urbana.

Torgerson, W.S. (1958). *Theory and Methods of Scaling.* Wiley: New York.

Concentration and Diversification of Sound Changes

T. Matsushita
Gifu, Japan

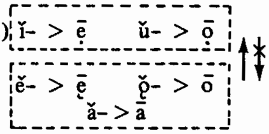
1. Introduction

At least two types of sound change are known to have taken place in stages: native changes and borrowing changes. A change in the rules of native phonology acting in a different direction from a change in loan phonology can increase the complexity of the phonetic system over the history of the language. This paper describes the direction of such sound changes using examples from the lengthening of native vowels in Middle English, Icelandic, and some High German dialects and of borrowings from French into English and from European languages into Japanese.

2. Historical evidence

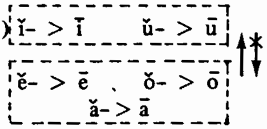
Vowel lengthening in open syllables took place over a wide area of Europe and can be regarded as movement toward a new paradigm. This process occurred in two stages in Middle English and Icelandic and in three in the High German dialects. The first stage in Middle English, the lengthening of non-high vowels, was complete by the middle of the 13th century. High vowels were later lengthened in Northern, North Midland, and East Anglian dialects.

(1) OE bă̄can > ME bāken 'bake'
OE mĕte > MR mę̄te 'meat'
OE þrŏte > ME þrǭte 'throat'
OE y̆fel > ME ę̄vel 'evil'
OE sŭmor > ME sọ̄mer 'summer'

(2)


The lengthening process in Icelandic took place in the 16th century in two stages with high vowel lengthening following that of non-high vowels, but not *vice versa*.

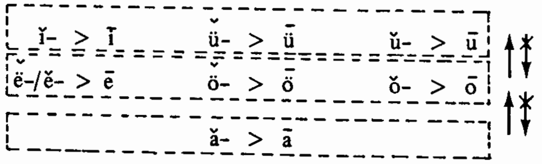
(3) ON sălr > ModIce. sālur 'hall'
ON gĕta > ModIce. gēta 'to get'
ON lĭfa > ModIce. līfa 'to live'
(Sommerfelt 1962: 82-3)

(4)


Though the outputs for these two languages differ somewhat in phonetic detail, the order of application to the inputs is the same; lengthening applied first to the non-high vowels and then to the high ones.

A similar lengthening took place in certain High German dialects (e.g. Lower Alemmanic) in the 14th century, but required three distinct stages – lengthening of first low, then mid, and finally high vowels.

(5)
ă- > ā Adel 'nobility', tragen 'bear', Wagen 'wagon'
ë̌- > ē kleben 'adhere', Rebe 'vine', weben 'weave', Leder 'leather'
ĕ- > ē heben 'heave', edele 'noble', legen 'lay'
ĭ- > ī Biber 'beaver', biben 'shake', wider 'against'
ŏ- > ō loben 'praise', Boden 'ground', Vogel 'bird'
ö̌- > ȫ ölen 'oil', Öler 'Ölmüller', Vögellin
ŭ- > ū Jugend 'youth', Tugend 'virtue', Jude 'Jew'
ǚ- > ǖ Jüdel 'Judenkind', mügen 'may'
(Eis 1950: 64-5)

(6)


In all three cases, the rule of native phonology which converted short, stressed vowels in open syllables to long ones expanded its scope in a definite order: from only non-high vowels at the first stage (German required an intermediate stage for mid vowels) to all vowels at the final stage. Once some members of a class underwent a change, the rest tended to follow by analogy even though they possessed less phonetic motivation to submit to the change.

3. Evidence from loans

The situation is, however, different in loan phonology. Initially, the manner of borrowing tends to be uniquely specified; that is, all members of a certain class undergo the same change when borrowed. This class may later be subdivided, permitting a plurality of changes, or later borrowings may adopt a different rule.

Perhaps the most well-documented example is French loanwords in English. Stressed vowels in open syllables were lengthened in the words that entered Middle English from Old or Norman French.

(7)
a; cape, plate(s), maten, cave, male, fame, blame
e; degre 'degree', cete 'whale', chere 'cheer'
i; deliten 'delight', quiten 'requite', arive(d) 'arrive(d)'
o; note, robe, trone, noble, close, cote 'coat'
u; dute 'doubt', rute 'route', spuse 'espouse', gout
ü sputin 'dispute', escusen 'excuse', üse(d) 'use(d)'
(Based on Behrens 1886)

(8)
OF ĭ- > ME ī OF ǚ- > ME ǖ OF ŭ- > ME ū
OF ĕ- > ME ē OF ŏ- > ME ō
OF ă- > ME ā

Later borrowings from Modern French into Modern English continued to do so – whenever possible – but added on- or off-glides to make up for the loss or change of some of the Middle English vowels.

(9)
a; façade [fə*sá*:d, fæ–; F. fasad]
e; bouquet [A, bu(:)kéi, _́ _/ B. bəukéi; F. bukɛ]
i; fatigue [fətí:g; F. fatig]
o; bureau [bjúərou], _ _́ ; F. byro]
u; recoup [rikú:p; F. *recouper* rəkupe]
ü; perdu [pə:djú:; F. pɛrdy]
(Skeat 1891, 173-6)

(10)
F. ĭ- > E. ī F. ǚ- > E. jū F. ŭ- > E. ū
F. ĕ-/ɛ̆- > E. ei F. ŏ-/ɔ̆- > E. ou
F. ă- > E. *ā*

The initial rule was simple – 'lengthen stressed vowels in open syllables' – but intervening developments worked against this rule and forced a restructuring. The net effect was that the loan phonology became more complex, not simpler.

(11)
ModF > ModE
OF > ME
Vowel lengthening ⇄ { i) Vowel lengthening
ii) Diphtongization
iii) Vowel lengthening with glide-prefixation

The same thing happened with the epenthetic vowels Japanese inserts between the consonants in Dutch, Portuguese, and English loanwords to make them conform to its phonetic paradigm of open syllables – schematically,

$$\# C_1LV_1C_2\ (C_3\ \ldots) > \# C_1V_2rV_1^{(1)}C_2V_3\ (C_3\ \ldots)$$

where C_1 denotes the velar and palato-dental stops (*k*, *g*, *t*, and *d*) and L, liquids (Japanese has only *r*). Historically, there are three distinct stages. 16th century loanwords (from Portuguese) have both forward and backward assimilation of the epenthetic vowels to the original V_1 (or its Japanese equivalent V_1').

(12) Po. Christo > J. k*i*ris*i*to 'Christ' 1591
Po. Christão > J. k*i*ris*i*tan 'Christian' 1587 (Arakawa 1967)

Borrowings between the middle of the 18th century and the middle of the next (mostly from Dutch and Portuguese) have only backward assimilation of the first epenthetic vowel; the quality of the second one is determined, not by forward assimilation, but by the preceding consonant (typically, *i* or *u*, but *o* after *t* or *d*).

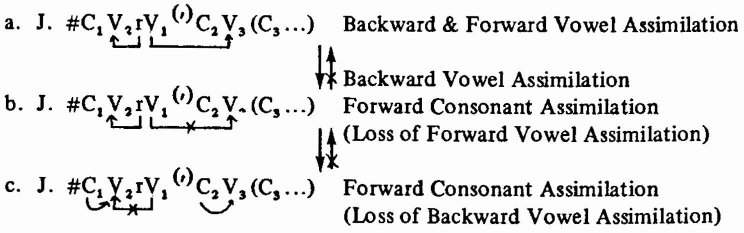
(13) Po. Christo > J. k*i*ris*u*to 1600
Po. Christão, Du. Christen > J. k*i*ris*u*tean 1713 (Saito 1967)

In the most recent period, all such vowel assimilation has disappeared in favour of consonant assimilation.

(14) E. (Jesus) Christ > J. (ziizasu) k*u*rais*u*to
E. Christian > J. k*u*ris*u*tyan

Diagramatically,

(15) Po./Du./E. $\#C_1LV_1C_2(C_3\ldots)>$



a. J. $\#C_1V_2rV_1{}^{(r)}C_2V_3(C_3\ldots)$ Backward & Forward Vowel Assimilation

Backward Vowel Assimilation

b. J. $\#C_1V_2rV_1{}^{(r)}C_2V_{\wedge}(C_3\ldots)$ Forward Consonant Assimilation

(Loss of Forward Vowel Assimilation)

c. J. $\#C_1V_2rV_1{}^{(r)}C_2V_3(C_3\ldots)$ Forward Consonant Assimilation
(Loss of Backward Vowel Assimilation)

4. Conclusion

Though the manner of borrowing into Japanese was once uniquely determined, drastic processes like Forward and Backward Vowel Assimilation of epenthetic vowels have gradually given way to Forward Consonant Assimilation, which represents a weaker process, but a more complicated rule.

Thus, native phonology and loan phonology seem to differ in the power of analogy to promote rule generalization. The examples cited suggested that, for native phonology, once some members of a class submit to a change, analogy works to include those with less phonetic motivation, while loan phonology has the option of restructuring borrowed classes in ways not always predictable by analogy.

References

Arakawa, Soubei (1967). *Gairaigo Jiten* (Dictionary of Loanwords in Japanese). Tokyo: Kadokawa Shoten.

Behrens, Dietrich (1886). Beiträge zur Geschichte der französischen Sprache in England. *Französische Studien V.* 2. Leipzig: O.R. Reisland.

Eis, Gerhard (1950). *Historische Laut- und Formenlehre des Mittelhochdeutschen.* Heidelberg: Carl Winter.

Saito, Shizuka (1967). *Nihongo-ni Oyoboshita Orandago-no Eikyo* (Influence of Dutch on Japanese). Tokyo: Shinozaki Shorin.

Skeat, Walter W. (1891). *Principles of English Etymology. Second Series: The Foreign Element.* London: Oxford University Press.

Sommerfelt, Alf (1962). *The Development of Quantity as Evidence of Western European Linguistic Interpretation. Diachronic and Synchronic Aspects of Language.* The Hague: Mouton, 81-86.

Stylistic Variation in R.P.

S. Ramsaran
London, England

1. Aim

The best-known works on stylistic variation in phonology have been concerned with accents which are susceptible to influence from outside. Typically, a regional accent is modified in a formal situation to display characteristics of a prestige norm. My research was designed to discover what happens as formality varies in English in the case of R.P., a non-regional accent which would not be expected to undergo modification towards some external norm.

2. Subjects and speech context

The subjects whose speech was studied formed a linguistically homogeneous group. All were educated beyond secondary level; none had speech which showed any features of identifiable regional accents. In order to ensure that the variation observed would be strictly confined within the single accent, the subjects were recorded in conversation with interlocutors whose accent was also R.P. Since the basis of the investigation was empirical, there was no elicitation and no experimenter was present (as is the case in studies by Labov (1966) and Trudgill (1974), for instance). The presence of an investigator surely introduces an additional variable; and if the situations are to be controlled, there must be as few variables as possible.

Clearly the context which affects style is partly linguistic and partly non-linguistic. If the interlocutors are unknown *or* feel a lack of common ground, the social situation is a formal one. This is the type of non-linguistic context which gives rise to *formal* speech. If the interlocutors know each other *and* feel that they share common ground, the social situation is a casual one. This is the type of non-linguistic context which gives rise to *mixed* speech. So, in the text from the casual situation in particular, it is useful to investigate the variation caused by *linguistic* context. This was done by carrying out a close-context analysis of the type suggested by Joos (1968).

3. Method

There were six primary subjects, two men and four women, whose spontaneous conversation was recorded in casual and formal situations. To achieve the latter situation, two strangers would be introduced socially and then left alone together. At that stage they were unaware that their speech was being recorded and were solely concerned with making polite conversation. The same speakers were recorded in casual situations in familiar surroundings talking to sister, wife or friend. In these cases the subjects knew that they were being recorded but the first ten minutes of each recording were discarded and analysis never started until channel cues (see Milroy and Milroy 1977) indicated that the participants had relaxed into natural casual conversation. Each recording lasted for something between half an hour and an hour. One, in which the subject wore a radio microphone transmitting to a recorder elsewhere in the building, lasted for six hours. Altogether about twenty hours of speech were analysed, including some close-context analysis in the case of the radio-transmitted recording. A total of just over one and a quarter hours of these texts was transcribed narrowly for a detailed analysis of segmental and prosodic features.

Three of the subjects were recorded in monologue (broadcast talk, sermon and seminar paper) but this was discovered to belong to a different genre from conversation, as is maintained by Abercrombie (1963) when he talks of spoken prose. Monologue does *not* represent the most formal end of a speech formality cline. When, therefore, formal and casual speech are contrasted here, what is referred to is the spontaneous conversation of formal and casual situations.

4. Phoneticians' assumptions

It was expected that the data would illustrate textbook statements concerning the increased frequency of weak forms, linking /r/, assimilations and elisions as the speech becomes more casual. It was also expected that the occurrence of these features would increase with pace.

5. Analysis

The majority of *pauses* in casual speech are unfilled and occur at clause boundaries. Formal speech shows an increase of filled and mid-clause pauses.

Intonation patterns show no marked distribution in casual speech. Formal speech shows some concentrations of one pattern (perhaps a stretch of several tone units with fall-rise nuclear tones or a stretch with rising nuclear tones). Casual speech has more level nuclear tones than formal speech. Casual speech has longer tone units (with longer preheads and tails) and contains fewer stressed syllables than formal speech.

Owing to that fact, *weak forms* may be replaced by strong forms more often in formal speech. The alternation of strong and weak forms, however, is entirely regular in both styles: weak forms occur unless the grammatical word is stressed. Since stresses are more frequent in the tone units of formal speech strong forms may occur more often.

Linking /r/ is frequent in all styles of speech. Its occurrence is of no stylistic significance.

Elisions (which increase with pace) are less rule-bound in casual speech than in formal speech where they are almost entirely regular (e.g. alveolar plosives interconsonantally, initial /h/ in unstressed non-initial pronouns etc.). Casual speech contains unpredictable elisions, such as those of /l/ and /ð/ in 'Well that's all right because you...' : [we˞ æts ɔ:ˏɹaIʔ/ bIkəʒu: ...]

Assimilation shows no correlation with pace. It is more frequent in casual speech than in more formal speech.

Pace fluctuates more in casual speech than in the most formal speech. Since pace is relevant to the occurrence of elisions, perhaps the rates that were measured should be mentioned: the slowest pace measured was 89 sylls/min (3.1 sylls/sec or 7.6 segs/sec) and the fastest was 324 sylls/min (5.4 sylls/sec or 13.4 segs/sec).

Plosive release was one of the most stylistically significant variables. Casual speech had more inaudible releases than formal speech.

The types of figures involved are as follows: the speech of three speakers was examined. Of the 431 voiceless plosives occurring either prepausally or in a two-stop cluster, 395 were inaudibly released, i.e. 'usually' something like 8% of plosives are audibly released. When, then, it is found that 38% of plosives are audibly released in a formal text, it would be fair to say that this feature is marked for style.

6. Summary

No feature is unique to a variety though some features may be distributionally marked. Since assimilations, elisions and linking /r/ occur frequently in all situations whether formal or casual, it seems even more important than has hitherto been believed that foreign speakers of English should use these features whenever they speak English. Contrary to expectations, of these three features it is only elision that bears any definite relation to pace. Nor is there any clear relationship between pace and the degree of the formality of the situation in which the speech occurs.

These observations lead to the conclusion that in R.P. there is no shift from one distinct style to another but that gradual variation is to be seen as inherent in a unitary system.

References

Abercrombie, D. (1963). Conversation and spoken prose. *English Language Teaching* **18.1**: 10-16.

Joos, M. (1968). The isolation of styles. In: Fishman, J.A. (ed.) *Readings in the Sociology of Language*. The Hague: Mouton, pp. 185-91.

Labov, W. (1966). *The Social Stratification of English in New York City.* Washington D.C.: Center for Applied Linguistics.

Milroy, L. and Milroy J. (1977). Speech and Context in an Urban Setting. *Belfast Working Papers in Language and Linguistics*, **2: 1**. Belfast: Northern Ireland Polytechnic.

Trudgill, P. (1974). *The Social Differentiation of English in Norwich.* Cambridge: Cambridge University Press.

The Relative Importance of Vocal Speech Parameters for the Discrimination of Emotions

R. van Bezooijen and L. Boves
Nijmegen, the Netherlands

1. Introduction

In the experimental literature on vocal expressions of emotion two quite independent mainstreams may be distinguished, namely research which focuses on the description of emotional expressions in order to establish the characteristic features of distinct emotional categories, and research which examines the extent to which subjects are able to distinguish emotions from one another when vocally expressed. In the present contribution an effort is made to link the two approaches by comparing the results of a multiple discriminant analysis based on auditory ratings of vocally expressed emotions to the results of a multidimensional scaling analysis based on the outcome of a recognition experiment in which the same emotional expressions were used as stimuli. The aim of the comparison was to gain insight into the relative importance of various vocal speech parameters for the discrimination of emotions by human subjects.

2. Method

2.1. Speech material

The data base comprised 160 emotional expressions, namely 8 (4 male and 4 female) speakers × 2 phrases × 10 emotions. The speakers were students, native speakers of Dutch, between 20 and 26 years of age. None of them had had any professional training in acting. The phrases were 'two months pregnant' (/tve: ma:ndə zvɒŋər/) and 'such a big American car' (/zo:n γro:tə ɑmerika:nsə o:to:/). The emotions were the nine emotions included in the emotion theory developed by Izard (1971), i.e., disgust, surprise, shame, interest, joy, fear, contempt, sadness, and anger, plus a neutral category.

2.2. Auditory ratings

The recordings were randomized per speaker and per phrase, and rated by six slightly trained judges on 13 vocal parameters, i.e., pitch level, pitch range, loudness/effort, tempo, precision of articulation, laryngeal tension, laryngeal laxness, lip rounding, lip spreading, creak, harshness, tremulousness, and whisper. To collect the ratings use was made of preprinted successive interval scales. Every utterance could be given only a single rating on each scale, which means that the scores represent some kind of perceptual average. All scales were considered as absolute except the pitch level scale, which was effectively split up into separate scales for male and female speakers. The scales of lip rounding and lip spreading were mutually exclusive, i.e., only one of the two scales could be rated at any one time. The same was the case for laryngeal tension and laryngeal laxness.

2.3. Recognition experiment

The same 160 stimuli which were auditorily described - now randomized separately for male and female speakers - were offered to 24 male and 24 female students attending a Teachers' Training College in Nijmegen. The mean age was 20 years and 6 months, ranging from 18 to 28 years. The subjects were seated in a language laboratory and listened to the recordings via headphones. They were asked to indicate on a rating sheet which out of the ten emotional labels fitted best each of the emotional portrayals they heard.

3. Results and discussion

First, we will present the outcomes of a multiple discriminant analysis based on the auditory ratings. Next, the outcome of a multidimensional scaling analysis based on the results of the recognition experiment will be given. Finally, the results of the two analyses will be compared.

Before the auditory scores were subjected to a multiple discriminant analysis, a number of statistical analyses were carried out in order to make sure that the ratings were reliable, and that the various parameters were relevant to the aim of the study and informative.

The reliability of the means of the scores was assessed by means of the so-called Ebel-coefficient (Winer, 1971, p. 286). It appeared that the coefficients ranged from .86 for precision of articulation to .95 for pitch level, values which may be considered to be satisfactorily high. Therefore, in the subsequent analyses use was made of the mean of the ratings of the six transcribers on each of the 13 parameter scales for each of the 160 stimuli. In order to assess whether the scores on the scales varied as a function of the emotions expressed, the scores on each of the scales were subjected to separate analyses of variance with three fixed factors each, namely sex of speaker, phrase, and emotion (level of significance = 5%). It appeared that the factor sex of speaker was significant for only one scale, and that the factor phrase was significant for only two scales. However, the effect of the factor emotion was significant for all scales, except for lip rounding. Since we were not interested in a parameter which apparently had not been systematically used to differentiate between emotions, lip rounding was excluded from the further analyses.

In addition, product-moment correlations were computed in order to examine how the different parameters were related. The highest correlation, that between loudness/effort and laryngeal tension, was .71, which means that only half of the variance in one parameter was accounted for by the other. The correlations among the rest of the variables were considerably lower. On the basis of these results we decided not to discard any more parameters.

The mean scores for each of the 160 emotional utterances on each of the 12 parameter scales retained were used as input for a multiple discriminant analysis. A stepwise procedure was chosen, i.e., a method in which the discriminating variables are selected for entry into the analysis on the basis of their discriminating power. The objective of the analysis was to attain an optimal separation of the ten groups of 16 utterances per emotion by constructing a restricted set of discriminant functions or dimensions which are linear combinations of the original variables.

With three dimensions, the first one accounting for 41% the second one for 22%, and the third one for 18% of the variance in the original variables, 62.5% of the 160 utterances were assigned to the right emotional category. This is only 4.5% less than the percentage of correct responses yielded by the recognition experiment with human subjects. With 10 categories to choose from, which corresponds with an accuracy to be expected by chance of 10%, this means that the accuracy of both the statistical and human classification are well beyond chance expectation. In Figure 1 the positions of the group centroids in the discriminant space spanned by the first and second dimensions are presented.

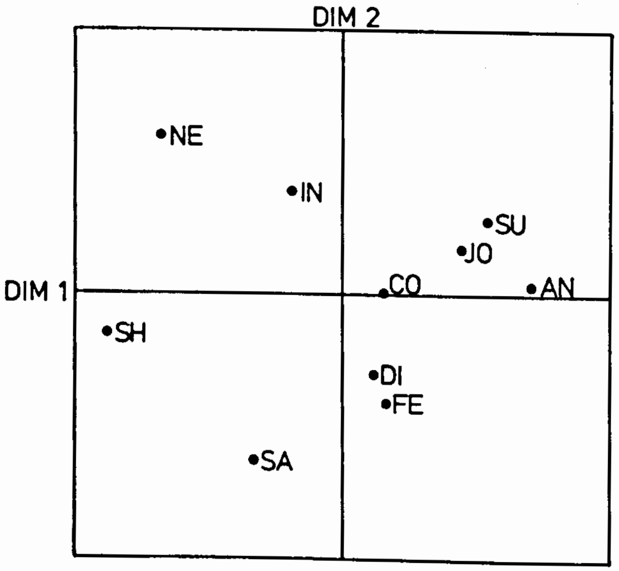


Figure 1. Positions of ten emotions in a two-dimensional space resulting from a multiple discriminant analysis on auditory ratings.

The positions of the different emotions along the first dimension presented in Figure 1 clearly suggest this dimension to be called a dimension of activation, shame, neutral, and sadness being the most passive emotions and joy, surprise, and anger being the most active ones. An interpretation in terms of level of activation is made even more plausible by the fact that the three parameters which correlate most highly with the first dimension are loudness/effort, laryngeal tension, and laryngeal laxness (r = .73, .66 and .54, respectively df = 158, $p < .001$). The only other correlation worth mentioning is that of pitch range with r = .35 (df = 158, $p < .001$).

The second dimension presented in Figure 1 is more difficult to name. If it were not for the extreme position of neutral and the neutral position of anger, it could be interpreted as an evaluative dimension, having unpleasant emotions like sadness and fear at its negative end and the pleasant emotions of joy and interest at the positive end. Another interpretation is suggested when we note that the positive end of the dimension is characterized by the least 'emotional' and at any rate the least impairing emotions, whereas the negative end shows a concentration of highly inhibiting emotions. The second dimension did not correlate very highly with any of the discriminating variables. The highest correlations were those with laryngeal laxness, pitch range, creak, and loudness (r = .49, .44, -.36, and .31, respectively, df = 158, p = < .001). The combination of much creak, high tension, low loudness, and narrow pitch range fits in with our previous interpretation of the second dimension in terms of emotional inhibition.

The third dimension extracted in the discriminant analysis is not considered here since we did not succeed in giving it a meaningful interpretation. It had disgust at the negative extreme, fear at the positive extreme, and neutral somewhere in the middle. It correlated most highly and almost exclusively with pitch level and tempo (r = .74 and .43, respectively, df = 158, $p < .001$).

Thus, from the distribution of the emotions along the three discriminant functions, and also from the correlations between these functions and the 12 vocal speech parameters, it appears that statistically speaking the most powerful variables for the separation of emotions are those which are related to level of activation. Are there any indications that the same would hold for the discrimination among emotions by human subjects?

In order to assess whether level of activation was an important criterion in the classification of emotional utterances by human subjects as well, the confusion data resulting from the recognition experiment were subjected to multidimensional scaling. Multidimensional scaling is a dataprocessing technique which has been especially designed to represent similarity between objects - and confusions among emotions could be interpreted as such - in terms of proximity in an n-dimensional space, thereby providing insight into the nature of the dimensions or stimulus characteristics that underlie the similarity judgments that subjects have emitted. In our case, the confusion data were processed by means of the MINISSA- program developed by E. Roskam and M. Raaijmakers from the University of Nijmegen, Holland, and

J. Lingoes from the University of Michigan, USA. Since asymmetrical input data were not permitted, the confusion data were first made symmetrical (Klein, Plomp, and Pols, 1970).

The combination of the Euclidean metric and two dimensions resulted in a stress of .11, which, according to Wagenaar and Padmos (1971), is significant at the 5% level. In order to enhance the comparability with the discriminant functions, the two axes resulting from the multidimensional scaling were rotated orthogonally in such a way as to maximally approach the positioning of the group centroids in the discriminant space as depicted in Figure 1. In Figure 2 the resulting configuration is presented.

4. Conclusion

Comparison of the two-dimensional spaces presented in Figures 1 and 2 shows that the projections of the ten emotions on the first dimension are quite similar: in both configurations neutral, shame, and sadness are situated towards the negative end; disgust, contempt, and fear towards the middle; and joy, surprise, and anger towards the positive end. The only notable level of activation has been one of the main stimulus characteristics on which the subjects have based their categorization decisions. In general the dimension of level of activation plays indeed a central role in the discrimination among emotional expressions. This also appears from the fact that in a recognition experiment in which groups of Taiwanese and Japanese adults judged the same emotional utterances, the same dimension emerged as well. On the basis of this outcome it could be hypothesized that the vocal parameters related to level of activation, i.e., loudness/effort, laryngeal tension, laryngeal laxness, and pitch range are not only important for the separation of emotional expressions in a statistical sense, but also in connection with the classificatory behavior of human subjects.

On the other hand, the projections of the ten emotions on the second dimensions presented in Figures 1 and 2 are very different from one another. Not much can be said, therefore, about the relative importance of the other vocal speech parameters for the discrimination amongst emotional expressions by human subjects.

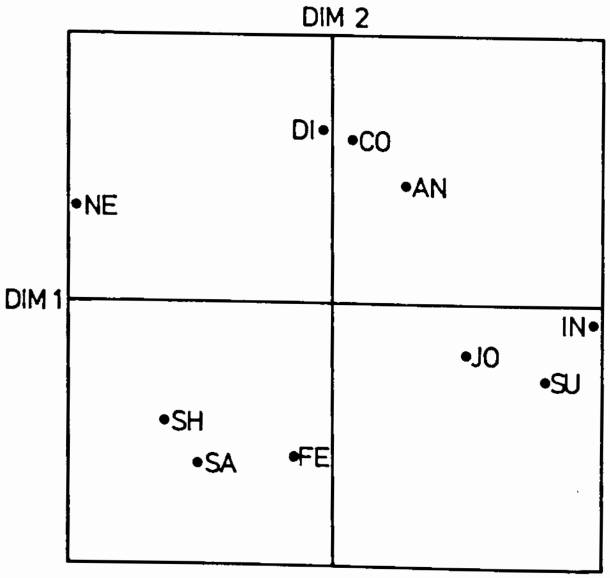


Figure 2. Positions of ten emotions in a two-dimensional space resulting from multidimensional scaling on recognition scores.

References

Izard, C.E. (1971). *The face of emotion.* New York: Appleton Century Crofts.

Klein, W., Plomp, R., and Pols, L.C.W. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America,* **48**, 999-1009.

Wagenaar, W.A. and Padmos, P. (1971). Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology,* **24**, 101-110.

Winer, B.J. (1967). Statistical principles in experimental design. *Psychometrica,* **32**, 241-254.

A Cross-Dialect Study of Vowel Perception in Standard Indonesian

E. Van Zanten and V.J. Van Heuven
Leyden, the Netherlands

1. Introduction

Shortly after World War II Indonesia gained its independence. One of the governmental acts was to impose one variety of Malay as the official standard language, Bahasa Indonesia, throughout the archipelago. As a result the Indonesians now speak the standard language (almost as a second language) as well as a local vernacular, which though obviously related to the standard language may differ from it in many respects. As a case in point consider the vowel systems of the three vernaculars that are dealt with in the present paper:

Toba Batak (5)	Javanese (6)	Sundanese (7)
i u	i u	i y u
e o	e ə o	e ə o
a	a	a

The 6-vowel system of Standard Indonesian (disregarding diphthongs) is equal to that of Javanese. Toba Batak lacks the central mid vowel, whereas Sundanese has two central vowels, viz. one mid and one high(er).

In our study of the Indonesian vowel system we are interested in the acoustic *and* perceptual properties of the monophthongs, and the possible influence of the regional substrate on the subjects' performance. The present paper is confined to a perceptual experiment only.

Our method is based on early work by Cohen, Slis and 't Hart (1963) on the perceptual tolerances of Dutch vowels, which was later successfully extended to contrastive vowel studies by e.g. Schouten (1975) and Hombert (1979). In these latter studies subjects were presented a large number of synthesized isolated vowel sounds, regularly sampled from a vowel space essentially defined by F_1 and F_2. Their task was to label each vowel sound in terms of one of the vowels of their language. This method proved sensitive enough to reveal differences between the internal representation of British-English vowels of native speakers and that of advanced Dutch learners of English (Schouten, 1975). Hombert (1979) successfully applied the method to the description of the vowel systems of a number of - strongly related - African Bantu languages. The issue at stake in our present study is whether the labelling method will also reliably reflect differences between the vowel systems of speakers of a standard language with different dialect backgrounds.

2. Method

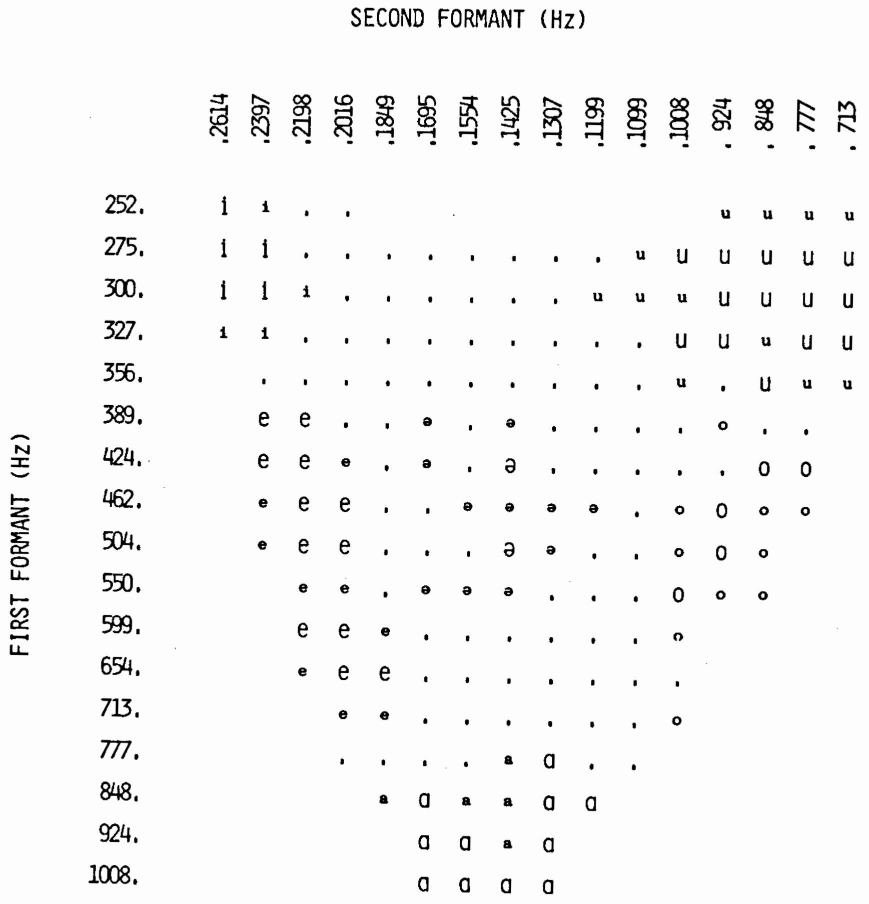
On the basis of an acoustic pilot study (Van Zanten and Van Heuven, 1982), realistic formant ranges were defined for Indonesian vowels spoken in isolation. A set of 188 monophthongs were then produced with a Fonema OVE IIIb speech synthesizer whose parameter values were controlled by a DEC PDP11/03 computer. All vowels were given a 350 ms duration including linear onset and offset portions (in dB) of 50 and 100 ms, respectively. During the steady state portion voicing was set at maximum intensity. F_1 and F_2 were systematically varied in steps of 9%, i.e. 3 times the Just Noticeable Difference commonly reported for F_1 and F_2 centre frequency changes (Flanagan, 1955; Mermelstein, 1978; Nord and Sventelius, 1979), sampling the acoustic vowel space in the way indicated in Fig. 1. F_4 and F_5 centre frequencies were set at 3500 and 4000 Hz, respectively, for all vowels; bandwidths B1-B3 were set at mid-range values. The centre frequency of F_3 equalled that of $F_2 + 600\ H_2$, with a minimum of 2460 Hz. Two tapes were prepared containing the set of 188 stimuli, preceded by a series of practice items, in counterbalanced random orders.

Three groups of Indonesians with different regional backgrounds participated in the experiment: 4 Toba Batak, 5 Javanese, and 4 Sundanese listeners. All the subjects had completed a university education in their own country, and had only recently arrived in the Netherlands to enroll in a postgraduate program at the University of Leyden. They participated on a voluntary basis, and were paid for their services.

Subjects were instructed to label each vowel stimulus as one of the six monophthongs of Standard Indonesian (forced choice), and to rate each token along a scale of acceptability: 1 (good), 2 (poor, but easily identifiable), 3 (unacceptable and hardly identifiable).

3. Results

After having tried out various weighting procedures, the responses were finally analysed such that 'good' tokens counted twice, 'poor' tokens once, and 'unacceptable' tokens were eliminated altogether. Figure 1 plots the data for the three listeners groups (panel A: Toba Batak; panel B: Javanese; panel C: Sundanese). Areas of preference were defined containing only those stimulus points that were identified as one particular vowel in at least 50% of the responses (small letters), and in at least 75% (large letters). To facilitate the exposition, summary statistics are given in Table I, specifying the number of stimulus points (absolute and relative) contained by each eara of preference.

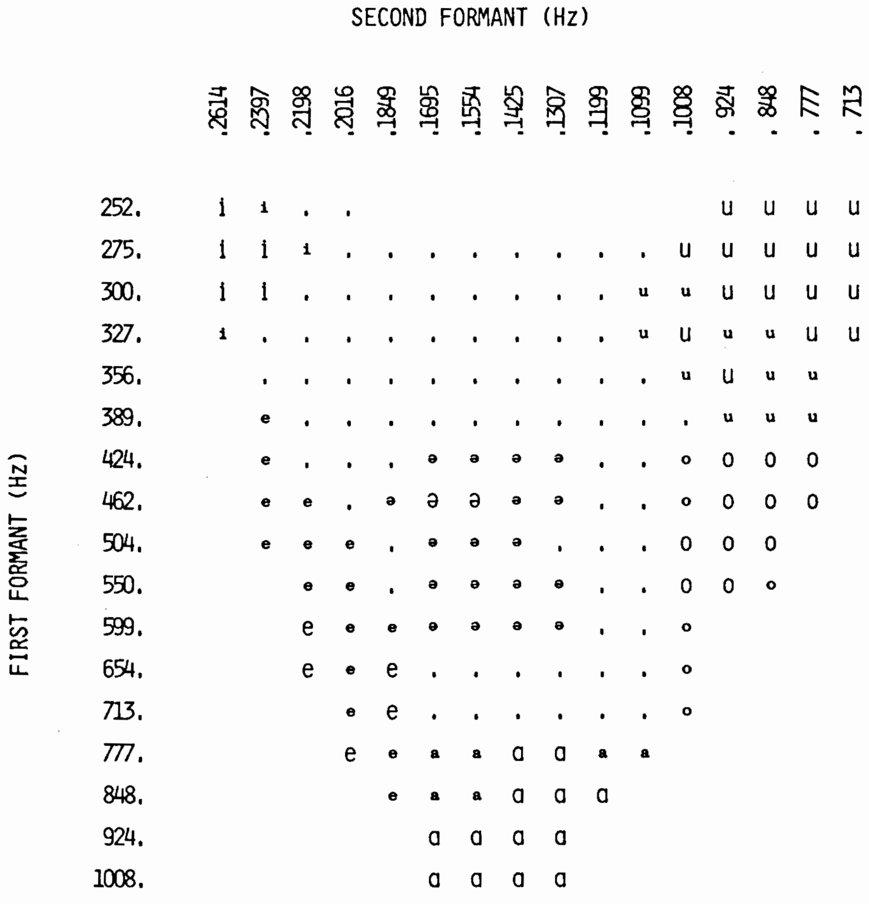


PANEL A: TOBA BATAK LISTENERS

Figure 1a. Distribution of responses to labelling test for three groups of listeners. Panel A: Toba Batak, N = 4, as a function of F_1 and F_2.

4. Conclusions and Discussion

There are several conspicuous differences across the three listener groups in the locations and sizes of the preferred vowel areas, specifically in the way the central region of the vowel space is divided over the competing vowel phonemes. Typically, the /ə/ area is small for the Toba Batak group, intermediate for the Javanese, and largest for the Sundanese. Conversely, the area associated with /u/ is large for the Toba Bataks, intermediate for the Javanese, and smallest for the Sundanese.



PANEL B: JAVANESE LISTENERS

Figure 1b. Distribution of responses to labelling test for three groups of listeners. Panel B: Javanese, N = 5, as a function of F_1 and F_2.

These differences in the distribution of the responses obviously reflect properties of the subjects' regional substrate languages. Remember that the Toba Batak dialect has no central vowel, which explains why /ə/ is the least favoured response category for the Bataks. Also, its area of dispersion is highly irregular, and only 2 out of 188 stimulus points are identified as /ə/ in more than 75% of the responses.

For listeners with a Javanese background, an /ə/ dialect, the preferred area for /ə/ is appreciably larger, and the responses are more regularly distributed. For the Sundanese group, having a background dialect with two central vowels, the preferred /ə/ area is larger still, and, perhaps more

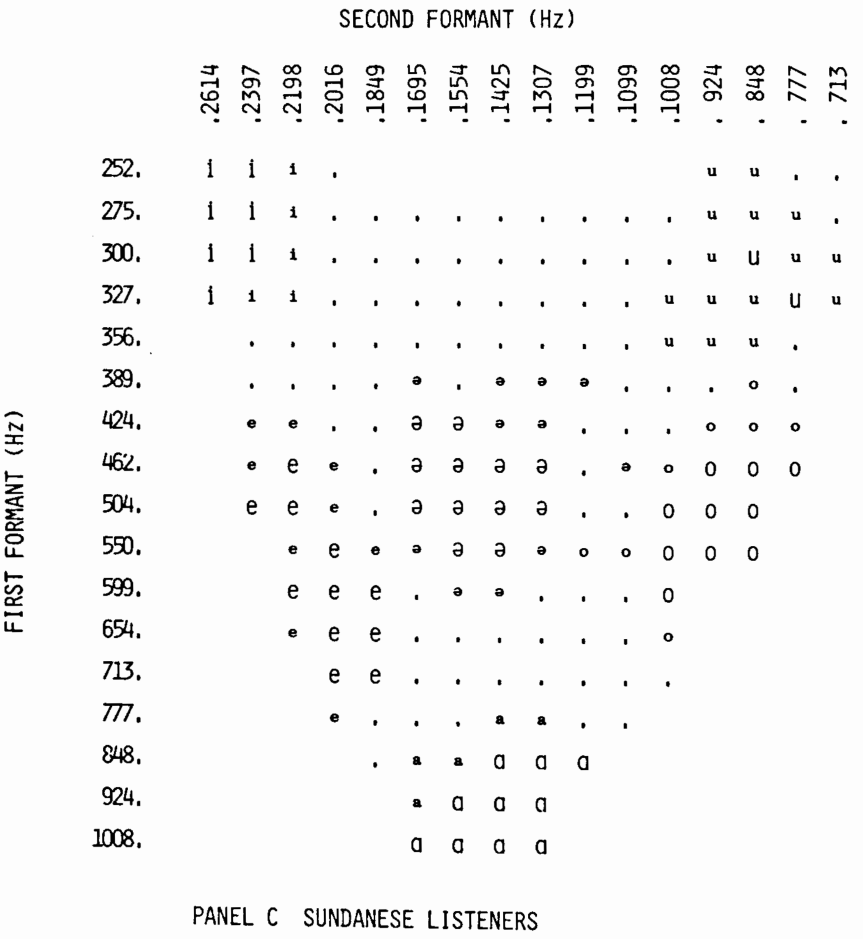


Figure 1c. Distribution of responses to labelling test for three groups of listeners. Panel C: Sundanese, N = 4), as a function of F_1 and F_2. Large letters represent stimuli identified as indicated by the phonetic symbol in at least 75% of the responses (after weighting, see text); small letters represent vowels identified with at least 50% agreement.

Table I. Number of stimulus points identified as one particular vowel in at least 50% (75%) of the responses, expressed absolutely and relatively, per vowel per group of subjects (4 Toba Bataks, 5 Javanese, 4 Sundanese)

Vowel	50% agreement			75% agreement		
	Bataks	Javanese	Sundanese	Bataks	Javanese	Sundanese
/i/	9 (5%)	8 (4%)	12 (6%)	5 (3%)	5 (3%)	7 (4%)
/e/	22 (12%)	20 (11%)	20 (11%)	12 (6%)	5 (3%)	11 (6%)
/a/	15 (8%)	19 (10%)	15 (8%)	11 (6%)	13 (7%)	11 (6%)
/o/	15 (8%)	20 (11%)	18 (10%)	5 (3%)	11 (6%)	10 (5%)
/u/	25 (13%)	25 (13%)	17 (9%)	14 (7%)	17 (9%)	2 (1%)
/ə/	13 (7%)	20 (11%)	23 (12%)	2 (1%)	2 (1%)	12 (6%)
unlabelled	89 (47%)	76 (40%)	83 (44%)	139 (74%)	135 (72%)	135 (72%)

importantly, here the distribution of especially /u/ is much more restricted: only 17 stimulus points are identified as /u/ with more than 50% agreement (against 25 for the other dialect groups), and only 2 with more than 75% (agaihst 14 and 17 for the Batak and Javanese listeners respectively). Presumably, the high(er) central vowel (which was not a response option open to the subjects) 'pushes back the /u/ boundary', i.e. precludes /u/ responses to stimuli with F_2 values larger than 1000 Hz, whereas the preferred /u/ area extends to 1100 Hz for listeners with a Javanese background (1 neutral vowel), and even 1200 Hz for Toba Batak listeners (no central vowel).

As a final observation we would like to point out a difference in overall performance on the part of the Javanese, as opposed to the other groups: it appears that the properties of the stimulus points that could not be adequately identified (i.e. with at least 50% agreement) tends to be smaller for the Javanese (40%) than for either the Bataks (47%) or the Sundanese (44%).

Summing up then, we have shown that the perceptual method of charting a vowel system proved sensitive enough to reflect influences of the regional substrates of listeners when asked to identify vowels in terms of the categories given by their common national language.

It also demonstrates that speakers of a vernacular that is most similar to the standard language (in terms of the inventory of monophthongs) are in a better position to reach high agreement (or: consistency) in the identification task than speakers whose background dialect has a (marginally) richer or poorer inventory.

Finally, we advocate a wider use of the perceptual method outlined here to the study of vowel systems under conditions where sophisticated laboratory equipment is not available for spectral analysis. The test tape we have prepared can be administered (to a large number of subjects in parallel, if necessary) in half an hour, and provides a wealth of easily interpretable and surprisingly stable data.

Acknowledgements

Work supported in part by a grant from the Foundation for Linguistic Research, funded by the Netherlands Organisation for the Advancement of Pure Research (ZWO) under project nr 17-21-20; we thank Jos. F.M. Vermeulen for technical assistance.

We acknowledge the support of the Dept. of Phonetics at the University of Utrecht by putting equipment at our disposal.

References

Cohen, A., Slis, I.H. and Hart, J. 't (1963). Perceptual tolerances of isolated Dutch vowels. *Phonetica*, **9**, 65-78.

Flanagan, J.L. (1955). A difference limen for formant frequency. *Journal of the Acoustical Society of America*, **27**, 613-617.

Hombert, J.-M. (1979). Universals of vowel systems: the case of centralized vowels. In: E. Fischer-Jørgensen, J. Rischel, N. Thorsen (eds.). *Proceedings of the Ninth International Congress of Phonetic Sciences*, Vol. II. Institute of Phonetics, Copenhagen, 27-32.

Mermelstein, P. (1978). Difference limens for formant frequencies of steady state and consonant-bound vowels. *Journal of the Acoustical Society of America,* **63**, 572-580.

Nord, L. and Sventelius, E. (1979). Analysis and prediction of difference limen data for formant frequencies. *Phonetic Experimental Research at the Institute of Linguistics University of Stockholm, Report I: Experiments in speech perception; Phonetlcs Research Seminar 1978-1979*, 24-37

Schouten, M.E.H. (1975). Native-language interference in the perception of second-language vowels, Doct. Diss. University of Utrecht.

Zanten, E. Van and Heuven, V.J. Van (1983). A phonetic analysis of the Idonesian vowel system: a preliminary acoustic study. *NUSA, Linguistic Studies in Indonesian and other Languages in Indonesia,* **15**, 70-80.

Tendencies in Contemporary French Pronunciation

I. Zhghenti
Tbilisi, USSR

1. Introduction

At the present stage of development the French language is undergoing an emancipation in all directions, and above all in phonetics. There are numerous cases where orthoepic norms are gradually becoming optional, changing into tendencies rather than stable pronouncing norms. The problem of the correlation between a norm and a tendency of pronunciation is of major importance for foreigners learning French outside France. Hence the goal of my research is to study the pronunciation tendencies which are quite obviously revealing themselves in the pronunciation of standard French.

To achieve the goal I have studied different sounds both separately and in groups, both on the phonetic and the phonemic levels, e.g. nasal vowels and their oral counterparts, semivowels, /ə/ caduc, the opposition /ɑ-a/ and /œ̃-ɛ̃/ and also the variphones of the phoneme /r/ (apical and uvular).

2. Method

The sound composition of French was studied on the basis of the phonological distribution of the given phonemes. The speech material used consisted in the recordings of 24 French people in Paris 1972. The phonemic analysis was based on listening tests and questionnaires. The experimental research was carried out by applying spectral analysis, synthesis, oscillography, X-ray analysis, using the facilities of the laboratories of the Moscow Institute of Foreign Languages, the Tbilisi State University as well as the Paris Institute of Phonetics (in 1982).

My aim was to detect articulatory shifts, state their direction, and show their effect on the timbre of vowels.

3. Results and discussion

The spectrum and the X-ray analysis revealed that the back sound /ɑ/ has shifted forward to a centralized position. The degree of this shift is so considerable that instead of the former back vowel /ɑ/ we can now speak of the centralized /a/:. Since the tongue position is more fronted, the centralized /a/: and the front /a/ show almost similar configurations for both the oral resonators and the palatopharyngeal cavities.

It was likewise established that the difference in the articulation of /œ̃/ and /ɛ̃/ is vanishing, leading to the dephonologization of /œ̃/ which points to a tendency towards open articulation, towards delabialization and towards the reduction of the number of nasal vowels to three.

The fact that /ɑ/ and /œ̃/ undergo dephonologization shows that although a minimal physical distinction remains in the opposition pairs of /ɑ-a/ and /œ̃-ɛ̃/, the distinctive function in speech is erased, or else it is transferred from one phoneme to the other: in our case from /ɑ/ to /a/ and from /œ̃/ to /ɛ̃/.

The frequency statistics of the occurrence of the French vowels allow us to suppose that the reason why /ɑ/ and /œ̃/ are vanishing lies in the overloading of these vowels with grammatical functions which entails their frequent repetitions in different contexts, the confusion of their articulation with that of the other member of the opposition and eventually the complete disappearance of these oppositions. For example the functional overloading of the phoneme /œ̃/ is seen in several grammatical functions: the indefinite article 'un' the numeral 'un franc' the indefinite pronoun 'pas un', 'l'un', 'un entre mille' the noun 'un', 'l'un deux', 'il est arrivé un hier', also 'pas un n'est venu', 'un que je plains de tout mon coeur' etc. Comparing the frequency of occurrence of /ɑ/ and /a/, the proportion is 8 to 1 in favour of the front vowel /a/. The phoneme /ɑ/ also has several grammatical functions: those of a preposition, a verb ('avoir', 3rd p. sg., 'il a'), an adverb 'il est là', etc. Thus the overloading with the grammatical functions leads to an articulatory similarity with the opposition counterpart and, eventually to the loss of the opposition /ɑ-a/.

The normative course of contemporary French maintains the co-existence of the apical /r/ and the uvular /r/ or 'r grasseyé'. The replacement of the apical /r/ by the uvular /r/ has been registered in various West European languages: French, German, Swedish, Dutch, Norwegian and others. This transformation of the variphones of the phoneme /r/ introduces new problems to linguists.

The analysis of the recorded speech material suggests that in some non-northern regions of France the apical /r/ is still observed as a relic form of pronunciation. As for the uvular /r/, we can confidently state that it has established itself in pronunciation in the north of France and in Paris.

In the languages of those West European countries where the uvular /r/ is pronounced there is a tendency towards the front pronunciation of vowels. This fact suggests that the uvular /r/ appears in those languages that are rich in vowel phonemes, with front vowels prevailing over back vowels, i.e. in the vocalic-type languages tending towards front articulation. The physiological reason could be that the tip of the tongue is too busy pronouncing the front vowels, so the trill has to be pronounced elsewhere.

4. Conclusion

On the basis of the phonetic experiments carried out, I can sum up by saying that there are the following tendencies in contemporary French pronunciation:

1. in the vowel system:
   a. There is a tendency towards front articulation, as seen in the disappearance of the opposition /ɑ-a/ to the advantage of /a/ i.e. in the frontward shift of the sound /ɑ/.
   b. A tendency towards open articulation as seen in the loss of the opposition /œ̃-ɛ̃/ to the advantage of /ɛ̃/.
   c. A tendency towards delabialization, e.g. the sound /œ̃/ is transformed into the sound /ɛ̃/, besides there is a trend to confuse the sounds /œ/ and /ə/ caduc, where the neutral /ə/ is frequently replaced by /œ/.
2. in the system of consonants:
   A tendency to stabilize the 'r grasseyé' variphone which ousts the apical /r/ into a position of a mere relic form.

Section 16
Phonetics and Phonology

On the Uses of Complementary Distribution

A. Liberman
Minnesota, USA

> The Snow Queen said to me: 'If you assemble this word, you will be your own master and I shall give you the whole world and a pair of skates.'
>
> H.-K. Andersen

1. The problem stated

The concept of complementary distribution (CD) played an outstanding role in all brands of classical phonology. It is implicitly present in de Saussure's *Course*, and it is a cornerstone of American descriptive linguistics, which did not receive any impulses from de Saussure. Trubetzkoy made wide use of it in the opening chapters of *Grundzüge der Phonologie*. In glossematics, the commutational test is centered on the same concept. The 20th century has witnessed the rise and fall of naive distributionalism, but the concept of CD has not developed since it became current, and today its range of action is defined as vaguely as it was fifty years ago.

If, together with pre-generative phonology, we assume that the sound string is made up of phonemes and that phonemes are realized in allophones, we shall arrive at the trivial conclusion that the allophones of one phoneme stand in CD. This statement follows from the definition of the allophone and needs no proof. The real problem is whether the statement about allophones of one phoneme standing in CD can be reversed, that is, whether segmental elements occurring in mutually exclusive positions are thereby allophones of the same phoneme. The most famous example of this type was discussed by Trubetzkoy and has become a locus classicus. Trubetzkoy observed that in Modern German *h* and ŋ stand in CD (*h* is word-initial prevocalic, and ŋ is word-final postvocalic) and yet represent different phonemes.

2. Present weakness of the theory

Theory of CD is marred by three weaknesses: position is taken for an unequivocal concept, which it is not; the elements whose distribution interests the phonologist are referred to simply as sounds, which is insufficient and misleading; the moment in the overall process of decipherment at which CD comes in is unspecified.

CD presupposes mutually exclusive positions. Usually, *position* means environment and is defined in phonotactic terms (for instance, between vowels, before voiceless consonants, after *s*, word-finally, etc.). It is enough to add prosodic factors for the picture to become very complicated. Thus, in a language that allows only schwa in the unstressed syllable, schwa will turn out to be an allophone of all stressed vowels. If *syllabic* and *nonsyllabic* are positional factors, *i* and *j* are always allophones of one phoneme by definition.

The example of *h*:ŋ is looked upon as unique because Trubetzkoy discovered two *phonemes* in CD. If he had searched among allophones, he would have found any number of them belonging to different phonemes and still standing in CD. Consider the situation in Russian. In this language there are palatalized and nonpalatalized consonants. Before the palatalized consonants all vowels are fronted. It follows that not only [a] before a palatalized consonant (conventionally designated as [ä]) is in CD with [a] before a nonpalatalized consonant but *all* the fronted vowels - [ä], [ö], [ü], [e] - stand in CD with *all* the retracted variants, for example, [ä] and [o]. Why are only [a] and [ä] felt to be allophones of one phoneme, rather than [a] and [ö]? According to Trubetzkoy, [a] and [ö] do not share a unique set of distinctive features and therefore belong to different phonemes. This is a correct but tautological answer; besides, there is no way of knowing the distinctive features of every phoneme before the phonemes themselves have been isolated.

3. Zinder's approach

Another approach to this problem goes back to Ščerba but is mainly known from Zinder's work. In Zinder's opinion, only such sounds constitute one phoneme as stand in CD *and* can alternate within the same morpheme; [a] and [ä] fulfill both conditions [a] and [ö] do not: compare [dal] 'gave' (singular) and ['däl'i] 'gave' (plural) (/l/ and /l'/ are independent phonemes). That [a] and [ä] arise automatically, depending on the quality of the postvocalic consonant, is obvious because [dal] *becomes* [däl'] under clearly defined circumstances and [a] is in CD with [ä]. Since [dal] never becomes [döl'] under similar circumstances, there is no need to connect [a] and [ö]. It is the morpheme and not position that serves as the generator of allophones and provides a natural limit for subphonemic alternations.

Zinder's rule is correct: if two sounds stand in CD and alternate within one morpheme, their alternation must indeed be caused by the changing phonetic environment, so they are allophones of the same phoneme. All Russian vowels followed by palatalized consonants are in CD with all vowels standing before nonpalatalized consonants; in any language, all vowels before *n*, *m* are nasalized and are in CD with non-nasalized vowels, etc., but the unity of the phoneme is achieved through the alternating morpheme. And yet, Zinder's rule cannot be applied in the search for the phoneme, and as a tool of phonological discovery it is as useless as the rule formulated by Trubetzkoy. In order to work with this rule, we must have the entire speech sequence segmented; for instance, we must know that *dal* is [d-a-l], that *dal'* is [d-ä-l'], and so forth. But if we are at the stage of assembling phonemes, if the nondiscreet current of speech has not yet yielded phonemes, transcriptions like [d-a-l] do not exist. Phonological analysis begins with morphological segmentation. From alternating morphemes we obtain phonemes as bundles of abstract distinctive features, then discover their phonetic correlates, and finally recognize the allophones of each phoneme. The only possible order of phonological discovery is this: morphemes--phonemes with their distinctive features--allophones. Allophones can never be obtained before phonemes, and all attempts to reverse the sequence and first to obtain allophones in order to assemble them into phonemes is self-deception. The same conclusion can be reached by the purely logical analysis of the rules proposed by Trubetzkoy and Zinder.

Here are these rules once more. Trubetzkoy: two sounds belong to the same phoneme if they stand in CD and possess the same unique set of distinctive features. Zinder: two sounds belong to the same phoneme if they stand in CD and alternate within one and the same morpheme.

4. Conclusion

Both Trubetzkoy and Zinder, though they have only started their search for the phoneme, already operate with the concept of the *same phoneme*. But what can they know about phonemes before phonemes have emerged? How did they arrive at their rules? Evidently, they know what the *same phoneme* is (this is all the more unexpected in Zinder's case, for he does without distinctive features). CD as a tool of discovery is worthless. At the stage at which it could have been profitable, before the emergence of segmented and fully characterized phonemes, it cannot yet be applied, and at the final stage, when the phonemes have been obtained, its function is modest: it either emphasizes the fact that the allophones of one phoneme must have the same distinctive features (a conclusion that follows from the definition of the allophones as a contextual realization of the phoneme and the phoneme as a bundle of distinctive features) or brings out the decisive role of the morpheme in phonetic segmentation (but this too is trivial by the time the search is over). Only in historical phonology, with its emphasis on letters, Zinder's rule can be put to good use.

On the Correlation of Phonetic and Phonemic Distinctions

A. Steponavičius
Vilnius, USSR

1. Introduction

Following the tradition of functional linguistics, the basic principles of phonological analysis amount to the distinctions between (1) paradigmatics and syntagmatics, (2) segmentics and prosody, (3) phonology and phonetics, and (4) language and speech. From the structural point of view the dichotomy between paradigmatics and syntagmatics is of primary importance. Despite all possible variations in views and disagreements concerning some details, the concepts of paradigmatics and syntagmatics no longer provoke any great controversy among functionalists. Paradigmatics is defined as language structure based on associative relationships, and syntagmatics as language structure based on linear (and simultaneous) relationships. Patterns, or models, of paradigmatic sound structure may be constructed on the basis of distinctive and nondistinctive sound features, phonemes and their phonetic realizations, prosodemes and their phonetic realizations, oppositions, series and correlations. Patterns of syntagmatic sound structure may be otherwise called patterns of the sound structure of words. The constituent parts of the sound structure of words are the following:

1. the sound and phonemic composition of words and morphemes;
2. the phonotactic rules of phoneme distribution, neutralization and syllabification;
3. the prosodic structure of words.

On the other hand, the distinction between phonetics and phonology remains at present the domain of most intensive investigations and numerous controversies. Generally speaking, this distinction may be defined in terms of the dichotomies between language and speech, paradigmatics and syntagmatics. Phonology lies in the domain of language, but not speech, and has both paradigmatic and syntagmatic aspects. Phonetics, on the other hand, lies both in the domain of speech and language (in that it is the level of both indiscrete material speech sounds and discrete 'sound types' of language) and, just like phonology, has both syntagmatic and paradigmatic aspects. The discrimination of speech sounds, language sounds and phonemes may contribute, among other things, to a better understanding of the development of the concept 'phoneme'. As is known, Baudouin de Courtenay tried to find psychophonetic explanations why actually different sounds may be perceived as identical entities, making a distinction between sound as mere phonation, which is transitory and variable, and a phoneme as a psychological equivalent of sound, which is permanent and invariable. In other words, speech sounds are opposed here to what covers both language sounds and phonemes, viz., language sounds and phonemes are not differentiated by Baudouin de Courtenay. This, naturally, in no way belittles his contribution to the development of the theory of the phoneme. The idea of discriminating phonemes from sounds prevails in Sapir's works. Nevertheless, the lack of proper discrimination between language sounds and phonemes continues whenever phonemes are defined as classes of sounds, or as sounds distinguishing one utterance from another, etc.

2. Phoneme and Feature Definitions

A real breakthrough was marked by the development of the theory of the phoneme when Roman Jakobson with his co-workers, first of all N. Trubetzkoy, introduced the definition of the phoneme as a cluster, or bundle, of DFs. This established the concept of the phoneme as a member of phonematic oppositions. Since then the progress of phonology has depended first and foremost on the theory of DFs. At the present stage of phonological investigations there are available exhaustive inventories of phonetic features, expressed in articulatory, acoustic or perceptual terms (cf. the works by Peter Ladefoged) and a number of systems of DFs, such as those by Jakobson, Fant and Halle, or Chomsky and Halle, with all the possible variations and modifications. Among the problems which demand special attention one may point out the need of further elaborating the systems of DFs by way of correlating phonemic distinctions with phonetic ones. Though on the whole we may consider those systems of DFs better in which DFs have clearly stated phonetic correlates, DFs need not necessarily be directly related to phonetic features. The lack of direct correspondences between phonemic and phonetic features ensues first of all from the binarism and hierarchical ordering of DFs and oppositions (naturally, if we accept these principles of phonemic analysis). Jakobson has greatly influenced and even determined the linguistic thinking of phonologists by demonstrating the possibility of presenting all types of oppositions and features as binary, and it must be conceded that, even purely hypothetically, the binary structure of DFs seems most plausible; DFs as elementary units of the phonological structure must be characterized by most elementary relationships, and binary oppositions are the most elementary of all possible relationships. The practice of phonological analysis, moreover, has fully confirmed this, as the most exact definitions of phonemes, their most consistent classifications and hierarchical order seem to be those which are expressed in terms of binary features. Besides, Trubetzkoy's system of DFs and oppositions may be freely integrated into binary systems, though, naturally, with some modifications of the former. Trubetzkoy's distinction between multilateral and bilateral oppositions preserves its significance in that it reflects the degree of closeness of relationships of phonemes, viz., the most close relationships in the case of bilateral oppositions and more loose relationships in the case of multilateral oppositions. Secondly, in terms of proportional and isolated oppositions we can express the degree of paradigmatic integration of phonemes. Reinterpretation, however, is indispensable in the case of Trubetzkoy's gradual, privative, and equipollent oppositions. As has been demonstrated by Jakobson, Halle, Fant, Chomsky and others, the distinctions between vowels of different tongue-height may be expressed in binary features. The notion of graduality, just like those of privativeness and equipollence, is useful, though, when we want to indicate the physical implementation of sound features. From the point of view of their physical nature, binary features may be termed privative when they are based upon the presence and absence of the same sound property, gradual when they present different gradations of the same property, and equipollent when they are represented by two physically different and logically equivalent properties. What I am aiming at is the fact that though DFs may be presented as universally binary, they are nevertheless based upon different relations of phonetic features. It may be noted here that for the sake of consistency of phonological analysis a positively expressed feature and the respective negatively expressed feature should be considered as two different features and not the same feature with the plus and minus values. The set-up of the hierarchies of DFs must be such that oppositions of a higher rank comprise oppositions of a lower rank. It follows from this that subclasses of different classes of phonemes are not structurally and functionally identical and must be set up independently, irrespective of the possible identity of the anthropophonic nature of their DFs. Thus phonetic features, distinctive for one set of phonemes, may be nondistinctive for another (cf. voice in sonorants, or the occlusiveness of nasal sonorants).

3. Correlates of Features

The same DFs may have different phonetic correlates, and, on the other hand, the same phonetic features may be realizations of different DFs. Phonologists and phoneticians have always been fully aware of the fact that what is referred to as a single DF is actually a complex of articulatory and acoustic parameters. Besides, this complex may be different in the realization of different phonemes of the same series. Thus the labial series may consist of purely labial and labio-dental articulations. In some cases of consonantal features referring to place of articulation the exact points of articulation are phonologically essential. Thus, in English in the series of fricative nonstop obstruents there are three kinds of apicals whose contrasts may be expressed in the most natural way as *dental* (/θ ð/) vs. *postdental* (*nondental*) (/s z š ž/), with the postdentals further contrasting as *alveolar* (/s z/) vs. *postalveolar* (*nonalveolar*) (/š ž/). The contrast of the English /l/ with /r/ may also be

expressed by means of the features *alveolar* vs. *postalveolar*. In many other cases different, though adjacent points participate in the production of the same local series. In English the phonetically pharyngeal breath /h/ constitutes the same series of nonlabials with the dorsal consonants. In Lithuanian and Russian there are two apical series; in the case of obstruents these differ phonetically as dental and postdental, and in the case of sonorants, as alveolar and postalveolar. For the sake of economy of description these two pairs of features may be reduced to a single pair and termed arbitrarily *front*, or *advanced* vs. *back*, or *retracted*. How the same phonetic features may serve as realizations of different phonemic features may be illustrated by the distinctions *gliding* (*diphthong*) vs. *non-gliding*, *long* vs. *short*, and *checked* vs. *free*. In the case of the correlation of contact (*checked* vs. *free*) the checked vowels are the marked members of the opposition, which are realized as short monophthongs, and the free vowels are the unmarked members, which are realized as diphthongs and long vowels.

A Distinctive Feature Based System for the Evaluation of Segmental Transcription in Dutch

W.H. Vieregge, A.C.M. Rietveld, and C.I.E. Jansen
Nijmegen, the Netherlands

1. Introduction

However extensive the literature on transcription systems may be, it remains astonishing to see that data on inter- and intrasubject reliability are almost completely lacking. One of the major problems in the assessment of reliability is that it requires a system with which differences between transcription symbols can be assigned numbers corresponding to the distances between the transcription symbols, or rather corresponding to the distances between the segments that the transcription symbols stand for. Preferably, these distances should be defined articulatorily rather than auditorily, since the training in the use of transcription symbols is largely articulatorily based as well.

For the construction of a system in which the distances between the Dutch vowels are numerically expressed, enough experimental data may be found in the literature (e.g. Nooteboom, 1971, 1972; Rietveld, 1979). The available data with respect to the Dutch consonants appear to us less satisfactory. Spa (1970) describes the Dutch consonants by means of 16 distinctive features. One of our main objections against Spa's system is that the front-back dimension - a dimension which is crucial for the classification and the adequate use of transcription symbols - is only implicitly represented by the features [cor], [ant], [high], [low], and [back]. Moreover, the validity of Spa's system was not experimentally tested. We therefore decided to develop a new consonant system for Dutch with a heavier emphasis on articulation. The validity of this system was assessed by means of an experiment in which subjects were asked to make dissimilarity judgments on consonant pairs.

2. The vowel system

From the data in the literature (Eijkman, 1955; Moulton, 1962; Nooteboom 1971, 1972; Rietveld, 1979; Booij, 1981; Schouten, 1981) - data which to a great extent have been tested experimentally - the following characteristics of the Dutch vowels may be established:

The 15 vowel allophones [i, y, e, ø, ɛ, a, ɑ, ɔ, o, u, Y, I, ə, œ, U] can be subdivided into long, halflong, and short. Before [r, R] [i:, y:, e:, ø:, a:, o:, u:] are long; in foreign words [ɛ:, œ:, ɔ:] are long. The remaining vowels are short in these positions. When not before [r, R], [e., ø., a., o.] are halflong, the rest is short. In our system long = 1, halflong = 2, short = 3.

From Rietveld (1979), it appears that 'the proprioceptive articulatory dissimilarities can be predicted quite satisfactorily by using a traditional vowel scheme and giving extra weight to differences on the front/back dimension' (1979: 88). This statement only pertains to the nine vowels examined by Rietveld, namely [i, e, ɛ, y, ø, u, o, ɔ, ɑ]. We assume that this finding applies to all Dutch vowels. Thus, for the front/back dimension we have used a weight factor of 2, resulting in front = 2, central = 4, and back = 6.

Finally, by adding two values for rounded/unrounded (rounded = 1, unrounded = 0) and four values for the high/low dimension (high = 4, high/mid = 3, mid/low = 2, low = 1) all Dutch vowels may be distinguished.

Following Moulton, diphthongs are considered as vowel + vowel sequences, the second vowel being non-syllabic allophonically.

Table I (upper half) shows the dissimilarity matrix which results from assigning the above values on the dimensions distinguished to all Dutch vowels. The distances thus established can be used to express differences in the choice of transcription notations for vowels numerically. Examples of maximal differences (numbers 9 and 10) are [ɑ-ø:, ɑ-i:, ɛ-u: ɑ-y:]; examples of minimal differences (numbers 1 and 2) are [i-y, e:-ø, ɛ-i].

Table I. Dissimilarity matrix for all Dutch vowels (upper half) and for all Dutch consonants (lower half). * = Consonants used in the experiment.

	a:	e:	o:	ø:	i:	y:	u:	œ:	ɛ:	ɔ:	a·	e·	o·	ø·	ɑ	ɛ	ɔ	U	I	Y	œ	ə	i	y	u								
*p	-	4	5	5	5	6	6	2	3	4	1	5	6	6	4	5	6	7	6	6	4	4	7	8	8								a:
*b	1	-	5	1	1	2	6	4	1	6	5	1	6	2	8	3	8	7	2	6	6	4	3	4	8								e:
*t	2	3	-	4	6	5	1	3	6	1	6	6	1	5	5	8	3	2	7	5	5	5	8	7	3								o:
*d	3	2	1	-	2	1	5	3	2	5	6	2	5	1	9	4	7	6	3	5	5	5	4	3	7								ø:
c	3	4	3	4	-	1	5	5	2	7	6	2	7	3	9	4	9	8	3	5	7	5	2	3	7								i:
*k	5	6	3	4	2	-	4	4	3	6	7	3	6	2	10	5	8	7	4	4	6	6	3	2	6								y:
g	6	5	4	3	3	1	-	4	7	2	7	7	2	6	6	9	4	3	8	4	6	6	7	6	2								u:
*f	3	4	3	4	6	6	7	-	3	2	3	5	4	4	6	5	4	5	6	4	2	2	7	6	6								œ:
*v	4	3	4	4	7	7	6	1	-	5	4	2	7	3	7	2	7	8	3	7	5	5	4	5	9								ɛ:
*s	4	4	2	3	5	5	6	1	2	-	5	7	2	6	4	7	2	3	8	6	4	6	9	8	4								ɔ:
*z	5	4	3	2	6	6	5	2	1	1	-	4	5	5	3	4	5	6	5	5	3	3	6	7	7								a·
ʃ	4	4	4	5	3	5	6	3	4	2	3	-	5	1	7	2	7	6	1	5	5	3	2	3	7								e·
ʒ	5	4	5	4	4	6	6	4	3	3	2	1	-	4	4	7	2	1	6	4	4	4	7	6	2								o·
*x	7	8	5	6	4	2	3	4	5	3	4	3	4	-	8	3	6	5	2	4	4	4	3	2	6								ø·
ɣ	8	7	6	5	5	3	2	5	4	4	3	4	3	1	-	5	2	3	6	6	4	4	7	8	4								ɑ
χ	6	7	4	5	5	3	4	3	4	2	3	4	5	1	2	-	5	6	1	5	3	3	2	3	7								ɛ
ʁ	7	7	5	4	6	4	3	4	3	3	2	5	4	2	1	1	-	1	6	4	2	4	7	6	2								ɔ
*m	3	2	5	3	6	8	7	4	3	5	4	5	4	8	7	7	6	-	5	3	3	3	6	5	1								U
ɱ	4	3	4	4	7	7	6	3	2	4	3	6	5	7	6	6	5	1	-	4	4	2	1	2	6								I
*n	5	4	3	2	6	6	5	4	3	3	2	5	4	6	5	5	4	2	1	-	2	2	3	2	2								Y
ɲ	6	5	6	6	3	5	4	7	6	6	5	4	3	5	4	6	5	3	4	3	-	2	5	4	4								œ
*ŋ	8	7	6	5	5	3	2	7	6	6	5	6	5	3	2	4	3	4	4	3	2	-	3	4	4								ə
*l	5	4	3	2	6	6	4	4	3	3	2	5	4	6	5	5	4	4	2	3	4	5	-	1	5								i
ɫ	7	6	5	4	6	4	3	6	5	5	4	7	6	4	3	3	2	6	5	4	3	3	2	-	4								y
r	4	3	2	1	5	5	4	3	2	2	1	4	3	5	4	4	3	3	2	1	4	4	1	3	-								u
ɾ	5	4	3	2	6	6	5	4	3	3	2	5	4	6	5	5	4	4	3	2	5	5	2	4	1	-							
*R	6	5	4	3	5	3	2	5	4	4	3	6	5	3	2	2	1	5	4	3	4	2	3	1	2	3	-						
*w	4	3	4	3	7	7	6	3	2	4	3	6	5	7	6	6	5	3	2	3	6	6	3	5	2	3	4	-					
ʋ	3	2	5	4	6	8	7	4	3	5	4	5	4	8	7	7	6	2	3	4	5	7	4	6	3	4	5	1	-				
*j	7	6	5	4	4	4	3	6	5	5	4	5	4	4	3	5	4	6	5	4	3	3	4	4	3	4	3	3	3	-			
*h	7	8	5	6	6	4	5	4	5	3	4	5	6	2	3	1	2	8	7	6	7	5	6	4	5	6	3	6	8	6	-		
ʔ	6	5	4	3	5	3	2	7	6	6	5	8	7	5	4	4	3	7	6	5	6	4	5	3	4	5	2	6	7	6	3	-	
	p	b	t	d	c	k	g	f	v	s	z	ʃ	ʒ	x	ɣ	χ	ʁ	m	ɱ	n	ɲ	ŋ	l	ɫ	r	ɾ	R	w	ʋ	j	h	ʔ	

3. The consonant system

From the literature the following data with respect to Dutch consonants may be established:

On the basis of the results of a multidimensional scaling analysis Van den Broecke (1976: 120) states that 'there is some evidence to believe that there are a number of different inner speech dimensions, at least 4, possibly 5, employed in similarity evaluations on inner speech stimuli'. The dimensions which Van den Broecke found allow for current phonetic interpretations in terms of place and manner of articulation features. Several systems developed for English consonants also make use of features of place and manner of articulation (Singh, 1976). Therefore, we decided to make use of these features too. As for place of articulation, our system differs from the system proposed by Spa, since there this feature was ony implicitly represented.

On the basis of the above considerations we selected the following distinctive features to distinguish the 32 Dutch consonants which may appear in a narrow transcription: the feature of place of articulation (bilabial = 1, labiodental = 2, dental/alveolar = 3, palato-alveolar = 4, palatal = 5, velar = 6, uvular = 7, glottal = 8) and seven binary features, i.e. voicing, nasality, continuity, glide, laterality, fricative, and flap.

3.1. The experiment

3.1.1. Method

In order to allow comparison with the results of the first experiment conducted by Van den Broecke (1976), and also in order to restrict the number of stimuli, from the 32 consonants given in Table I, lower half, a subset of 18 consonants was selected for use in the experiment namely [p b, t, d, k, f, v, s, z, x, j, l, r(R), h, m, n, ŋ, w]. These consonants are indicated with an asterisk in Table I. Twenty-five first year speech therapy students were presented with these consonants pairwise in medial word position (cf. Van den Broecke: in isolation); they were asked to rate each pair on articulatory dissimilarity (cf. Van den Broecke: on dissimilarity represented perceptually by means of 'inner speech') on a 10-point scale (10 = maximal dissimilarity, 1 = minimal dissimilarity). The stimulus material consisted of $(18^2 - 18)/2 = 153$ word pairs which differed as little as possible, containing the same number of syllables, and exhibiting, with a few exceptions, the same stress pattern. The stimuli were offered in random order on paper. After the experiment was over, the subjects were asked to indicate which /r/-realizations they used because in Dutch /r/ may be realized both as [r] and [R]. In the instructions it was emphasized that during the rating the whole articulatory apparatus should be taken into consideration. As the subjects had just started their training as speech therapists, they had acquired no more than a negligible amount of phonetic knowledge.

3.1.2. Results and discussion

In order to gain insight into the dimensions underlying the dissimilarity judgments of the subjects, multidimensional scaling was carried out. Input to the program (ALSCAL, W. Young, Y. Takane, R.J. Lewyckyl, 1977) were the means of the dissimilarity scores. In Table II the stress values, the random stress values, and the correlations between the ultrametric distances and the dissimilarities are given for 1, 2, 3, 4, 5, and 6 dimensions (Euclidean metric).

Table II. Values of three statistics as a function of the number retained dimensions in an MSCAL (Euclidean metric).

Dimension	1	2	3	4	5	6
stress	38.6	26.4	18.1	11.2	7.7	5.1
stress random	47.3	27.9	18.5	13.0	9.6	-
correlation	.80	.81	.85	.91	.95	.97

As may be expected, the stress and correlation coefficients are most favorable in the six-dimensional solution. However, since the sixth dimension was not phonetically interpretable, we opted for the five-dimensional solution for which the stress and correlation coefficient are also satisfactory.

The positions of the 18 consonants in the five-dimensional solution are graphically presented in Figures 1a, 1b, and 1c. As may be seen, the first dimension clearly represents the front-central-back continuum: at the one end we find labials dentals, and alveolars, in the center the palatal [j], and at the other end the velars, uvulars, and glottals. The second dimension is less easy to name. Globally, if it were not for the position of [l], it looks like a +cont/-cont dimension. More specifically, the classes of fricatives, plosives, and nasals + laterals may be distinguished along it. Ignoring the position of [ŋ], and taking into account the overlapping area in which labials and the [h]

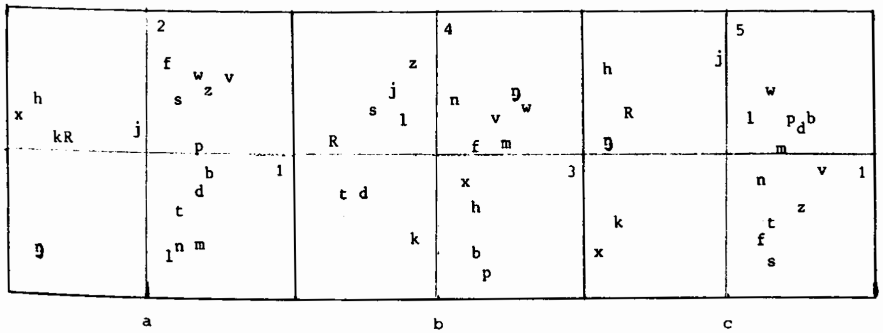


Figure 1. (a, b, c) Positions of 18 Dutch consonants in a five-dimensional solution.

are positioned the third dimension may be interpreted as a +lab/-lab dimension. With the exception of the positions of [z] and [s], the fourth dimension may be labeled +son/-son. Finally with the exception of the position of [h]and [p], the fifth dimension may be interpreted as a voiced/unvoiced dimension.

The correlation between the mean dissimilarity scores yielded by the experiment and the distances as defined in our consonant system is not very high (r = .61). Moreover, the positions of the consonants in the five-dimensional solution as depicted in Figures 1a, 1b, and 1c suggest that in the consonant system the feature of place of articulation should be split up in less than eight categories, and that the manner of articulation features should be changed. Therefore, our consonant system was revised in three different ways. In the first revised version the eight categories of place of articulation were maintained whereas the manner of articulation features were changed (deletion of the features +nas/-nas, +glide/-glide, and +fric/-fric); in the second version the place of articulation features were reduced to three categories, and the manner of articulation features were changed in a different way than in the first version (deletion of +cont/-cont and introduction of a dimension with at the one extreme plosive/fricative and at the other nas/lat), in the third version the place of articulation feature was reduced to five categories, and the manner of articulation features was changed in the same way as in the second version.

Of the three revised versions of the consonant system, the third one - the dissimilarities of which are given in Table I, lower half, - correlated most highly with the dissimilarities obtained in the experiment (r = .75, r = .64, and r = .58 for the third, second, and first revised version, respectively). The third version of our consonant system (SysV) was further evaluated by also relating it to the first experiment conducted by Van den Broecke (ExpB)and the consonant system developed by Spa (SysS).

The correlation between ExpV and ExpB was found to be fairly high (r = .80). From this it may be deduced that partly similar and partly differing criteria have been used by the subjects to judge the stimuli, the differing criteria probably having to do with the differences in experimental set-up. (Recall that in ExpB the stimuli were offered in isolation whereas in ExpV they were offered in medial word position. Recall also that in ExpB the subjects were asked to judge the stimuli auditorily whereas in ExpV they were asked to judge the stimuli articulatorily).

The correlation between SysV and SysS was considerably lower (R = .57). This may be the result of the fact that, as was said before, the features of SysV were more phonetically oriented than the ones in SysS. It furthermore appeared that the correlation of SysV with ExpV was significantly higher than the correlation of SysV with ExpB (r = .75 and r = .55 respectively; t = 5.19, p < .01, df = 16), and that the correlation of SysS with ExpB was higher (but not significantly so) than the correlation of SysS with ExpV (r = .65 and r = .61, respectively; t = 0.914, p>.10, df = 116). These differences could very tentatively be interpreted in terms of differences in experimental set-up and structure between the two consonant systems: SysS is a better predictor of the auditorily based dissimilarity judgments of Van den Broecke, and SysV is more successful in predicting our articulatory based dissimilarity judgments. The choice of either of the two systems will therefore depend on the purpose it has to serve.

Acknowledgements

We would like to express our thanks to drs. R. van Bezooijen for her great help in preparing the final version of this paper. We are also grateful to dr. M.P.R. Van den Broecke for making the score data from his first experiment available to us.

References

Booij, G.E. (1981). *Generatieve fonologie van het Nederlands.* Utrecht: Het Spectrum.

Broecke, M.P.R. van den (1976). *Hierarchies and rank orders in distinctive features.* Assen: van Gorcum.

Broecke, M.P.R. van den, Stoop, I. (1982). Perceptual dimensions in psychological consonant features. *PRIPU, Vol. 7, 2,* Utrecht.

Eijkman, L.P.H. (1955). *Phonetiek van het Nederlands.* Haarlem: F. Bohn N.V.

Klahr, D. (1969). A Monte Carlo investigation of the statistical significance of Kruscal's nonmetric scaling procedure. *Psychometrica,* **34**(2), 319-330.

Moulton, W.G. (1962). The vowels of Dutch: phonetic and distributional classes. *Lingua,* **11**, 294-312.

Nooteboom, S.G. (1971). Over de lengte van korte klinkers, lange klinkers en tweeklanken van het Nederlands. *Nieuwe Taalgids,* **64**, 396-402.

Nooteboom, S.G. (1972). *Production and perception of vowel duration, a study of durational properties of vowels in Dutch.* Utrecht: Diss. University of Utrecht.

Rietveld, A.C.M. (1979). Judgments on the articulatory similarity of Dutch vowels. *IFN-Proceedings,* **3**, 79-88.

Schouten, M.E.H. (1981). Het verschil tussen 'bot' en 'bod' - een vergeefse speurtocht. *De Nieuwe Taalgids,* **74**, 6, 537-546.

Singh, S. (1976). *Distinctive features, theory and validation.* Baltimore: University Park Press.

Spa, J.J. (1970). Generatieve fonologie. *Levende Talen,* **266**, 191-204.

Young, F.W., Takane, Y., Lewyckyi, R.J. (1977). *ALSCAL-4: Dynamic core version, version 4.01D.*

Section 18
History of Phonetics

The Study of Vowel Quality by Early Arab and Muslim Phoneticians

M.H. Bakalla
Riyadh, Saudi Arabia

1. Introduction

In a previous paper of mine entitled 'The contribution of the Arabs and Muslims to the study of vowel length' (Bakalla, 1978), I tried to present the analysis of vowel quantity of Classical Arabic as it was originally given in the early Arabic sources stretching over more than 10 centuries beginning with the 7th Century A.D. In the present paper I will try to present the analysis of vowel quality in early Arabic sources. No attempt is made here at exhaustive coverage of this phenomenon in the literature of the Islamic countries or through the Islamic periods. Rather a bird's eye view of the salient features is selected from various sources with references to their authors. It is also beyond the scope of this paper to trace the origin and development of the present phenomenon regarding vowel quality and related matters; such treatment deserves a separate paper.

2. Vowel system

The first point to be discussed here is the triangular concept of the vowel system of Arabic, in particular Classical Arabic. Arabic was analysed as having three basic vowels. They are the long vowels [a:], [i:], [u:]. They were commonly called *ḥurūf*. The same term was also given to the consonants of Arabic. Thus the term refers to the written representation of the vowels and consonants. Ibn Sina or Avicenna (died 428 A.H./1037 A.D.) differentiates between the two categories by calling the vowel *muṣawwit* and the consonant *ṣāmit* (Ibn Sīnā 1963:42). The use of the two terms indicates the contrast between 'vocalic' and 'nonvocalic' nature of speech elements. Literally *muṣawwit* means sound or sonorous sound while *ṣāmit* means silent or non sonorous sound.

3. Long vs. short vowels

In the 10th Century A.D. Ibn Jinnī (1954:19f) drew the distinction between the long vowels and the short vowels as is clearly stated in the following:

> 'The short vowels *ḥarakāt* are parts of the long vowels *ḥurūf*

> *madd wa līn*, namely, the *ʔalif* [a:], the *wāw* [u:], and the *yāʔ* [i:]. Just as those *ḥurūf* are three, so are the *ḥarakat* three in number. These are the *fatḥah* [a], the *kasrah* [i], and the *ḍammah* [u]. Early grammarians used to call the *fatḥah* the small *ʔālif*, the *kasrah* the small *yāʔ*, and the *ḍammah* the small *wāw*. In this regard they were correct.'

Here Ibn Jinnī seems to consider the quality as the distinguishing factor between long and short vowels for he continues to state:

> 'To demonstrate the fact that short vowels are but parts (or fractions) of the corresponding long vowels *ḥurūf*:- If you lengthen (*ʔa* ʃba^{c}ta) the short vowel, the corresponding long vowel will occur as a result. For example, a in ^{c}*amr*, if lengthened, ā will occur and you will say ^{c}*āmr*; likewise, i in ^{c}*inab* will become ī as in *^{c}*īnab*; and so u in ^{c}*umar* will become ū as in *^{c}*ūmar*. Had the short vowels not been parts of these long counterparts, the long vowels would have not been anticipated or been produced from them." (Ibn Jinnī 1954:20).

The term given to the lengthening of vowels is ʔ*i* ʃbāʕ. A long vowel is called *ḥārf madd wa līn* where *madd* refers to lengthening and *līn* is special quality of the vowel for being produced without any hindrance or obstruction.

4. Quality differences

The early Arab and Muslim phoneticians were aware of the differences in quality between the vowels themselves. Ibn Jinnī (1954:8) states:

> 'The *ḥurūf* or sounds which are produced with the articulators held apart are three [a:], [i:], and [u:]. The most open and the softest of them is [a:]. Further the sound which occurs in [a:] is different in quality from the sounds [i:] and [u:]. Conversely the sound which occurs in [i:] is different in quality from [a:] and [u:]. The reason for the differences is that in each of the three cases the vocal tract has a different configuration.'

Ibn Sīnā (1963:49) describes the articulation of the vowels as follows:

> 'The long vowel [a:] and the short vowel [a] are produced by a smooth emission of the air-stream without any interference. The long vowel [u:] and the short vowel [u] are produced with little interruption of the air-stream at, and narrowing, of the lips, along with slight gentle propulsion upwards on the way out. The long vowel [i:] and its sister short vowel [i] are produced with little interruption and narrowing of the point of articulation, along with a gentle propulsion downward.'

5. Consonants and vowels

In their description of the phonological or phoneme inventory of Arabic, the early Arab and Muslim phoneticians distinguished between the two main types of sounds or phonemes including consonants and vowels. The first type is called basic and includes the long vowels discussed above, namely [a:], [i:] and [u:]. The corresponding short vowels are already parts of the respective long ones. There were other varieties of vowels which were discussed in the early literature. Of relevance here are the two long vowels [e:] known as *ʔalif ʔimālah* or '*umlaut*' and [ɑ:] known as *ʔalif tafxim* or 'emphatic or velarized' [a:]. According to Ibn Jinnī (1954:51) they are both acceptable and readily used in the standard language. As for *ʔalif ʔimālah*, he mentions the vowel which is between *ʔalif* [a:] and *yāʔ* [i:] as in *ʕālim* → [*ʕe:lim*], *x̄atam* → [*xe:tim*]. As for *ʔalif tafxim* Ibn Jinnī gives the following description:

'*ʔalif tafxīm* is that which is between [a:] and [u:] as in their pronunciation of *salām alayk* and *qāma Zayd*. This is why the symbol [u:] was given in the writing of *ṣalāt zakāt* and *ḥayāt*, because [a:] was pronounced in the direction of [u:].'

6. Variants

There is another class of largely short vowels which are not normally marked in writing. Ibn Jinnī (1954:58f) considers them as a consequence of vocalic harmony or vowel assimilation *tajānus aṣ-ṣawt*. Hence they were treated as secondary sounds or variants. Some of these are [ɨ], [ʉ], [ɛ] and [æ]. These varieties and others are treated in more detail in Bakkala (1982).

7. Conclusion

In conclusion, the contribution of the Arab and Muslim phoneticians are both numerous and interesting. This paper has only pointed out some of the salient features in one small area of their analysis of the vowel system. It is meant to point to the problems rather than solve them. Definitely further investigation is necessary before a final assessment can be made.

References

Bakalla, M.H. (1978). The contribution of the Arabs and Muslims to the study of vowel length. In: *Study of Sound, Bulletin of the Phonetic Society of Japan,* **11**, 197-202.

Bakalla, M.H. (1982). *Ibn Jinnī: An Early Arab Muslim Phonetician. An Interpretive Study of His Life and Contribution to Linguistics.* London and Taipei: European Language Publications.

Ibn Jinnī (1954). *Sirr Ṣinā^{c} at al-I^{c}rāb.* Vol. 1. Edited by Mustapha al-Saqqa et al. Cairo: Mustapha al-Halabi Publishers.

Ibn Sīnā (1963). *Arabic Phonetics: Ibn Sīnā's Risalāh on the Points of Articulation of Speech Sounds*. Translated by Khalil I. Semaan. Lahore: Sh. Muhammad Ashraf.

Semaan, Khalil I. (1968). *Linguistics in the Middle Ages. Phonetic Studies in Early Islam*. Leiden: E.J. Brill.

Sībawaihi (1968). *Al-Kitāb*. Baghdad. (Reprint).

The Birth of Experimental Phonetics in Spain. Pere Barnils and his Barcelona Laboratories (1913-1933)

J. Julià
Barcelona, Spain

Pere Barnils (1882-1933) can be considered the major figure in Catalan linguistics during the first third of this century and the leading experimental phonetician in Spain for the same period of time. He introduced experimental phonetics in this country by founding its first laboratory to develop that new linguistic science together with the edition of its outstanding journal of phonetics. Barnils set up the basis of Catalan phonetics as a science and he was the first to apply experimental techniques in order to correct speech and voice disorders.

He had studied for three years at the German University of Halle an der Saale (1908-1911), where he took his doctorate in Romance Linguistics in June 1912, before he had the chance of meeting l'Abbé Rousselot in Paris.

Pere Barnils was in Paris (November 1911 - May 1912), sponsored by the Barcelona provincial council ruled by the great Catalan nationalist Enric Prat de la Riba so that he could follow Gilliéron's lectures at the École de Hautes Études and above all those of Rousselot, whose worldwide reputation had reached Barnils in Halle. Our Catalan linguist learnt how to make artificial palates and to interpret kymographic tracings at the Institut Catholique and at the modest laboratory of the Collège de France, close to La Sorbonne, directed by Rousselot and assisted by his best known disciple Joseph Chlumský of Prague. The interest in experimental techniques aroused 'chez l'Abbé' continued throughout his career. In 1913, once back in Barcelona, he edited the first journal of Catalan linguistics, which turned out to be the first periodical of modern linguistics that was published in Spain: *Butlletí de Dialectologia Catalana.*

At the same time he started to organize what would become the first Laboratory of Experimental Phonetics in that country (1913-1921). It was set up at the Institut d'Estudis Catalans in Barcelona and began to work satisfactorily in 1915 and full-time from mid 1916 up to the end of 1917.

In April 1914 Barnils represented Catalonia at the First International Congress of Experimental Phonetics held in Hamburg and was elected as the representative of Spain to the newly born International Association of Phonetics. For a few weeks he worked with Panconcelli-Calzia at his laboratory set up at Hamburg Colonial Institute so that the young Catalan phonetician could get acquainted with the recent experimental techniques developed in Europe.

During World War One, which broke out the following summer, Barnils' laboratory became accidentally the leading center in experimental phonetics in Western and Southern Europe. He developed it into a major scientific center of research with an interdisciplinary approach: in addition to the studies of Catalan descriptive phonetics by experimental means, he worked with laryngologists and speech therapists and welcomed all sort of researchers – native as well as foreign – whose interests were related in some way or other to phonetics.

In the summer of 1917 his magnificent first - and last - volume of *Estudis Fonètics* appeared, which included the works carried out at the laboratory, such as studies on the articulations of /k/ and /g/ in Majorcan Catalan, vowel nasality, alveolar roll vibrations and articulatory force in voiceless plosives (to cite only those by Barnils on Catalan phonetics). Unfortunately he had to leave his experimental work because of both political reasons and conflicts with the management. As a result his first laboratory was ultimately closed down in 1921, after three years of inactivity.

His interest in the disorders of speech and voice led to the founding of his private 'Laboratori de la Paraula' (1914-1933) and his taking charge of the Barcelona School for Deaf-mutes (1918-1930) to whom he devoted almost all of the rest of his life. There he founded his third Laboratory of Experimental Phonetics and edited its journal *La Paraula* (1918-1921). At the moment of his death Barnils was the editor of his third phonetics journal: *El Parlar* (1931-1932), while he directed his private 'Laboratori' and School for the deaf-mutes, which he had just founded. Although they were short-lived periodicals they attained a high degree of quality which was up to the standard of similar works in Europe and America.

Barnils was an active participant in international congresses: shortly before his death he read papers on speech pathology at the Second International Congress of Linguists (Geneva, 1931), at the First ICPhS (Amsterdam, 1932) and at the Fifth International Congress of Logopedy and Phoniatrics (Vienna, 1932).

We must also point out that our phonetician always showed great interest in popularizing the phonetic sciences, practically unknown at the time, by means of newspaper and encyclopedia articles, which together with his phonetic journals and other works made him the greatest contributor ever to Catalan experimental phonetics.

Other Barnils' phonetic publications include: 'Sobre fonetica catalana. Vocals' (*Bolletí del Diccionari de la Llengua Catalana*, Palma de Mallorca, 1911); 'Études de prononciations catalanes à l'aide du palais artificiel' (*Revue de Phonétique*, Paris, 1912); 'Études de prononciations catalanes' (*Ibid.*, 1913); 'Les consonants dites semi-sordes' (*Estudis Romànics, I*, Barcelona, 1916); 'Sobre la quantitat de les vocals tòniques' (*Ibid.*); 'Comentaris a un fragment recitat' (*Estudis Fonètics*, Barcelona, 1917); 'Notes fonètiques disperses: I-Sobre el parlar judeo-espanyol. II-Sobre el grec de Corfú' (*Ibid.*; *Treballs realitzats durant l'any 1917* (Barcelona, 1918); 'A propòsit de l'aparell de Franz X. Rötzer' (*La Paraula*, Barcelona 1918); 'Un archivo fonográfico' (*Ibid.*); 'El timbre de la vocal neutra del català' (*Ibid.*, 1919); 'Un donativo a nuestra escuela: el Phonoscope Lioret' (*Ibid.*, 1920); 'Pathégrafo' (*Enciclopedia Espasa-Calpe*, Barcelona, 1920); 'Fonética' (*Ibid.*, 1924); 'Sordomudo' (*Ibid.*, 1927); *El Laboratorio de la Palabra* (Barcelona. 1928); *Defectes d'el parlar* (Barcelona, 1930); 'Articulacions alveolars condicionades' (*Miscel.lània A.M. Alcover*; Palma de Mallorca, 1932).

Section 19
Phonetics and Foreign Language Teaching

Production and Perception of English Vowel duration by Dutch Speakers of English

B.A.G. Elsendoorn
Eindhoven, the Netherlands

1. Introduction

There is no doubt that the English produced by Dutch native speakers generally falls short of what English native speakers normally do. This results in a non-native accent, which is very likely due to native-language interference.

A number of studies have paid attention to qualitative differences between the two languages (Schouten, 1975) as well as to intonational differences (Elsendoorn, 1979; Willems, 1982). A third factor that might contribute to the perception of non-nativeness could be the phonologically based dissimilarities in vowel durations between the two languages. A well-known example of this is the voicing contrast of final obstruents in English, which is accompanied by varying the duration of the preceding vowel, due to the fact that there is no such voice contrast in final Dutch obstruents.

Present-day teaching methods in the Netherlands concentrate on vowel and consonant quality and pay little attention to durational differences. The Dutch student of English is at best advised to realize the contrast of pairs such as e.g. 'beat-bead' by simply prolonging the vowel duration and is hoped to be made conscious of this by endless listen-and-repeat drills. As this method does not seem to yield satisfactory results, the need was felt for a systematical description of differences in English vowel durations produced by native speakers of English and of Dutch, to investigate their influence on acceptability and to examine the relationship between the production and perception of vowel durations by Dutch speakers of English.

As has been demonstrated by Sheldon and Strange (1982), this relationship can be very complex. They showed that, contrary to what is commonly assumed, incorrect perception of second-language phonemes does not necessarily imply incorrect production of these phonemes.

The first part of this report gives a survey of systematic differences between English vowel durations produced by native speakers of English and of Dutch. The second part describes the influence of these differences on acceptability judgments and in the third part we will go into the relationship between production and perception of English vowel durations by Dutch speakers of English.

2. Production

Data have been obtained for English vowels, essentially spectrally similar to Dutch vowels and produced in isolated CVC words. Measurements were performed with the aid of a computer programme (Van den Berg, 1980). It appeared that the contrast between intrinsically short and long vowels was much greater for English native speakers (ENS) (N=5) than for Dutch speakers of English (DSE) (N=16). Another remarkable fact turned out to be the contrast in vowel duration as a function of the following consonant. As was expected vowel durations of DSE were all shorter than ENS durations when the vowels preceded a voiced obstruent; the DSE duration for vowels followed by a voiceless obstruent, however, was longer than the ENS duration in the same environment in all cases. No approximation of ENS durations was found with Dutch speakers of English parallel to an increasing familiarity with English, as evidenced by number of years of training.

In mono- and polysyllabic words embedded in sentences differences between the two groups of speakers were much reduced, but still present in most cases. It appeared that position of the word in the sentence contributed most to variations in vowel durations.

3. Perception

To test the acceptability of non-native vowel durations for native speakers two perception tests were presented to English native speakers (N=20). In one test vowel duration in isolated monosyllabic CVC words was varied, in the other sentence material was used in which the variable was the duration of the vowel in the last word. In both experiments subjects had to fulfil a word recognition task as well as judge the acceptability of pronunciation. The duration of the vowel in both experiments was set at either average ENS or DSE duration derived from the production data mentioned previously. In the word perception test words were used that had been read out by an English native speaker and a Dutch speaker of English. A combination of these two variables (viz. speaker and duration) resulted in four different stimulus types. In the sentence perception test only ENS produced material was used.

The word perception test showed that acceptability judgments were nearly identical for the two durational varieties within one speaker condition. The ENS produced stimulus words were, however, judged to be significantly more acceptable than the DSE varieties. The DSE produced stimuli led to a great number of incorrect identifications as well: those words that should have ended in a lenis obstruent were perceived as ending in a fortis consonant, regardless of whether vowel duration equalled average ENS or DSE duration. In the sentence perception test those stimuli containing the ENS duration were judged significantly more acceptable than those containing the DSE vowel duration; incorrect vowel duration did not often lead to incorrect recognition.

4. Relation between speech production and perception

Nooteboom (1972) showed that speakers are aware of durational structures in their mother tongue and fully capable of accurately reproducing them in a perception task. It has been demonstrated that speakers possess some inner criterion that they can use to match vowel durations in their mother tongue. This relationship need not be exactly the same for vowel durations in a second language. To examine the relation between the inner representation of vowel durations and their production an experiment was designed in which subjects were asked to adjust the vowel duration of an English CVC-word in such a way that it matched what they thought would be a correct native English duration. By means of a blind knob subjects were able to vary vowel duration continuously between 0 and 600 ms (for a description of the computer program used see Van den Berg, 1982). Additionally they had to read out the words used in the experiment.

Results indicated a high correlation between produced and adjusted durations for ENS and DSE in their respective mother tongues. Contrary to this, there appeared to be a sharp contrast between DSE production and adjustment of vowel duration on English words depending on the voice feature of the final consonant. In the case of words ending in a fortis obstruent adjustment data were similar to produced durations, i.e. longer than what is generally produced/adjusted by ENS. In the case of lenis obstruents, however, adjusted DSE duration was also longer than ENS duration, whereas, according to produced durations, it was expected to be shorter. A plausible explanation might be that subjects' responses were influenced by the pronunciation of the final obstruent. It seems that this voicing characteristic triggers the subconscious knowledge which DSE seem to have about this phonological phenomenon in English, although it is not actually realized in their production.

It also appeared from this experiment that Dutch speakers of English seem to refine their ideas about vowel duration in English as they become more proficient, since standard deviations of adjustment decrease with growing familiarity with and knowledge of the English language.

Acknowledgement

The research reported on here was carried out at the Institute of Phonetics, University of Utrecht, under grant number 17-21-12 of the Foundation for Linguistic Research, funded by the Dutch organization for the Advancement of Pure Research (Z.W.O.).

References

Berg, C.G. van den (1980). A speech editing system. *Progress Report of the Institute of Phonetics Utrecht* **5** (1): 87-91.

Berg, C.G. van den (1982). A computer programme to manipulate the duration of speech signals. *Progress Report of the Institute of Phonetics Utrecht* 7(1): 58-60.

Elsendoorn, B.A.G. (1979). A contrastive analysis of pitch movements of Dutch and English interrogative sentences. *Progress Report of the Institute of Phonetics Utrecht* 4(2): 4-30.

Nooteboom, S.G. (1972). Production and perception of vowel duration. Doctoral dissertation, Utrecht.

Schouten, M.E.H. (1975). Native-language interference in the perception of second-language vowels. Doctoral dissertation, Utrecht.

Sheldon, A. and W. Strange (1982). The acquisition of /r/ and /l/ by Japanese learners of English: evidence that speech production can precede speech perception. *Applied psycholinguistics* 3: 243-261.

Willems, N.J. (1982). *English intonation from a Dutch point of view*. Foris, Dordrecht.

Detection of Foreign Accentedness

J. E. Flege
Birmingham, U.S.A.

1. Introduction

Humans derive several important kinds of information from the speech signal in addition to referential meaning. One of these is 'accentedness,' a realization by the listener that his interlocutor differs in social, geographical, or linguistic background. This study examines the ability of American listeners to detect accent in the speech of French-English bilingual talkers. Previous studies have revealed two important aspects of accentedness. First, it lessens intelligibility, at least in non-optimal (e.g. noisy) conditions (Lane, 1967; Lehtonen and Leppanen, 1980). Native speakers can identify more of the words produced by fellow native speakers than by non-native speakers. Similarly, the non-native speaker is able to identify more words produced by native speakers of the target language than by other non-natives, including those who share the same mother tongue (Johansson, 1978). As children mature they become better able to detect departures from the phonetic norms which uniquely characterize their native dialect (Elliot, 1979; Scovel, 1981). The ability of L2 learners to detect accent and to authentically pronounce the target language gradually improves, although L2 learners may never match native speakers of L2 in either respect (Flege, 1980; Flege and Port, 1981; Scovel, 1981). All of these developments suggest a slow evolution in the internal phonetic representation of speech sounds and the ability to detect departures from them.

2. General Method

English phrases were produced by 8 monolingual native speakers of American English and 8 native speakers of French, all women. The French speakers were women 28-48 years of age who had lived in Chicago an average of 13 years at the time of the study and who all spoke English with an obvious French accent in the author's estimation. No attempt was made to control the native dialect of the French speakers. Four were from Paris, two from Belgium, one from Annecy, and one from St. Etienne. The English phrases examined in this study were either read from a list or produced in utterance initial position during a spontaneous story. It was hypothesized that the second, more demanding task would result in less authentic English pronunciation by the French speakers since it presumably required greater attention to the content of the speech being produced as opposed simply to its phonetic form.

A waveform editing program was used to isolate increasingly shorter portions of the digitized speech signals and to cross splice portions of words. Stimuli in the five related experiments to be reported here were normalized for RMS intensity before being converted to analog form with 12-bit resolution, filtered, and presented to subjects (listeners) binaurally at a comfortable listening level. The subjects, all native speakers of American English, were seated in a sound booth before a response box with two buttons. Stimulus randomization, presentation, and data collection were all run under the real-time control of a small laboratory computer (PDP 11/34).

2.1. Experiment I

Two replicate productions by each speaker of the phrases 'Two little birds' and 'Two little dogs' from both of the two speaking conditions (Isolated Phrase, Spontaneous Story) were presented separately for forced choice identification as 'native' (American) or 'non-native' (French). The subjects in this experiment were 10 Americans with a mean age of 32 years who had backgrounds in phonetics, linguistics, or French. Of these subjects, 3 spoke French and 4 indicated 'frequent' exposure to French-accented English.

The subjects were easily able to identify the phrases produced by the non-native speakers as 'non-native.' There was no difference between speaking conditions. In the Isolated Phrase condition the French women were correctly chosen as 'non-native' 91% (1087) of the time; in the Story condition 88% of the time (1056 judgments). There were less than 1% incorrect identifications of American women as 'non-native.'

The phrases produced by most of the individual French Speakers were correctly identified more than 90% of the time. One French speaker's production of 'Two little dogs' was accepted as 'native,' but it should be noted that her language background was correctly identified in 'Two little birds,' probably due to the deletion of /r/ in 'birds.' This suggests that a specific sound may continue to pose difficulty for a language learner, even in the face of a generally close approximation to the phonetic norms of a foreign language. It also suggests that the 'distortion' (or omission) of a single phonetic 'segment' may cue foreign accent.

2.2. Experiment II

Next, the first syllables of the phrase length stimuli used in the first experiment (/tu/) were presented to subjects for paired comparisons. The subjects' task was to determine which member of the pair had been produced by the 'non-native' (French) speaker. The 10 subjects (mean age, 28 years) were sophisticated in that they had training in phonetics or spoke French.

Overall, the subjects correctly chose the /tu/ produced by the French speakers as 'non-native' 95% (5447) of the time. Syllables produced by the Americans were incorrectly chosen only 5% (313) of the time. There was again no difference between the Isolated Phrase and Story conditions, so the remaining experiments examined only speech taken from the Isolated Phrase condition.

2.3. Experiment III

Several measures were taken to generalize the findings of experiments I and II. The 10 subjects (mean age 22 years) in this experiment were students having little or no experience with French or French-accented English. In addition to /tu/ the /ti/ syllable from 'TV reception' and 'TV antennas' was also presented for identification as 'native' or 'non-native.'

Overall, syllables produced by the French speakers were correctly identified 76.7% (2945) of the time. The Americans were incorrectly identified as 'non-native' only 22.3% (857) of the time. Both the frequency of correct identifications and rejections differed significantly from chance ($p < .01$).

Systematic debriefing after the experiment revealed that subjects could not identify the native language of the French speakers. Thus their ability to detect accent probably resulted from sensitivity to departures from English phonetic norms rather than a tacit knowledge of the phonetic characteristics of French-accented English.

2.4. Experiment IV

The phonetic differences that might cue accent were now restricted to the domain of a single phonetic 'segment' (i.e., /t/, /u/, or /i/). One set of hybrid syllables was created by electronically splicing the many /t/'s produced by native and non-native speakers onto a single, good exemplar of an English vowel (/i/ or /u/). Another set was created by splicing the multiple /i/ and /u/ vowels produced by native and non-native speakers onto a single good exemplar of English /t/ (edited from /ti/ and /tu/ syllables, respectively). The American-American and French-American hybrid stimuli were presented in pairs. The subjects' task was once again to identify which of the two stimuli per trial sounded 'non-native.'

Differences sufficient to cue accent resided in just the /t/, /i/, and /u/ segments. The correct recognition rates were: 63% (/t/ + constant /i/); 71% (/t/ + constant /u/); 66% (constant /t/ + /i/); and 69% (constant /t/ + /u/). All were significantly above chance levels ($p < .01$).

2.5. Experiment V

The paired comparison method was again used in the final experiment to examine whether differences in just one portion of a phonetic 'segment' might cue accent. The first 30 ms (± 1 ms) of /tu/ was presented to 9 students in Speech-Language Pathology, all of whom had some training in phonetics. Although these '/t/-burst' stimuli were not recognizable as speech sounds, subjects were instructed to use their own pronunciation of /t/ as a standard for determining which of the two stimuli per trial had been produced by a non-native speaker. No other training feedback, or familiarization was given.

In 68.7% (3164) of the cases the /t/-bursts produced by the French speakers were correctly identified as 'non-native' ($p < .01$). Eight of the 9 subjects performed at above-chance rates. Of the 8 French talkers, 6 were identified above chance ($p < .01$).

3. Discussion

This study shows that listeners are able to detect accent on the basis of small differences in the quality of /u/ or /i/, or in the place of articulation for /t/. One might simply assume this finding to be inevitable, since children learning English and French as native languages learn to produce /i/, /u/, and /t/ according to the phonetic norms of those languages. However, the acquisition process probably never involves the direct comparison of the kind of small phonetic differences that distinguished the native French and English speakers in this study. Our subjects might also have learned to 'filter out' the phonetic differences that distinghuish cognate sounds in French and English since they are not relevant to phoneme identity in either language. These results demonstrate that adults do possess the ability to detect within-category (subphonemic) differences between language varieties, and to use this information in detecting foreign accent.

The present results suggest that listeners develop phonetic category prototypes against which to judge specific speech sounds. Although we have no direct evidence concerning the articulatory differences distinguishing the /t/ produced by the native French and English talkers, it is mostly likely a difference in the width and place of tongue contact. The French /t/ was sufficiently far from the range of auditory properties acceptable as /t/ that it was rejected as an English sound. An important question for future research is how 'accentual' information and the information cuing category identity are processed during speech perception. A recent study indicated that subjects' responses to stimuli changed as a function of whether the stimuli were presented as 'Spanish' or 'English' (Elman, Diehl and Buckwald, 1977). A 'prototype' model of accent detection fails to predict such a finding since it presumes that information relevant to phonetic category identity must be processed prior to assessment of 'accent' (i.e., departures from phonetic norms for a particular sound category). One possibility raised by this finding is that accentual information is processed integrally along with information leading to the identification of a phonetic category.

References

Elliot, L. (1979). Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using material with controlled word predictability. *J. Acoust. Soc. America*, **66**, 651-653.

Elman, J., Diehl, R. and Buchwald, S. (1977). Perceptual switching in bilinguals. *J. Acoust. Soc. America*, **62**, 971-974.

Flege, J. (1980). Phonetic approximation in second language acquisition. *Lang. Learning*, **30**, 117-134.

Flege, J. and Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Lang. and Speech*, **24**, 125-146.

Johansson, S. (1978). *Studies in Error Gravity.* Gothenberg Studies in English, **44**. Department of English, Univ. of Gothenberg.

Lane, H. (1963). Foreign accent and speech distortion. *J. Acoust. Soc. America*, **35**, 451-453.

Lehtonen, J. and Leppanen, H. (1980). Observations on the effect of background noise on the understanding of English by Finnish students. *Papers on Speech Res.*, **2**, 108-116. Inst. Finnish Lang. and Communication, Univ. of Jyvaskyla.

Scovel, T. (1981). The recognition of foreign accents in English and its implications for psycholinguistic theories of language acquisition. In: J.-G. Savard and L. Laforge (eds.). *Proceed. 5th Int. Cong. Int. Assoc. of Applied Ling.* Quebec: Laval Univ. Press.

Patterns of English Word Stress by Native and Non-native Speakers

J. Fokes, Z.S. Bond, M. Steinberg
Athens, U.S.A.

1. Introduction

While early impressionistic studies (Pike, 1945) defined stress as modification of pitch, loudness, and duration, later investigations, as summarized by Lehiste (1970), included acoustical measurements of fundamental frequency, amplitude, and length in order to determine the extent of their contribution to stress. Taking into account these acoustic parameters of speech sounds, stress has been investigated in both perceptual and production modes. Perceptually, loudness appears to be a weaker cue for stress than either duration or pitch. The perception of stress, however, does not seem to depend upon any one, or any combination of acoustic parameters. Speakers apparently have considerable latitude in the use of acoustic parameters. Lieberman (1967) found a higher percentage use of increases in frequency, amplitude, and length in stressed syllables by native English speakers on both initial stressed and second stressed syllables in words.

Non-native speakers of English either fail to use certain acoustic parameters or use them inappropriately in their speech patterns. Noteworthy is Adams' study (1979) in which she found similar performance by both native and non-native speakers for the production of stressed syllables but not for unstressed syllables in nursery rhymes and contextual speech.

This study was designed to compare the acoustical characteristics of fundamental frequency, amplitude, and length of English stressed and unstressed syllables in speech of native and non-native speakers and to interpret the findings in terms of word stress.

2. Method

2.1. Subjects

Two groups of subjects participated. Three were monolingual native American English speakers attending a midwestern university. Six were non-native English speakers enrolled in an English pronunciation class. Though their background in English differed, the students had adequate command of English to enroll in the University, according to the scores made on the Michigan Test of English Proficiency. Each spoke natively a different language: Chinese (Taiwan), Hausa, Japanese, Korean, Persian, and Spanish (Colombia). All were male.

2.2. Materials

Five types of test words were selected: 1) prefixed words with second syllable stress, such as *confess*; 2) the same words with an *-ion* suffix, e.g. *confession*; 3) and 4) words which change stress pattern upon suffixation, such as *confirm* and *confirmation*; 5) words of similar phonetic shape but initial syllable stress, such as *conquer*. There were 25 different words, five of each type.

2.3. Procedure

Each subject recorded the prepared word list three times, in three different orders. For each token, the peak frequency and amplitude of the first two syllables were measured, using a Voice Identification Pitch Analyzer (PM 100). The length of the syllables was measured from the amplitude trace. Tokens which could not be segmented were measured from spectrograms (approximately 15% of the corpus). Reliability was obtained on ten items from the two researchers who measured the acoustic parameters. There was 90 percent agreement for frequency within three Hz or less, 100 percent agreement for amplitude within two dB or less, and 70 percent agreement for length within 10 msec or less.

Average fundamental frequency, amplitude, and length were computed for each measured syllable for each of the five word types. Data from the three Americans were combined for comparison with each of the measures for the non-native English speakers.

3. Results

3.1. Fundamental frequency

The Americans exhibited the expected fundamental frequency differences distinguishing stressed from unstressed syllables in the prefixed two and three syllable words (*confess*, *confession*, *confirm*). The Americans tended to produce syllables with secondary stress and reduced syllables, as in *confirmation*, with essentially the same fundamental frequency. In words such as *conquer*, the unstressed syllable was produced with a somewhat higher average fundamental frequency than the stressed syllable.

All non-native speakers but the Persian used fundamental frequency in a way similar to the American pattern for the prefixed one and two syllable words, as shown in the first three comparisons in Figure 1, panel a. The stressed syllable was associated with a higher fundamental frequency than the unstressed. The Japanese speaker, however, tended to show a greater relative pitch difference between the two syllables than the other speakers.

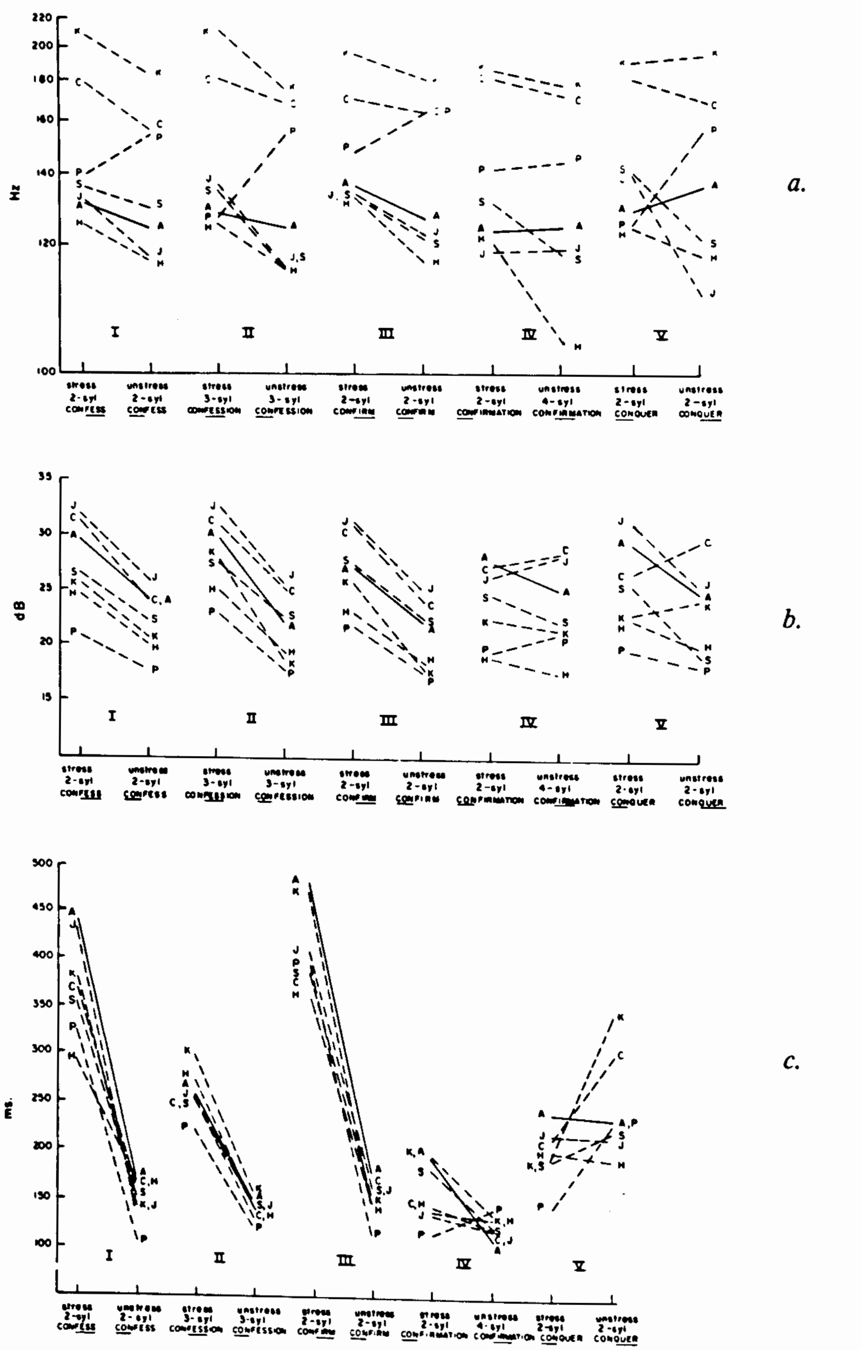


Figure 1. Comparison of stressed and unstressed syllables in the speech of native and non-native speakers. Speakers are identified by letter: A-American, C-Chinese, H-Hausa, J-Japanese, K-Korean, P-Persian, and S-Spanish. Panel a: Change in frequency (Hz) from stressed to unstressed syllables; Panel b: Change in amplitude as measured in decibels (dB); Panel c: Change in length as measured in milliseconds (ms).

The Persian speaker consistently used a higher fundamental frequency with unstressed than the stressed syllables, contrary to the American pattern. For words of the type *confirmation* and *conquer*, both the Spanish and the Hausa speakers used an extensive frequency drop from the stressed to unstressed syllable (Comparisons IV and V in Figure 1).

3.2. Relative amplitude

The American speakers and all the non-native speakers seemed to employ similar relative amplitude differences for the two and three syllable prefixed words, in that the amplitude of stressed syllables was higher than that of the unstressed syllables. These data are included in the first three comparisons of Figure 1, panel b. For the first two syllables of words such as *confirmation*, three of the non-native speakers (Persian, Chinese, Japanese) produced the second syllable with an increase in amplitude, rather than with the decrease found for native speakers. For words such as *conquer*, the Korean and the Chinese speakers tended to produce both syllables with equal amplitude.

3.3. Length

As in their use of amplitude, the non-native speakers were very similar to the Americans in using length as a correlate of stress for the two and three syllable prefixed words, as shown in the first three comparisons of Figure 1, panel c. Although Americans made the reduced syllable shorter than the initial syllable in *confirmation*, only the Korean and the Spanish speaker did so as well. The other speakers tended to make both syllables approximately equal in length. For words such as *conquer* the Americans as well as the Japanese and Hausa speaker used syllables of nearly equal length. The Persian, Chinese, and to a lesser extent, the Spanish speaker, prolonged the second syllable so that its length exceeded that of the first.

4. Discussion and Conclusion

The six non-native English speakers were able to control appropriately the acoustical parameters associated with lexical stress, although the Persian speaker was less successful than the others in producing appropriate fundamental frequency differences. They typically produced words such as *confess*, *confirm*, and *confession* with appropriate fundamental frequency, amplitude, and length differences between the stressed syllable and the prefix, though on occasion with somewhat exaggerated values.

The non-native speakers were considerably less successful in controlling the acoustical parameters in words such as *confirmation*. They produced a full rather than a reduced vowel in the second syllable which, probably as a consequence, caused them to use inappropriate fundamental frequency, amplitude and length relationships.

Words such as *conquer* were also difficult for non-native speakers, although these words follow the typical initial syllable stress pattern for English nouns. The non-native speakers might have been influenced by the stress rules for prefixed words; clearly, they were not employing the appropriate acoustical correlates of stress.

Although Adams' (1979) study differed from this one in a number of respects, her findings were similar: non-native speakers of English tend to use appropriate acoustic parameters for stressed syllables but have difficulty producing unstressed syllables appropriately.

Although there are obvious limitations because of the single subject per language, two suggestions can be made. Continued investigations of stress patterns on multisyllabic words in citation as well as contextual speech need to be undertaken. Secondly, the instructional value of this research is that effective training of non-native speakers should contain practice not only in the use of stressed and unstressed syllables but in the use of the corresponding full and reduced vowels.

Acknowledgements

Support for portions of this project was provided by the Ohio University Baker Fund.

References

Adams, Corinne (1979). *English Speech Rhythm and the Foreign Learner.* The Hague, Paris and New York: Mouton Publishers.

Lehiste, Ilse (1970). *Suprasegmentals.* Cambridge, Massachusetts and London, England: The M.I.T. Press.

Lieberman, Philip (1967). *Intonation, Perception and Language.* M.I.T. Research Monograph No. 38. Cambridge, Massachusetts: The M.I.T. Press.

Pike, Kenneth L. (1945). *The Intonation of American English.* Ann Arbor: University of Michigan Publications, *Linguistics*, Vol. 1.

Korrektiver Ausspracheunterricht auf auditiver Basis

H. Grassegger
Graz, Austria

1. Einleitung

Im Ausspracheunterricht kommt dem *auditiven* Aspekt eine grundlegende – weil im Lernprozeß zeitlich vorgeordnete – Rolle zu. Das zeigen Lautsubstitutionen im Fremdsprachenerwerb, die auf charakteristische Hörfehler zurückzuführen sind. Solche Hörfehler haben wenigstens *eine* Ursache in dem Erlebnis phonetischer Ähnlichkeit, das den Hörer dazu bringt, Zuordnungen des fremdsprachlichen Lautes zu Elementen seines muttersprachlichen Systems vorzunehmen. Auf der artikulatorischen Seite erschienen diese Zuordnungen wieder als Lautsubstitutionen.

Wenn man die prinzipielle Bedeutung des Konzepts der auditiven Ähnlichkeit als eine (sicherlich nicht einzige) Ursache für lautliche Interferenzerscheinungen anerkennt, erhebt sich die Frage, wie auditiv-phonetische Ähnlichkeit gemessen werden kann, bzw. welche Schlußfolgerungen aus den Ähnlichkeitsurteilen für einen korrektiven Ausspracheunterricht zu ziehen sind.

In einer ersten Studie zu diesem Problemkreis wurde versucht, die auditive Ähnlichkeit eines stimmlosen dentalen Lateralfrikativs mit einer Reihe von möglichen Substitutionen im Urteil deutscher Hörer zu erfassen. Die Anregung dazu kam von einer Darstellung der englischen und deutschen Substitutionen für den walisischen Lateralfrikativ [¢] (Ternes, 1976). In Hör- und Transkriptionsübungen wurde nämlich festgestellt, daß deutsche Hörer das walis. Phon mit [çl], englische Hörer hingegen mit [θl] wiedergeben. Vom Standpunkt der artikulatorischen Nähe aus wären auch die Kombinationen [sl], [ʃl] bezw. [fl] (vgl. den Eigennamen *Fluellin* aus walis. *Llewllyn*) mögliche Substitutionen. Daß diese nicht gewählt werden, läßt sich z.T. mit phonotaktischen Gesetzmäßigkeiten erklären (so ist z.B. anlautendes [sl] im Deutschen nicht zulässig), beruht aber wahrscheinlich auch auf der Irrelevanz der artikulatorischen Nähe für Lautsubstitutionen.

2. Methode

Für das Experiment wurden aus einem Demonstrationsband zu phonetischen Transkriptionsübungen 14 Lautpaare zusammengeschnitten, die jeweils den dentalen Lateralfrikativ und eines der sieben folgenden Phone enthielten: [θ,s,ʃ,f,l,λ̥,ç]. Die Laute wurden so angeordnet, daß jeweils 7 Paare den Lateralfrikativ als ersten (POSITION A) bzw. als zweiten Laut (POSITION B) enthielten. In ähnlicher Weise wurde ein weiterer für jede Position aus je 5 Paaren bestehender Test hergestellt, in welchem die Vergleichsstimuli Lautkombinationen mit einem stimmhaften Lateral waren: [θl, ʃl,sl,fl,çl]. Diese insgesamt 24 Testpaare wurden in randomisierter Form zu einem Testband zusammengestellt und 22 Hörern (Studenten deutscher Muttersprache mit geringer phonetischer Vorbildung) vorgegeben. Die Vpn. sollten ihr Urteil über jedes Testpaar [ɬ]/[. . .] bzw. [. . .]/[ɬ] skalar von vollkom men unähnlich (0) bis maximal ähnlich (10) eintragen. Die Ergebnisse sind in Abb. 1 für Position A und in Abb. 2 für Position B graphisch dargestellt.

3. Resultate

Jeder Vergleichsstimulus ist durch eine Säule repräsentiert, deren Höhe den durchschnittlichen Ähnlichkeitsgrad zwischen 0-10 angibt. Die Differenzen der Durchschnittswerte wurden mittels t-Test auf ihre Signifikanz geprüft. Steht eine Lautsäule isoliert, so ist die Differenz ihres Ähnlichkeitswertes zu den benachbarten Lauten signifikant. Weisen die Differenzen keine signifikanten Unterschiede auf, werden die betroffenen Lautsäulen zu einem Block zusammengefaßt. Diese Blöcke stellen also Ähnlichkeitsklassen von Lauten dar, die alle untereinander ähnlich sind.

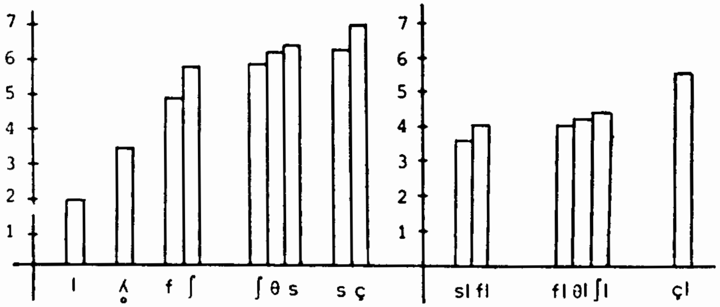


Abb. 1. Urteile deutscher Hörer über die Ähnlichkeit von Einzellauten bzw. Lautkombinationen zum Lateralfrikativ [ɬ] (Position A: [ɬ]/[. . .]).

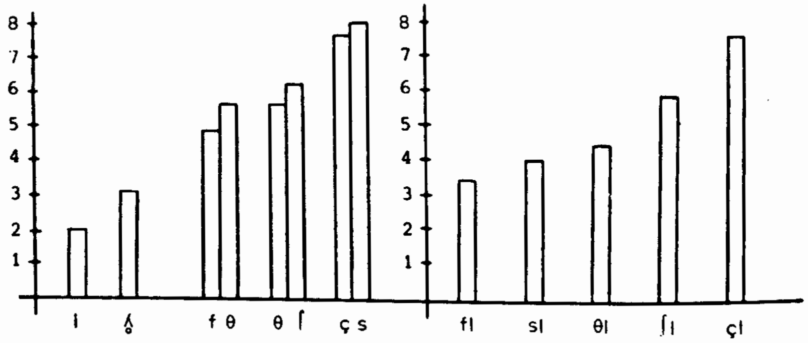


Abb. 2. Urteile deutscher Hörer über die Ähnlichkeit von Einzellauten bzw. Lautkombinationen zum Lateralfrikativ [ɬ] (Position B: [. . .]/[ɬ]).

Dadurch kann ein und derselbe Laut (z.B. [ʃ] in Position A) zwei Ähnlichkeitsklassen angehören. Im Hinblick auf die noch nicht näher bestimmten auditiven Qualitäten der (signifikant unterschiedlichen) Ähnlichkeitsklassen ist daraus zu schließen, daß der Laut [ʃ] zwei auditiv relevante Eigenschaften besitzt, von denen er eine mit [f], die andere mit [s] und [θ] teilt.

Aus den Abbildungen ist ersichtlich, daß in beiden Positionen die Laterale [l,λ̥] die geringste Ähnlichkeit zu walis. [ɬ] aufweisen, wahrend [s,ç] and der Spitze der Ähnlichkeitsskala liegen. Dieses Ergebnis steht im Einklang mit der beobachteten deutschen Substitutionsrealisation ([çl]), welche den palatalen Frikativ enthält. Gegen eine Ersetzung des Ziellautes [ɬ] durch eine Kombination mit [s] sprechen anscheinend trotz der auditiv-phonetischen Ähnlichkeit des [s] mit [ɬ] phonotaktische Gründe; diese schränken die Wahl der nach dem Kriterium der auditiven Ähnlichkeit corhandenen Substitutionsmöglichkeiten ein.

Bei den Lautkombinationen weist, wie nach den Beobachtungen von Ternes (1976) zu erwarten, [çl] in beiden Positionen die größte Ähnlichkeit zum walis. Lateralfrikativ auf. Bemerkenswert ist die in Position A und Position B stark unterschiedliche Klassenbildung. Signifikanzprüfungen ergaben, daß nur in Position B die Werte *aller* Lautkombinationen signifikant verschieden sind. Überdies sind die absoluten Ähnlichkeitswerte in Position B für [sl], [ʃl] und [çl] signifikant höher als in Position A. Auch bei den Einzellauten liegen in Position B Vergleichsstimuli mit hohen Ähnlichkeitswerten (also [ʃ,s,ç]) signifikant über den entsprechenden Ähnlichkeitsgraden in Position A. Die Ursache dieser unterschiedlichen Bewertung ist darin zu suchen, daß der subjektive Ähnlichkeitsgrad für die einzelnen Substitutionsmöglichkeiten offensichtlich von der Anordnung der zu vergleichenden Laute abhängig ist (anordnungsspezifische Ähnlichkeit). Wenn der bekannte (≡in der AS der Vpn. vorhandene) Laut an erster Stelle des Vergleichspaares (≡ Position B) steht, sind die Ähnlichkeitswerte wesentlich höher als im umgekehrten Fall. Dies gilt wenigstens für jene Stimuli die die ausgangssprachlichen Frikativa [s, ʃ,ç] enthalten, während die Beurteilung der nicht ausgangssprachlichen Laute [θ,λ̥] bzw. der Lautkombination [θl] keine hinsichtlich der Position zum Lateralfrikativ (A oder B) interpretierbaren Unterschiede aufweist. Auch bei den Vergleichslauten [l], [f] bzw. bei der Lautkombination [fl] lassen sich keine signifikanten, positionsbedingten Unterschiede feststellen, was wohl auf die im Vergleich zu [ɬ] geringe bzw. nicht vorhandene Geräuschintensität von [f,fl] bzw. [l] zurückzuführen ist.

Für die Vergleichsstimuli [s, ʃ,ç] bzw. [sl, ʃl,çl] läßt die unterschiedliche Bewertung in den beiden Positionen aber auf eine kategorielle Perzeption der Teststimuli schließen. Die bei bekannten Lauten automatisch und daher 'zielsicherer' ablaufende Kategorisierung führt dazu, daß die Bewertung des zweiten (fremden) Lautes viel exakter vorgenommen werden kann, da der Bezugspunkt fixiert ist. Ein zuerst dargebotener unbekannter Laut ist wegen der nicht-automatisierten Vorkategorisierung schwerer zu speichern; das Ähnlichkeitsurteil über den als zweiter Stimulus einlangenden AS-Laut fällt dementsprechend 'unsicherer' aus, was sich in den vergleichsweise geringeren Ähnlichkeitswerten niederschlägt.

4. Schlussbemerkung

Für ein gezieltes Hördiskriminationstraining ergibt sich daraus die didaktische Konsequenz, bei kontrastiven Übungspaaren zuerst das bekannte AS-Phon und dann erst das ZS-Phon zu präsentieren. Welche Lautpaare zur auditiven Einübung eines bestimmten ZS-Phons zweckmäßigerweise überhaupt eingesetzt werden sollen, ergibt sich wiederum aus der Ähnlichkeitshierarchie innerhalb der von AS-Sprechern produzierten Substitutionslaute bzw. -lautkombinationen.

Die Erstellung eines Programms für das Hördiskriminationstraining zu einem bestimmten 'Problemlaut' der ZS hat also in drei Stufen zu erfolgen:

1. Diagnostische Erfassung der AS-Substitutionen, welche die AS-Sprecher bei der Konfrontation mit dem ZS-Laut produzieren. Bei dieser Bestandsaufnahme müssen auch die erst in jüngerer Zeit starker beachteten Einflüsse von dialektalen und soziolektalen Hörgewohnheiten zusätzlich verwertet werden.

2. Die Aufstellung einer Ähnlichkeitshierarchie dieser Substitutionsmöglichkeiten zum betreffenden ZS-Laut mit Hörern der AS-Sprache (bzw. der ausgangssprachlichen Variante). Da der vorliegende Test nicht so sehr unmittelbar didaktische Zielsetzungen für deutsche Hörer verfolgte, wurden auch die im dt. Lautinventar nicht vorkommenden Stimuli [λ̥,θ,θl] aufgenommen; jene diskutierten Testergebnisse, die für hördiskriminatorische Übungen relevant sind, lassen sich aber auch ausschließlich an Hand solcher Stimuli gewinnen, die zum Lautbestand der deutschen Sprache gehören.

3. Herstellung von Übungsprogrammen, in denen die Gliederung des Übungsmaterials und die Schwerpunktsetzung auf der Basis der gewonnenen Ähnlichkeitshierarchien erfolgen. Unter Berücksichtigung des Phänomens der anordnungsspezifischen Ähnlichkeit werden wenigstens in den ersten Trainingsphasen bei Kontrastübungen die AS-Laute *vor* den ZS-Lauten zu präsentieren sein.

Reference

Ternes, E. (1976). *Probleme der kontrastiven Phonetik.* Hamburg, S. p. 31 ff.

Phonic Transfer: The Structural Bases of Interlingual Assessments

A. R. James
Amsterdam, the Netherlands

1. Introduction

For a number of years now, dissatisfaction with the power of 'predictive' or even 'diagnostic' structural phonologically-based comparisons of native language (NL) and target language (TL) to explicate the occurrence and above all the variability of TL pronunciation forms (cf., e.g. the attempts of Ritchie (1968) and Michaels (1974) to analyse the problem within a feature framework) has led to increasing attention being paid to extra-linguistic determinants of L_2 pronunciation variation: predominantly those relating to verbal task and level of proficiency. Much of this dissatisfaction with structurally based models of explanation stems from i) the inherent restrictions imposed by the descriptive frameworks adopted and, ii) partly deriving from this, a simplified view of the dynamics of second language production and acquisition. The assumption being that learner TL linguistic behaviour will manifest to a greater or lesser degree the structural properties of the NL phonology (whether specified in terms of phonemes, features or, latterly, syllable structure primes) and/or the physically, predominantly acoustically, established phonetic properties of the NL. It is the purpose of the present paper, in an examination of a persistently knotty, but typical, problem in the analysis of second language segmental production, to suggest ways in which both descriptive frameworks and, concomitantly, structural perspectives on L_2 sound learning and performance need to be broadened in order to account adequately for the observed data.

2. Structural bases

2.1.

Within present frameworks, the relations between on the one hand observable structural determinants of TL pronunciation behaviour and on the other observed variability in the pronunciation of TL forms are difficult to capture. While accepting that all manner of psycho-, socio- and extra-linguistic factors must clearly co-determine patterns of sound system acquisition and production with their inherent variability, any attempt to establish a defining link *between* structural properties of both NL and TL, *and* such variability must in the first instance come to terms with the structural 'what' that is to be produced and acquired in a TL. In contrast to phonologists and phoneticians working on child language, researchers in foreign language acquisition have been strangely reluctant to address this issue. Production and acquisition of a TL sound system involves command of a) a phonological structure, b) a phonetic structure, and c) an articulatory structure (cf. James 1983). In the most general terms, these structures may be characterized as, respectively, a) the systematic arrangement of elements defined in terms of their relative distinctiveness; b) the systematic arrangement of physically measurable properties associated with such elements; and c) the systematic arrangement of motor configurations associated in turn with such properties and elements.

2.2.

The proximity of any L_2 utterance to a target norm may be assessed with regard to the degree of proficiency in each or all of these structural components of the sound system, which involves not only command of the elements themselves but also of the patterns of association obtaining between the forms of the different components: e.g. for any given 'phoneme', there must be an associated set of phonetic properties and an associated articulatory configuration. Thus for /s/, the structural associations may take the form

$$/s/ \sim \begin{bmatrix} \text{alveolar} \\ \text{fricative} \\ \text{etc.} \end{bmatrix} \sim \begin{Bmatrix} \text{tip-blade raising} \\ \text{tip-blade advancing} \\ \text{etc.} \end{Bmatrix}$$

the latter being expressable in terms of, for example, the articulatory parameters of Hardcastle (1976) or Ladefoged (1980). The direct structural influences of the NL may be present to a greater or lesser degree within these structural components and the associative links between them. The degree to which they are present is a product of the foreign language learner's *own assessment* of the relatedness of TL and NL structures. NL structures are not automatically projected willy nilly onto TL forms. Of course relatability judgements fluctuate: diachronic patterning of gradual approximation to TL forms as shown, e.g., by Dickerson (1975), is reflective of increasingly negative values of relatability assessment, which in turn corresponds to developing proficiency in the TL. Synchronically, such assessments are crucially mediated by structural factors and variation itself by pressures of the suprasegmental context. Sound elements which are perhaps most directly available to cross-linguistic comparison are those which constitute the phonological 'alphabet' of a language, i.e. the minimal concatenative elements of lexical entries.

3. Interlingual assessments

3.1.

In the light of the discussion so far, consider now the explication of a well-known pattern of Dutch English pronunciation which has resisted analysis within traditional frameworks of reference; i.e. the observation that English [ð] is predominantly produced as a dental stop or tap word- or syllable-initially and as an alveolar fricative finally. In its segmental inventory Dutch has [t̪ d̪] (dental stops) as well as [s̠ z̠] (post-alveolar fricatives). Intuitively, it would appear that this distribution of non-target variants of [ð] in Dutch English may be related to some kind of positional strength hierarchy as, eg., proposed by Hooper (1976), whereby in syllable-initial position, i.e. a 'strong' position, a 'strong' consonant - here a stop - might be more likely to occur than a 'weaker' one (e.g. a fricative), and vice-versa in syllable-final position as a 'weak' environment. However, why does German English for instance have primarlily [s] or [z] for [ð] in either position, when German has both dental-alveolar stops and fricatives?

3.2.

The assessment procedure of a Dutch speaker of English with regard to [ð] may be reconstructed as follows:

phonological:	gross distributional relatedness to a number of C types in the NL		
phonetic:	place – dental/alveolar	~	NL [t̪ d̪ s̠ z̠]
	manner – fricative	~	NL [s̠ z̠]
articulatory:	tip advancing, raising, etc.	~	NL [t̪ d̪]

The suprasegmental context in which the sound is embedded in actual production then determines which potential form of the segment is realized, i.e. conditions the source of the phonic transfer. However, within a hierarchically ordered phonological framework the structural status itself of a given sound unit is determined by properties of the suprasegmental context (cf., e.g., James, 1982). Consistent with 'natural' interpretations of phonological structure (cf. Linell, 1982) and current 'non-translational' models of speech production (cf. Fowler et al., 1980), structural features of the suprasegmental context may be seen to be directly 'realized' - as various types of linguistic constraints - in speech performance. One such linearly effective constraint on the speech syntagma derives from the phonological analysis of syllables as comprising the constituents onsets and rhymes, phonetically characterized by respectively 'consonantal' and 'vocalic' feature values and articulatorily by respectively closing and opening gestures of the vocal tract (cf. also

Fowler, 1980). Globally, given phonoactic structure and degree of phonological 'strength' (James, 1982), the constraint may be expressed as maximal pressure for occlusion at syllable onsets, maximal pressure for aperture at rhyme (nuclei). Another such constraint, relevant to languages such as English, Dutch and German, derives from the phonological analysis of syntagmatic units of speech into 'heads' and 'clitics' i.e. proclitics and enclitis (cf. Knowles, 1974), phonetically characterized by local rate of utterance (cf. also Crompton, 1981) expressable as accelerando to head, rallentando on head and enclitic, and articulatorily by velocity of gesture. At syllable level, the constraint may be expressed as maximal pressure for high velocity at onsets, maximal pressure for low velocity at rhymes (nuclei and codas).

If, on the basis of relatedness assessments - (phonological), phonetic or articulatory -, there is a form of the NL available meeting these constraints, it will be transferred. In the case of Dutch English, the articulatory configuration associated wiih NL [t̪ d̪] is commensurate with the requirements of suprasegmental position syllable-initially (i.e. at onset), therefore a [t̪] or [d̪] will be transferred, whereas in syllable-final position the choice of a fricative articulation is suprasegmentally more motivated, the produced form thus reflecting the phonetic association made between place and manner properties of TL [ð] and NL 'equivalents', an alveolar fricative being transferred. In the first example, the structural basis of phonic transfer is articulatory, in the second, phonetic. In German English on the other hand, association is restricted to that of phonetic structure. Standard German [t d] involves an articulatory configuration of tip and blade raising and advancing, [s z] blade raising and advancing (Wängler, 1974; Lindner, 1975), thus there is no articulatory structure within the NL available to meet the requirements of the suprasegmental context. However, in all cases the suprasegmental context will exert an influence on the 'manner degree' of segments initially and finally via the constraints of tract occlusion and articulatory velocity. Thus, syllable-initially in unstressed syllables the target [ð] in Dutch English is observed as a dental tap [ɾ̪], indicating that the intrinsically faster closure rate of a tap articulation (as opposed to a stop) is necessary for the completion of occlusion in conditions of high utterance rate as associated with unstressed syllables in phrase proclitic position.

4. Conclusion

The present analysis hopes to show that the linguistic explication of TL pronunciation data necessitates a more differentiated view of the structural determinants of phonic transfer than has seemed to be possible within received phonological and phonetic frameworks of reference.

References

Crompton, A. (1981). Phonetic representation. Unpubl. paper, Univ. of Nottingham.

Dickerson, L.J. (975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly* **9/4**, 401-407.

Fowler, C.A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics* **8**, 113-133.

Fowler, C.A., Rubin, P., Remez, R.E. and M.T. Turvey (1980). Implications for speech production of a general theory of action. In: *Language Production. Volume 1. Speech and Talk.* B. Butterworth (ed.), Academic Press: London, 373-420.

Hardcastle W.J. (1976). *Physiology of Speech Production.* Academic Press: London.

Hooper, J.B. (1976). *An Introduction to Natural Generative Phonology.* Academic Press: London.

James, A.R. (1982). The suprasegmental loci of segmental variants in Dutch English. Verslag van de 158e Vergadering van de Nederlandse Vereniging voor Fonetische Wetenschappen, 10-17.

James, A.R. (1983). Transferability and dialect phonology: Swabian English. In: *Dialektphonologie und Fremdsprachenerwerb.* A. James, B. Kettemann (eds.), Narr: Tübingen, 164-188.

Knowles, G. (1974). The rhythm of English syllables. *Lingua* **34**, 115-147.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language* **56**(3), 485-502.

Lindner, G. (1975). *Der Sprechbewegungsablauf. Eine phonetische Studie des Deutschen.* Akademie-Verlag: Berlin.

Linell, P. (1982). The concept of phonological form and the activities of speech production and perception. *Journal of Phonetics* **10**, 37-72.

Michaels, D. (1974). Sound replacements and phonological systems. *Linguistics* **126**, 69-81.

Ritchie, W. (1968). On the explanation of phonic interference. *Language Learning* **18**, 183-197.

Wangler, H.-H. (1974). *Atlas deutscher Sprachlaute.* Akademie-Verlag: Berlin.

Die Syllabisch-Akzentologischen Modelle der russischen Substantive

E. Jasová
Banska Bystricá, Czechoslovakia

1. Einleitung

Unser Beitrag stellt eine linguistische Beschreibung der syllabisch-akzentologischen Beziehungen der russischen Substantive und ihre Anwendung für den Unterricht in der russischen Sprache vor. Diese linguistische Beschreibung bietet dem Lehrbuchautor und dem Lehrer ein System der zahlreichsten und in der Frequenz höchsten Substantivmodelle, die für Übungen zum Spracherwerb verwendet werden können. Zu den Grundproblemen des Unterrichts in der russischen Sprache als Fremdsprache in der slowakischen bzw. tschechischen Schule gehört nach unserer Auffassung auch die Lehre von der Betonung, die durch einen Komplex von phonetischen und morphologischen Eigenschaften gekennzeichnet ist. Das typische Merkmal der russischen Sprache ist: a) die freie Betonung, die auf die erste bis letzte Silbe des Wortes fallen kann und b) die veränderliche Betonung, die sich in einem Paradigma des Wortes verschieben kann, im Unterschied zur slowakischen oder tschechischen Sprache, in denen die Betonung an die erste Silbe des Wortes gebunden ist. Der Hauptgegenstand unserer Untersuchung ist die syllabisch-akzentologische Beziehung der russischen Substantive unter Zugrundelegung des Frequenzwörterbuches der russischen Sprache (Častotnyj slovar' russkogo jazyka, 1977). Diese wird an einer begrenzten Zahl von Substantiven (unabgeleiteten und abgeleiteten, einfachen und zusammengesetzten, 'einheimischen' und 'fremden') unter zwei Aspekten untersucht. Sie werden zusätzlich in einer Übersicht nach Silbenzahl und sinkender Frequenz dargestellt.

2. Untersuchungsaspekte

2.1. Der syntagmatische Aspekt

Wir stellen die Distribution der freien Betonung vom Standpunkt der betonten Silbe und der Silbenzahl des Substantivs nach der Grundform des Wörterbuches (Nom. Sing.) dar. Wir legen die absolute und relative Frequenz fest. (Vgl. die Tabelle Nr. I).

Tabelle I.

	Substantive			
	Anzahl		Frequenz	
	Absolut	%	Absolut	%
1. silb.	411	10,44	42 715	17,08
2. silb.	1 395	35,44	99 745	39,89
3. silb.	1 069	27,16	61 502	24,59
4. silb.	522	13,26	24 281	9,71
5. silb.	377	9,58	16 246	6,50
6. silb.	132	3,35	4 942	1,98
7. silb.	25	0,64	561	0,22
8. silb.	5	0,13	75	0,03
Total	3 936	100,-	250 067	100,-

Wir sehen, daß in der russischen Sprache die zahlreichsten (1 395, d.i. 35,44%) und auch die in der Frequenz höchsten (99 745, d.i. 39,89% aus der Gesamtzahl von 3 936 Substantiven) die *zwei*silbigen Substantive sind. An zweiter Stelle stehen nach der Zahl die *drei*silbigen Substantive (1069, d.i. 27,16%) und auch nach der Frequenz (61 502, d.i. 24,59%) usw. Die Anzahl und auch die Frequenz der *sechs-*; *sieben-* und *acht*silbigen Substantive sinkt deutlich ab. Vom Standpunkt der grammatischen Kategorie des Geschlechts (männlich, weiblich, sächlich) stehen nach Zahl und Frequenz die Substantive männlichen Geschlechts (1 721, d.i. 43,72%, und 107 195, d.i. 42,86%) an erster und die Substantive weiblichen Geschlechts an zweiter Stelle. Die Substantive sächlichen Geschlechts haben die niedrigste Anzahl, aber nach unserer Untersuchung eine relativ hohe Frequenz. Unter dem syntagmatischen Aspekt stellen wir nach dem Wörterbuch 14 grundlegende syllabisch-akzentologische Modelle nach Anzahl und Frequenz fest. Diese haben die Betonung auf derselben Silbe bzw. demselben Morphem im ganzen Paradigma. Das ist die 'feste' Betonung. Sie zeichnen sich nach der Akzent-Theorie von V. Straková (1978) durch ihre akzentologische Relevanz aus. Im ganzen Paradigma variiert folglich nur die Zahl unbetonter Silben bei den flektierten Morphemen. Bei *zwei*silbigen Substantiven beobachten wir in der Distribution der Betonung ein bestimmtes Gleichgewicht. Die Zahl der Substantive mit dem syllabisch-akzentologischen Modell 'betont/unbetont' (x́ x) beträgt 51,04%, und die Frequenz ist 51,25%. Bei dem syllabisch-akzentologischen Modell 'unbetont/betont' (x x́) beträgt die Zahl 48,96%, und die Frequenz 48,75%. Bei den *drei*silbigen Substantiven tritt in der Zahl von 51,45% und auch in der Frequenz von 49.03% des syllabisch-akzentologischen Modells 'unbetont/betont/unbetont' (x x́ x) deutlich hervor. Bei den *vier*silbigen Substantiven beeindrucken nach der Zahl (48,47% und 38.31%) und auch in der Frequenz (60,73% und 27,49%) die zwei Modelle 'unbetont/betont/unbetont/unbetont' (x x́ x x) und 'unbetont/unbetont/betont/unbetont' (x x x́ x). Wir beobachten bei den russischen Substantiven die Tendenz, die mittlere Silbe zu betonen. Dieselbe Tendenz bemerken wir auch bei den *fünf*silbigen

Substantiven. In der Zahl von 79,58% und auch in der Frequenz von 80,11% tritt das Modell 'unbetont/unbetont/betont/unbetont/unbetont' (x x x́ x x) auf.

2.2. Der paradigmatische Aspekt

Wir erforschen: a) die syllabische Variabilität von Wortformen des Paradigmas und b) die Bewegung der Betonung in den Paradigmen der Substantive. Die syllabische Variabilität von Wortformen der Paradigmen der russischen Substantive ist durch die Flexion bedingt. Die russische Sprache, ähnlich wie die slowakische bzw. die tschechische, gehört zu dem flektierenden Typ von Sprachen. Auf Grund der Flexion unterscheiden wir: a) die isosyllabisch-akzentologischen Modelle 'śkóla, śkóle, śkólu...' (x́ x) 'betont/unbetont' und b) die heterosyllabisch-akzentologischen Modelle 'śkól' (Gen. Plur., eine Silbe betont und zwei Silben unbetont), (x́ x x). Vom Standpunkt der Bewegung der Betonung in den Wortformen der Substantive beobachten wir in den isosyllabisch-akzentologischen Modellen: a) die Betonungs-Identität 'ruká, rukí, ruké...', (x x́), und auch b) die Betonungs-Differenz 'rukí' (Gen. Sing.) und 'rúki' (Nom. Plur.), (x x́) und (x́ x). Auf Grund des Silben-Prinzips stellen wir in den Paradigmen der Substantive mit veränderlicher Betonung die grundlegenden syllabisch-akzentologischen Modelle fest, in denen die Betonung in der Mehrzahl der Wortformen der Substantive auf derselben Silbe oder demselben Morphem liegt. Die Identität der Betonung bildet die Grundlage im Paradigma, außer in jenen Fällen, wenn der Singular in Opposition zum Plural steht, 'travá' (Sing.) und 'trávy' (Plur.). Unter diesem Aspekt beobachten wir, daß die Substantive mit fester Betonung fünfmal öfter als die Substantive mit veränderlicher Betonung (3 295 zu 641) vertreten sind, jedoch die Substantive mit fester Betonung nur eine zweimal höhere Frequenz aufweisen als die Substantive mit veränderlicher Betonung (68,93% zu 31,07%). (Vgl. Tabelle Nr. II). Die veränderliche Betonung ist besonders für *ein-*, *zwei-* und *drei*silbige russische Substantive charakteristisch. Bei einem mehrsilbigen Substantiv, d.i. im Kompositum, ist in der russischen Sprache die veränderliche Betonung selten. Unter dem paradigmatischen Aspekt unterscheiden wir 16 grundlegende syllabisch-akzentologische Modelle.

Tabelle II.

	Substantive			
	Anzahl		Frequenz	
	Absolut.	%	Absolut.	%
Feste Betonung	3 295	83,71	127 374	68,93
Veränderl. Betonung	641	16,29	77 693	31,07
Sa.	3 936	100,–	250 067	100,–

3. Beschluß

Unsere vorläufigen Forschungsergebnisse stellen nur einen Teil der Gesamtproblematik der syllabisch-akzentologischen Beziehungen der russischen Sprache vor. Der Ansatz einer neuen Methodik des Unterrichts in der russischen Sprache erfordert eine Weiterführung der linguistischen Analyse und Beschreibung der syllabisch-akzentologischen Beziehungen auch anderer Wortarten (z.B. der Verben, Adjektive usw.) und eine Synthese dieser Probleme.

Literaturhinweise

Častotnyj slovar' russkogo jazyka (1977). Pod red. L.N. Zasorinoj, Moskau.

Straková, V. (1978). Ruský přizvuk v přehledech a komentářich, Praha.

Timing of English Vowels spoken with an Arabic Accent

F. Mitleb
Irbid, Jordan

1. Introduction

This study is intended to provide empirical evidence concerning the extent to which the temporal properties of vowels in English with an Arabic accent resemble native vs. target-language norms. Phonemically Arabic differs from English in that it possesses only three long vowels /ii,aa,uu/ and their short counterparts while English has many more vowels some of which are quite similar to the six Arabic vowels.

A recent experimental study of the Jordanian dialect of colloquial Arabic by Mitleb (1981) has shown that stop voicing does not significantly affect either stop timing in final position in monosyllables or vowel duration. In English, however, it is well known that vowels are much longer before voiced consonants than before voiceless ones (Peterson and Lehiste, 1960) and that closure duration of English voiced consonants in post-stressed position is shorter than that of their corresponding voiceless counterparts (Lisker, 1957).

Arabic, unlike English, possesses a phonotactic constraint that limits monosyllabic words to CVVC and CVCC syllable types (Swadesh 1937). This implies that short vowels /i,u,a/ contrast phonemically with their long counterparts /ii,uu,aa/ (Al-Ani, 1970) but that final single (short) consonants occur only after the long vowels and geminate (long) consonants only occur in syllables with short vowels (Al-Ani, 1970). Thus vowel length and consonant length are not independent in monosyllabic words but are confounded. In English, however, they may appear either after lax (short) or tense (long) vowels. That is, phonological length of a vowel is independent of the final consonants in closed monosyllables (Peterson and Lehiste, 1960). Moreover, in spite of the fact that both languages possess the phonemes /t/ and /d/, American English has an optional rule that generally changes an underlying /t-d/ contrast into apical flap [ɾ] as in *writer* and *rider* (Chomsky, 1964). In American English, this rule also applies across word boundaries to wordfinal /t/ and /d/ as in sentences like 'put it away' and 'read a book'.

Current proposals such as the Contrastive Analysis Hypothesis claim that this interference from the native to foreign language is primarily at the abstract level of phonological or phonetic segmental features rather than at the lower levels such as phonetic implementation. Thus, within the framework of Contrastive Analysis, difficulties that face second-language learners are attributed to differences in phonemic inventories, allophonic membership in phonemic inventories, distributional differences of phonemes and syllable structure differences between the first and second language (Lado, 1957). Accordingly this theory would predict (1) that Arabs should have difficulty in 'unlearning' their syllable structure constraints to produce novel English syllable types, and (2) that Arabs would not learn the optional rule of flapping characteristic of American English in post stressed position. However, it seems that Contrastive Analysis makes no prediction about the difficulty for Arabs to produce the phonetic implementation rules of English. This is because Contrastive Analysis deals only with transcriptions based on traditional articulatory features as basic data. Phonetic transcriptions however, disregard the physical properties of speech sounds under the assumption that such properties are 'supplied by universal rules' (Chomsky and Halle, 1968:295). Yet, recent phonetic studies have shown, for example, that the effect of voicing on preceding vowel duration is not an absolute universal but rather a language-specific variable, (Port, Al-Ani, and Maeda, 1980). This suggests that non-segmental differences exist between the temporal structures of languages that must be accounted for in the analysis of each language (Port, Al-Ani, and Maeda, 1980).

Methods

2.1. Stimulus Materials

The following 12 real or possible English minimal pairs were chosen for this study (beat, bead, bit, bid, bait, bade, bet, bed, boot, booed, *but, *bud, boat, bode, bought, bawd, bot, bod, butt, budd, bat, bad, bite, bide). A list of sentences was prepared on 3×5 cards in a quasi random order. Subjects were instructed on the cards to read the asterisked words to rhyme with *foot* and *could*, that is, /bʊt/ and /bʊd/. Three tokens of each test word embedded in the carrier sentence 'He says ____ again and again'.

2.2. Subjects

Two groups of seven speakers each served in this experiment: an American group and a Jordanian group. The Americans were all male graduate students of linguistics at Indiana University at the age 24-30 and came from a variety of regions of the country. The Jordanians were all male native speakers of Arabic (Jordanian dialect) aged 25-30. The Jordanians had been in the United States for over two years.

2.3. Recordings and Analysis

A total of 72 sentences were read by each subject from cards at normal speaking tempo and recorded on a Revox A 700 tape recorder. Wide-band spectrograms were made for each token measuring vowel duration and final apical duration.

3. Results

3.1. Vowel Duration

a. Voicing Effect. The ratios of vowel duration in a voiceless environment to that in a voiced one are .80 for Americans (F(1,3)=157, p<.001) and .91 for Arabic-accented English (F(1,3)=16.13, p<.001). Due to the smaller size of the voicing effect on vowel duration exhibited by the Jordanians, the interaction of group and voicing is also significant (F(1,6)=27.6, p<.001). Moreover, these two groups differ significantly with respect to the overall vowel duration (F(1,4)=298.34, p<.001) which are much shorter (m=138 msec.) for the Jordanian than for American (m=184 msec.).

b. Tensity or Length Eftect. The ratios of lax/tense vowel average overall were .85 for Americans and .76 for native Arabic speakers when speaking English. The effect of tensity is significant for both Americans (F(1,3)=47.7, p<.001) and Arabs (F(1,3)=89.9, p<.001) independently. Furthermore, the greater tensity effect by the Arab speakers than Americans is also significant (interaction of tensity and group (F(1,6)=5.60, p<.02).

c. Vowel Height Effect. The percentage difference between all high vs. all low vowels is the same for the two groups. It is significant for both Americans (F(1.3 = 35.8, p<.001) and Arabs (F(1,3)=15.9, p<.001) and amounts overall to 14%. As for group and vowel height interaction, however no significant effect is found (F(1,6)=1.70, n.s.);

3.2. Apical Stop Closure Duration

a. Voicing Effect. The ratios of t/d are 1.31 for Americans (F(1,3)=48.79, p<.001) and 1.13 for Arabs (F(1,3)=404, p<.045). Thus due to the difference in ratio between these two groups of about .21, the interaction of group and voicing is also found to be significant (F(1,6)=6.18, p<.009).

b. Flapping of Apical Stops. A flap is defined here as having a closure of 40 msec. or shorter than this.[1] Both American controls and Arabs make a good percentage of the total number of English t/s and d/s as apical flap according to the above criteria. The proportion of flapped t/s and d/s is not noticeably different for the two groups. Americans flap about 1/2 of the t/s and 4/5 of the d/s, while Arabs flap 1/2 of the English t/s and 2/3 of the d/s. Chi-square tests on the proportion of flapped t/s and d/s show that American-English is not significantly different from Arabic-accented English (χ^2 (1) = 0.67, n.s.).

c. Vowel Tensity Effect. There is no significant duration difference between consonant closure following tense vowels and that following lax ones for either the American group (F(1,3)=.075, n.s.) or the Arab group (F(1,3)=-198, n.s.). However, the overall consonant closure duration difference between these two groups amounts to 4.5 msec. (F(1,4)=13.87, p<.001). Yet, the interaction of group and tensity falls short of significance on consonant closure duration.

[1] Zue and Lafferriere (1979) also considered 40 msec. or shorter as their criterion for flapped apical.

4. Review of Results and Conclusions

To clarify the results of our experiment, the results from the Arabic experiment reported in Mitleb (1981) are juxtaposed to those of the English experiment in Figure 1.

In this display, the durations of vowels before /t/ are on the vertical axis and those before /d/ are on the horizontal axis. A diagonal line is drawn indicating points where the vowel before /d/ equals the vowel before /t/. Thus, the nearer a point is to this line, the smaller is the difference between a vowel before a voiced consonant and before a voiceless one. Examining this figure, we note that Arabic exhibited a weak voicing effect on vowel duration. However, Arabic-accented English showed a modest durational difference between vowels as a function of voicing which is stronger than Arabic and weaker than American English.

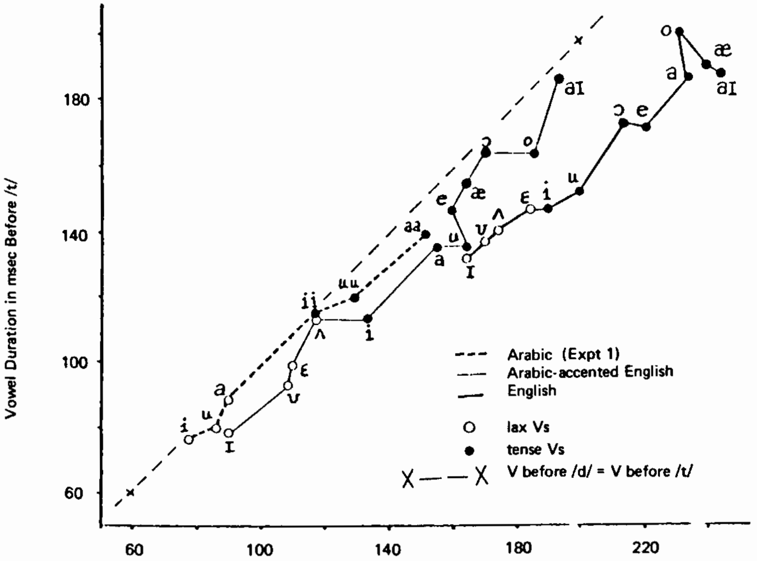


Figure 1. Duration of various vowels before /d/ and /t/ in ms.

Thus, we may conclude that Arabs produce a reduced version of the English voicing effect, and an exaggerated version of the English vowel effect on vowel duration, and exhibit overall vowel durations that are mid-way between Arabic and American English. This conclusion supports, on empirical grounds, the contention of Flege (1981) that foreign-accented speech is an 'approximative system'. However, these results overall run counter to the view that second language learners when faced with new phoneme, allophone, allophonic distribution or syllable type do 'transfer' the structure of the native phonological system in producing the target one (Lado, 1957), since there is no evident interference of 'monosyllabic types' of Arabic on the English production of Arabs. The Jordanians in this experiment flapped /t, d/ in the manner reported for American English. Nevertheless, Arabs seem to use Arabic short and long vowel timing for English lax and tense vowels. Although they lengthen vowels overall, they do not lengthen them enough to match Americans. These results are congruent with earlier studies (Flege and Port, 1981) which proposed that the pronunciation interference from the native language to the target language occurs primarily at the level of phonetic implementation rather than at the level of phonological features and phonotactics.

Overall, then, none of the cases of potential interference from the native language to the target language examined in this study that could be specified in straightforward segmental terms - like phonotactic constraint or allophonic rules - give evidence of posing particular difficulties for our subjects. Yet our data do provide some evidence that phonetic implementation-level differences between languages are a source of interference from the native language into the second language. Thus, our results on foreign accent appear to provide support for the hypothesis that differences at the segmental levels of phonological and phonetic elements between languages are easier to overcome than differences at the temporal implementation level for an adult language learner.

Acknowledgements

I would like to thank Robert Port for the extensive assistance in completing this study. This research was supported by the Yarmouk University, Irbid, Jordan.

References

Al-Ani, S. (1970). *Arabic Phonology.* The Hague: Mouton.

Chomsky, N. (1964). Current issues in linguistic theory. In: J. Fodor and J. Katz (eds.), *The Structure of Language.* Englewood Cliffs: Prentice Hall, 50-118.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English.* New York: Harper and Rowe.

Flege, J.E. (1981). Phonetic approximation in second language acquisition. *Language Learning, Vol.* **30**, No. 1, 117-133.

Flege, J. and Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech,* **24**, 125-146.

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers.* Cambridge: M.I.T. Press.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language,* **33**, 42-49.

Mitleb, F. (1981). *Segmental and Non-segmental Structure in Phonetics: Evidence from Foreign Accent.* Unpublished Ph.D. dissertation, Department of Linguistics, Indiana University.

Peterson, G., and Lehiste, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America* **25**, 175-184.

Port, Al-Ani and Maeda, (1980). Temporal compensation and universal phonetics. *Phonetica,* **37**, 235-252.

Swadesh, M. (1937). The phonemic interpretation of long consonants. *Language,* **13**, 1-10.

English Intonation from a Dutch Point of View

N.J. Willems
Eindhoven, the Netherlands

1. Introduction

This paper presents a short account of an experimental-phonetic investigation, described in full in Willems (1982), which attempts to characterize and describe the melodic (intonational) aspects of non-nativeness in English pitch contours produced by Dutch native speakers, with the aim of developing a course of melodic pronunciation precepts for Dutch students of English.

Up to now the teaching of British English intonation in the Netherlands has either been neglected or largely restricted to repeated imitation (drill-method). One of the very drawbacks of this drill method is that students are incapable of generating new instances, since they will not acquire any insight into underlying rules.

Instead of drill methods it could be more profitable to make students conscious of intonational structures of the target language by providing them with an explicit experimentally based description in the form of a rule based intonation course and by training them in analytic listening to pitch phenomena.

Our working method is largely based on the pioneering intonation research of the 'Dutch school' (Cohen and 't Hart, 1967), which describes quasi-continuous pitch contours as sequences of discrete stylized pitch movements (straight-line approximations), the perceptual relevance of which has been established in listening tests.

2. Acoustic measurements of fundamental frequency curves

An extensive comparison was made between about 600 instrumentally analysed fundamental frequency curves produced by a dozen speakers of either language, who were asked to read out an English prose text. Electroglottographic signals were recorded and were subsequently analysed by means of a computerized F_0-analysis program. The following properties of the pitch movements were established: direction, magnitude, slope, duration and position of the pitch movement with respect to vowel onset.

The analysis showed that pitch contours in English can adequately be described by means of three parallel declination lines (low-mid-high). For most contours a full range of 10 to 14 semitones was found.

The Dutch native speakers showed in their English the following major systematic deviations from the corresponding manifestations of pitch movements produced by the native speakers of English:

1. The Dutch native speakers (DNS) replaced more than one tenth of the falls produced by the English speakers (ENS) by a simple rise.
2. The size of the pitch movements produced by the DNS was considerably smaller in most cases.
3. DNS tended to exaggerate a final rise.
4. A gradually rising pitch movement (inclination) on unaccented words preceding a fall was often lacking in the case of the DNS.
5 By and large the ENS started an utterance on the mid level, whereas DNS showed a clear tendency to start at the low level.
6. Following a prosodic boundary ENS often resumed a contour by a virtual jump from the low level to the mid level (reset). The DNS tended to use the low level again.

3. Perception experiments

In these tests we have tried to avoid interferences from deviations other than those in pitch by asking subjects to evaluate only sentences produced by English native speakers with artificially imposed variations in pitch contours.

In our first perception test we used a synthetic speech sample. English listeners were asked to assess the acceptability of stepwise variations in magnitude of the excursion and position of the pitch movement in the syllable with respect to vowel onset. The outcome of this experiment showed that the preferred size for the averaged excursion of English pitch contours should be at least twice as large (12 semitones) as the standard excursion used for Dutch (4-6 ST). The position of the most common English pitch movement, a prominence-lending fall, was found to be early to mid in the syllable (0-150 ms after vowel onset). In a second perception test the perceptual relevance of the deviations brought to light by the instrumental analysis was established by experimentally manipulating the original pitch contours.

These contours were stylized and superimposed on utterances produced by the native speakers of English by means of LPC-resynthesis. The outcome of this experiment showed that linguistically naive English native speakers were quite capable of judging the acceptability of native language pitch contours and of making clear distinctions between correct and incorrect contours: two groups participating in our experiments - students and office employees - agreed overwhelmingly and scored in a very consistent way. Moreover the English native speakers considered the stylized versions of original pitch contours to be very acceptable, which means that the stylization method is also suited to describe British English intonation.

Deviations in pitch movements which were nearly always considered to be unacceptable were a reverse direction (mainly replacement of the characteristic English pitch falls by rises), too small an excursion, and a combination of the two. The English mid level may be taken as a 'neutral' starting and restarting point for the majority of the contours and appears to be a striking characteristic of English intonation. Nevertheless deviations of the Dutch native speakers at this point, who tend to start and continue at the low level, did not appear to be conspicuous in all cases. Also the omission of a gradual rise (inclination) was not always considered to be unacceptable by the English subjects.

In a final experiment we corrected the pitch movements produced by a Dutch speaker of English according to a few provisional precepts mainly with respect to direction, excursion and restarting level. Results showed high acceptability scores for the corrected contours, suggesting the potential effectiveness of the pronunciation precepts.

4. Discussion: towards an intonation course

From our experiments, specifications for the realization of the most frequent 'tone group' in English, tone 1 (Halliday, 1970), have come to the fore. The 'main' pitch accent (nucleus) is realized by means of a full fall varying from about 8 to 20 semitones, or, in our standardized three level system, by a fall of 12 ST. This fall must start early in the syllable. figure 1 presents a few possible realizations of tone 1, with one and with two accents.

Our results suggest that continuation of this study will allow the design of an intonation course for British English. In contrast with most existing English intonation courses, the instructions in this course would be based on experimental evidence and as such stand a better chance of representing an explicit and consistent survey of the language's melodic structures. Moreover the notational system of straight-line contours is straightforward and easy to comprehend (cf. figure 1). The results obtained by van Geel (1981) with such a course for Dutch intonation by Collier and 't Hart (1981) proved most promising. His subjects - laryngectomees using an electrolarynx with semi-automatic pitch control - were made aware of pitch contours in their native language (Dutch) by means of this course and were found to be quite able to produce acceptable pitch contours by triggering the built-in pitch movements at the right moment.

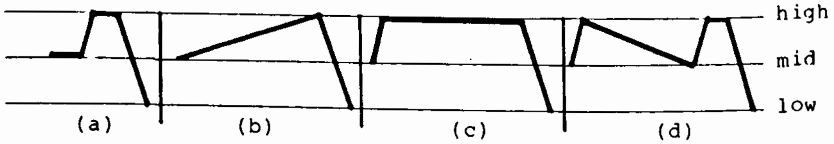


Figure 1. Examples of possible realizations of tone 1: (a) one accent: a fall with an optional half rise (b) one accent: a fall with a preceding gradual rise (c) two accents: half rise and fall; high declination in between (d) two accents: half rise, followed by an optional half rise and a fall, and a gradual fall in between.

The success of the stylization method for English gives reason to have every confidence in a similarly profitable effect of such a course for learners of English in general.

Acknowledgements

This study was carried out at the Institute of Phonetics, Utrecht University, and was funded by its Faculty of Arts.

References

Cohen, A. and J. 't Hart (1967). On the anatomy of intonation. *Lingua,* **19**, 177-192.

Collier, R. and J. 't Hart (1981). *Cursus Nederlandse intonatie.* Acco, Leuven.

Geel, R.C. van (1981). Intonatie in kunstlarynxspraak. *Logopedie en Foniatrie,* **53**, 315-319.

Halliday, M.A.K. (1970). *A course in spoken English: intonation.* Oxford University Press, Oxford.

Willems, N.J. (1982). *English intonation from a Dutch point of view.* Foris, Dordrecht.

Section 20
Speech Pathology and Aids for the Handicapped

Intonation Patterns in Normal, Autistic and Aphasic Children

C. Baltaxe, J.Q. Simmons and E. Zee
Los Angeles, USA

1. Introduction

Current knowledge of prosody and language disordered children remains limited. Studies with normal children show that *first*, early in language acquisition, prosodic development is more advanced than phonological, syntactic, and semantic development. Early prosodic units seem to fulfill a facilitative function perceptually and productively and to constitute 'frames' for other units of language (Bruner, 1975). *Second*, these early prosodic 'frames' appear to be more stable than the segmental dimensions accompanying the prosodic contours (Menn, 1979). *Third*, there is some evidence that control of fundamental frequency develops first, timing second and segmental contrasts last (Allen and Hawkins, 1980). *Fourth*, children's knowledge of the prosodic system is interdependent with their knowledge of other levels of language and may not reach adult refinement until about the age of twelve (Cutler and Swinney, 1980). Prosody may also be an important variable in children with development disorders of language. When prosody is impaired, its facilitative function may be disturbed, affecting other levels of language. Conversely, when phonological, syntactic, semantic or pragmatic development is delayed or disturbed, prosody may also be affected.

2. Experiment

The present study examined the intonation contours in normal, autistic, and aphasic children. The subjects consisted of six normal, five autistic, and six aphasic children, matched as closely as possible for socio-economic class, sex, and mean length of utterances as a measure of psycho-linguistic age (MLU for all groups: 1.45 to 4.46 morphemes). The children ranged in age between 2-0 and 4-0 years for the normal subjects, 4-6 and 12-2 years for the autistics and 4-5 to 12-2 years for the aphasics. Simple, neutral, declarative utterances of the subject-verb-object variety, produced spontaneously under controlled conditions, were examined for the prosodic characteristics and markers listed below. These markers were chosen because they seemed to capture important aspects of the intonation contour and to provide a useful basis for comparison among the groups.

1. *Frequency range*, used to express the intonation contour of the utterances studied;

Table I. Percentage of occurrence for expected markers of intonation contour for simple declaratives in normal, aphasic, and autistic children (group results)

	Numbers of utterances	P_1		P_2		P_3		Terminal fall	Declin. effect		Covariations
		F_0	I	F_0	I	F_0	I	F_0	F_0	I	F_0/I
Normals	N = 38	32	30	26	25	33	37	34	15	14	26
	100%	84%	79%	68%	66%	87%	97%	89%	39%	37%	68%
Autistics	N = 47	31	33	21	21	43	42	36	5	6	16
	100%	66%	70%	45%	45%	91%	89%	76%	11%	13%	34%
Aphasics	N = 43	14 (31)	31 (33)	22	13	23	25	31	17	12	11
	100%	45%	94%	51%	30%	53%	58%	72%	39%	28%	25%

P_1 = Peak_1 (subject), P_2 = Peak_2 (verb), P_3 = Peak_3 (object); F_0 = peak fundamental frequency; I = peak intensity.

2. *terminal fall*, generally associated with the declarative mode in Standard English (O'Shaughnessy, 1979);
3. *intonation contour of the utterance* as characterized by a series of pitch obtrusions expected on the stressed vowels of the utterance in subject-verb-object positions (Martin, 1982);
4. *declination effect* or the tendency of pitch to drift downward over the declarative intonation group (Cooper and Sorenson, 1981);
5. *covariation of frequency and intensity* over the declarative intonation contour (Lieberman, 1967).

The utterances analyzed were processed from an Ampex tape-recorder through a fundamental frequency and intensity meter and an Oscillomink. Results were based on acoustic measurements of Oscillomink tracings, consisting of duplex oscillogram, fundamental frequency, intensity, waveform, and timemarking. Fundamental frequency range was determined by further statistical analysis. Terminal fall, intonation contour, declination, and covariation of frequency and intensity were reported in percentages of actual occurrence of these markers in expected positions (Table I - group results). Five Hz was chosen as the minimum necessary difference to identify the existence of the above markers for speakers with a narrow frequency range. Table II provides mean and standard deviations of values of peak F_0 (Hz) to illustrate declination effect for each of the three key stressed syllables (P1,P2,P3) for each individual subject. Examples of prosodic contours for each subject group are presented in Figure 1.

Table II. Declination effect. Mean and standard deviations of values of peak F_0 (Hz) for each of three key stressed syllables in simple declaratives (S-V-O) in normal, aphasic and autistic children

	Subjects	P_1 (F_0)		P_2 (F_0)		P_3 (F_0)	
		X	S.D.	X	S.D.	X	S.D.
Normals	Subj. 1	248	94.1	230	86.4	232	91.2
	Subj. 2	249	96	232	95	333	132
	Subj. 3	170	147	165	143	167	145
	Subj. 4	257	110	215	79	212	83
	Subj. 5	214	144	195	130	192	129
	Subj. 6	198	111	178	101	170	97
Autistics	Subj. 1	298.1	96.2	252.2	87.2	277.2	99
	Subj. 2	232.5	77	205	70.9	211.1	74.6
	Subj. 3	313.5	115	267	102.7	257.6	166.79
	Subj. 4	208.5	93.88	205.7	91.0	248.7	110
	Subj. 5	217.2	82.5	201.2	81.6	204.4	77.15
Aphasics	Subj. 1	251.6	111.99	262.5	83.89	243.3	77.73
	Subj. 2	256	226.78	200	173	176	153
	Subj. 3	176	72.2	175	77.6	171.6	70.5
	Subj. 4	266	149	244	143.6	280	166.7
	Subj. 5	208	67.5	206	67.4	200	67.0

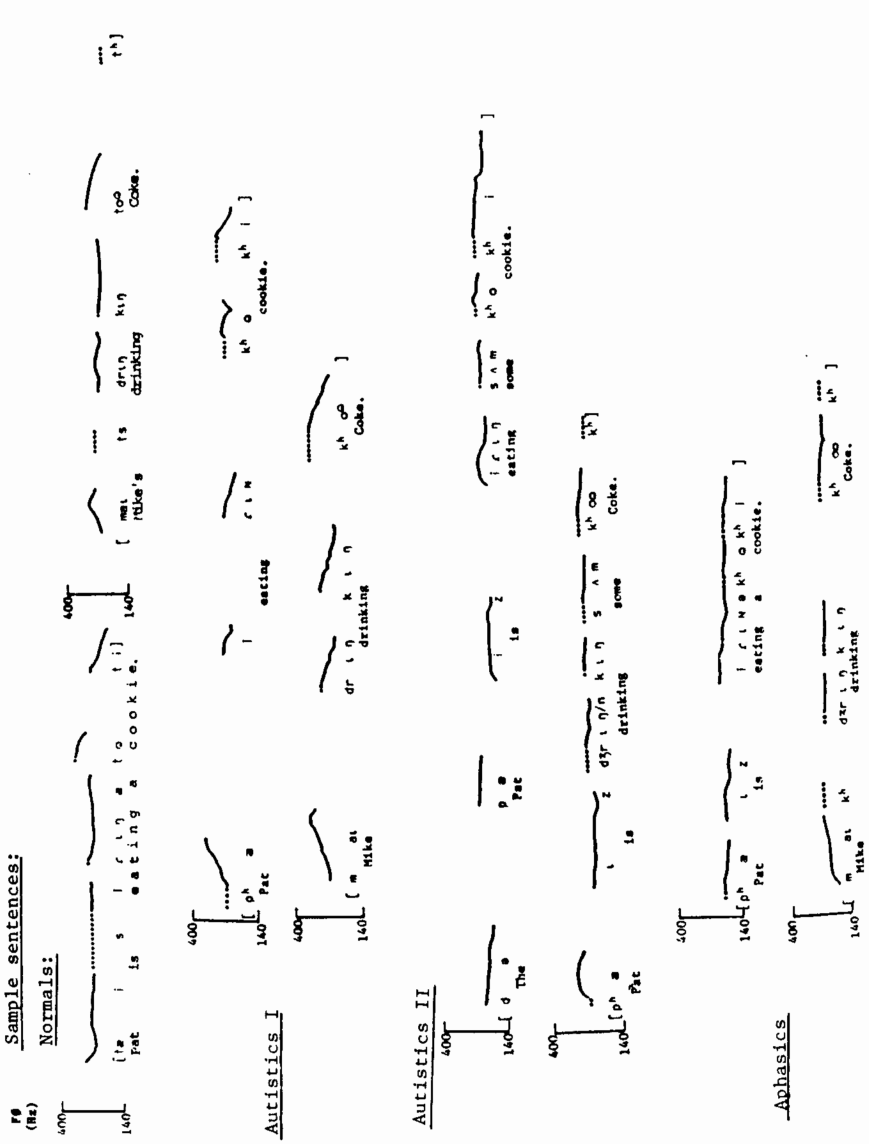


Figure 1. Sample of sentences of normal, autistic and aphasic children.

3. Results

3.1. Frequency range

A comparison of the fundamental frequency ranges across the three groups showed that the normals had the greatest range (122.67 Hz, SD 63.63), followed by the autistics (96.79 Hz, SD 24.78) and then the aphasics (69.69 Hz, SD 25.26). Differences between normals and aphasics, and autistics and aphasics were significant at the .005 level. Frequency ranges for individual autistic subjects were either highly exaggerated or very narrow. Subjects with a narrow range were not significantly different from the aphasics.

3.2. Terminal fall.

Terminal fall occurred with the highest percentage for the normals, followed by the autistics, and then the aphasics. Individual profiles (Table II) showed terminal fall in all but one of the normal group, while only three autistics and two aphasics produced fall consistently.

3.3. Intonation contour

As defined by expected occurrence of pitch obtrusions on stressed vowels in subject (P1), verb (P2) and object (P3) positions. In all groups the highest percentage of pitch obtrusions occurred on P3, followed by P1, then by P2. The aphasics differed as a group in that P3 was followed by P2. A similar rank order was also seen when intensity perturbation was considered for the normal and the autistic group, while for the aphasics P1 ranked first, P3 second and P2 last. When both frequency and intensity contour were considered, stressed vowels were marked by pitch obtrusion, intensity obtrusion, or both. Table II shows considerable within and between subject variability.

3.4. Declination effect.

Declination occurred with the highest percentage for the normals, second for the aphasics, and third for the autistics (Table I). When considering declination for fundamental frequency and intensity, covariation of the two parameters was not always present in the individual profiles and considerable within and between subject variability was evident. Declination between P1 and P2 was most consistent, while only a few of the subjects showed the expected declination for all three stressed syllables of the utterances (Table II).

3.5. Covariation of frequency and intensity over the declarative intonation contour.

The normals again showed the highest percentage of covariation, followed by the autistics, then the aphasic group. However, individual profiles of the autistics and aphasics showed that some subjects lacked covariation altogether (Table II).

In summary, despite considerable within and between subject variability, some of the markers examined appear more stable and consistent than others. The language deficient subjects generally showed less stability and greater individual variation. Differential impairment was also seen with respect to individual markers, individual groups and individual subjects. Both frequency and intensity thus appear to be important prosodic markers in the speech of young children. The linguistic salience of these parameters may differ for individual children or groups. Covariation of frequency and intensity expected for mature patterns thus may be the result of maturational factors and learned behavior. Although the frequency parameter has been characterized as initially the most stable, when broken down into the individual markers described, some appear more stable than others. It may be that such markers develop earlier, while others, less stable and consistent, might be more dependent on maturational factors, learned behaviors, and other linguistic development. For the abnormal groups, such markers may then exhibit particular vulnerability.

Acknowledgement

This research was supported by USPHS NS 16479-2, NH 30897, and MCT 927-12/ MCH 4-434700-31886.

References

Allen, G., and Hawkins, S. (1980). Phonological rhythm: Definition and development. In: Yeni-Komshian, Kavanagh, and Ferguson (Eds.) *Child Phonology.* New York: Academic Press.

Bruner J. (1975). The ontogenesis of speech acts. *Journal of Child Language,* **2**, 1-19.

Cooper, W. and Sorenson, J. (1981). *Fundamental frequency in sentence production.* New York: Springer Verlag.

Cutler, A., and Swinney, D. (1980). Development of the comprehension of semantic focus in young children. Paper presented to the Fifth Boston University Conference on Language Development.

Lieberman, P. (1967). *Intonation, perception, and language.* Cambridge: M.I.T. Press.

Martin, P. (1982). Phonetic realization of prosodic contours in French. *Speech Communication,* **I**, 3-4.

Menn, L. (1976). *Pattern, control, and contrast in beginning speech: A case study in the development of word form and word function.* Unpublished doctoral dissertation University of Illinois, Urbana-Champaign.

O'Shaughnessy, D. (1979). Linguistic features in fundamental frequency patterns. *Journal of Phonetics,* **7**, 119-145.

Phono-articulatory Stereotypes in Deaf Children

L. Handzel
Wroclaw, Poland

1. Analysis

Analysis of the acoustic features of the syllable, as a context-defined phono-articulatory act, was made in a group of 30 children, aged 8-13 years with impaired hearing, residual hearing and deafness. The range of their hearing is illustrated by the collective audiograms (Fig. 1). The control group was identical as to the age and number of children. Each of the examined children uttered by heart an interrogative and affirmative sentence that differed only by the grammatical form of the first word and the purpose indicated by the intonation characteristic of this kind of sentence. Each sentence consisted of three bisyllabic words.

1. 'Budes zitra doma?' (in Engl.: 'Will you be home tomorrow?')
2. 'Budu zitra doma.' (in Engl.: 'I will be home tomorrow.')

The tone pitch, as well as the variability in the pitch of the basal tone of the voice, was recorded on tape while the relative intensity of the speech tones were monitored on the screen of a cathode oscilloscope. The acoustic spectrum of the utterances was registered by a Visible Speech Sonograph. The results from the acoustic analysis, verified statistically, show the mean values and standard deviations of the acoustic phenomena in each syllable. The results are exemplified by collective diagrams involving all groups under examination. The acoustic patterns in the phono-articulatory acts of the control group have not, of necessity, received a separate characterization, having served only as a point of reference for the groups with hearing pathology, Fig. 2.

In children with impaired hearing the tone pitch of the basal voice is characterized by alternating descending and ascending courses, with interrogative utterances showing a typical elevation in the first syllable of the sentence; only by its flattening does it differ from that of children with normal hearing. There is also a difference in the standard deviation within the hearing-impaired group, i.e. a range of 6 tones as compared with a ½ tone range in normal hearing. In children with residual hearing and deafness, such regularity is indeed lacking; instead, there is always a higher tonal pitch in the first syllable of each word, Fig. 3.

By analogy to the patterns of the basal tone in children with impaired

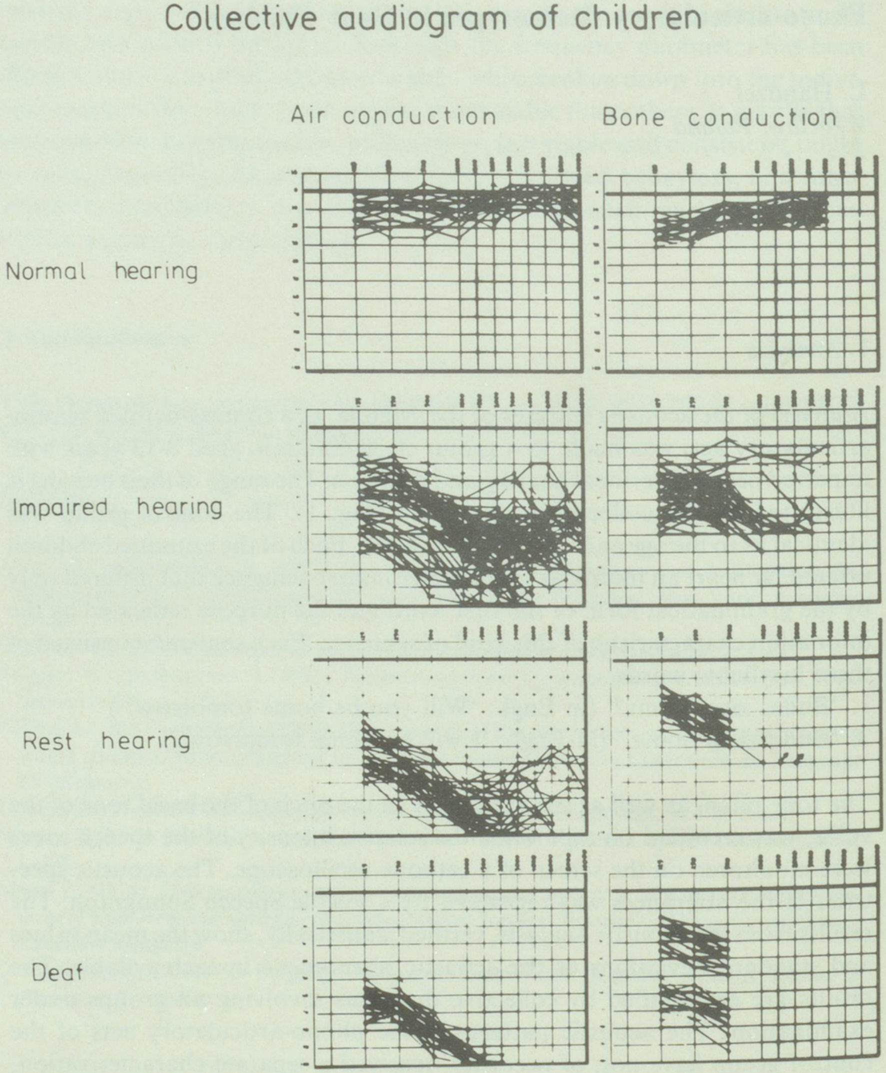
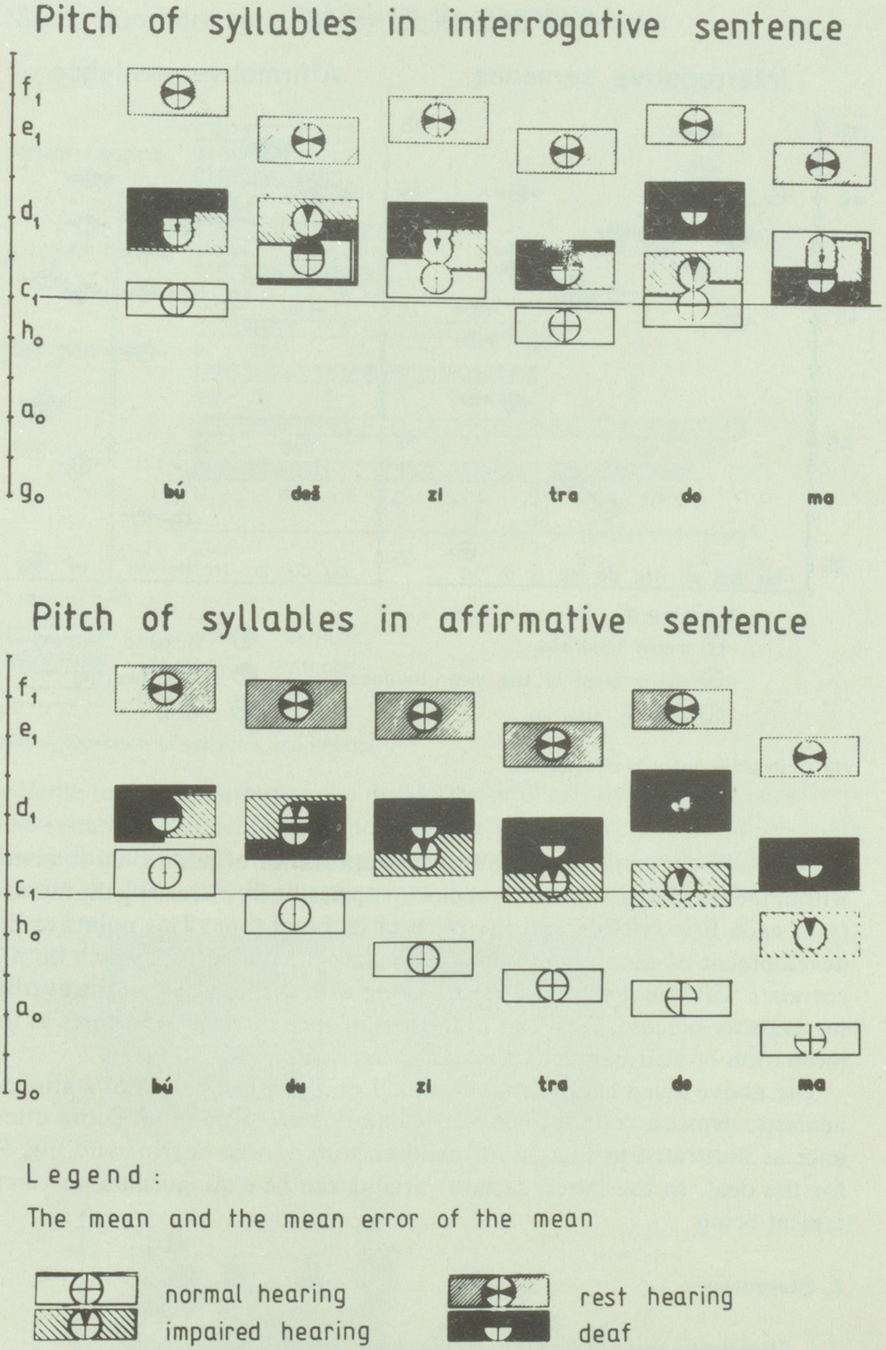


Fig. 1. Audiograms of 4 groups of children: with normal, impaired and residual hearing and deaf children.

hearing, the values of the mean intensity of individual syllables show fewer differences as compared with those for normal hearing. In children with hearing rests and deafness, the higher position of the basal tone of each first syllable is accompanied by an increase in intensity. This is also true for the next to last syllable.

The obviously fluent passage from one syllable to another, as observed in children with normal hearing, is disturbed in children with impaired hearing by the appearance of short intervals between the words. The intervals show higher values in children with impaired hearing, the highest ones occurring in



Pitch of syllables in interrogative sentence

Pitch of syllables in affirmative sentence

Fig. 2. Tone pitch of the basal voice.

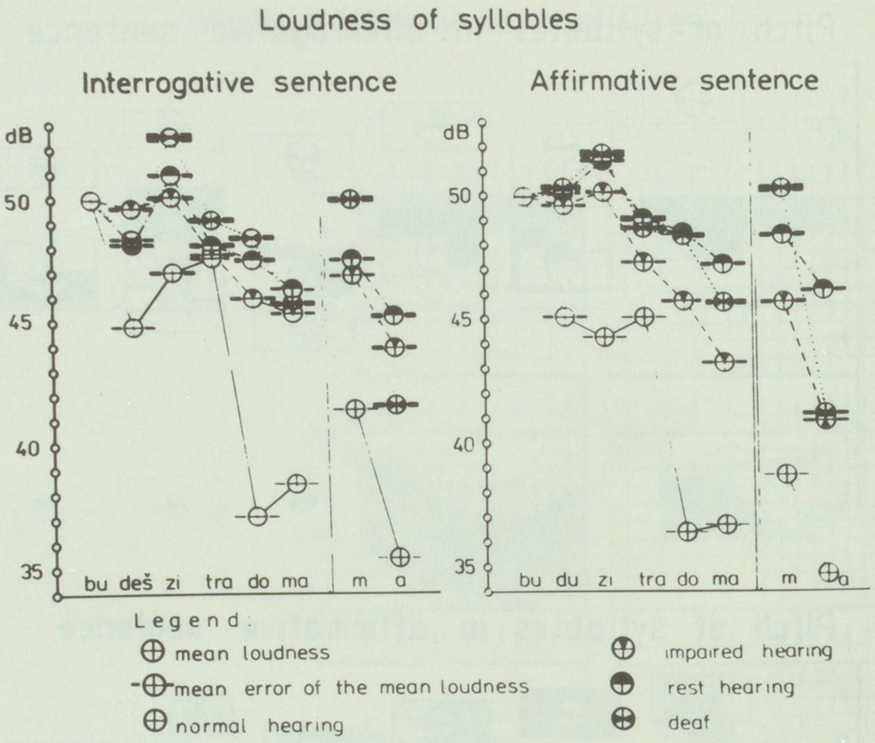


Fig. 3. Relative intensity of syllables.

the deaf. In the latter two groups the appearance of additional intervals within the word emphasizes the acoustic expression by increasing the intensity of each first syllable and the pitch of its basal tone. This points to the development of autonomic phono-articulatory syllabic acts which do not correlate with the text of utterance. Along with the tendency of intraverbal intervals to increase, they lead to the loss of such distinctive features as the intonation of affirmative and interrogative sentences.

The above given characteristics, based on the mean values of statistical analysis, depict a certain phono-articulatory stereotype of an entire utterance as illustrated in Fig. 5a for children with normal hearing and Fig. 5b for the deaf. In the latter, certain variants can be distinguished, the most typical being:

2. Stereotypes

2.1. Monotono-monodynamo-monorhythmic stereotype.

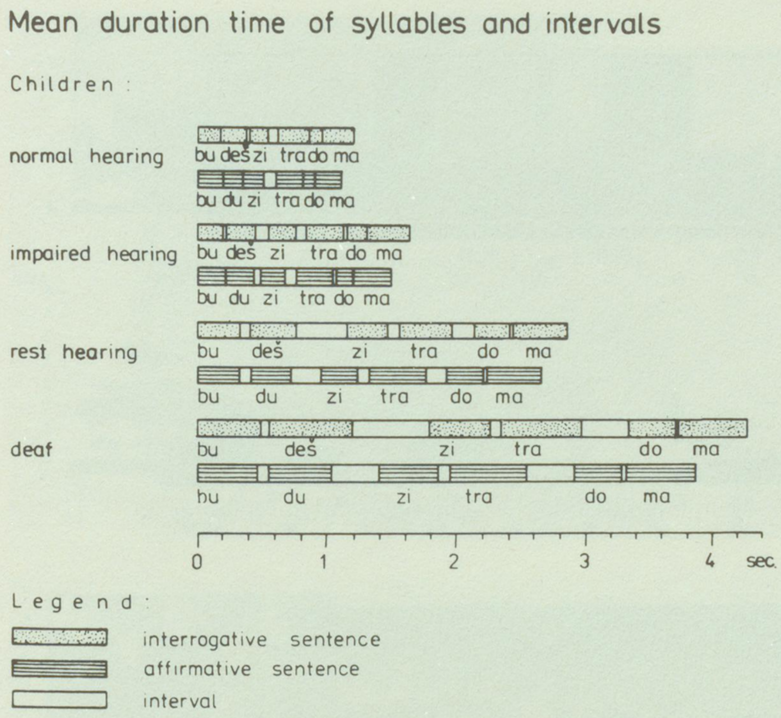


Fig. 4. Duration of syllables and intervals.

Here two variants can be singled out: In Fig. 7a, the striking feature is the repetition of the same pitch of the basal tone, the intensity and duration of consecutive syllables. The clear-cut, equal articulatory passages from one syllable to another attest to the development of one stereotype for consecutive syllables, independent of the syllabic composition. Fig. 7b differs from the variant 7a only by the smaller articulatory handicaps and smaller distortion in the formants. Both variants of the monotono-monodynamo-mono-

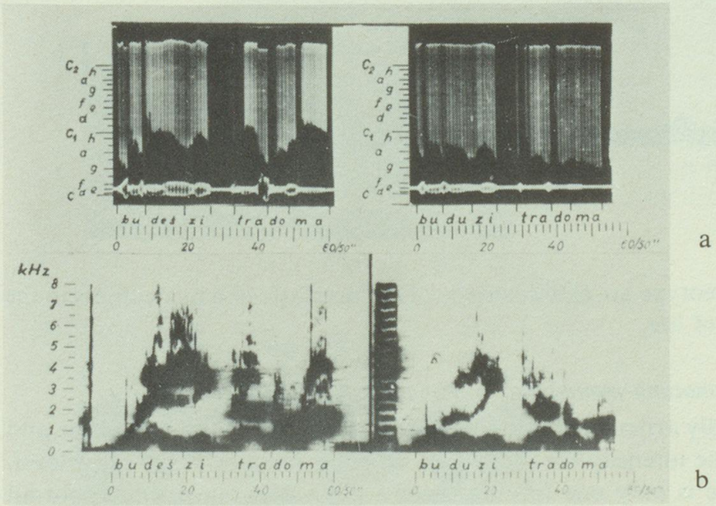


Fig. 5.

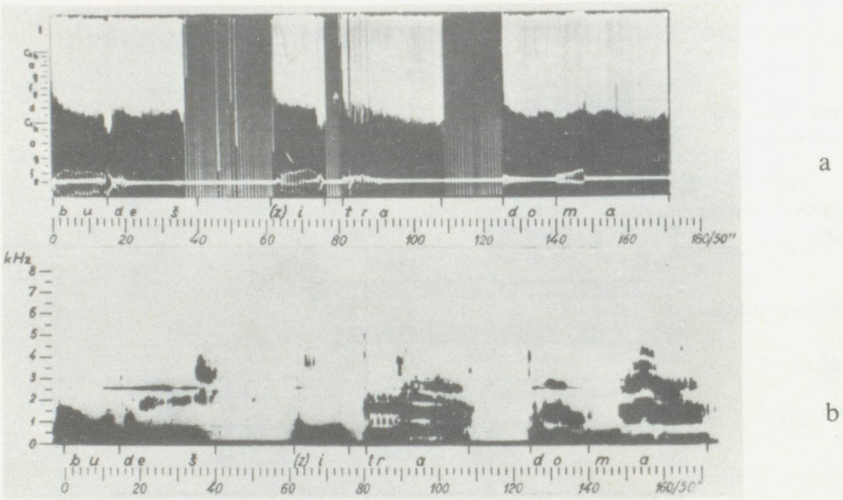


Fig. 6.

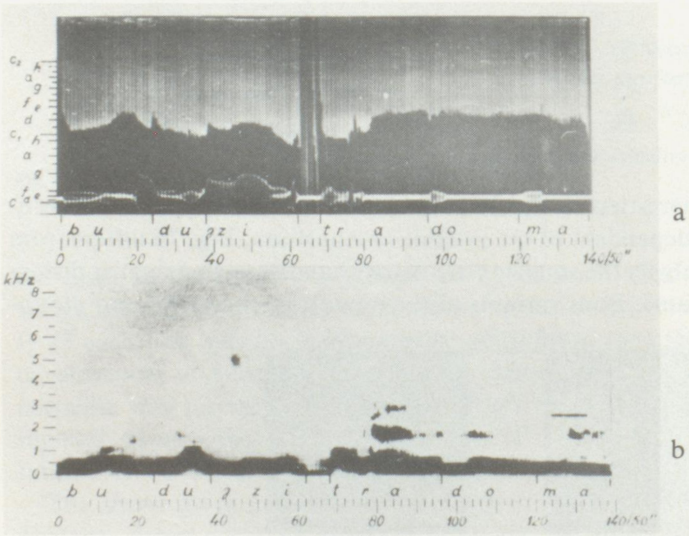
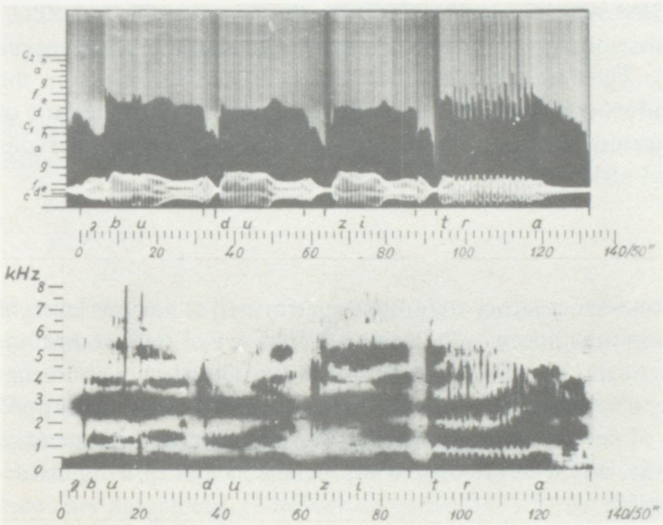


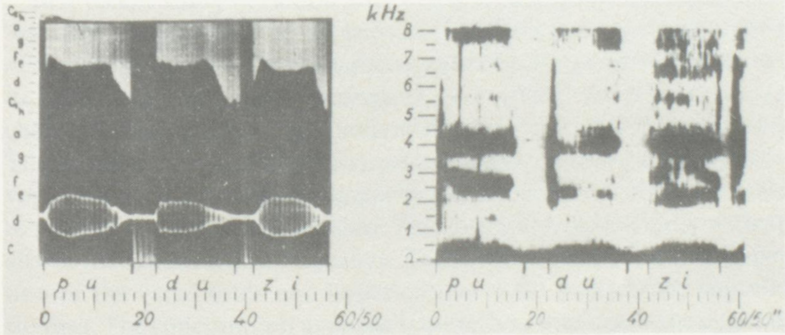
Fig. 7.

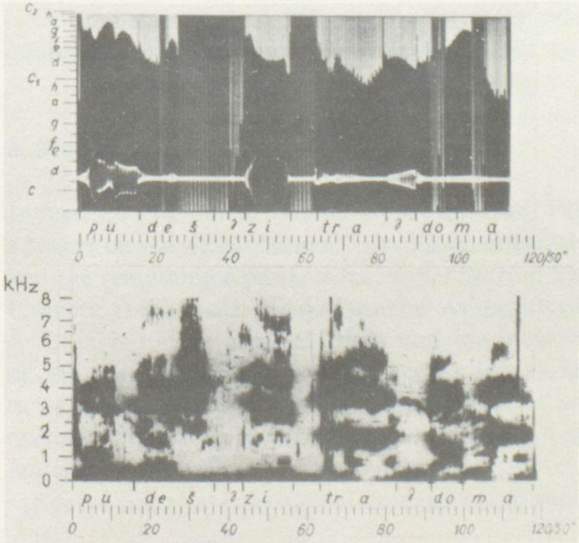
rhythmic stereotype are characteristic of children affected by deafness in the first months of life.

2.2. Quasi-monotono-monodynamo-monorhythmic stereotype.

While the badly articulated syllables merge, the pitch of the basal tone and the level of the intensity for individual syllables remains at the same level. Usually, there is only one formant with a pitch and range similar for all vowels (Fig. 6). The quasi-monotono-monodynamo-monorhythmic stereotype is typical of children with congenital deafness.







2.3. Disruption of the phono-articulatory structure.

Conspicuous oscillations up to one octave of the basal tone pitch can be observed (Fig. 8). They are parallelled by considerable variations in the intensity of individual syllables. This type of disruption is characteristic of children with congenital deafness coexisting with pathology of the nervous system and mental afflictions.

3. Conclusion

Studies on the phono-articulatory stereotypes performed at various levels of the language system in children with hearing defects are of importance not only from a theoretical point of view, they also lead to detection of abnormalities and thereby to selection of appropriate treatment of the handicapped and development of new rehabilitation methods. This means enrichment of the diagnostic array, as far as the degree and persistence of hearing disturbances are concerned.

Acoustic Measurement of Voice Quality in Dysphonia after Traumatic Midbrain Damage

E. Hartmann and D.v.Cramon
Munich, Federal Republic of Germany

1. Introduction

The advantage of acoustic measurement of voice disorders in clinical investigation and treatment has often been emphasized.

Basic categories of pathological voice quality have been established by auditory judgement. However, often there is little agreement on these categories regarding their relation to the various vocal dysfunctions.

In a recent study (Cramon and Vogel, in press) the perceptual categories 'breathy,' 'rough' and 'tense' proved to be sufficient to describe the features of central dysphonia. This is substantiated by assignment of these categories to three basic laryngeal settings (Laver, 1980). Breathy voice quality is characterized by insufficient adduction of the vocal folds during phonation. The acoustic correlate is an increase in the spectral noise components. Rough voice quality is caused by irregularities in the vibration of the vocal folds, due to rigidity or asymmetric tension. This correlates acoustically with aperiodicity of the fundamental frequency. Tense voice quality is due to hyperadduction of the vocal folds and hypertension in the walls of the supralaryngeal cavities. The concomitant acoustic feature is an increase in the upper harmonic components in the spectrum.

Our main aim was to differentiate the above mentioned pathological voice qualities acoustically and to make available an index for the degree of severity of these pathologies.

2. Method

14 male and 10 female patients were examined. 18 of them had suffered from a severe closed head trauma with subsequent traumatic midbrain syndrome and the remaining 6 patients had suffered from a cerebro-vascular accident. Phonetic testing had revealed general symptoms of dysphonia in all patients, such as reduced range of pitch and loudness, increased pitch level and deviant voice quality. They were chosen from a larger sample so as to form three groups of equal quality. The presence of peripheral lesions could be excluded following phoniatric examination. 10 males and 7 females, age-matched, served as control group.

Speech material consisted of the cardinal vowels /u, o, a, e, i/ uttered twice by each subject. The recorded items were digitized and the fundamental periods were tracked using the autocorrelation method. Subsequently the voiced segments of each vowel were subjected to spectral analysis.

On the basis of these data 5 parameters were computed. The parameter 'Time lag of pre-exhalation' Tb was defined as the duration of the pre-exhalation noise. A noise energy threshold was used to determine automatically the beginning of turbulent airflow, and the first fundamental period signalled the end of the pre-exhalation segment. The parameter 'Fundamental period perturbation' PP measured the average percentual deviations of consecutive fundamental periods from a smoothed trend line. In the frequency domain, the parameters A1 and A5 measured the relative amount of spectral energy in the frequency bands from 1 to 5 KHz and from 5 to 10 KHz respectively. V5 was calculated as the spectral variance in the range from 5 to 10 KHz, after band-pass filtering of the spectral function. The described parameters were computed for each of the vowels and were subsequently averaged.

3. Results

The results for the parameter Tb indicated discrimination of the group of breathy voices from each of the other groups, which was highly significant. The breathy voices revealed Tb values between 40 and 120 msec. Only one of the control speakers showed a notable pre-exhalation of 10 msec duration, 3 of the tense voices reached this upper limit and little increase of Tb was observed for 5 subjects with rough voices.

The parameter PP differentiated significantly the group of rough voices from each of the other groups. Rough voices exhibited PP values between 1.6 and 2.6 %. These results are lower than those measured in peripheral lesions. Three tense voices with increased PP may have additional roughness, thus a slightly harsh voice. One tense voice demonstrated PP values even below the control range. It showed a vibration mode which may be characterized as 'mechanical.' Comparing Tb and PP values of some subjects, we may assume a hoarse voice type, if both parameters exhibit high values.

The scatter diagram of the spectral parameters A1 and A5 displayed the control group at the lower left corner with low A1 and A5. The values for breathy and tense voices tended to shift up and/or to the right with tense voices (except for one subject) showing a lower A5 than the breathy voices. Discriminant analysis separated breathy and tense voices significantly from each other as well as from the control group. The rough voices lay in the range of the normal voices in most of the cases.

The particular distribution of spectral energy for breathy voice quality reflects an overall increase of noise energy with a predominance in the 5 to 10 KHz range. Thus the parameter A5 seems to be an additional parameter for classifying breathy voice quality. A significant correlation of parameter Tb and A5 supported this assumption. The particular form of the spectral distribution of tense voice quality may be due to an increase in higher harmonic components with a predominance in the 1 to 5 KHz range. The tense voice with high A5 may be a compound voice type with additional breathiness.

The parameter V5 turned out to be an additional acoustic correlate of the perceptual category 'tense voice.' However, classification was only reliable for male tense voices. Tense voice quality seems to be produced not only by laryngeal but also by supralaryngeal settings. Overall tension with hard surfaced vocal tract walls may improve the resonance and narrow the formant bandwidths. This may cause an increased variance of the spectral amplitude.

In a follow-up study the voices of 13 patients with dysphonia after traumatic midbrain damage were measured at 2 different stages. (Stage I: 1-12 weeks after mutism; Stage II: more than 24 weeks after mutism). Additionally to the discussed parameters we calculated the mean fundamental frequency MF_0.

2 subgroups of patients were established. The first subgroup consisted of 5 patients who revealed spectral parameters in the normal range at stage I. The second subgroup consisted of 8 patients who initially showed abnormally high spectral parameter values.

At stage II the first group exhibited an increase of the spectral parameters A1 above the range of the control speakers/ whereas the parameter A5 remained almost constant in the normal range. This particular relation of A1 and A5 indicated evolving of tense voice quality.

The second group exhibited a decrease in A1 and A5, which had been particularly high at the initial stage. This indicated a decrease of breathy voice quality. Both groups showed decreasing values of the parameters MF_0 and Tb. These findings were in accordance with the auditory judgement: a decrease in initially high pitch and breathiness, and evolving of tense voice.

The group means of PP and V5 lay in the normal range at both stages. These parameters only indicated individual changes.

4. Conclusion

In summary we found objective acoustic measures for the auditory categories of pathological voice quality, which proved to be appropriate to separate the components of compound voice types. Moreover, different stages in the process of phonatory recovery could be described quantitatively.

References:

Cramon, D., von and Vogel, M. (in press). *Zentrale Stimmstoerungen*. Erlangen: Perimed-Verlag.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

A Contribution to the Phonological Pathology of Speech Structure in Children with Impaired Hearing

A. Jarosz
Wroclaw, Poland

1. Procedure

Examinations were carried out in 12 pupils (4 girls, 8 boys) in the 5th form of the school for children with impaired hearing in Wroclaw; their age ranged between 12 and 13 years, except for 2 persons aged 14 and 15 years. In phoniatric and neurologic examinations the children exhibited a moderate to high degree of hearing impairment, a mixed type being present in 4 and a nervous one in 8 of them (see collective audiogram, Fig. 1). The utterances of the children were arranged so as to include naming a number of objects, events etc. and telling picture stories connected with everyday situations known to each child. The taperecorded linguistic material was analysed phonologically; the essential phenomena concerning vowel phonemes alone are given below:

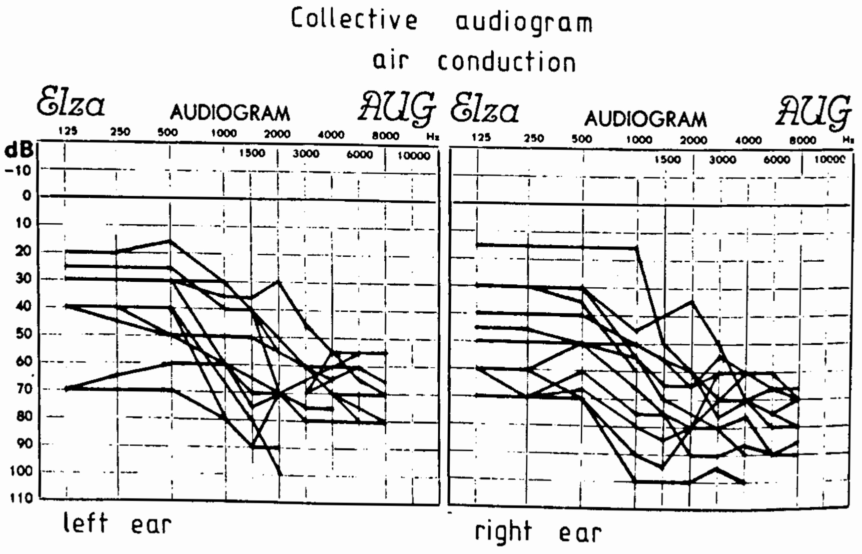


Fig. 1. A contribution to the phonologic pathology of speech structure in children with impaired hearing.

2. Analysis

2.1. Substitutions

Substitution of phoneme /y/ for phoneme /i/ after alveolar consonants from group š* which are replaced in each position by those of group ś**, exceptionally of group s*** (e.g./taleźe/ – instead of /taleže/: in Engl. plates; /uśtećka/ – instead of /xustečka/: in Engl. a handkerchief; /seb'ei/ – instead of /gžeb'en/: in Engl. a comb; /źabafk'i/ – instead of /zabafk'i/: in Engl. toys; /deś pada/ – instead of /dešč pada/: in Engl. it is raining):
- /śima źelaśko/ – instead of /tšymą želasko/: in Engl. she holds an iron;
- /ćeb'eń na vu̯ośi/ – instead of /gžeb'eń na vu̯osy/: in Engl. a haircomb;
- /kotek pśiśu̯a/ – instead of /kotek pšyšet/: in Engl. a cat has come;
- /xu̯opći pu̯aka u̯odny/ – instead of /xu̯opčyk pu̯ače gu̯odny/: in Engl. a little hungry boy has cried.

2.2. Coexistence of forms such as:

- /ʒ́efčynka/ or /ʒ́efcynka/ or /ʒ́efćinka/ – instead of /ʒ́efčynka/: in Engl. a little girl;
- /pacy/ or /pači/ – instead of /patšy/: in Engl. she watches;
- /zobacy/ or /zobaći/ – instead of /zobačy/: in Engl. she will see.

2.3 Neutralizations

Absence of /i/ by /y/ substitution; no example was recorded. In the Polish literature on speech disorders in all types of deficient hearing, the pronunciation of 'i' instead of 'y' is given as the most frequent deformation of vowels. It is suggested that these vowels are readily mistaken, as they differ but slightly in their articulation and the position of the tongue is not visible because of the mandibular-maxillary approachment. The regularities given under 1-3 seem, however, to deny such a mechanism. They would rather prove that these abnormalities result from a scheme that finds its justification in the phonological system of the language.

Phonological oppositions are subject to neutralizations. One of them is the neutralization of the opposition /i/:/y/ after prepalatal consonants (group ś among others), since they can only be followed by 'i' and also by the alveolars (among others by the s group) invariably followed by 'y'. In the children with various degrees of deficient hearing that have failed to pronounce consonants from group s and replace them permanently by those from group ś, according to the linguistic rules, 'i' occurs instead of 'y'. If the group š is mistaken for group s only sporadically, the language of children with impaired hearing invariably shows the vowel 'y' after the latter.

3. Conclusions

1. In children with various degrees of deficient hearing phonological abnormalities of the vowel system are likely to manifest themselves only as an apparent disorder of the phoneme sequence admissible in Polish language.
2. Verification of the above hypothesis on strictly selected material can be of importance from a pragmatic point of view and for modification of rehabilitation procedures.
3. A separate problem to be dealt with is that of phoniatric and neurologic studies which, in addition to socio-environmental factors, would elucidate the anatomo-physiological conditionings of speech disorders in the children under examination.

* consonants from group š: š, č, ž, ǯ
** consonants from groups ś: ś, ć, ź, ʒ́
*** consonants from group s: s, c, z, ʒ

A Tactual 'Hearing' Aid for the Deaf

K-E. Spens and G. Plant
Stockholm, Sweden; Sydney, Australia

1. Introduction

Much work aimed at using the sense of touch, to at least partly overcome the effects of a profound hearing loss, is going on in several research facilities throughout the world. Almost all authors have shown that it is possible to convey acoustic information, including speech, via the skin. Spens (1980) in a comparative study of different tactile systems, has shown that the information conveying capacity of a tactile 'hearing' aid is significant even if the processing scheme is rather simple and the number of channels (vibrators) is restricted to one. One possible conclusion from that result is that it may not be worthwhile to wait for the development of the optimal tactile 'hearing' aid but rather to design a tactile aid which at least fulfills the two very important characteristics of: 1) being capable of conveying some important acoustic information and 2) being conveniently wearable and cosmetically acceptable. Wearability has so far been a feature which has not received highest priority. In order to make daily use of the aid possible, however, and to be able to collect long term evaluative data this is a necessary feature.

2. Description of the aid

The aid looks just like a conventional body-worn hearing aid (Fig. 1). It has a built-in microphone and uses batteries which last for about a week. The aid's circuitry, however, is very different from that of a conventional hearing aid. The aid is designed to extract the intensity (loudness) variations of sound and convert them into vibratory patterns. The aim of the processing system is to present an unambiguous representation of the syllabic patterns in speech. These patterns are felt via a small vibrator which is connected to the aid by a thin cord. The vibrator has been especially designed to match the vibratory capacities of the skin while using very little power. Most users of the aid prefer to wear the vibrator on their wrists although some use it hand-held while one subject has it mounted into an ear mould. This does not mean that he hears the signal but rather feels it through the skin in contact with the ear mould. Another option currently being investigated is to mount the vibrator onto a ring which is worn on the finger. One obvious advantage with the tactile aid is that there is no feedback problem. This is a critical factor when

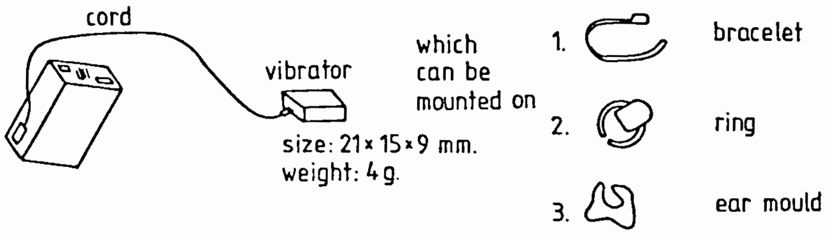


Fig. 1. The 'hearing' aid for deaf persons and different ways to stimulate the skin.

we take into consideration that many deaf persons have rejected hand-held bone conductor aids on the grounds that the feedback from such aids is potentially embarrassing to the hearing impaired person and disturbing to hearing persons.

3. Some preliminary results

At this time, approximately 15 deaf persons have had some experience with the aid for periods ranging from 3 to 20 months. Most of these are now wearing the aid on a daily basis. All but one of them has an acquired hearing loss. The prelingually deaf subject has a background of oral training and does not use sign language. Six of the subjects received special training for 2 hours weekly over a period of 12 weeks. This training was administered at a training center for deaf adults and was provided by 2 experienced teachers.

It is difficult to give a quantitative measure of the aid's effectiveness but the continued use of the aid by users indicates that they receive some net benefit from its use. Net benefit is defined as the sum of the positive and negative aspects of the aid. The positive aspects appear to be the perception of useful acoustic information while the negative aspect includes the inconvenience of wearing a technical aid. It appears that the aid's potential lies in three main areas:

1. a. The awareness of sounds in the environment.
   b. The detection of warning signals in the environment.
2. As a supplement to lipreading.
3. To provide information enabling the deaf person to monitor his/her own voice.

Both objective and subjective tests have been administered in an attempt to gain information on the advantages and disadvantages of the aid. The subjective evaluation consisted of a questionnaire which sought to obtain the deaf person's own perception of his/her handicap with and without the vibrotactile aid for the areas; environmental sounds, lipreading and voice monitoring. The scores obtained for the questionnaire are given in Table I. The scoring method used a 5-point scale ranging from 0-4 points, where 0 point corresponds to a situation of no perceived handicap. Thus the higher the score obtained the higher the subjects' own rating of his/her handicap. It can be seen that in the perception of environmental sounds and, to a lesser extent in lipreading, the subjects feel that the aid provides benefits. For the monitoring of the subjects' own voice, however, it appears that there are no differences between the aided and unaided conditions.

Table I. Subjective rating of the handicap in the aided and unaided condition (N = 6). Results are given in % of the maximum value.

	Unaided	Aided
Environmental sounds	60,5	36,6
Lipreading	57,4	53,5
Monitoring of own voice	66,7	70,0

A further indication of the aid's value can be found in the written comments of those deaf persons currently using the aid. These include:

'The vibrator helps me to perceive more with less effort'.

'People say to me "You seem to be more confident" '.

'The rhythmic information helps me, especially when I am talking to persons who are difficult to lipread'.

'I enjoy being able to feel environmental sounds.....it helps me feel more in contact with the environment'.

'When you see a movie, theatre or TV the aid helps by telling me about those background sounds gunfire, cars, music etc. which are not captioned'.

'As far as environmental vibrations, it feels like a part of myself. (i.e., the subway door closing, or traffic noise, or the sound of the vacuum cleaner plus assisting me in my lipreading.)'

'The rhythm conveys the voice quality. A ring master uses one sort of rhythm while a teacher uses another, and an auctioneer sounds different to a news reader'.

'All of these things I take for granted until I take the aid off at night'.

Efforts to objectively evaluate the effectiveness of the aid was done for environmental sounds and for lipreading.

To measure the aid's ability to transmit environmental sounds, 2 sets of 5 common environmental sounds were presented. The sounds were first presented in a set order to familiarize the subjects with the test material. The sounds were then readministered in a random order and the subjects were asked to indicate which of the 5 had been presented. The mean scores obtained for both tests were around 80% correct which is a level well above that which could be attributed to chance (20%).

The subjects' lipreading ability with and without the vibrotactile aid was measured using two methods. Fig. 2a presents the results obtained in a lipreading test by normally hearing persons artificially deafened by the

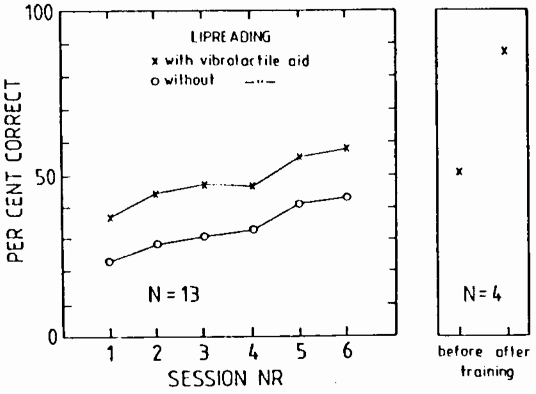


Fig. 2. a. Lipreading of sentences by artificially deafened normal hearing subjects with and without the vibrotactile aid. b. Lipreading of sentences by deaf subjects with the vibrotactile aid before and after training.

presentation of masking noise through headphones for the duration of the testing. These results highlight the value of the vibratory signal as a supplement to lipreading even with minimally trained subjects. Fig. 2b presents results obtained on a lipreading test by those deaf subjects who received special training. Both of these tests, however, used sentence materials and it was thought that a better measure of the aid's effectiveness would be obtained if material was used which more closely resembled every-day communication. The 'Tracking Technique' (De Filippo and Scott, 1978) was felt to most closely meet these demands. The method involves the reading of portions of a text to the deaf person and then asking them to repeat exactly what was said. Any deviations from the text are unacceptable and the phrase or sentence is repeated until it is repeated perfectly. This is a very difficult task which provides an extremely beneficial form of training. The material is presented for a set period of time, in this case 5 minutes aided and 5 minutes unaided, and the number of words correctly identified is calculated. This is then used to derive the number of words per minute perceived. Results obtained using this method with the training group are presented in Fig. 3. The results again show an improvement in the aided condition. It should be noted, however, that these are average results for the group and do not show individual differences between subjects. Some of the subjects appear to derive far more benefit from the aid than others. In all cases, however, the scores obtained with the aid were at least as high as those obtained in the unaided condition. All of these results indicate that the aid provides a positive support for lipreading.

These results indicate that the vibratory aid described in this paper provides useful information which help overcome, at least partially, some of the problems resulting from a profound hearing loss. Further research is needed which attempts to maximize the information available through a wearable vibrotactile aid.

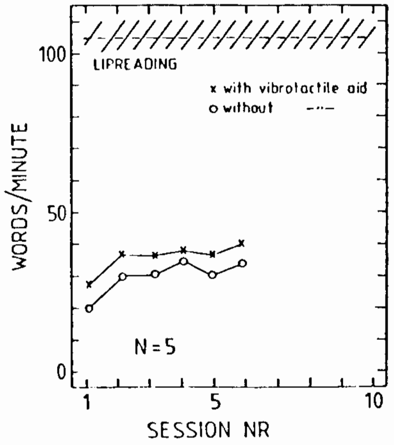


Fig. 3. Lipreading using the tracking technique by deaf subjects with and without the vibrotactile aid.

Acknowledgements

We would like to express our sincere appreciation to Birgit Ossian-Grubbstrom, Director of the pedagogical section of the Audiological Department of Södersjukhuset, and Birgitta Rollven, Teacher of the Deaf for providing the training. We also want to thank our subjects for sharing their experiences with us, thereby giving us a more complete knowledge of the use of the vibrotactile aid by deaf persons.

The project has been supported by the Swedish Board of Technical Development (project 79-5148).

References

De Filippo, C.L. and Scott, B.L. (1978). A method for training and evaluation the reception of ongoing speech. *J. Acoust. Soc. Am.* **63**, 1186-1192.

Spens, K-E. (1980). Tactile speech communication aids for the deaf: a comparison. *STL-QPSR* 4/1980, 23-39.

Verbal Apraxia in Cases of So-called Broca's Aphasia in the Light of Electromyographic Studies in Vowel Articulation

B. Weiss
Wrocław, Poland

1. Procedure

Studies on the bioelectric activity of the muscles involved in the speech act are scarce in the pertinent literature. None deals with the bioelectric activity of the articulatory muscles in aphasia.

Examinations were made in 3 men at the age of 50-60 years, with high education, after 1 year rehabilitation; they exhibited articulatory disorders due to apraxia; 3 normal persons served as control.

A DISA electromyograph attached, by means of collodion, to the left and right side of the upper lip were used for examination; bioelectric activity was recorded with repeating with and without visual control and while reading in the following systems: V1, V2, V3; V1, V4, V5; Cn, Vn; Cn, Vn, Cn; Cn, Vn, Cn, Vn. The material obtained will be exemplified by data concerning the expression of the systems: Vi, V2, V3 = a, e, i and V1, V4, V5 = a, o, u.

a - in Engl. corresponding to a e.g. in car
e - in Engl. corresponding to e e.g. in pen
i - in Engl. corresponding to i e.g. in sea
o - in Engl. corresponding to o e.g. in pot
u - in Engl. corresponding to u e.g. in put

2. Results

In the healthy and normally speaking subjects bioelectric activity is similar in repetition and reading (Fig. 1). With vowels a, e the activity is absent and with i only slight (Fig. 2). Activity appears before the articulation of o, u passing to the articulation phase of the syllables. It appears and fades away gradually reaching its peak at the articulation of u. The interferential record is low.

Considering the electromyograms from all 3 trials, the following regularities can be observed:

1. Fig. 3. The bioelectric activity from the left side of the musculus orbicularis oris is higher on repetition with visual control. Right side interferential records are low or intermediate as with a. The left side is clearly increased, particularly with u. Volley discharges, basal line oscillations are visible. This record reflects an action of 'searching' the articulatory system.

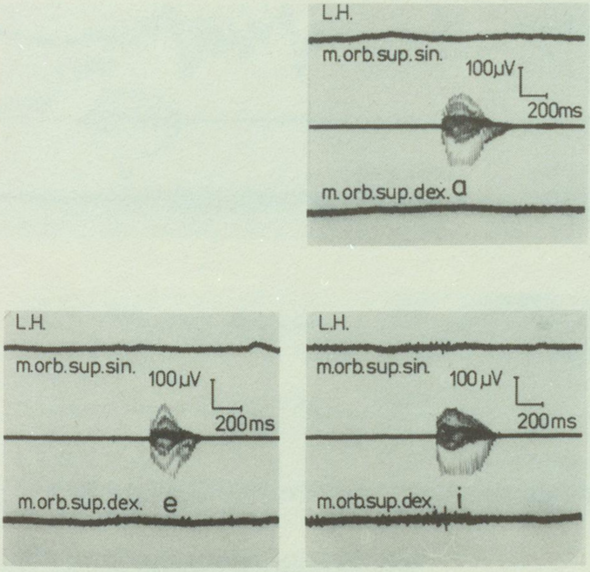


Fig. 1.

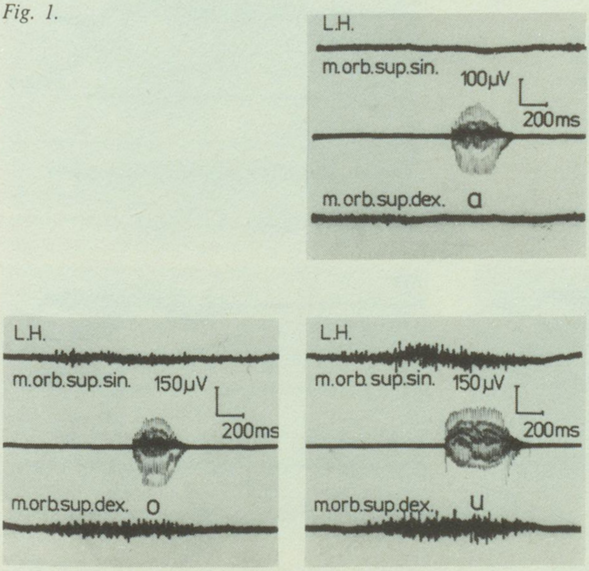


Fig. 2.

Striking is also the short time of the muscle activity before the production of vowel o.

2. Fig. 4. Bioelectric activity invariably occurs with the production of a, e, i; the interferential records are low, irregular, especially with i, on the left side.
3. Fig. 5. There is a tendency for the activity from the left and right side of the m. orbicularis oris to become uniform on repetition without visual control. This uniformity is to be seen bilaterally. The activity of the muscle lasts longer and fades away only after the production.

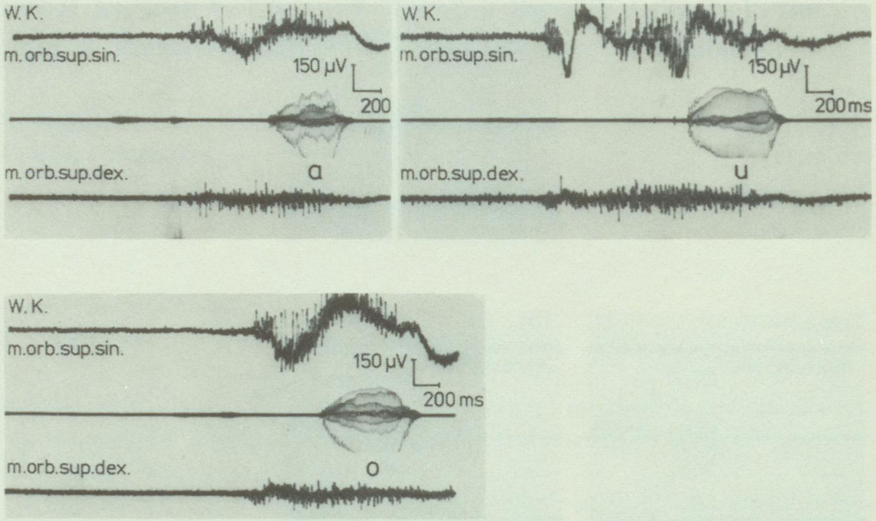


Fig. 3.

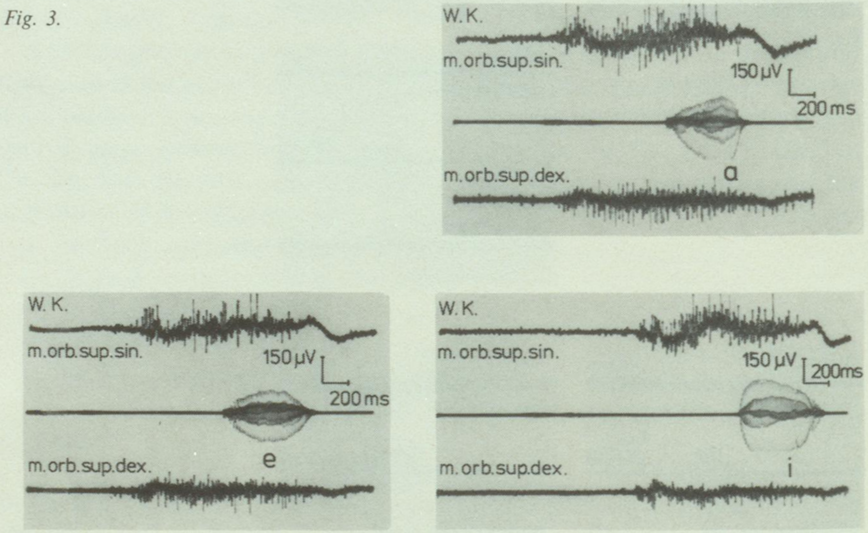


Fig. 4.

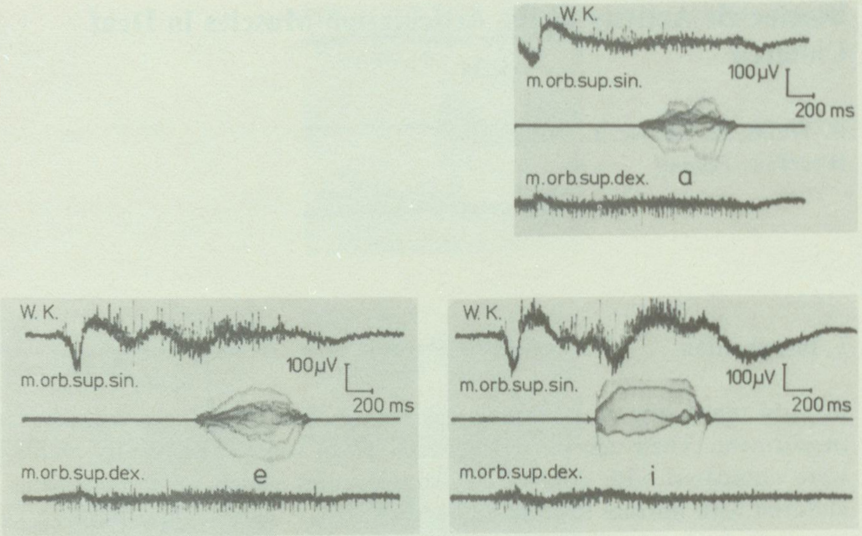


Fig. 5.

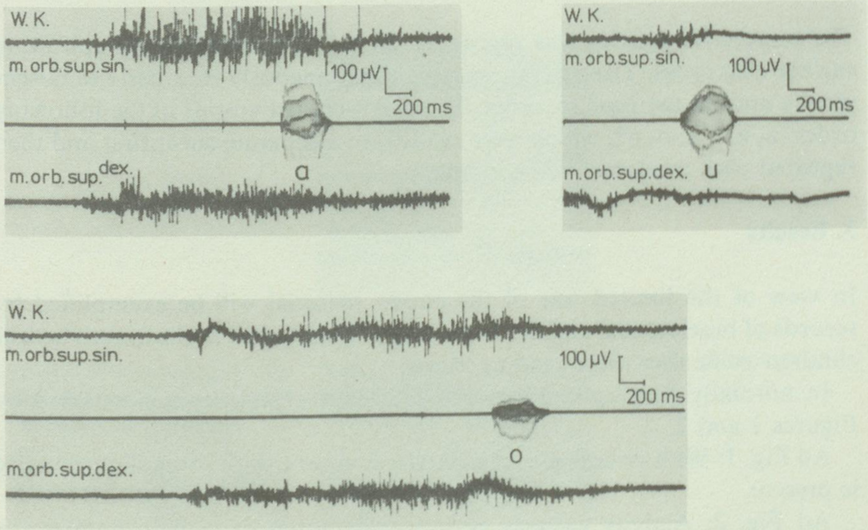


Fig. 6.

4. Fig. 6. A great variability of the bioelectric activity is observed on reading. With a, a volley-like interferential record is obtained on the left and a group discharge on the right side. During expression of o there is a long lasting activity of the muscle (2200 ms); with u, a left side rudimentary record with a tendency for rhythmization is obtained.
5. All the above given examples also indicate the time of bioelectric activation before the expression of the vowels; it ranges, on average, between 500 and 2200 ms; the shortest on repetition with visual and auditory control and the longest one on reading, though repetitions tend to reduce it. Regularities pointing to abnormalities which in turn reflect the complexity of the problem under observation can by no means be regarded as binding conclusions. They do, however, prove the necessity of continuing studies the effectivity of which, considering the pathomechanism of the disorder itself and methods of its rehabilitation, is hardly predictable. Nonetheless, it appears that rehabilitation of patients with speech production disorders in aphasia due to apraxia should be based on administration of visual articulatory models under auditory control with concomitant demonstration of written patterns.

Bioelectric Activity of the Articulation Muscles in Deaf Children

B. Weiss, A. Jarosz, L. Handzel
Wrocław, Poland

1. Introduction

M. orbicularis oris of the mouth was examined in 10 children with hearing impairment. Their age ranged between 13-14 years. The results obtained were compared with control electromyograms which were made with 3 children with normal hearing and speech.

2. Method

The bioelectric activity was registered on a DISA electromyograph using surface electrodes. The utterances were simultaneously recorded oscilloscopically and on the tape recorder. The test included vowels in the following order: a, e, i, a, o, u*, which were read from a separate sheet first and then repeated with moderate voice intensity.

3. Results

In view of the limited size of the paper, material will be exemplified by records of bioelectrical activity from the circular muscle of the mouth in deaf children while they were reading alone.

In normally hearing children the bioelectrical activity is illustrated by Figures 1 and 2.

Ad Fig. 1: With vowels e, a, the activity is absent, with i only slight activity is present.

Ad Fig. 2: Activity appears before articulation of o, u to involve the articulation phase itself; it appears and vanishes gradually, and reaches its maximum with articulation of u. In the deaf children desorganization of the bioelectrical activity with its pronounced enrichment is a predominant feature at the time of reading the vowels, this variability being clear-cut mainly before their articulation. This may be either a synchronic activity in the form

* a - in Engl. e.g. corresponding to a in car,
e - in Engl. e.g. corresponding to e in pen,
i - in Engl. e.g. corresponding to i in sea,
o - in Engl. e.g. corresponding to o in pot,
u - in Engl. e.g. corresponding to u in put.

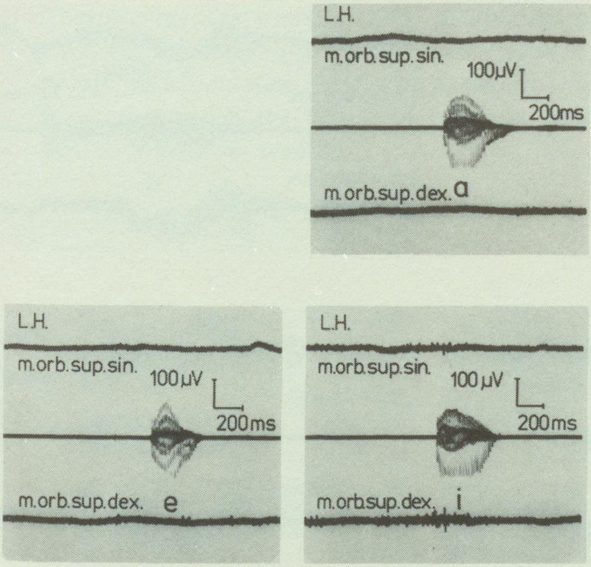


Fig. 1.

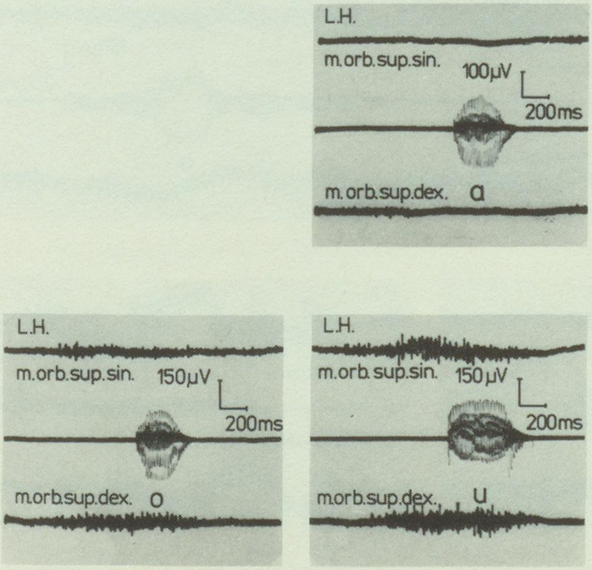


Fig. 2.

of volleys or potential groups (Fig. 3), or asynchronic (fig. 4) when alternative continuous activity in the form of interferential or indirect record corresponds with a low interference, or electrical silence. During articulation itself the bioelectric activity is basically synchronic but very rich, as compared with that of hearing persons. It lacks, however, an explosive character. Interferential records prevail, especially in the first articulation phase.

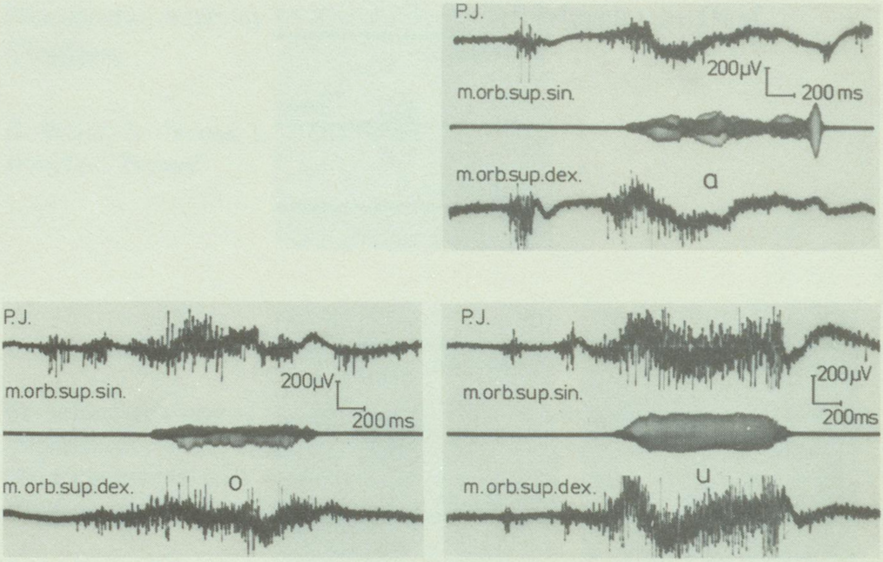


Fig. 3.

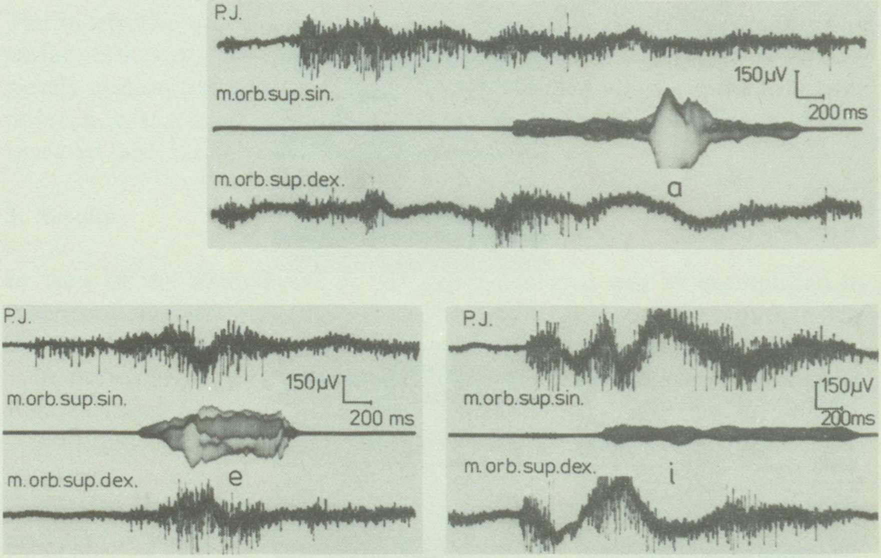


Fig. 4.

4. Conclusions

1. The bioelectrical pattern of the circulatory muscle of the mouth in deaf children's reading task differs from that in children with normal hearing, with no relationships existing between them.
2. The pattern itself is highly differentiated; though some regularities can be distinguished in prearticulation and articulation phases.

3. The bioelectrical activity in the prearticulation phase of deaf reading children, which during repetition tends to vanish or reappear immediately before the articulation itself, with a loss of its variability, might be due to superposition on the action of speaking of an additional intellectual operation such as reading.

Section 21
Hearing Pathology and Speech

Hearing Impairment and the Perception of Speech Sounds

C. Gélinas-Chebat
Montréal, Canada

1. Epistemological issues

Multi Dimensional Scaling (MDS) methods commonly used in phonetics impose two basic assumptions:
1. *Symmetrical Distances*
2. *Homogeneous Samples*

to enable the analyst to develop perceptual maps of stimuli, the dimensions of which are assumed to represent the attributes along which each stimulus is compared (Shepard, 1972).

These assumptions constitute two heavy constraints upon empirical results:
1. the number of times stimulus 'i' is perceived 'j' by experimental S's is not necessarily equal to the number of times stimulus 'j' is perceived 'i' by the same subject: imposing that d(ij) = d(ji) implies that the specific nature of the stimuli is significantly altered.
2. assuming that S's are similar obviously precludes any further analysis of S's characteristic to explain dispersion of perception: moreover analyses of perceptual diversity do *not* use the same class of data as joint space configurations (respectively quadrant I and IV in Coombs' classification of data: Coombs, 1964).

2. 'Nearest neighbour'

Two French statisticians (Chandon and Pinson, 1981) have proposed a method encompassing these two heavy assumptions. In (very) short, this technique is included in the proximity analysis paradigm. When matrices of occurrence Mij are formed for two S's, A and B, a distance between A and B is calculated: this distance is the summation of the difference of observations elevated to square, in each cell (i,j) of the matrix for individual A and its homolog for B.

$$dAB = \sum_{MAT} \sum_{RON} \sum_{COL} (Cij\ (A) - Cij\ (B))^2$$

This method enables us to stress the *individual differences* of S's (instead of assuming sample *homogeneity*) and real distances between stimuli (without artificial symmetrization).

3. Research objectives

1. Since they assume sample homogeneity, researchers using MDS methods cannot take into account inter-individual differences: for a given group of individuals having similar audiometric test results, the whole perceptual configuration is necessarily attributed to common audiometric characteristics (Bilger and Wang, 1976). Not only the perceptive configuration is altered by artificial symmetrization, but the assumed homogeneity involves a very heavy assumption: auditive perception is strictly a peripherical phenomenon. This paper shows that individuals with similar audiometric patterns do not perceive similarly identical stimuli.
2. More specifically, the assumed relation between audiometric scores and perceptive behavior leads to the conclusion that hearing impaireds hardly perceive acute sounds (e.g. Schultz and Kratt, 1971; Pascoe, 1975; Barth and Chulliat, 1980). We challenge this specific relation with a distinctive feature analysis.

4. Study

12 French speaking adults whose better ear's threshold was at least 30 dB, were presented 36 monosyllables of CV type (both fricatives and occlusives) combined with vowels /a, i, u/.

These monosyllabic words were recorded by a male voice and presented 5 times in a random order to the individual subject through headphones; the sound level was adjusted by the individual subject to a comfortable level during a familiarization period.

Each subject had to cross the word which he/she believed to have heard among a choice of six words, printed in a form handed to the subject. Each word presentation was preceded by a warning light.

Data: The data were organized in 72 square matrices (six for each of the 12 S's as shown in the next table, table I).

The 'i-j' cell of the square matrix is the number of times the word 'i' has been perceived as the word 'j'.

5. Results

a. Interindividual audiometric differences

The 'nearest neighbour' technique, based upon the calculation of a distance from one individual to the other, allows us to construct a tree of proximity between individuals as shown in the next figure (figure 1).

Table I. Results subjects # 1

	P	T	K	B	D	G		P	T	K	B	D	G		P	T	K	B	D	G
PAS	2	2	1	0	0	0	PIE	5	0	0	0	0	0	POU	4	0	1	0	0	0
TA	0	5	0	0	0	0	TI	0	5	0	0	0	0	TOUX	0	5	0	0	0	0
CAS	0	0	5	0	0	0	QUI	1	0	4	0	0	0	COU	0	0	5	0	0	0
BAS	0	0	0	5	0	0	BIS	0	0	0	5	0	0	BOUE	0	0	0	5	0	0
DA	0	0	0	0	5	0	DIT	0	0	0	0	3	2	DOUX	0	0	0	0	5	0
GARS	0	0	0	0	0	5	GUY	0	0	0	0	1	4	GOÛT	0	0	0	0	0	5
	F	S	CHV		Z	J		F	S	CHV		Z	J		F	S	CHV		Z	J
FA	3	2	0	0	0	0	FIT	5	0	0	0	0	0	FOU	5	0	0	0	0	0
SA	0	5	0	0	0	0	SI	0	5	0	0	0	0	SOU	0	5	0	0	0	0
CHAT	0	0	5	0	0	0	CHIE	0	0	5	0	0	0	CHOU	0	0	5	0	0	0
VA	0	0	0	5	0	0	VIE	0	0	0	5	0	0	VOUS	0	0	0	5	0	0
ZA	0	0	0	0	5	0	ZI	0	0	0	0	5	0	ZOU	0	0	0	0	5	0
JAS	0	0	0	0	0	5	GIT	0	0	0	0	0	5	JOUE	0	0	0	0	0	5

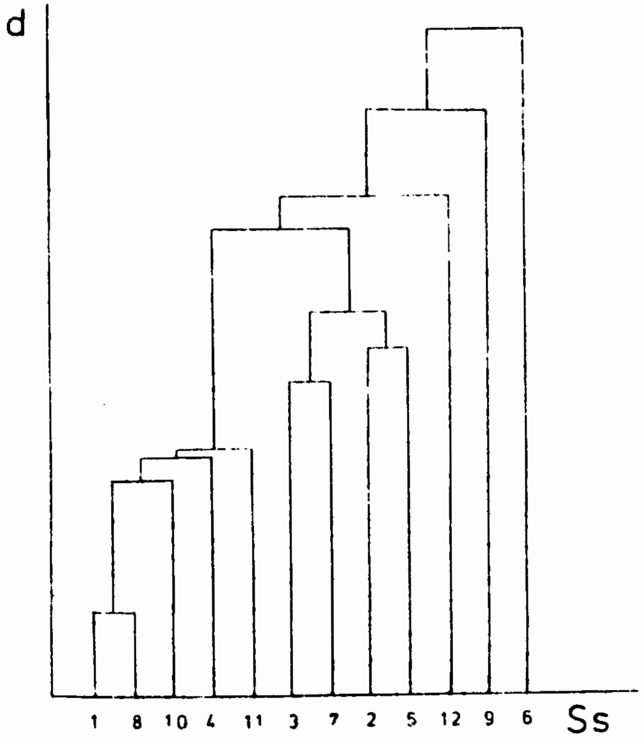


Figure 1. The tree of proximity with the nearest neighbour analysis.

This technique enables us to classify our impaired S's with the data of the perceptual test, without having to take into account their audiometric data.

A close examination of these audiological data shows that there is *no evident relationship with the classification based on the perception tests.* Individuals classified as similar for their perception of words have no comparable audiological characteristics in terms of threshold and discrimination. On the contrary some individuals with similar audiological data were classified in different groups through the nearest neighbour technique and the perceptual data.

b. Interindividual perceptual differences

The second way we used to analyze our data was a 'distinctive features' approach. Through this technique we tested the assumption that acute sounds are least successfully by the hearing impaired.

Our data were organised in two by two matrices as shown in Table II. The row represents the presence of feature 'F' on the first row and the absence of this feature 'F' on the second row, for the word emitted. The columns represent, for the word perceived, the presence of the feature 'F' on the first column and the absence of this feature 'F' on the second column.

Table II. A Distinctive Feature Matrix

	F	NF	
F	N-x	x	N
NF	y	M-y	M

We have chosen a classification as proposed by Jakobson and Halle (1967)
[p,b,f,v] as 'GRAVE'
[t,d,s,z] as 'ACUTE'
[k,g,ʃ,ʒ] as 'COMPACT' rather than non-grave/non-acute.

We also have pooled the 'VOICED' data which means that [b,v,d,z,g and ʒ] are in the same group.

For each row we were able to fix thresholds beyond which we can state the results are not randomly distributed according to chi-squared tables. Thus with the error data (x,y) we draw a graph as shown next figure (figure 2).

Here we have 4 areas but only 3 deserve attention:

- area 1 where there is no confusion:
  'F' is perceived as 'F'
  'NF' is perceived as 'NF'
- area 2 where there is no confusion on 'NF' but where 'F' is significantly perceived as 'NF'
- area 3 where there is no confusion on 'F' but where 'NF' is significantly perceived as 'F'

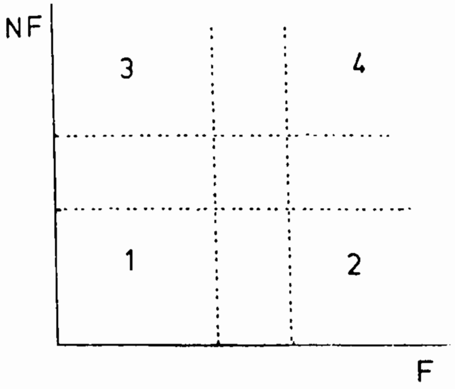


Figure 2. Graph of the distinctive features analysis.

A thorough examination of our data shows that not all features do bring about the same confusion. The next figure (Figure 3) describes the hierarchy of confusion.

The less successfully perceived sounds are grave and not acute, but there

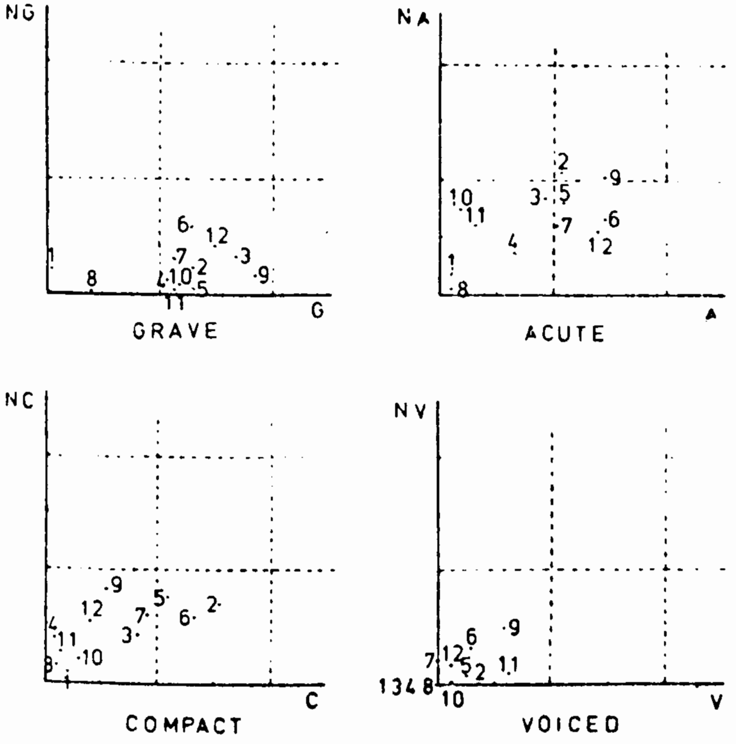


Figure 3. Results of the distinctive features analysis.

sounds are less successfully perceived than the compact one. The voiced sounds appear always to be recognized as voiced.

The different groups established by the nearest neighbour analysis could also be identified in this distinctive features analysis and shows the relationship between those two techniques.

6. Conclusion

The two techniques used (i.e. 'nearest neighbour' and 'distinctive feature') enable phoneticians to classify individuals when using perceptual data only.

Those techniques are superior to the MDS (multidimensional scaling analysis) since the researchers are not compelled to use big samples and do not force data into a statistical technique by artificial symmetrization.

We are inclined to think that audiological data cannot predict perceptual behavior mainly because hearing impaired individuals develop personal strategies of compensation since perception is a *central cognitive process* whereas audiological data concern physiological and peripheral processes.

References

Barth, S., Chulliat, R., (1980). Perception auditive des fricatives par les déficients auditifs. *Actes des XIe journées d'études sur la parole*, le groupement des acousticiens de langue française. Strasbourg, 18-24.

Bilger, R.C., Wang, M.D. (1976). Consonant Confusions in Patients with Sensorineural hearing Loss. *Journal of Speech and Hearing Research (J.S.H.R.)*, **19**(4): 719-748.

Chandon, J.L., Pinson, S. (1981). *Analyse typologique theories et applications*. Paris: Masson, p. 254.

Coombs, C.H. (1964). *A Theory of Data*. New York: John Wiley and Sons Inc.

Danhauer, J.L., Singh, S. (1975). *Multidimensional Speech Perception by the Hearing Impaired. A Treatise on Distinctive Features*. Baltimore: University Park Press, p. 130.

Jakobson, R., Halle, M. (1967). Phonology in Relation to Phonetics. In: *Manual of Phonetics*. Amsterdam: Kaiser, L., (ed.), North Holland Publishing Co., 216-251.

Lafon, J.C. (1972). Perception phonétique au seuil d'audition. In: *Papers in Linguistics and Phonetics to the Memory of Pierre Delattre*. The Hague: Valdman Ed., Mouton: 287-307.

Pascoe, D.P. (1975). Frequency Responses of Hearing Aids and Their Effects on the Speech Perception of hearing Impaired Subjects. *The Annals of Otol., Rhino. and Laryng* Supplement **23**: 84.

Schultz, M.C., Kratt, A.W. (1971). Lack of Perceptual Reality of the Phoneme for hearing-handicapped Children. *Language and Speech*, **14**: 178-186.

Shepard, R.N. (1972). Psychological representation of speech sounds. In: E.E. David Jr. and P. Denes (eds.) *Human Communication: A Unified View*: McGraw Hill, New York 67-113.

Accents and their Differential Influence on the Perception by Normally Hearing and Hearing-impaired Subjects

D. Günzburger
Utrecht, the Netherlands

1. Introduction

Although the importance of prosodic cues in speech perception has been established beyond any doubt, relatively little can be found in the literature about the influence of prosodic factors on the perception of speech by persons afflicted with a hearing deficiency. The aim of the experiment to be reported on was to investigate whether prosodic cues are of comparable influence on the speech decoding process of normally hearing and hearing-impaired persons. The variables investigated are emphasized vs. non-emphasized words, position of sentence accent, position of word and speaker variation.

2. Method

In the assessment of the hearing level and its effect upon speech perception it is not enough to use tests which reflect mainly the functioning of the ear and the peripheral-neural parts of the system in response to simple stimuli such as pure tones or clicks for the following reasons: a patient's ability to hear and understand speech may not only be differentially affected by the specific nature of the deficiency, but it will also depend upon other variables, such as for example age at onset of the hearing loss, degree of linguistic attainment and amount of auditory training, all of which reflect the degree of exposure to normal language and are different in quality from those variables concerned simply with the acoustic input side of the system.

Bench and Bamford (1979) fervently plead for an advantage of sentences over words in speech audiometry tests. They mainly derive their arguments from the simple fact that 'because sentences are more than mere strings of words, perception of words in isolation is not necessarily a good predictor of the perception of sentences, which constitute the material of everyday speech' (1979: 17). An additional advantage of sentences over words is that they also allow investigation of the time domain, since they are of sufficient duration to permit alteration of the temporal characteristics of speech. The same holds of course for various other prosodic parameters.

On the strength of these arguments we decided that our test material should consist of sentences. For the construction of the test sentences the following criteria were kept in mind: occurring words should be in common use and checked for frequency of occurrence in natural language, sentences should be syntactically simple and more or less equal in length so as not to be a burden on the memory and, of course, they should be semantically acceptable. As to the number of sentences we decided - on the basis of various trial runs - that 17 sentences would achieve an acceptable balance between reliability and test duration.

We selected our speakers - one male and one female - on the grounds of good intonation, clear articulation and good voice quality. Both were native speakers of Dutch. Sentences were read out in three versions, the difference consisting in sentence accent assignment. Recordings were made using high-quality recording equipment. For masking purposes speech noise was recorded on a parallel track. Based on a pilot study S/N ratio for normally hearing subjects was decided to be -2 dB and for hearing impaired subjects +5 dB so as to achieve an average score of approximately 50% correct for either group.

In all 106 subjects participated in the test; 53 were pupils of various secondary schools for hearing impaired children; their ages ranged from 12 to 17 years and they all had a congenital hearing loss. Hearing impaired subjects took the test individually. The 53 normally hearing subjects were either undergraduate students of the Department of English of Utrecht University or students of the Dutch Academy for Tourism, none of whom had any self-reported hearing deficiencies. They took the test in a language laboratory equipped with headphones. Subjects of each group were randomly assigned to an experimental condition (male or female speaker, early, middle or late sentence accent). They received oral and written instructions and were asked to write down whatever they heard after every test item, even if it were only a fragment of the whole sentence.

3. Analysis and Results

The first two sentences were considered practice items and were therefore not incorporated in the analysis of results. The remaining 15 sentences were analysed as to influence of accent (emphasized vs. non-emphasized words), position of accent in the sentence (early, middle or late) and position of the word in the sentence. Results were quantified in terms of number of correctly perceived words. Figure 1 shows the results.

Although S/N ratios for normally hearing and hearing impaired subjects had been based on pilot experiments, we have to accept the fact that hearing impaired subjects' scores were lower than the hoped for 50% limit and normally hearing listeners' results were lower still. We therefore did not consider it justified to carry out interpopulation comparisons as to absolute values of test results, but we want to make the following observations:

- in all conditions, i.e. both subject categories, both speakers and all accent positions, accented words are perceived better than unaccented words.
- hearing impaired listeners score highest on the early accent position condition and lowest on the late accent position condition. Normally hearing subjects score highest on the middle accent position; this again holds for the male and the female speaker condition.
- SDs are in all conditions smaller for unaccented words than for accented words.
- SDs are on the whole smaller for normally hearing subjects than for hearing impaired subjects; this holds for both the male and the female speaker condition.

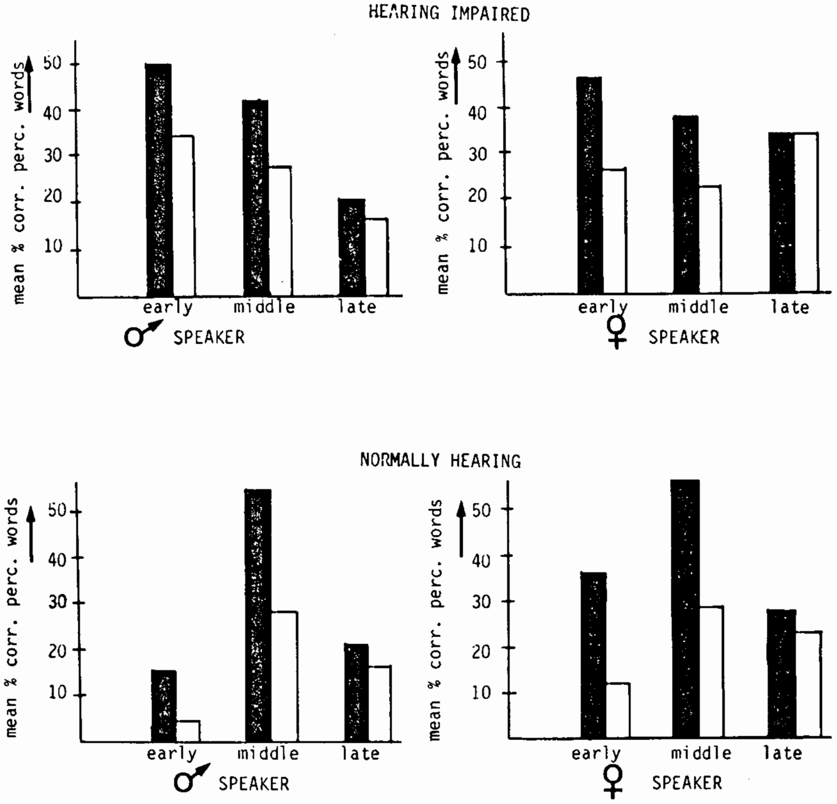


Fig. 1. Mean percentage of correctly perceived words per condition. ■, accented words; □, unaccented words,

Since early, middle and late accent positions are defined in terms of *accented* words only, it seemed necessary to investigate the influence of word position in the sentence in general on the percentage of correctly perceived words by hearing impaired and normally hearing subjects in both the male and female speaker condition. Results of this analysis can be seen in figure 2.

Results of a breakdown of means for all *unaccented* words made us observe the following:

- both listener groups show relatively high scores for word positions early in the sentence.

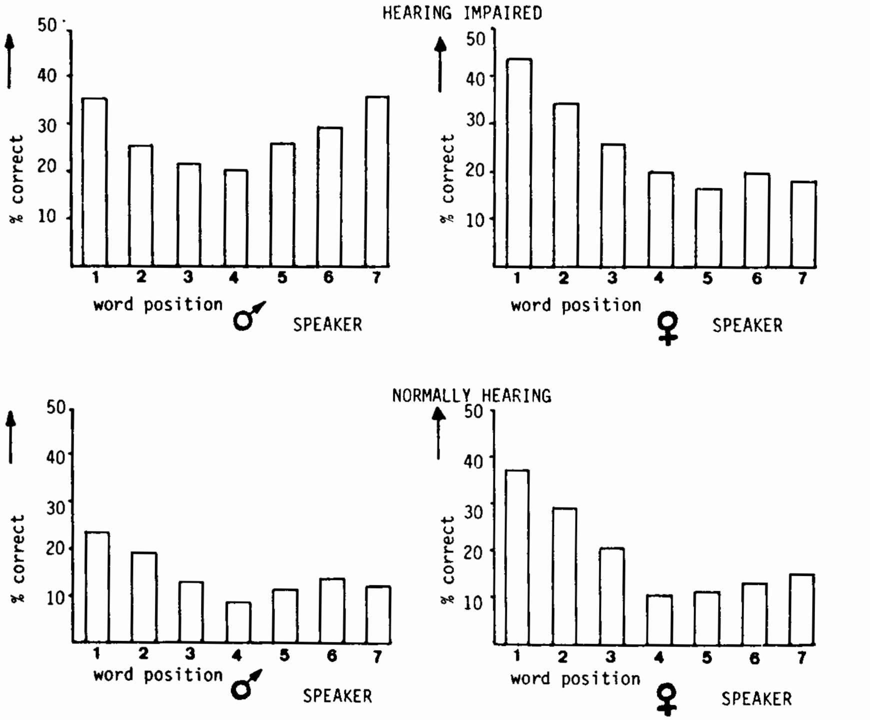


Fig. 2. Mean correctly perceived words per word position for either male or female speaker condition.

- SDs are smaller for normally hearing listeners than for hearing impaired listeners; this difference is more marked in the male speaker condition than in the female speaker condition.
- (corollary): the degree of variability of responses is related to the absolute number of correct responses; thus we must assume that a high percentage of correct scores is caused by a small number of extremely good responses.

4. Conclusions

As to test scores of unaccented words we see a declining discrimination line for *both* groups of subjects indicating a higher percentage of correctly perceived words at the beginning of the sentences. This can most probably be explained by the fact that the amount of information is highest at the beginning of an utterance, or sentence and that the listener has trained himself to focus his attention on that part of the sentence. However when looking at the influence of *position of accent* we see different results for the two populations: hearing impaired listeners show - as was also the case for unaccented words - the highest scores on the early accent position, whereas normally hearing listeners score highest on the middle accent position. In other words: the general structure of test results remains the same for hearing impaired listeners whether looking at accented or unaccented words, whereas this is clearly not the case for the normally hearing population.

It is also known from the literature that hearing impaired persons generally have great difficulties in detecting which of a limited number of words in an utterance is emphasized, whereas normally hearing people are accurate and highly consistent in making decisions as to what parts in a sentence are stressed (Risberg and Agelfors, 1978; Lea, Medress and Skinner, 1972). These observations corroborate our findings and our main conclusion is then that accent has a less differentiating function in the decoding process of speech by hearing impaired listeners than by normally hearing persons.

Overall percentage correct responses is somewhat higher in the female speaker condition than in the male speaker condition (31% vs. 27%). An attempt at generalizing these findings would not be warranted in this particular case because of the lower speech rate of our female speaker as compared with the male speaker. The influence of speech rate on the perception of normally hearing and hearing impaired listeners will be discussed in the related paper by Vingerling.

References

Bench, J. and Bamford, J. (eds.) (1979). *Speech-hearing tests and the spoken language of hearing-impaired children.* Academic Press, London, New York.

Lea, W.A., Medress, M.F. and Skinner, T.E. (1972). *Prosodic aids to speech recognition.* Univac Report PX 7940.

Risberg, A. and Agelfors, E. (1978). On the identification of intonation contours by hearing impaired listeners. *STL-QPSR 2-3*: 51-61.

Auditory Evaluation of the Speech of Deaf Children

P. Janota
Prague, Czechoslovakia

1. Introduction

An important aspect of the successful socialization of hard-of-hearing or deaf persons is the quality of their speech. The hearing environment which deaf adolescents enter after finishing school judges the quality of their speech not only from the point of view of its intelligibility, but also from the more complex point of view of 'social acceptability' or 'general acceptability' of their speech as a means of communication. The striking departure of their speech from the orthoepic norm, its unnaturalness and decreased intelligibility are caused not only by deviations in articulation, but to a great extent by an abnormal use of suprasegmentals as well; this may lead to difficulties in communication with normally hearing people not used to the peculiar character of the speech of the deaf and consequently to impaired social contact or even isolation.

2. Procedure

The question has arisen whether the term 'social acceptability' or 'general acceptability' could be used as an efficient criterion for judgments on the quality of speech. Therefore a listening test was prepared in which the task of the listeners was to give their judgments using this term on 50 speech samples of hard-of-hearing and profoundly deaf Czech children, pupils of the fourth (9-10 years) and ninth (14-15 years) forms of several special schools for children with impaired hearing. All the speech samples were recorded within a period of several months and the same technical equipment was used. The samples were extracts of an identical text: several short sentences were worked out to be sufficiently easy for the children to understand; additionally these sentences contained all the Czech vowels, a representative sample of Czech consonants, several examples of assimilation, consonant clusters, various forms of intonation, etc. The samples were used without repetition, i.e. they contained the speech of 50 different children; the duration of the test was 19 minutes.

The whole test was presented to several groups of listeners - in the first part of the investigation these were 30 students of Czech from the Charles University Philosophical Faculty, 40 students of special pedagogy from the

Charles University Pedagogical Faculty and a special group of 30 teachers of the deaf. The first two groups were thus 'naive listeners' - none of them was in regular contact with the speech of the deaf, whereas the third group consisted of expert listeners, very well acquainted with the speech of the deaf and of hard-of-hearing children through daily contact with them.

The task of the listeners was to evaluate the social acceptability of speech in the items of the test. The criterion was described shortly as 'the acceptability of speech: (1) as a means of communication with normally hearing people, (2) its general quality, (3) its special character and (4) its deviation from normal speech, in addition to its intelligibility.'

The listeners were asked to mark the items by means of normal rating scale 1-5, currently used in Czech schools. The use of the scale values was not prescribed or explained, but before the test the participants listened to six additional items for orientation and as a minimum training.

3. Results

The computations of the results showed very good agreement of judgments (chi-square tested) within the two students groups and within the group of teachers. But there was a clean-cut and significant difference in the judgments of students vs. teachers: the teachers were less rigorous in judging the speech samples. This result could be expected - experienced teachers are able to estimate the degree of hearing loss of the speakers and they can better appreciate the child's effort. This may also explain a slightly greater dispersion in the ratings of teachers compared with those of students. A more important finding, however, becomes evident when both the teachers' and the students' judgments are arranged according to their ranks: the rank order of both arrays is nearly identical (Spearman's rank correlation coefficient $R = 0.99$), i.e. though the teachers were less strict in their ratings, there are nearly no ordinal differences in comparison with the students' ratings.

It is evident that there is a causal relation between hearing loss and quality of speech. Therefore the hearing loss was calculated in dB as the average value of losses for the better ear on the speech frequencies in pure-tone audiograms, and then these calculated losses were correlated with the ranking assigned to each individual sample. A correlation coefficient $R = 0.65$ was found; thus the correlation was positive and significant, but rather weak. In other words: quality of speech is not rigidly determined by hearing loss, a child with a certain amount of hearing loss can develop better or worse speech within considerably broad limits. Indeed, among the speech samples several examples can be found where in spite of identical ratings of social acceptability the differences in hearing loss amount to 30 dB and vice versa. Besides, the correlation analysis yields data, by means of which several cases can be singled out in which the quality of speech deviates significantly from the expected value, i.e. children needing special attention.

In the second part of the investigation the identical test procedure was extended to more qualities of speech of the deaf: besides general acceptability', four other aspects of speech were tested: articulation, rhythm, intonation and pitch aberration from what would be expected as normal. The listeners, two other groups (71 total) of students with elementary knowledge of phonetics but without any experience with the speech of the deaf, were asked to concentrate on these particular aspects of the speech samples and to try to ascribe values on the rating scale to them. Again this task proved feasible, with very good agreement (chi-square tested) within and between groups of listeners. In addition these results were in good agreement (chi-square tested) with the results of the first stage of investigation.

The main aim of this test, however, was an analysis of the relation between the investigated aspects and expecially the correlation of these aspects with social acceptability as well as their relation to hearing loss. All the essential coefficients of correlation are summed up in Table I and Table II.

Table I. Matrix of coefficients of rank order correlation R between aspects: articulation, rhythm, pitch abberation, intonation and social acceptability

	Accept.	Inton.	Pitch	Rhythm
Articulation	0.98	0.90	0.77	0.89
Rhythm	0.93	0.96	0.80	
Pitch	0.83	0.86		
Intonation	0.95			

Table II. Matrix of coefficients of rank order correlation R between aspects: articulation, rhythm, pitch abberation, intonation, social acceptability and hearing loss

	Articul.	Rhythm	Pitch	Inton.	Accept.
Hearing loss	0.72	0.66	0.63	0.67	0.73

4. Discussion

All the coefficients are positive, significant and have rather high values. The strongest correlation is found, as can be expected, between social acceptability and articulation (0.98), but immediately after this value comes the coefficient for intonation (0.95) and the only slightly lower value for rhythm (0.93). These data demonstrate the considerable importance of good intonation and rhythm for a sufficient acceptability of speech. Here again a rather weak (0.73) correlation, though somewhat stronger than in the previous experiments, can be found between hearing loss and social acceptability.

Besides the general tendencies of the experimental group stated above, the data thus obtained may be used to characterize particular deaf individuals.

First, the values of correlation between hearing loss and social acceptability of speech may be used to identify those deaf children whose quality of speech is either considerably better or considerably worse than might be expected on the basis of their degree of hearing loss.

Second, the speech of each child can be characterized by the numerical values for the five evaluated aspects of his or her speech. These individual data may be expressed in the form of simple graphs in which the relative value of each aspect is apparent.

5. Conclusions

The results of the described test have shown that

1. the aspect 'social acceptability' is a usable criterion in evaluating the quality of the speech of the deaf;
2. the aspects 'articulation,' 'rhythm,' 'intonation,' and 'pitch aberration' can be tested by the same procedure;
3. the auditory evaluation of these aspects allows us to show experimentally, among other things, the close relation between the social acceptability of speech and its intonation and rhythm;
4. this kind of test may be performed with groups of listeners without any special training.

   In addition, it is hoped that the described tests may add to the knowledge of the attitudes of non-expert listeners toward the speech of the deaf; these attitudes are to be respected in the education of the deaf.

On the Possibility of Tactile Categorical Perception

H.G. Piroth, H.G. Tillmann
Munich, Federal Republic of Germany

1. Introduction

If trials to develop devices enabling cutaneous transmission of speech have failed so far, the reason could well be that no adequate method has been found to transfer information via the skin with a velocity comparable to the values known for the auditory mode (e.g. Kirman, 1973). Furthermore, to evaluate the general possibilities for tactile transmission of speech it is necessary to compare the psychological phenomena underlying tactile perception with compatible effects in auditory speech processing. Since one of the most important effects discussed in this connection is 'Categorical Perception' (e.g. Repp, 1982) a series of tests was carried out to determine the possibility of establishing tactile phenomenal categories that also fulfil the conditions for categorical perception.

2. Stimuli and Procedure

The stimuli were constructed, produced, and presented under computer control (PDP 11/50) with the system SEHR-1 for electrocutaneous stimulation, consisting of two hardware and two software components: a 16-channel stimulus generation device producing current-controlled a.c.-impulses of the form in Fig. 1 (the impulse frequency is variable along the physiologically important range from 100 to 500 Hz), an interface to the computer, a MACRO-library for controlling the interface, and a FORTRAN-library for producting and presenting the test stimuli (Tillmann and Piroth, 1983).

Our test sequences were based on the 'rabbit phenomenon' (Geldard and Sherrick, 1972), which in its standard version is elicited when m tactile taps are presented with a constant temporal interval Δt to each of n equidistantly and linearly ordered loci. Each single tap was stimulated by three impulses repeated at a rate of 5 ms (cf. Fig. 1). Rabbit sequences with 9 taps at three different loci were used. The temporal interval separating the taps was varied to obtain two continua, each consisting of 8 stimuli. The intertap interval varied from sequence to sequence in steps of 20 ms. In continuum I a range from 5 to 145 ms was chosen, to enclose also intervals too short to be perceived as isolated taps, but eliciting the apparent movement phenomenon (Sherrick and Rogers, 1966). In continuum II the intertap interval varied from 145 to 285 ms, to cover the point where the rabbit phenomenon breaks down and single taps at the veridical loci are perceived. As single rabbits could be discriminated by the parameter of total length, two different stimulus forms have been tested. In the first case the overall duration of each sequence was held constant: the stimuli were repeated for 3.6 s in continuum I and for approximately 6.4 s in continuum II. In this case the end locus of the stimulus sequence could not be held constant. Therefore a second stimulus form was realized by shortening the sequences to the amount necessary to make each sequence end at the same locus as it started at. 3 members and 1 student of the institute participated in the experiments, only 3 had previous experience of electrocutaneous stimuli. The stimuli were presented via a matrix containing 3 pairs of circular gilded brass electrodes, 9 mm in diameter each. The smallest distance between the two electrodes of a pair was 1 mm. The three electrode pairs were arranged linearly in such a way that the distance from the mid locus to each of the edge loci was 6 cm. The matrix was applied on the dorsal part of the left forearm, the distal electrode pair being placed about 3 cm from the wrist.

As there is an interference between stimulus intensity and the parameters that constitute the rabbit phenomenon - temporal interval and distance between loci - (Cholewiak, 1976), each test was preceded by a calibration procedure with the distal electrode as reference point. After the absolute threshold and the threshold of annoyance had been determined the intensity of the referent stimulus was set to the absolute threshold plus 60% of the difference to the threshold of annoyance. Then the subjects received the reference stimulus in alternation with one of the two proximal stimuli in

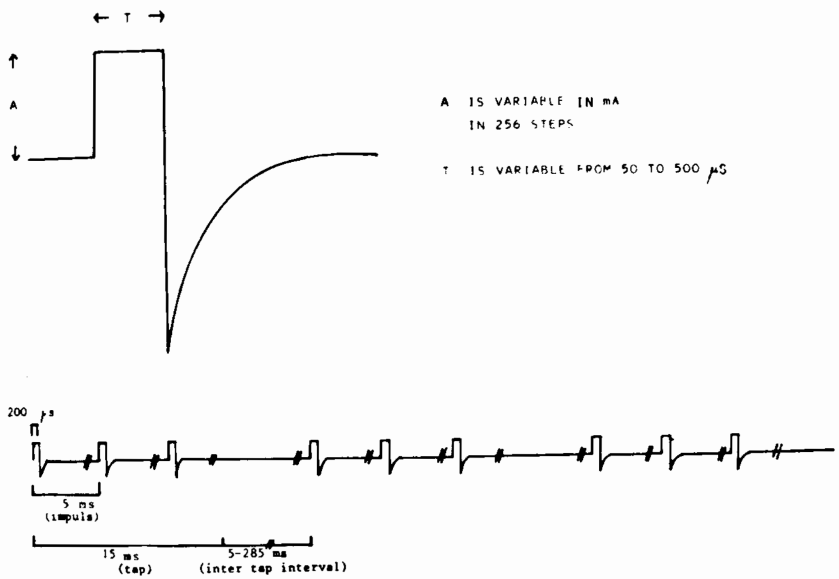


Fig. 1. Form of the stimuli used.

order to match the magnitude at these skin loci to that of the reference locus. It should be noted that the calibration was done with the taps used in the experiment that followed immediately. The stimuli described above were presented (in randomized order) in an identification test with an interstimulus interval of 5 s, in a one-step discrimination test with an interstimulus interval of 5 s between the pairs, and (in several cases) in a two-step discrimination test under the same conditions. In all tests each stimulus appeared three times.

In the case of identification the subjects were instructed to identify one of four categories: (i) 'continuous movement' (no single taps are perceived), (ii) 'rhythmically structured rabbit' (taps are perceived at more than 3 loci between the edge stimuli in a sequence having an intrinsic rhythmic structure), (iii) 'constant rabbit' (taps are perceived at more than 3 loci but separated by equal intervals and distances), (iv) 'single taps' (more than 1 tap is perceived at the same locus and only 3 loci are perceived at all). Additionally, the subjects were informed that not all of the categories had to appear in the same test. Before the test was executed, the stimuli were presented once in randomized order. The complete series of tests was carried out first using continuum I and then continuum II.

3. Results and Discussion

Fig. 2 and Fig. 3 show the results of identification for continua I and II and the respective one-step discrimination results. Both stimulus forms (constant duration vs. constant starting and ending points) are mixed in order to prohibit a significant evaluation according to those critical parameters. Fig. 4 and Fig. 5 show the same identification values, but categories (ii) and (iii)

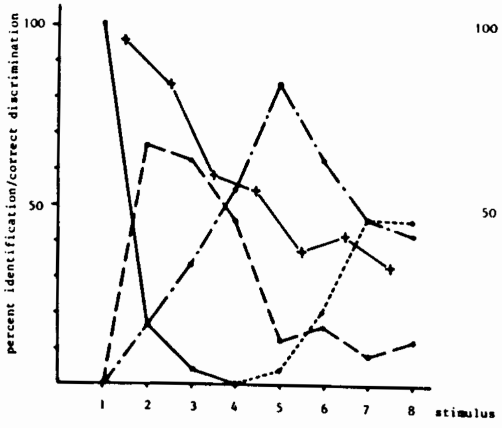


Fig. 2. Results for continuum I; ———: category (i), — —: category (ii), —·—: category (iii), ---: category (iv); *———* discrimination.

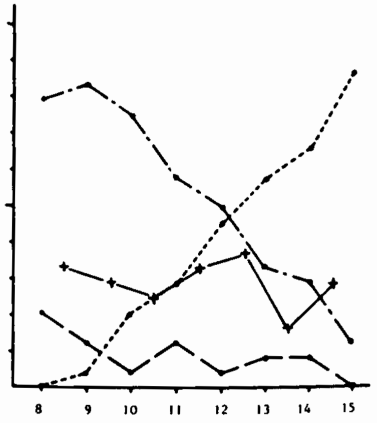


Fig. 3. Results for continuum II.

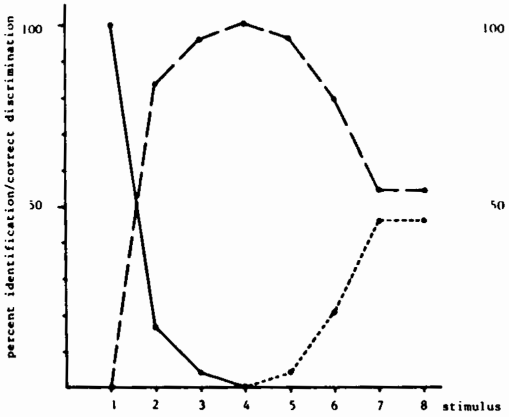


Fig. 4. Results for continuum I; ———: category (i) — —: categories (ii) and (iii), ---: category (iv).

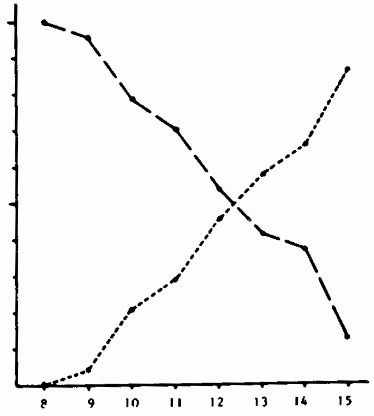


Fig. 5. Results for continuum II.

are added together to form one category 'rabbit.' Only the category switch from 'apparent movement' to 'rabbit' at stimulus 1.66 (corresponding to an interval of 18.33 ms) shows steep boundaries with an adequate discrimination peak (95.83%). Discrimination decays continuously and reaches chance without being increased at category switches. Even two-step discrimination was not significantly better and is therefore omitted from the results. Concerning stimulus 8 which was part both of continuum I and II an important uncertainty is revealed. Whereas in continuum I categories (iii) and (iv) reach approximately 50%, category (iii) has a starting value of 79.17% in continuum II, and even category (ii) is present at chance level throughout the continuum. The most general conclusion to be drawn from the results is that the responses (i) - (iv) accumulate in certain sections of the continuum. A natural case of categorical perception only exists between 'apparent movement' and 'rabbits.' The identification of rhythmically structured and constant rabbits is much poorer. On the other hand, this additional category switch (cf. Fig. 2) could be established without any training phase simply by once exposing the subjects to the stimuli and by giving a phenomenal description of the categories. A 100% peak, of course, can only be reached if both categories are taken together (Figs. 4 and 5). In the context of the two continua, stimulus 8 is judged quite differently. This indicates relative uncertainty in category assignment which could probably be reduced by a learning phase: after having been presented in continuum I, stimulus 8 received 100% identification in continuum II. Slow taps show a much less clear category switch and therefore seem to be less adequate for the establishment of categorical perception.

References

Cholewiak, R.W. (1976). Satiation in Cutaneous Saltation. *Sensory Processes* **1**, 163-175.

Geldard, F.A., Sherrick, C.E. (1972). The Cutaneous 'Rabbit'. A Perceptual Illusion. *Science* **178**, 178-179.

Kirman, J.H. (1973). Tactile Communication of Speech: A Review and an Analysis. *Psychological Bulletin* **80**, 54-74.

Repp, B.H. (1982).Categorical Perception: Issues, Methods, Findings. Haskins Lab.: *Status Report on Speech Research SR-70*, 99-183.

Sherrick, C.E., Rogers, R. (1966). Apparent Haptic Movement. *Percept. Psychophys.* **1**, 175-180.

Tillmann, H.G. and Piroth, H.G. (1983). Das System zur elektrischen Hautreizung SEHR-1. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)* **17** (in press).

Dyslexia and Developmental Dysphasia; a Deficit in Processing Rapid Spectral Changes?

B. Pompino-Marschall and H.G. Tillmann
Munich, Federal Republic of Germany

1. Introduction

P. Tallal and coworkers have found that a group of developmental dysphasics suffer from a deficit in processing rapid spectral changes in the case of the formant transitions of stop-/a/-syllables, whereas synthetic 'stop'-/a/-syllables with unnaturally lengthened formant transitions with a duration of 82 ms were more readily discriminated than syllables with transitions of 30 ms (for review see Tallal and Piercy 1978). The same results were reported for dyslexics (Tallal 1980). In the framework of a German Research Council project on speech pathology we planned to develop a diagnostic test by exploiting this effect with a variety of phonetically controlled stimuli. As we failed to reproduce this effect in preliminary, rather simple discrimination tasks with manipulated naturally produced material and with synthetic material where normal and dyslexic subjects rather showed a tendency to react worse in the case of lengthened transitions, we devised a new testing procedure in order to be able to get a closer view of the resulting effects.

2. Stimuli and Procedure

Our stimuli consisted of two seven-step continua between /abam/ and /adam/ with 40 ms or 70 ms CV-transitions (cf. Fig. 1), synthesized with a program based on Klatt (1980). The critical CV-syllable was embedded in the context /a - m/ to control e.g. speech rate effects.

Formally our procedure resembles a paradigm that has been applied in experiments with young children (Wilson, 1978). A series of background stimuli is presented with a critical target stimulus randomly embedded. The subject has to react on the occurrence of the target stimulus. The background stimulus is one of the endpoints of an acoustically defined continuum, as originally used in categorical perception experiments, and the target is moved from the other endpoint of the continuum in the direction of the background. The discrimination threshold is determined by the target stimuli that produce no reactions. We wrote a computer program to run the following test. The subjects had to react by pressing a button on the occurrence of a target stimulus which was one in a sequence of five stimulus presentations. The normal interstimulus interval was 750 ms, the pause

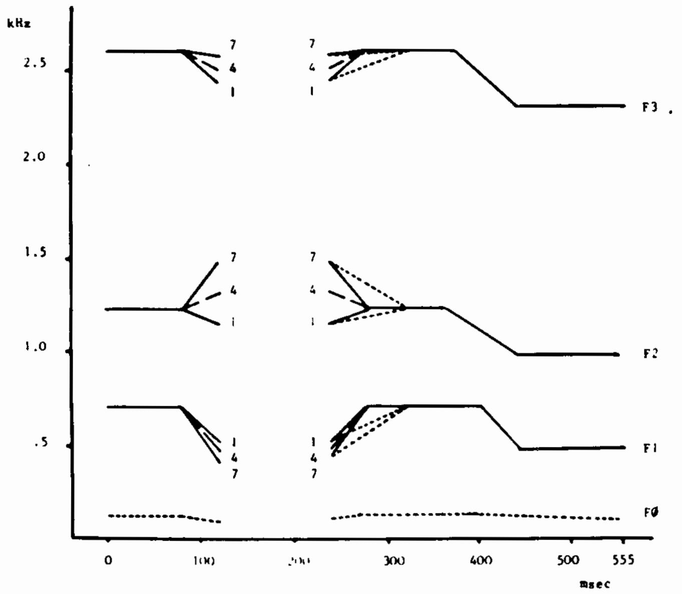


Fig. 1. Schematic sonagrams of the endpoint (and midpoint) stimuli used (1: /abam/, 2: /adam/; unbroken line: 40 ms transitions, dotted line: 70 ms transitions).

between the sequences 4 s. The first part of the experiment was the training phase in which the subject had to learn to discriminate the endpoint target from the complementary background. To indicate the target stimulus additionally a small lamp within the button was switched on and every correct answer was rewarded by playing a short melody. When in the second phase (without supplementary signals) the subject had reacted on the occurrence of the target stimulus three times consecutively, the actual test was started. Now, after a correct response, the target moved one step in the continuum towards the background. As soon as there was a missing reaction the target was moved backwards two steps, and so on. After five missing reactions the test was finished. Besides the discrimination treshold, the reaction times were also automatically measured. This testing was done through the two continua in both directions. The category boundary is defined by the midpoint between the two resulting thresholds (cf. Fig. 2). Eleven developmental dysphasics and eight dyslexics all aged between seven to thirteen took part in the experiments.

3. Results and discussion

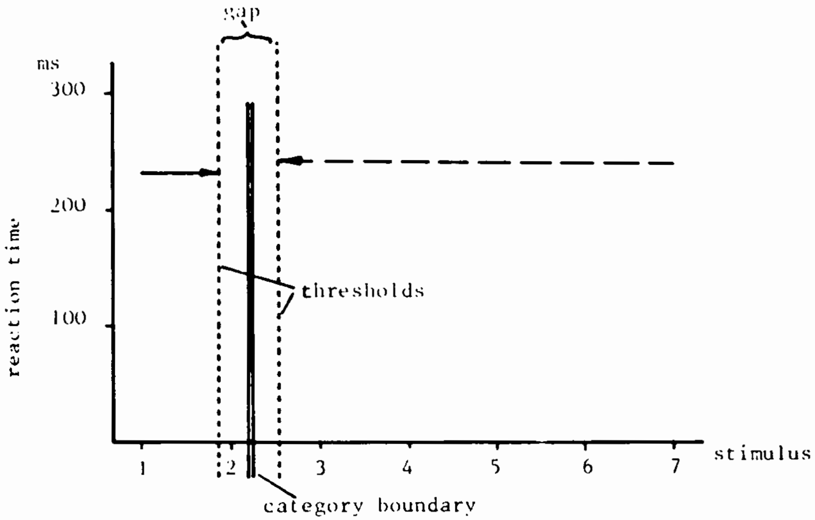


Fig. 2. Schematic representation of the measurements taken; → test with stimulus 7 as background, — → test with stimulus 1 as background.

Table I. Results of the discrimination threshold test

	Development dysphasics		Dyslexics	
	short transitions	long	short transitions	long
Boundary	3.24	2.9	3.23	3.15
Gap	1.62	1.04	-0.15	0.75
Reaction time	254	365	216	231

The results are shown in Table I. In accordance with the results of our preliminary tests we can see a tendency towards worse discrimination in the test with lengthened CV-transitions by the group of dyslexics as shown by a small gap between the thresholds (in contrast to a small overlap in the other test) and also reflected in the reaction time. There is no clear picture for the group of developmental dysphasics. Statistical analysis of the results revealed that there were no significant differences between the tests with short vs long transitions for either group with respect to reaction time, phoneme boundary and gap or overlap at the thresholds (partly due to the great inter- and intrapersonal variability in the responses). Nor were there any significant group differences with respect to phoneme boundary and gap or overlap in either test. Reaction times also showed no significant group differences in the test with short CV-transitions. However, contrary to the predictions of the Tallal hypothesis in the long transition continuum the developmental dysphasics were significantly slower than the dyslexics. We therefore conclude that the possible perceptual deficit in developmental dysphasics is not as simple as proposed by P. Tallal and coworkers.

Acknowledgements

This investigation was supported by the German research council (Grant TI 69/17).

We would like to thank D.B. Pisoni for making the program KLTEX available to us (supported by NIH Grant NS 12179). We are further indebted to the Max-Planck-Institut für Psychiatrie, Munich and the Arbeitskreis Legasthenie, Munich for making the subjects available.

References

Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am.* **67**, 971-995.

Tallal, P. (1980). Auditory temporal perception, phonics and reading disabilities in children. *Brain Lang.* **9**, 182-198.

Tallal, P. and Piercy, M. (1978). Defects of auditory perception in children with developmental dysphasia. In: Wyke, M. (ed.) *Developmental Dysphasia* (London etc.) 63-84.

Wilson, W.R. (1978). Behavioral assessment of auditory functions in infants. In: Minifie, F. and Lloyd, L. (eds.) *Communicative and Cognitive Abilities: Early Behavioral Assessment.* Baltimore.

Perception of English Contrastive Stress by Brain-damaged Adults

J. R. Solomon
Hayward, U.S.A.

1. Introduction

Like segmental features of language, prosody can be investigated in terms of both production and perception. With respect to the former, the literature is rather rich, consisting of a variety of research studies as well as clinical and other observations of both children and adults, and from a number of different languages. From such studies we gain the impression that prosodic features are early acquired by children and relatively resistant to loss in adults who have suffered brain damage, including people with fairly severe aphasia.

The literature on the perception of prosody is considerably smaller, particularly with reference to aphasia. There are only six studies in which the ability of aphasic subjects to comprehend prosodic information has been investigated - one of phonemic pitch accent in Japanese (Sasanuma et al. 1976), one of lexical tone in Thai (Gandour and Dardarananda 1983), one of correct vs. incorrect placement of syllabic stress in Rumanian (Mihăilescu et al. 1970), and four in English, including Blumstein and Goodglass' 1972 study of the perception of syllabic stress as a means of making syntactic distinctions between noun-verb pairs like *'transport* and *trans'port*, and semantic-syntactic distinctions between noun-noun phrase pairs like *'yellowjacket* and *yellow ' jacket*. More recently Baum and her colleagues (1982) found that compared to normal subjects, Broca's aphasics were unable to make distinctions between phonemically similar sentences on the basis of either sentential stress or juncture, or to profit from increased stress on functors in sentences.

The prosodic feature investigated in the current study is English contrastive stress. What contrastive stress is has been a vexed question (Schmerling, 1976); it is used here to mean those variations in pitch, loudness and length that have the effect of highlighting or emphasizing differences in meaning between phrases or sentences.

2. Materials and Methods

2.1. Stimuli

The stimuli for this study were two series of commands based on those found in the *Token Test* (De Renzi and Vignolo 1962), a widely used test of auditory comprehension in aphasia that holds extralinguistic or contextual cues to a minimum. The test uses a limited repertoire of five colors (blue, green, yellow, white and red), two shapes (circle and square), and two sizes (large and small). There are five parts to the test; in the current study, only Parts I and IV were chosen as models for the stimuli, Part I because it presents items of a level of complexity that would allow subjects with even severe impairments of auditory processing to perform at least some items correctly, and Part IV because it is challenging enough to be likely to elicit errors from subjects with relatively mild impairments.

The stimuli based on Part I of the *Token Test* had the following form:

I-1. Touch the red circle.
I-2. Touch the red *square.*
I-3. Touch the *blue* square.
I-4. Touch the *yellow circle.*
I-5. Touch the *green square.*

These commands were presented to subjects verbally in separate sets with two distinct readings. One of the readings used contrastive stress, emphasizing the differences in content between each succeeding sentence, as suggested by the italics in the examples above. There were 16 of these stimuli in this part of the study (the first command, because it contrasts with nothing, was a dummy item and was unscored). Precisely the same set of 16 stimuli was used for the other condition; with this reading (the Control condition), a uniform rising intonation was used at the end of each command, ignoring the inherent contrasts between commands.

The commands based on Part IV of the *Token Test* had the following shape:

IV-1. Touch the small yellow circle
and the small *blue* circle.
IV-2. Touch the large blue square
and the *small red* square.
IV-3. Touch the small red circle
and the *large white square.*

When each of the 15 items in this part of the study was presented in the Contrastive Stress condition, the first object noun phrase in each sentence was read with the same uniform rising intonation used at the end of the Control condition reading of the Part I-type stimuli. In the second NP, the words that contrast with words occupying the same syntactic position in the first NP were read with contrastive stress, as suggested by the italics in the examples above. Thus the domain of contrastive stress in the Part IV-type stimuli was the command itself (the second NP with reference to the first), whereas with the Part I-type stimuli it was the immediately preceding stimulus. When presented in the Control condition, the Part IV-type stimuli were read with a uniform rising intonation at the end of each object NP.

All the commands were presented by tape recording, including a screening test to determine whether the subject's auditory comprehension and visual and motor abilities were sufficient to identify separately each of the nine target words (the colors, shapes and sizes) used in the test and to perform the task in general; this also served the function of finding a comfortable listening level for each subject. Half of the subjects heard the sets of stimuli with the Contrastive Stress condition first (first the 16 Part I-, then the 15 Part IV-type stimuli); then they heard the identical stimuli read in the Control condition. The other half of the subjects heard them in the order Control, then Contrastive Stress.

2.2. Subjects

There were 42 subjects in the study, all of whom had had a single cerebrovascular accident of either the left or the right hemisphere. All of the subjects with left hemisphere lesions had received an independent diagnosis of aphasia, usually from a speech-language pathologist. Subjects with right hemisphere lesions were determined to be not aphasic, on the basis of their performance on the *Boston Diagnostic Aphasia Examination* (Goodglass and Kaplan 1972), which was part of the testing protocol for this study. None of the subjects had a history of alcohol abuse, mental retardation, senile or presenile dementia, psychiatric problems or significant hearing loss. All were right-handed, and all were native speakers of American English. There were 15 subjects with single lesions of the right hemisphere and 27 with single lesions of the left hemisphere. All subjects were neurologically stable at the time of testing, and all were a minimum of one month post-onset. Additional information on time post-onset, as well as other demographic information, is presented in Table I.

Table I. Demographic information

		Aphasics (lefts)	Non-aphasics (rights)
Number:		27	15
Sex:	Male:	18	9
	Female:	9	6
Age	Mean:	57.7	55.3
in years:	Range:	38–75	32–72
Education:	Mean:	13.2	12.9
in years:	Range:	8–20	9.5–19
Months post-onset:	Mean:	39.1	37.9
	Range:	1–217	6–120

3. Results and Discussion

Contrastive stress appears to make a difference in the ability of aphasic patients to comprehend relatively short verbal commands, such as those in Part I of the *Token Test*. In this study the 27 left hemisphere damaged (i.e. aphasic) subjects demonstrated a statistically significant difference in performance with the Part I-type stimuli (as measured by a system of weighted scoring devised for the study and expressed in Table II in terms of percent correct); this difference favored the Contrastive Stress condition over the Control condition ($p < .05$). The right hemisphere damaged (i.e. non-aphasic) subjects performed the Part I-type stimuli perfectly (or nearly so) in both conditions.

For the longer, Part IV-type stimuli, there were no significant differences in performance between the Contrastive Stress and Control conditions for either of the subject groups. The right hemisphere damaged subjects again performed essentially like normal subjects, making very few mistakes, randomly scattered. The aphasic subjects also performed with virtually the same level of correctness in the two conditions, but with more errors than the non-aphasic subjects.

One reason for this outcome may be that this task in itself (i.e. in terms of the information content of the segmental phonemes alone) was too difficult for the aphasic subjects to allow the difference in prosodic styles evident and significant with the Part I-type stimuli, to emerge. Evidence for this may be found in their lower mean level of correctness in both testing conditions for the longer vs. the shorter stimuli; both of these differences (84.8% vs. 89.8% for the Control condition, and 84.7% vs. 92.5% for the Contrastive Stress condition) were statistically significant.

Table II. Results: Weighted scores (percent correct)

		Aphasics (lefts)		Non-aphasics (rights)	
Stimulus type	Condition	Mean	S.D.	Mean	S.D.
Part I	Control	89.8	16.1	100	0.0
		**		NS	
	Stress*	92.5	12.0	99.8	1.0
Part IV	Control	84.8	15.3	99.6	0.8
		NS		NS	
	Stress*	84.7	14.6	99.7	1.0

* Contrastive stress.

** $p = .037$.

4. Conclusions

The implications of these findings seem to be that when given an auditory language processing task that is manageable, such as stimuli like those in Part I of the *Token Test*, aphasic individuals perform that task significantly better in the presence of contrastive linguistic stress than with a neutral reading of the same material. When the task is more difficult, as in Part IV of the *Token Test*, the advantage conferred by contrastive stress disappears; in fact, if the listener's processing capacities are already overburdened by segmental information, the addition of suprasegmental information may result in poorer rather than better performance.

Darley and his colleagues (1975:6) have defined prosody generally as 'all the variations in time, pitch, and loudness that accomplish emphasis, lend interest to speech, and characterize individual and dialectal modes of expression.' This definition suggests a peripheral role to prosody, implying that if prosodic features were not present in a message, there would be little or no effect (in English, at least) on the understanding of its semantic content. The findings of this study, on the other hand, suggest that prosodic features such as contrastive stress convey not only a speaker's affect and attitude but also part of the 'core' of this message, and that aphasic individuals (even when severely impaired) retain the ability to comprehend the linguistic as well as the paralinguistic information that prosody conveys.

References

Baum, Shari R., Joanne Kelsch Daniloff, Raymond Daniloff, and Jeffrey Lewis (1982). Sentence comprehension by Broca's aphasics: effects of some suprasegmental variables. *Brain and Language* **17**, 261-71.

Blumstein, Sheila, and Harold Goodglass (1972). The perception of stress as a semantic cue in aphasia. *Journal of Speech and Hearing Research* **15**, 800-806.

Darley, Frederic L., Arnold E. Aronson, and Joe R. Brown (1975). *Motor speech disorders.* Philadelphia: W.B. Saunders.

De Renzi, Ennio, and Luigi A. Vignolo (1962). The Token Test: a sensitive test to detect receptive disturbances in aphasics. *Brain* **85**, 665-78.

Gandour, Jack, and Rochana Dardarananda (1983). Identification of tonal contrasts in Thai aphasic patients. *Brain and Language* **18**, 98-114.

Goodglass, Harold, and Edith Kaplan (1972). *The assessment of aphasia and related disorders.* Philadelphia: Lea and Febiger.

Mihăilescu, Lucretia, M.I. Botez, and A. Kreindler (1970). Decoding of correct and wrong word stress in aphasic patients. *Rev. Roumaine Neurol.* **7**, 65-74.

Sasanuma, Sumiko, Itaru F. Tatsumi, and Hiroya Fujisaki (1976). Discrimination of phonemes and word accent types in Japanese aphasic patients. *Proceedings of the XVIth International Congress of Logopedics and Phoniatrics*, Interlaken, 1974, ed. by E. Loebell, 403-08. Basel: Karger.

Schmerling, Susan F. (1976). *Aspects of English sentence stress.* Austin: University of Texas Press.

Speech Rate and its Differential Influence on the Perception of Normally Hearing and Hearing-impaired Subjects

M. Vingerling
Utrecht, the Netherlands

1. Introduction

Prosody has been firmly integrated into phonetic research. Speech rate is an important prosodic cue, which has not been extensively investigated especially as far as the influence on intelligibility is concerned. As regards normally hearing listeners, it is generally accepted that as the rate of speech is increased beyond 'normal' limits, articulation deteriorates and thus affects intelligibility. Besides if speech is produced at a very high rate, it cannot be processed as fast as it is received, which results in a loss of speech information. As to hearing-impaired listeners, it is often assumed that slowing down the speech rate will improve intelligibility and it is common usage in clinical practice to speak at a significantly slower than normal rate when communicating with the hearing-impaired. The aim of the experiment to be reported on here was to investigate whether speech rate is of comparable influence on the speech decoding process of normally hearing and hearing-impaired listeners.

2. Method

The stimulus material used in this experiment consisted of 25 sentences, selected from the corpus as described by Plomp and Mimpen (1979). With regard to the arguments in favour of using sentences instead of words, the reader is referred to Gil-Günzburger and Vingerling (1981). The sentences were read out by two native speakers of Dutch: one male and one female, who had been instructed to speak at a rate they considered to be normal. The rate at which the sentences had been spoken was taken as a starting-point for further processing. By means of LPC the sentences were artificially expanded to 156% of the original duration (slow rate), and compressed to 64% of the original duration (fast rate). The resulting 50 sentences were offered to the original speakers with the instruction to repeat them trying to keep as closely as possible to the rate of speech of the sentence presented. In order to prevent ceiling effects in the case of the normally hearing listeners, the stimulus material was masked with noise. The type of noise used was speech noise i.e. noise with a spectrum that corresponded with the average of the speech spectra of both speakers. Based on a pilot study the S/N ratio was set at -1 dB

for normally hearing listeners and +7 dB for hearing-impaired listeners aiming at a target level of approximately 50% correct responses for the normal rate.

In all 168 subjects participated in the test. The hearing-impaired subjects (40) were pupils from various secondary schools for hearing-impaired children, their age ranged from 10 to 17 years and they all suffered from a congenital hearing loss of maximally 60 dB for the best ear. The normally hearing subjects (128) were pupils from the fifth and the sixth forms of various primary schools, their age ranged from 10 to 12 years and they had no self-reported hearing deficiencies. All subjects took the test individually. They received oral and written instructions and were asked to write down what they heard after every test sentence, even if this amounted only to fragments.

3. Analysis and Results

3.1. Hearing-impaired listeners

The results were analysed as to influence of speech rate and sex of the speaker. Table I gives the results per condition i.e. normal, slow and fast rate and per speaker for hearing-impaired listeners only.

This table shows that the differences in mean scores between the conditions normal and slow are very small in all cases i.e. per speaker and pooled for both speakers. Analyses of variance showed that the mean scores for the fast condition differed significantly from the mean scores for the normal and slow conditions ($p < .01$) with one exception: the difference in mean scores between the slow and the fast condition in the case of the female speaker proved to be insignificant. A post-hoc analysis (Scheffé) showed that the conditions normal and slow can be grouped together i.e. they do not differ significantly from each other in terms of mean scores (in all cases). The fast condition differed significantly from the normal and slow condition in the case of the male speaker and pooled for both speakers. In the case of the female speaker all conditions could be grouped together i.e. there was no significant difference in scores between the three conditions.

Table I. Mean correctly perceived words and standard deviations (in %) per condition, per speaker and pooled for male and female speaker (= total)

	Correct in % ♂		Correct in % ♀		Correct in % total	
Condition	$\bar{x}$	SD	$\bar{x}$	SD	$\bar{x}$	SD
Normal	59	38	64	30	62	33
Slow	63	35	59	30	61	29
fast	7	13	50	29	28	21

3.2. Normally hearing listeners

The results were analysed with regard to the influence of speech rate only. As to the influence of the sex of the speaker, the results of the normally hearing listeners will be compared with those obtained from the hearing-impaired listeners. Table II gives the results per condition i.e. normal, slow and fast rate, pooled for both speakers, for normally hearing listeners only.

Table II shows that the difference in mean scores between the conditions normal and slow is negligible whereas the mean scores for the fast condition appear to be considerably lower than those obtained in both the normal and the slow condition. Further analysis will be carried out and reported on later.

4. Conclusion

Test results show that the difference in scores (in mean % correctly perceived words) between the speech rates normal and slow are extremely small; this holds for both normally hearing and hearing-impaired listeners. Speaking at a fast rate appears to have dramatic consequences for both subject groups.

As regards normally hearing listeners the effect of speech rate on intelligibility has not been extensively investigated. Therefore, a more detailed analysis of the stimulus material used in this experiment seems warranted. This analysis might concentrate on the following two questions

1. What kind of factors play a role in the realisation of speech rate?
2. Are the factors responsible for the realisation of speech rate the same as those that are involved in the perception of speech rate and its influence on intelligibility?

As to hearing-impaired listeners, studies of the effect of time-compression and expansion of speech have been conducted in order to document the degenerative effects associated with various kinds of hearing loss and the process of aging (cf. Spitzer and Osborne, 1980). With regard to the influence of speech rate on intelligibility, it is commonly accepted to rely on mere assumptions. Seeing that the results of this experiment fail to support what is generally believed, viz. that slowing down the speech rate will improve intelligibility, further investigation into this matter seems feasible.

Table II. Mean correctly perceived words and standard deviations (in %) per condition, pooled for male and female speaker (= total)

	Correct in % total	
Condition	$\bar{x}$	SD
Normal	74	16
Slow	76	18
Fast	26	14

As regards the aim of this experiment, viz. to investigate whether speech rate is of comparable influence on the speech decoding process of normally hearing and hearing-impaired listeners, the results seem to indicate that there are no important differences.

References

Gil-Günzburger, D. and Vingerling, M. (1981). The influence of sentence accent on the perception of normally hearing and hearing-impaired listeners. *Progress Report Institute of Phonetics University of Utrecht* **6**(1): 3-23.

Plomp, R. and Mimpen, A.M. (1979). Improving the reliability of testing the speech-reception threshold for sentences. *Audiology* **18**(1): 43-52.

Spitzer, J.B. and Osborne, D.K. (1980). The effect of open- versus closed-set procedures in the perception of compressed speech. *Journal of Auditory Research* **20**(3): 169-172.

Section 22
Speech Errors

An Analysis of Pronunciation Mistakes and Second Language Strategy in the Case of Italian and Mandarin Speaking Learners of English

C. Fonda
Montreal, Canada

1. Introduction

Among the phonetic errors that a learner of English is bound to make, a faulty pronunciation of English obstruents is most conspicuous. What is not clear, however, is to what extent, if any, this may be attributed to a phonetic interference of the mother tongue. From a phonological point of view, English and Italian share the same stops although in English the range of the allophones is, in each instance, wider than in Italian. English differentiates its stops into aspirates and unaspirates, exploded and unexploded, glottals, taps and so on. Conversely, in Italian, stops, whether voiced or voiceless, do not offer such a gamut of phonetic variations. Indeed, their striking characteristic is that they are always exploded. This is the case because of Italian syllabic constraints which demand that all syllables be always stressed, although one syllable is more stressed than other ones. As a result, co-articulation, syllabicity, vowel coloring and the like, which play such an important role in English pronunciation, are of no consequence in Italian. Should we lend a careful ear to the anomalies of the peculiar accent of Italian speaking learners of English, we would quickly discover that there is an indiscriminate, universal use of exploded stops. This peculiarity is easily perceptible in the case of monosyllabic words ending in a stop as, for example, *big* and *spot* pronounced respectively /ˈbıh-gə/ and /ˈspɔh-tə/. Such shifts of course alter not only the phonetic quality of the stops involved but also the phonological structures of the words themselves.

This indiscriminate use of exploded stops is responsible for other phonetic changes, some of the most striking being the loss of aspiration of stops in syllable initial position, the absence of lateral and nasal plosion and the resulting loss of syllabicity in words such as *bottle* [ˈbɔtl], *kitten* [ˈkıtn], which the near-bilingual pronounces as [ˈbɔh-təl], and [ˈkıh-tən]. These peculiarities seem to suggest that the negative transfer that takes place in these instances does not operate at the level of the individual phoneme as, to mention one example, the replacement of aspirates with unaspirates, but it involves the entire phonological system. What appears to support this, is the reverse process whereby words which the near-bilingual borrows from English and incorporates into his everyday lexicon, undergo exactly the same process of phonetic adaption: plumber [ˈplom-bje-rə], garbage [ˈgae-rae-bıhčə]. Not without humour, Professor Clivio of York University has suggested that the use of such phonetically re-structured loan-words from English has brought into existence a new language which he called Italiese or Italish. Further evidence that the learner of English is transferring phonological constraints of his mother tongue rather than individual phonemes seems to be supported by the peculiar treatment of English intervocalic /s/ and dorso-velar /ŋ/. In Northern Italian dialects, voiceless /s/ in intervocalic position shifts into voiced /z/. In Southern dialects no shift occurs, while in the case of speakers who conform with the so-called Tuscan usage (Central Italy) the voicing of intervocalic /s/ is anything but regular or systematic. For example, /s/ is voiced in *Cosimo*, *rasente*, *presagio* but is voiceless in *cosi*, *rasoio*, *presa*. In English, unlike French where intervocalic /s/ is regularly voiced, the shift is not positional as witnessed by *pose* /powz/ and *result* /rɪzəlt/ where intervocalic /z/ coexists with intervocalic /s/ as in *case* /keɪs/ and *base* /beɪs/. As a result, Italian learners of English fail to display any consistency with respect to these English alternations. For example, my informants from Central Italy alternate the voicing of /s/ quite capriciously. Those of the North almost regularly voice intervocalic /s/ while those of the South seldom if ever do it. There is no doubt that this apparently erratic linguistic behaviour in my subjects which reflects the treatment of intervocalic /s/ in the Italian dialects of the North, Centre, and South, points to the presence of phonetic constraints which operate at the level of underlying phonemic structures.

English has three nasals: /n/, /m/, /ŋ/. Italian has also three nasals: /n/, /m/, and dorso-palatal /ñ/. In English, phonemes /n/ and /g/ do not cluster together in the same syllable. This accounts for the fact that the spelling in words such as *sing*, *singer* is nothing but a convention to represent phoneme /ŋ/. This spelling, of course, proves, in the case of the learners of English, extremely misleading in words such as, for example, *lingua*, *linguist* which, actually should be spelled ling-gua and ling-guist. All my informants pronounced mono-syllabic words such as *ring*, *sing*, as [rɪn] [sɪn] or [rɪng] [sɪng] but never as [rɪŋ] [sɪŋ]. However, words, with /ŋ/ in medial position such as *singer, lingua* were consistently pronounced as ['sɪn-gər], ['lin-gwə]. If we keep in mind that also in Italian /n/ and /g/ do not cluster together in the same syllable we can see that the negative transfer (NT) the learner is subject to is not simply a transfer from the mother tongue of individual phonemes which approximate the target sound but a more encompassing NT involving the underlying phonological structure of the Italian language.

2. Analysis

To test my assumption that phonetic NTs operate at the level of underlying phonemic structures and phonological constraints, I conducted some additional experiments using other languages. The following findings concern the extent of phonetic interference made by the first language (Mandarin) on the class of fricatives in the second language English. In column A of Table I are listed the English fricatives, in column B the corresponding sounds in Mandarin and in columns C, D, E the shifts, if any, undergone by each phoneme in word initial, medial and final position when pronounced by Mandarin speaking learners of English in the experiment.

We can see that fricatives which occur in both Mandarin and English, do not suffer any discernible shift. However, English fricatives which are non existent in Mandarin, show definite signs of phonetic interference except for the phoneme /v/ which is correctly pronounced in word initial and medial positions, probably because in both instances it occurs in a voiced environment.

When a shift occurs in word initial position, the same shift can be seen in word medial intervocalic position but not in word final position. This suggests that in these instances we are not dealing with NTs of individual phonemes whose use is then extended by overgeneralization buth with a negative transfer of much more overriding importance. Indeed, the inconsistencies of these shifts between final position and elsewhere can be easily accounted for if we take into account the fact that in the Mandarin morphological system a word or syllable is made up of an initial consonant, a final vowel sound, and the tone. For example,媽 (mother) can be transcribed as /mæ/ and中国(china) can be transcribed as /čuŋguo/, where each word has an initial consonant, a final vowel and a tone. Therefore, a Mandarin speaker

Table I Corresponding Mandarin phonemes according to environment

A English phonemes	B Mandarin phonemes	C Word initial position	D Word medial position	E Word final position
h	–	h	h	h
f	f	f	f	f
v	–	v	v	f
s	s	s	s	s
z	–	dʒ	dʒ	s
ʃ	ʃ	ʃ	ʃ	ʃ
ʒ	–	–	ʃ	tʃ
θ	–	t	t	f
ð	–	d or dʒ	d or dʒ	f

θ > t	word initial	think [tɪŋk]
	word medial	lethal [lɪtəl]
θ > f	word final	path [pæf]
ð > d or dʒ	word initial	that [daet] or [dʒæt]
ð > f	word final	bathe [beɪf]
v > f	word final	give [gɪf]
z > dʒ	word initial	zoo [dʒu]
	word medial	razor [reɪdʒə] and z > s word final like buzz [bəs]
ʒ > ʃ	word medial	illusion [ɪluʃən]
ʒ >tʃ	word final	garage [gərɑtʃ]

tends to divide English words into segments of a consonant followed by a vowel or a diphthong to make it coincide with the structure of Mandarin syllables. When this occurs, any change that takes place in word initial position is also repeated in word medial position. The different treatment of these phonemes in word final position simply conforms to the underlying morphological structure and phonetic constraints of Chinese, a language characterized by its large number of monosyllabic words.

Acknowledgement

For the experiment with Mandarin speaking learners of English, I am indebted to Mr. Jin Yang Tion, a student in my Honors Linguistics programme.

References

Breitenstein, P.M. (1978). The Application of Contrastive Linguistics. In: *English Language Teaching*, **33**(1): 21-26.

Clivio, G.P. (1978). 'When in Canada do as the Canadese do.' MacLean's, (February 6, 1978), p. 49.

Soudek, Leo I. (1977). Error Analysis and the System of English Consonants. In: *English Language Teaching*, **31**(2): 125-130.

List of Participants

Algeria

EMERIT, Ph.D., E.
Université d'Alger, Institut de Linguistique et de Phonétique, 20, Rue Chahid Khalef Mustapha, El-Blar

Argentina

GURLEKIAN, Dr. J.A.
Laboratorio de Investigaciones Sensoriales, Marcelo T. de Alvear 2202 - 4°p, 1453 Buenos Aires

Australia

ADAMS, Mrs. C.M.
University of Sydney, Dept. of Behavioural Sciences in Medicine, Sydney N.S.W. 2006

CLARK, Dr. J.E.
Macquarie University, School of English & Linguistics, North Ryde, Sydney 2113 New South Wales

DERMODY, P.
National Acoustic Laboratories, 5 Hickson Road, Sydney, N.S.W. 2000

MACKIE, Miss K.
National Acoustic Laboratories, 5 Hickson Road, Sydney, N.S.W. 2000

O'KANE, Mrs. Dr. M.J.
Canberra College of Advanced Education, School of Information Sciences, P.O. Box 1, Belconnen 2616

PALETHORPE, Mrs. S.
Macquarie University, Speech and Language Research Centre, North Ryde, 2113 Sydney

Austria

GRASSEGGER, Dr. H.
Universität Graz, Institut für Sprachwissenschaft, Mozartgasse 8, 8010 Graz

HOLZER, A.T.
Medizinisch-technische Schulen TIROL, Med.-Univ.-Klinik, Anichstrasse 35, 6020 Innsbruck

KOMMENDA, M.
Institut für Nachrichtentechnik, Technische Universität Wien, Gusshausstrasse 25, 1040 Wien

KUBIN, G.
Institut für Nachrichtentechnik, Gusshausstrasse 25/389, 1040 Wien

Belgium

COILE, B.M.J. VAN
Lab. voor Electronica en Meettechniek, R.U. Gent, St. Pietersnieuwstraat 41, Gent

COLLIER, Prof. Dr. R.
University of Antwerp, Prinsstraat 13, 2000 Antwerp

FRANCARD, Dr. M.
Université Catholique de Louvain, Collège Erasme, Place Blaise Pascal 1, 1348 Louvain-la-Neuve

HARMEGNIES, B.G.F.
Université De Mons, Service de Phonétique, Faculté de Psychopedagogie, Avenue du Champ de Mars, Chaville II, 7000 Mons

MERTENS, P.
Katholieke Universiteit Leuven, Afdeling Toegepaste Linguistiek, Faculteit Letteren & Wijsbegeerte, Blijde-Inkomststraat 21, 3000 Leuven

SCHOENTGEN, J.
Université Libre de Bruxelles, Institut de Phonétique, Avenue F.-D. Roosevelt 50, 1050 Bruxelles

VERLUYTEN, Dr. S.P.
Universiteit Antwerpen (UIA), Universiteitsplein 1, 2610 Wilrijk

WAJSKOP, Prof. M.
Université de Bruxelles, Institut de Phonétique, 50, Avenue F.-D. Roosevelt, C.P. 110, 1050 Bruxelles

WEI, V.
Free University Brussels

Brazil

ARAGAO, M.
Universidade Federal da Paraiba, Mestrado em letras, Conj. Humanistica, Bloco 5/VFPB, 58000 Paraiba

EVERETT, Dr. D.L.
State University of Campinas/SIL, Department de Linguistica - IEL, Unicamp, Campinas 13.100

MORAES, J.A.

SHOCKEY, Mrs. Prof. L.R.
Universidade Federal da Paraiba, CCHLA-DLEM-BLOCO 4, Cid. Univ., João Pessoa 58.000

WEISS, Ms. H.E.
Summer Institute of Linguistics, Caixa Postal 14-2221, 70349 Brasilia - DF

Canada

ESLING, Dr. J.H.
University of Victoria, Department of Linguistics, Victoria, B.C. V8W 2Y2

FONDA, Prof. C.
Concordia University, 7141 Sherbrooks Street West, Montreal H4B 1R6

GELINAS-CHEBAT, Prof. C.
Université de Québec à Montréal, C.P. 8888 Succ 'A', Montréal H3C 3P8

JACQUES, Prof. B.
Université du Québec à Montréal, C.P. 8888 Succ 'A', Montréal H3C 3P8

LAVOIE, G.
Université Laval, Département de Langues et Phonétique, Québec

MARTIN, Prof. P.J.
University of Toronto, Experimental Phonetics, 39 Queen's Park CR.E., M5S 1A1 Toronto/Ont.

MORIN, Prof. Y.Ch.
Université de Montréal, Département de Linguistique, Montréal H3C 3J7

OSTRY, Dr. D.J.
McGill University, 1205 Dr. Penfield, Montreal H3A 1B1

RIGAULT, Prof. A.
McGill University, Department of Linguistics, 1001 Sherbrooke West, Montreal H3A 1G5

SANTERRE, Prof. L.
Université de Montréal, 3200 Rue Jean-Brillant, Montréal/Québec

SAUNDERS, S.R.
Bell-Northern Research, 3 Place du Commerce, Nuns' Island, Verdun/Quebec H3E 1H6

SCHREY, Drs. E.F.
Ongwanada Hospital, 117 Park Street, Kingston/Ont. K7L 1J9

WARKENTYNE, M.J.
University of Victoria, Department of Linguistics, P.O. Box 1700, Victoria B.C. V8W 2Y2

Chile

BIANCHI, Ms. M.R.
University of Concepción, Department of Foreign Languages, P.O. Box 82-C, Concepción

MAGANA, Ms. J
University of Concepción, Department of Foreign Languages, P.O. Box 82-C, Concepción

China, People's Republic of

GUI, Prof. C.
Guangzhou Institute of Foreign Languages, English Department, Guangzhou

MAO-CAN, Lin
Chinese Academy of Social Sciences, Institute of Linguistics, 29, Xue Yuan Lu, Beijing

KAI-WAN, Z.
Sichuan University, Foreign Languages Department, Chengdu, Sichuan

ZONG-JI, WU
Chinese Academy of Social Sciences, Institute of Linguistics, 29, Xue Yuan Lu, Beijing

Czechoslovakia

DOHALSKA-ZICHOVA, Mrs. Dr. M.
Charles University, Institute of Phonetics, Nám. Krasnoarmejcu 2, 11000 Prague 1

JANOTA, Dr. P.
Charles University, Institute of Phonetics, Krasnoarmejcu 1, 11638 Prague 1

JAROSLAVA, Dr. P.
University of Brno, Dept. of Linguistics and Phonetics, Arne Nováka 1, 60200 Brno

JASOVÁ, Mrs. Dr. E.
Pädagogische Fakultät, Tajovského 40, Banská Bystrica

KULLOVÁ, Ph.Dr. J.
Charles University, Fac. of Philosophy, Hybernská 3, 11000 Prague - 1 - Nové Mesto

PALKOVÁ, Ms. Z., Ph.D.
Charles University, Dept. of Linguistics and Phonetics, Faculty of Philosophy, Krasnoarmejcu 2, 11638 Prague

VACHEK, Dr. J.
Nam. Jirino 2, Podebrad 18, Prague 3

Denmark

ANDERSEN, Prof. H.
University of Copenhagen, Njalsgade 96, 2300 Copenhagen
BASBOLL, Prof. H.
Odense Universitet, Nordisk Institut, Campusvej 55, 5230 Odense M
CHRISTENSEN, J.B.
University of Copenhagen, Dept. of English, 84, Njalsgade, 2300 Copenhagen S.
FISCHER-JORGENSEN, Prof. E.
Kongestien 45, 2830 Virum
GREGERSEN, Ms. K.
Odense University, Rasmus Rask Institute of Linguistics, Campusvej 55, 5230 Odense M
HERSLUND, M.
Københavns Universitet, Njalsgade 78-80, 2300 København S
HOLM, C.
p/o. Christiansholmes Tvaervej 23, 2930 Klampenburg
HUTTERS, Mrs. B.
Institut for Fonetik, Njalsgade 96, 2300 Copenhagen S
JORGENSEN, J.N.
Royal Danish School of Educational Study. Endruphborg, Department of Danish, 2400 Copenhagen NV
PEDERSEN, M.F.
Health Service of Copenhagen Schools, Sjaelør Boulevard 135, Copenhagen Valby 2500
THORSEN, O.
University of Copenhagen, Institute of Phonetics, Njalsgade 96, 2300 Copenhagen
THORSEN, Mrs. N
University of Copenhagen, Institute of Phonetics, Njalsgade 96, 2300 Copenhagen

Finland

AULANKO, R.T.
University of Tampere/Phonetics, P.O. Box 607, 33101 Tampere 10
DUFVA, M.I.
The University of Jyväskyla, Theoepolku bB 7, 40720 Jyväskyla
HEIKKINEN, Mrs. H.
University of Jyväskylä, Department of Phonetics, Seminaarinkatu 15, 40100 Jyväskylä
HURME, P.
University of Jyväskylä/Phonetics, 40100 Jyväskylä
KARJALAINEN, M.A.
Helsinki University of Technology, Acoustic Laboratorium, Otakaari 5 A, 02150 Espoo 15
LAHTI, Mrs. L.L.
University of Oulu, Institute of Nordic Languages, Kasarmintie 8, Snellmania, 90100 Oulu 10
LAINE, U.K.
Helsinki University of Technology, Acoustic Laboratorium, Otakaari 5 A, 02150 Espoo 15
LEHTONEN, J.
University of Jyväskylä, Department of Phonetics and Linguistics, 40100 Jyväskylä
IIVONEN, Prof. A.K.
University of Helsinki, Department of Phonetics, Vironkatu 1, 00170 Helsinki 17
NEVALAINEN, Ms. T.A.
University of Helsinki, English Department, Hallituskatu 11-13, 00100 Helsinki 10
NIEMI, Mrs. E.S.
University of Joensuu, Department of Swedish, 80101 Joensuu 10
NIEMI, E.J.
University of Joensuu, Department of Phonetics & General Linguistics, 80101 Joensuu

SOVIJÄRVI, Prof. A.A.I.
Helsinki University, Department of Phonetics, Liisankatu 16 D, 00170 Helsinki 17

France

ABRY, C.
Université III, Institut de Phonétique, B.P. 25, 38040 Grenoble Cedex
ALLIERES, J.
Université de Toulouse, 109 rue Vauquelin, Toulouse 31058 Cedex
AULOGE, J.Y.
Ecole Centrale de Lyon, 36 Avenue Guy de Collongue, 59130 Ecully
BACRI, Mrs. N.
Centre d'Etude des Processus cognitifs, Ehess-CNRS, 54 Boulevard Raspail, 25006 Paris
BOE, L.J.
University III, Institut de Phonétique, 38400 Saint Martin d'Heres
BOULAKIA, G.
Université Paris 7, Dpt. de Recherches Linguistiques, Laboratoire de phonétique de DRL. (U.P7.), 10 Rue Charles V, 75004 Paris
BOYSSON-BARDIES, Ms. B. de
C.N.R.S., Laboratoire de Psychologie, 54 bd Raspail, 75006 Paris
CARRE, R.
E.N.S. d'Electronique, 23 rue des Martyrs, 38031 Grenoble
CARTON, Prof. Dr. F.
Université de Nancy, B.P. 3397, 54015 Nancy
CHAFCOULOFF, M.
C.N.R.S., Université de Provence, Institut de Phonétique, 29 Av. R. Schuman, 13621 Aix en Provence
CHEVRIE, C.
Institut National de la Santé et de la Recherche Médicale, Laboratoire du Pr. Scherrer, Inserm U3, 47, Bd. de l'Hôpital, 75651 Paris Cedex 13
DELGUTTE, Dr. B.
Centre National d'Etude des Telecommunications, Route de Trégastel, B.P. 40, 22301 Lannion
DELYFER, Mrs. M.T.
Université de Bordeaux, Gironde
DESTOMBES, F.
IBM France, 36 Avenue Raymond Poincaré. 76116 Paris
DUEZ, Mrs. D.
Université de Provence, Institut de Phonétique, 29, Avenue R. Schumann, 13621 Aix Provence
DURAND, P.
Université de Provence, Institut de Phonétique, 29, Avenue R. Schumann, 13621 Aix en Provence
ESKENAZI, Ms. M.
L.I.M.S.I./CNRS, B.P. 30, 91409 Orsay
GOUDAILLIER, J.P.
Université René Descartes, Laboratoire de Phonétique, 104, Quai Clichy, 92110 Paris
GSELL, Prof. Dr. R.
Université de la Sorbonne Nouvelle, Institut de Phonétique, 19 Rue des Bernardins, 78005 Paris
HIND, A.
29 rue du Transvaal, Paris 20
HOMBERT, J.M.
Université de Lyon II, CRLS - Avenue de l'Université, 69500 Bron
KONOPCZYNSKI, Mrs. G.
Université de Besançon, Laboratoire de Phonétique, 25030 Besançon

LAGADEC, Ms. M.J.
Université de Lille III, B.P. 149, 59653 Villeneuve d'Ascq

LEFÈVRE, J.P.
CIT-Alcatel, Speech Lab. Processing, 10, bis rue Louis Lormand, 78310 La Verrière

LOCKWOOD, P.
C G E - Laboratoires de Marcoussis, Dpt. DEO-INF, Marcoussis 91460, Route de Nozay

LONCHAMP, F.
Université de Nancy II, Institut de Phonétique, B.P. 33-97, 54000 Nancy

MAEDA, S.
C.N.E.T., Dept. R.C.P., B.P. 40, 22301 Lannion

MARCHAL, Prof. A.
Institut de Phonétique, 79 Ave. R. Schuman, 13621 Aix en Provence

METTAS, Mrs. Dr. O.
CNRS, 32 Rue Charden-Lagache, 75016 Paris

MILLOT, Ms. C.
Université de Besançon, Laboratoire de Phonétique, UER Lettres, 30 Rue Jégevand, 25000 Besançon

MURILLO, G.A.
Institut de la Communication Parlée, 23, Rue des Martyrs, 38031 Grenoble Cedex

PERRIER, P.
Institut Communication Parlée, Grenoble France, 23 rue des Martyrs, 38031 Grenoble Cedex

PUECH, G.
Université Lyon II, CRLS - Avenue de l'Université, 69500 Bron

RIALLAND, Mrs. A.
C.N.R.S. Era 433 Paris, 19 rue de Bernardines, 75005 Paris

ROSSI, M.
Institut de Phonétique, Aix en Provence

ROSTOLLAND, Ing. D.
Laboratoire Physiologie Neurosensorielle, 15 rue de l'Ecole de Médecine, Paris 75270, Cedex 06

RYBAK, B.
Institut de Linguistique, 19 rue de Bernardines, Paris 75005

SAGART, L.
C.N.R.S., CRLAO, 54 Bld. Raspail, 75006 Paris

SCOTTO DI CARLO, Mrs. M.
Université de Provence, Institut de Phonétique, 29, Avenue Robert Schuman, 13621 Aix-en-Provence

SIMON, Prof. P.
Université des Sciences Humaines de Strasbourg, Directrice de l'Institut de Phonétique, 22, rue Descartes, Strasbourg 67084

SORIN, C.
CNET, Route de Trégastrel, 22301 Lannion

TESTON, B.
Université de Provence, Institut de Phonétique, 29, Av. Robert Schuman, 13621 Aix-en-Provence

TRAN, Ms. T.H.
Université Paris 7, 2, Place Jussieu, 75251 Paris Cedex

VAISSIERE, Mrs. J.
C.N.E.T., B.P. 40, 22301 Lannion

VIDAL, Mrs. M.M.
Université de Paris V, Institut de Linguistique, 12, rue Cujas, 75005 Paris

WENK, Dr. B.J.
University of Strasbourg, 25, rue du Soleil, 67084 Strasbourg Cedex

ZERLING, J.P.
Institut de Phonétique, 22 rue Descartes, 67000 Strasbourg Cedex

Germany, D.R.

MEHNERT, Dr. Ing. D.
Humboldt-Universität zu Berlin, Albrechtstrasse 22, 1040 Berlin

STOCK, Prof. Dr. E.
Universität Halle, WB Sprechwissenschaft, Klement-Gottwalds-Str. 3, 4020 Halle

Germany, F.R.

ALMEIDA, Dr. A.
Institut für Phonetik, Greinstrasse 2, 5000 Köln 41

BARRY, W.J.
Universität Kiel, Institut für Phonetik, Olshausenstrasse 40-60, 23 Kiel 1

BARLINER, Dr. A.M.
Universität München, Geschwister-Scholl-Platz 1, 18 München 22

BENDA, Ms. U. von
Max-Planck-Institut für Psychiatrie, Kraepelinstrasse 10, 8000 München 40

BIERFERT, H.M.A.
IKP, Universität Bonn, Poppelsdorfer Allee 47, 5300 Bonn

BRAUN, Mrs. A.
Forschungsinstitut für deutsche Sprache, 'Deutscher Sprachatlas', Wilhelm-Röpke-Strasse 6A, 3550 Marburg

DIESTELMANN, Dr. M.
Kinderzentrum München, Lindwurmstrasse 131, 8000 München

DOMMELEN, W.A. van
Institut für Phonetik, Olshausenstrasse 40-60 Haus N 50 d, 2300 Kiel

ENDRES, Prof. Dr. W.
Technische Hochschule Darmstadt, Hans-Busch-Institut, FBr. 18 Nachr. Techn., Merckstrasse 25, 6100 Darmstadt

GERSIC, Dr. S.
Universität Köln, Institut für Phonetik, Greinstr. 2, 5000 Köln 41

GLAVE, R.D.
Institut für Kommunikationsforschung und Phonetik, Poppelsdorfer Allee 47, 5300 Bonn

HARTMANN, E.
Max-Planck Institute for Psychiatry, Kraepelinstrasse 10, 8 Munich 40

HEIKE, Prof. Dr. G.
Universität Köln, Institut für Phonetik, Greinstr. 2, 5000 Köln 41

HESS, Dr. Ing. W.J.
Technische Universität München, Lehrstuhl für Datenverarbeitung, Postfach 202420, 8000 München 2

HOEQUIST, C.E. Jr.
Universität Kiel, Institut für Phonetik, 40-60 Olshausenstrasse, Haus N 50d, 2300 Kiel

HOOLE, M.A., P.
Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3/11, 8000 München 40

INDEFREY, H.H.
Technische Universität München, Lehrstuhl für Datenverarbeitung, Franz-Joseph-Strasse 38/1, 8000 München 40

KOHLER, Prof. K.J.
Universität Kiel, Institut für Phonetik, Olshausenstrasse 40-60, 2300 Kiel

KOSTER, Prof. Dr. J.P.
Universität Trier, Sprachenzentrum, Postfach 3825, 5500 Trier

KRIER, F.
Universität Kiel, Romanisches Seminar, Olshausenstrasse 40-60, 2300 Kiel

KOLAS, Dipl. Ing. W.
Ruhr - Universität Bochum, Lehrstuhl für Akustik, Postfach 102148, 4630 Bochum
KUNZEL, Dr. H.
Bundes Kriminalamt, Thaerstrasse 11, 6200 Wiesbaden
LADD, Dr. D.R.
Universität Giessen, Fachbereich Psychologie, Otto-Behaghel strasse 10 F, 6300 Giessen
LU, Prof. S.
Academia Sinica, Peking, Institut für Phonetik, Greinstrasse 2, 5000 Köln 41
MASTHOFF, H.
Universität Trier, Sprachzentrum, 5500 Trier
MÖBIUS, B.
Universität Bonn, Institut für Kommunikationsforschung und Phonetik, Poppelsdorfer Allee 47, 5300 Bonn 1
NEPPERT, J.M.H.
Universität Hamburg, Phonetisches Institut, Alsterglacis 3, 2000 Hamburg 36
NEY, Dr. H.
Philips GmbH, Forschungslaboratorium Hamburg, Postfach 540840, 2000 Hamburg 54
PETURSSON, Dr. M.
Universität Hamburg, Phonetisches Institut, Alsterglacis 3, 2000 Hamburg 36
PEZER, Ms. M.
Universität Köln, Institut für Phonetik, Greinstrasse 2, Köln 41
PIROTH, M.A.H.G.
Universität München, Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3, 8000 München 40
POMPINO-MARSCHALL, Dr. B.
Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3/II/VG, 8000 München 40
PORZIG, Mrs. U.M.
Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3/II/VG, 8000 München 40
REINIGER, Mrs. U.M.
Universität Frankfurt, Robert-Mayerstrasse 2-4, 6000 Frankfurt/M. 1
RIAL Y COSTAS, J.A.
Universität Freiburg, I.Br. Engl. Sem., 7800 Freiburg, I.Br.
RIGOLL, G.
Frauhofer-Institut für Arbeitswirtschaft und Organisation, Holzgartenstrasse 17, 7000 Stuttgart 1
RUHL, Dipl.-Ing. H.W.
Ruhr-Universität Bochum, Lehstuhl für Akustik, Universitätstrasse 150, Postfach 102148, 4630 Bochum 1
RUSKE, Dr. Ing. G.
Technical University of Munich, Lehrstuhl für Datenverarbeitung, Franz-Joseph Strasse 38, 8 München 40
SAUER, B.
BKA, Germany, BKA, K 1-2, 52 Wiesbaden
SCHÄFESKUPPER, Dr. Phil. P.
Max Planck Institut, Kraepelinstrasse 10, 8000 München 40
SCHÄFER-VINCENT, Dr. K.
Universität Kiel, Institut für Phonetik, Olshausenstrasse 40-60 Haus N 50 D, 2300 Kiel
SCHIEFER, Dr. Phil., Mrs. L.
Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3, 8000 München 40
SCHROEDER, Prof. Dr. M.R.
Drittes Physikalisches Institut, Buergerstrasse 42-44, 3400 Göttingen

SCHWEISTHAL, Dr. K.G.
Universität München, Institut für Phonetik, Schellingstrasse 3, 8000 München 40
SENDLMEIER, M.A., W.F.
Institut für Kommunikationsforschung und Phonetik, Poppeldorfer Allee 47, 5300 Bonn 1
SILVERMAN, Ms. K.E.A.
University of Giessen, Department of Psychology, Otto Behaghel Strasse 10, 6300 Giessen
SOMMER, Ing. Grad. R.J.
Universität München, Institut für Phonetik, Schellingstrasse 3, 8000 München 40
STRUBE, Dr. rer.nat. H.W.
Universität Göttingen, Drittes Physikalisches Institut, Bürgerstrasse 42-44, 3400 Göttingen
TATARU, Dr. Mrs. A.
Bergheimer Str. 26, 6900 Heidelberg
TERNES, Prof. Dr. E.
Universität Hamburg, Phonetisches Institut, Alsterglacis 3, 2000 Hamburg 36
THON, Dr. Ing. W.
Universität Kiel, Institut für Phonetik, Olshausenstrasse 40-60, 2300 Kiel 1
TILLMANN, Prof. Dr. H.G.
Universität München, Institut für Phonetik und Sprachliche Kommunikation, Schellingstrasse 3, 8000 München 40
TIMMERMANN, G.
Universität Kiel, Institut für Phonetik, Olshausenstrasse 40-60 Haus N 50 D, 2300 Kiel 1
TODTER, Mrs. R.
Institut für Phonetik, Olshausenstrasse 40-60 Haus N 50 D, 2300 Kiel 1
TURK, Dr. E.
Universität des Saarlandes, Fachrichtung 8.6, 6600 Saarbrücken 11
WILHELMS, R.
III Physikal. Institut Göttingen, Bürgerstrasse 41/44, 3400 Göttingen

Hong Kong

KWOK, Mrs. H.
University of Hong Kong, Dept. of English Studies and Comparative Literature, Hong Kong

Hungary

BOLLA, K.
Hungarian Academy of Sciences, Institute of Linguistics, I. Szentháromság u. 2. Pf. 19, 1250 Budapest
FÖLDI, E.
l'Université N., Lajos Kossuth, Debrecen
OLASZY, G.
Department of Phonetics, Institute of Linguistics, L 250 Pf. 19, Budapest
VALACZKAI, DR. L.
Józef Attila University Szeged, Egyetem utca 2, 6721 Szeged

Ireland

NI CHASAIDE, A.
University of Dublin, Centre for Language and Communication Studies, C.L.S.C., Trinity Vollege, Dublin 2

Israel

BONNAN, G.
Hebrew University Jerusalem, Department of English, Jerusalem
FRANK, Ms.J.
Tel-Aviv University, School of Communication Disorders, Teachers for the Deaf, Tel-Aviv
FREEDMAN, Mrs. B.
Bar Ilan University, Ramat-Gan, Tel Aviv
GOLDSTEIN, Prof. J.L.
Tel Aviv University, Faculty of Engineering, Ramat Aviv, Tel Aviv 69978
LAUFER, Ph.D.A.
Hebrew University, Hebrew Language Department, Phonetics Laboratory, Mt. Scopus, Jerusalem
TOBIN, Y. Ph.D.
Ben-Gurion University of the Negev, Foreign Literatures and Linguistics, P.O.Box 653, Be'er Sheva 84 105

Italy

AVESANI, C.
Scuola Normale Superiore, Piazza dei Cavalieri 7, 56100 Pisa
BENEDETTO, Ms. M.G. di
University of Rome, Institute of Electrical Communications, Via Eudossiana, 18, 00184 Rome
BORTOLINI, U.P.
Centro di Fonetica, Via Oberdan 10, 35100 Padavo
COLLA, Ms. A.M.
Elettronica San Giorgio, ELSAG S.p.A., Via Hermada, 6, 16154 Genova
FARNETANI, Mrs. Dr. E.
Centro Ricerche Fonetica, Via Oberdan 10, 35100 Padova
GIANNINI, Mrs. A.
Instituto Università Orientale, Fon. Sper., Piazza S. Giovanni Maggiore, 80134 Napoli
HURCH, Dr. B.
Università di Genova, Sezione di Germanistica, Piazza S. Sabine 2, 16124 Genova
MAROTTA, Mrs. G.
Scuola Normale Superiore - Pisa, Piazza dei Cavalieri, 56100 Pisa
PETTORINO, M.
Instituto Università Orientale, Fon. Sperim., Piazza S. Giovanni Maggiore, 80134 Napoli
SALZA, Dr. P.L.
CSELT, Via G. Reis Romoli 274, 10148 Torino
UGUZZONI, Mrs. Prof. A.
Università di Bologna, Instituto di Glottologia, Via Zamboni 38, 40126 Bologna
VAYRA, M.
Scuola Normale Superiore, Piazza dei Cavalieri 7, 56100 Pisa
VOGEL, Mrs. Dr. I.
University of Venice, Institute of Romance Philology, San Sebastiano, Venice

Japan

EGUCHI, S.M.D.
The University of Tsukuba, Institute of Human Science, Division of Audiology and Speech Science, Sakura-mura, Niihari-gun, Ibaraki-Ken, 305
FUJISAKI, Prof. Dr. H.
University of Tokyo, Department of Elect. Engineering, Bunkyo-ku, Tokyo, 113

HIRANO, Prof. M.
Kurume University, School of Medicine, Department of Otolaryngology, 67 Asahi-machi, Kurume, 830
HIROSE, Prof. H.
University of Tokyo, Research Institute of Logopedics and Phoniatrics, Faculty of Medicine, 7-3-1 Hongo Bunkyo-ku, Tokyo (113)
HOMMA, Y.
Osaka Gakuin University, Kishibe, Suita-shi, Osaka 564
HONDA, K.
Research Institute of Logopedics and Phoniatrics, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113
KAKEHI, Prof. H.
Kobe University, 1 Rokkodai Nada-ku, Kobe 657
KOBAYASHI, Prof. T.
Setsunan University, Ikeda-nakamachi 17-8, Neyagawa/Osaka 572
KORI, S.
Osaka University of Foreign Studies, 2734 Aomatani, Minoo-shi, Osaka 567
MASAKI, S.
Waseda University, Department Electrical Engineering, Shirai Laboratory, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo 160
MATSUSHITA, T.
Gifu University, Dept. of English, Faculty of Liberal Arts, Gifu 50Z
MORIGUCHI, T.
Meisel University, 33/Hodokubo Illno Shl, Tokyo 191
NAKANISHI, T.
Tokushima Joto High School, Nakatokushima, Tokushima 770
NISHIMI, Miss R.
R.N. Music Research Institute, 1-8-23 Kita-Senzoku, Ota-ku
OGAMINO, Y.
The University of the Sacred Heart, Hiroo 4-3-1. Shibuya-ku, Tokyo 150
SAKOW, Prof. W.A.
St. Margaret's College, 4-29-23 Kugayama Suginami-ku, Tokyo 168
SAWASHIMA, P.N.D., Prof. M.
University Tokyo, Faculty of Medicine, Research Institute of Logopedics and Phoniatrics, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113
SHIMIZU, PROF. K.
Nagoya Gakuin University, 1350 Kamishinano, Seto City, 480-12 Aichi-ken
SHIMOMURA, Prof. T.
Technical College of Osaka Prefecture, 26-12 Saiwai-cho, Neyagawa-shi, Osaka 572
SIMADA, Z.
Kitasato University, Department of Physiology, School of Medicine, Kitasato 1-15-1, Sagamihara, Kanagawa 228
SONODA, Y.
Kumamoto University, Faculty of Engineering, Kumamoto 860
WAIDA, T.
Osaka Women's University, Daisen-cho, Sakai, 590 Osaka
YANAGIDA, Dr. M.
Osaka University, Institute of Scientific and Industrial Research, Mihogaoka 8-1, Ibaraki, Osaka 657

Jordan

ANAMI, M.I.
University of Jordan, P.O. Box 13340, Amman

EL-HALEES, Prof. Y.A.K.
The University of Jordan, The English Department, Amman

MITLEB, Dr. F.M.
Yarmouk University, Irbid

Kuwait

AYOUB, Prof. A.R.M.
University of Kuwait, Dept. of Arabic, Faculty of Arts, P.O. Box 5969, Kuwait

Mexico

KELLER, Mrs. K.C.
Instituto Linguistico de Verano, Apartado 22067, 14000 Mexico D.F.

The Netherlands

BAART, J.L.G.
University of Leiden, Department of Linguistics, Phonetics Lab., Cleveringaplaats 1, Leiden

BAKKER, Ms. E.
University of Amsterdam, Department of English, Spuistraat 210, Amsterdam

BALEN, C.W. van
Koolmees 55, 2986 VC Ridderkerk

BALKOM, Drs. L.J.M. van
Institute for Rehabilitation Research, Zandbergseweg 111, 6432 CC Hoensbroek

BERG, C.G. van den
Rijksuniversiteit Utrecht, Fonetisch instituut, Trans 14, Utrecht

BLOOTHOOFT, G.
Rijksuniversiteit Utrecht, Institute of Phonetics, Trans 14, 3513 CR Utrecht

BONDER, Drs. L.J.
Universiteit van Amsterdam, Instituut voor Fonetische Wetenschappen, Herengracht 338, 1016 CG Amsterdam

BONNOT, J.F.
University of Amsterdam, Department of French, Singel 134, 1015 AG Amsterdam

BOOIJ, Prof. Dr. G.E.
Vakgroep Algemene Taalwetenschap, Postbus 7161, 1007 MC Amsterdam

BOT, Dr. K.L.J. de
Katholieke Universiteit Nijmegen, Institute of Applied Linguistics, Erasmusplein 1, 6500 HB Nijmegen

BOUMA, H.
Institute for Perception Research IPO, P.O. Box 513, 5600 MB Eindhoven

BOVES, Drs. L.
Nijmegen University, Institute of Phonetics, P.O. Box 9103, 6500 HD Nijmegen

BOXELAAR, Ir. G.W.
Universiteit van Amsterdam, Instituut voor fonetische wetenschappen, Herengracht 338, 1016 CG Amsterdam

BRADLEY, Ms. D.
Max-Planck-Insituut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

BROECKE, Dr. M.P.R. van den
University of Utrecht, Institute of Phonetics, Trans 14, 3513 ER Utrecht

BROKX, Dr. Ir. J.
St. Radboudziekenhuis, Audiologisch Centrum, Philips van Leydenlaan 15, Postbus 9101, 6500 HB Nijmegen

BROWN, C.M.
Max-Plack-Institut für Psycholingstik, Berg en Dalseweg 79, 6522 BC Nijmegen

BUITING, Drs. H.J.A.G.
Katholieke Universiteit Nijmegen, Institute of Phonetic Sciences, Erasmusplein 1, 6525 HT Nijmegen

COHEN, Prof. Dr. A.
Rijksuniversiteit Utrecht, Fonetisch Instituut, Trans 14, 3515 ER Utrecht

CONINX, Dr. F.
Instituut voor Doven, Theerestraat 42, 5271 GD St. Michielsgestel

CRANEN, B.
Katholieke Universiteit Nijmegen, Institute of Phonetics, P.O. Box 9103, 6500 HD Nijmegen

CRUL, Dr. T.A.M.
University Nijmegen, St. Radboudziekenhuis, Afd. KNO, Audiological Center, Nijmegen

DECKERS, C.M.H.A.
Slavic Seminar/institute for Phonetic Sciences, Amsterdam

DUPUIS, M. Ch.
University of Leiden, Phonetics Laboratory, P.O. Box 9515, 2300 RA Leiden

DIJK, Drs. J.S.C. van
University of Amsterdam, Institute of Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam

EIJSBERGEN, F. van
Digital. Sound. Corp-Nemimij BV, Laan Copes van Cattenburg 76-78, Den Haag

ELSENDOORN, Drs. B.A.G.
Instituut voor Perceptie Onderzoek, Den Dolech 2, 5600 MB Eindhoven

ERP, Mrs. A.J.M. van
Instituut voor Fonetiek Nijmegen, Erasmusplein 1, Nijmegen

EVERAERT, M.
Stichting Z.W.O., Instituut voor A.T.W., Trans 14, Utrecht

FAGEL, Drs. W.P.F.
Universiteit van Amsterdam, Instituut voor fonetische wetenschappen, Herengracht 338, 1016 CG Amsterdam

FESTEN, Dr. J.M.
Vrije Universiteit Amsterdam, Afdeling Geneeskunde, v.d. Boeckorststraat 7, 1081 BT Amsterdam

FRAUENFELDER, U.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

GEEL, Drs. R.C. van
Rijksuniversiteit Utrecht, Instituut voor Fonetiek, Trans 14, 3513 ER Utrecht

GELDER, Dr. L. van
University of Amsterdam, Academical Medical Centre, Phoniatric Department of Ear-Nose-Throat Clinic, Meibergdreef 9, 1105 AZ Amsterdam

GOEMAN, Drs. A.C.M.
P.J. Meertens-Instituut K N A W, Keizersgracht 569-571, 1017 DR Amsterdam

GRAAF, Dr. T. de
Instituut voor Fonetische Wetenschappen, Grote Rozenstraat 31, 9712 TG Groningen

GREVEN, Mrs. Dr. A.J.
ENT Clinic foniatric, Free University, De Boelenlaan 1117, Amsterdam

GÜNZBURGER, Mrs. Dr. D.
Rijksuniversiteit Utrecht, Fonetisch Instituut, Trans 14, 3513 ER Utrecht

GUSSENHOVEN, C.H.M.
Universiteit Nijmegen, Instituut Engels-Amerikaans, Erasmusplein 1, 6300 HD Nijmegen

GVOZDANOVIC, Dr. J.
Universiteit van Amsterdam, Slavisch-Seminarium, Spuistraat 20, 1012 VT Amsterdam

HART, J. 't
Institute for Pereption Research IPO, Postbus 513, 5600 MB Eindhoven

HENDRIKS, Ir. J.
Dr. Neher Laboratorium PTT, St. Paulusstraat 4, Postbus 421, 2260 AK Leidschendam

HERMANS, Drs. B.J.H.
Universiteit van Amsterdam, Instituut voor Neerlandistiek, Herengracht 330-336, Amsterdam

HERPT, Dr. L.W.A. van
Universiteit Amsterdam, Instituut Fonetische Wetenschappen, Herengracht 338, 1016 EG Amsterdam

HEUGTEN, Ir. B. van
Dr. Neher Laboratorium PTT, Postbus 421, 2260 AK Leidschendam

HEUVEN, Dr. V.J. van
University of Leyden, Department of Linguistics, Phonetics Laboratory, P.O. Box 9515, 2300 RA Leiden

HEIJINK, Ir. R.J.
Dr. Neher Laboratorium PTT, Postbus 421, 2260 AK Leidschendam

HILLCOAT, A.
Floating Point Systems, Antwerpseweg 9-201, 3802 PB Gouda

HORST, Ir. E.H.J. van der
Dr. Neher Laboratorium PTT, P.O. Box 421, 2260 AK Leidschendam

ICKENROTH, J.
Rijksuniversiteit Utrecht, Instituut voor toegepaste Taalkunde, Wilhelminapark 11/12, Utrecht

JAMES, A.R.
Universiteit van Amsterdam, Engels Seminarium, Spuistraat 210, 1012 VT Amsterdam

JANSEN, C.I.E.
Katholieke Universiteit Nijmegen, St. Radboudziekenhuis, Afd. Stem- en Spraakstoornissen K.N.O. Kliniek, Nijmegen

JANSONIUS-SCHULTHEISS, Drs. K.
Instituut voor Fonetische Wetenschappen, Herengracht 338, 1016 CG, Amsterdam

KAPPNER, E.O.
Instituut voor Fonetische Wetenschappen, Herengracht 338, 1016 CG Amsterdam

KERKHOVEN, W.J.
Floating Point Systems, Antwerpseweg 9-201, 2803 PB Gouda

KLAASSEN-DON, Mrs. Drs. L.E.O.
TNO, Institute for Perception, Kampweg 5, 3769 DE Soesterberg

KLOP, Dr. J.W.
Mathematisch Centrum, Kruislaan 413, 1098 SJ Amsterdam

KOOPMANS-VAN BEINUM, Mrs. Dr. F.J.
Institute of Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam

KOSTER, C.J.
Free University, Amsterdam

KRUYT, Mrs. Drs. J.G.
Instituut voor Nederlandse Lexicologie, W.N.T., Postbus 9515, 2300 RA Leiden

LEATHER, J.H.
Universiteit van Amsterdam, Spuistraat 210, 1012 VT Amsterdam

LEVELT, Prof. W.J.M.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

LOOTS, Mrs. M.
Tilburg University, Hogeschoollaan 225, 5037 GC Tilburg

MAASSEN, Drs. B.
University Nijmegen, Department of Experimental Psychology, Montessorilaan 3, 6525 GC Nijmegen

MARCUS, Dr. S.M.
Institute for Perception Research - IPO, Postbus 513, 5600 MB Eindhoven

MARSLEN-WILSON, Dr. W.D.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

MAURER, Drs. F.X.F.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

MEINSMA, G.L.
Institute Phonetic Sciences Amsterdam, Herengracht 338, 1010 CG Amsterdam

MENS, L.H.M.
Max-Planck-Instituut Nijmegen, Kath. Universiteit Nijmegen, Montessorilaan 3- room A-05.15, Nijmegen

MEULEN, Drs. Sj. van der
Academisch Ziekenhuis Utrecht, Afdeling Foniatrie, Catharijnesingel 101, Utrecht

MEIJERS, G.
Tilburg University, Subfaculty of Letters, P.O. Box 90153, 5000 LE Tilburg

NESPOR, Mrs. Dr. M.
Universiteit van Amsterdam, Italiaans Seminarium, Spuistraat 210, 1012 VT Amsterdam

NIEBOER, G.L.J.
University of Amsterdam, Institute of Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam

NOOTEBOOM, Prof. Dr. S.G.
Institute for Perception Research, Postbus 513, 5600 MB Eindhoven

NOSKE, Drs. R.G.
University of Amsterdam/ZWO, Instituut voor ATW, Spuistraat 210, 1012 VT Amsterdam

ODE, Mrs. C.
Institute for Applied Linguistics, Keizersgracht 143, 1015 CK Amsterdam

OLDE KALTER, A.L.M.P.
B.S.O./Buro voor Systeemontwikkeling, Koningin Wilhelminalaan 3, 3503 RH Utrecht

OS, Mrs. Drs. E.A. den
Rijksuniversiteit Utrecht, Fonetisch Instituut, Trans 14, Utrecht

OTTEVANGER, Miss Drs. I.B.
Utrecht University, Fonetisch Instituut, Trans 14, Utrecht

PABON, Drs. J.P.H.
Rijksuniversiteit Utrecht, Fonetisch Instituut, Trans 14, Utrecht

PEETERS, Drs. W.J.M.
Rijksuniversiteit Utrecht, Inst. voor Duitse taal- en letterkunde, Muntstraat 4, 3512 EV Utrecht

PLOMP, Prof. Dr. Ir. R.
Institute for Perception TNO, Postbus 23, 3769 ZG Soesterberg

POLS, Prof. Dr. Ir. L.C.W.
Amsterdam University, Institute of Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam

POVEL, Dr. D.J.
University of Nijmegen, Psychological Laboratory, Montessorilaan 3, 6525 HR Nijmegen

RASCH, Dr. R.A.
Instituut voor Muziekwetenschap, Drift 21, 3512 BR Utrecht

REENEN, Dr. P.T. van
Free University of Amsterdam, General Linguistics, De Boelelaan 1105, Amsterdam

RIETVELD, Drs. A.C.M.
Katholieke Universiteit Nijmegen, Institute of Phonetics, P.O. Box 9103, 6500 HD Nijmegen

RIJNJA, Ms. C.M.A.
Mariaziekenhuis, Dr. Deelenlaan 5, 5042 AD Tilburg

RINGELING, J.C.T.
Utrecht University, English Department, Oudenoord 6, 3513 ER Utrecht

RITSMA, Prof. Dr. R.J.
University Hospital Groningen, Institute of Audiology, P.O. Box 30.001, 9700 RB Groningen
SCHAFFERS, Ir. F.J.
Philips Telecommunicatie Industrie BV, P.O. Box 32, 1200 JD Hilversum
SCHEFFERS, Ir. M.T.M.
Institute for Perception Research IPO, P.O. Box 513, 5600 MB Eindhoven
SCHOUTEN, Dr. M.E.H.
Rijksuniversiteit Utrecht, Engels Instituut, Oudenoord 6, 3513 ER Utrecht
SCHRIJNEMAKERS, Drs. J.M.
Technische Hogeschool Twente - Enschede, Afdeling KAB; K. 332, Postbus 217, 7500 AE Enschede
SCHUTTE, Dr. H.K.
Rijksuniversiteit Groningen, Department of Phoniatrics and Logopedics of the Clinic for ORL., P.O. Box 30.001, 9700 RB Groningen
SLIS, I.H.
University of Nijmegen, Institute of Phonetics, Erasmusplein 1/Postbus 9101, 6500 HD Nijmegen
SMOLENAARS, Drs. A.J.
University of Amsterdam, Psychological Laboratory, Weesperplein 8, 1018 XA Amsterdam
SPITZ, Mrs. L.R.
Instituut van Fonetische Wetenschappen, Amsterdam
STELT, Mrs. Drs. J.M. van der
University of Amsterdam, Institute of Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam
STROOP, Dr. J.
Universiteit Amsterdam, Instituut voor Neerlandistiek, Herengracht 506, Amsterdam
TANS, Drs. N.
Dr. Neher Laboratorium PTT, Paulusstraat 4, 2264 XZ Leidschendam
TERKEN, J.
Instituut voor Perceptie Onderzoek, Postbus 513, 5600 MB Eindhoven
TROMP, Drs. H.R.C.
Dr. Neher Laboratorium PTT, Postbus 421, 2260 AK Leidschendam
TYLER, Mrs. Dr. L.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen
VEEN, Ms. M.Q.
Utrecht University, Institute of Phonetics, Trans 14, Utrecht
VIEREGGE, Prof. Dr. W.H.
Katholieke Universiteit Nijmegen, Instituut voor Fonetiek, Erasmusplein 1, 6500 HD Nijmegen
VINGERLING, Mrs. Drs. M.
Rijksuniversiteit Utrecht, Fonetisch Instituut, Trans 14, Utrecht
VLUGT, Drs. M.J. van der
Instituut voor Perceptie Onderzoek, Den Dolech 2, 5612 AZ Eindhoven
VOGTEN, L.M.
Instituut voor Perceptie Onderzoek, Den Dolech 2, 5612 AZ Eindhoven
WEENINK, Drs. D.J.M.
University of Amsterdam, Institute of Phonetic Sciences, Herengracht 338, Amsterdam
WELTENS, Drs. H.H.G.
Katholieke Universiteit Nijmegen, Instituut voor Fonetiek, Erasmusplein 1, 6500 HD Nijmegen
WEMPE, A.G.
Institute for Phonetic Sciences, Herengracht 338, 1016 CG Amsterdam
WIENEKE, Dr. Ir. G.H.
Academisch Ziekenhuis Utrecht, Foniatrie University Utrecht, Postbus 16250, 3500 CG Utrecht

WILLEMS, Ir. L.F.
Inst. for Perception Research, P.O. Box 513, 5600 MB Eindhoven
WILLEMS, Dr. N.
Institute for Perception Research (IPO), P.O. Box 513, 5600 MB Eindhoven
WITKAM, Ms. E.
Opril Grote Markt 12, 4461 AK Goes
WOUTERS, Drs. H.A.L.
Universiteit van Amsterdam, Instituut voor Fonetische Wetenschappen, Herengracht 338, Amsterdam
WIJNEN, F.N.K.
Rijksuniversiteit Utrecht, Psychologisch Laboratorium, Utrecht
ZANTEN, Mrs. Drs. E.A. van
Leiden University, Phonetics Lab. Dept. of Linguistics, Cleveringaplaats 1, Leiden
ZELLE, Ing. H.W.
Philips-Elcoma Div. CAB, Gebouw BA55, Postbus 218, 5600 MD Eindhoven
ZWITSERLOOD, Mrs. Drs. P.
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79, 6522 BC Nijmegen

New Zealand

BAUER, Dr. L.
Victoria University, Dept. of English, Private Bag, Wellington, New Zealand

Nigeria

WILLIAMSON, Mrs. Prof. K.
University of Port Harcourt, Dept. of Linguistics of African Languages, P.M.B. 5323, Port Harcourt

Norway

ANDRESEN, B.S.
University of Bergen, English Institute, P.O. Box 23, 5014 Bergen
FOLDVIK, A.K.
University of Trondheim, Department of Linguistics, 7055 Dragvoll
KLOSTER-JENSEN, Prof. M.A.
Baerumveien 235, 1340 Bekkestva
LYSNE, Mrs. A.K.
University of Bergen, Engelsk. Inst., Postbox 23, 5014 Bergen
SIVERTSEN, Mrs. Prof. E.
University of Trondheim, Department of English, 7055 Dragvoll
SLETHEI, K.
University of Bergen, Department of Linguistics and Phonetics, Sydnesplass 9, 5000 Bergen
VANVIK, A.
Oslo University, Phonetics Department, Oslo 3

Poland

AWEDYK, W.
Adam Mickiewicz University, Institute of English, Marchlewskiego 124/126, 61-874 Poznán
BASZTURA, C. Ph.D.
Technical University of Wroclaw, 1-28, ul. Wybizeze Wyspianskiego 27, 50-157 Wroclaw
GORSKA, Mrs. E.
Adam Mickiewicz University, Institute of English, Marchlewskiego 124/126, 61-874 Poznan

GRZYBOWSKI, Ph.D.S.
Pedagogical University Bysgoszcz, ul. Chodkiewicza 30, 85-064 Bysgoszcz
HANDZEL, D.M. D.PH., Prof. L.
Samodzielna Pracownia Foniatrii, Medical Academy, ul. Chalubinskiego 2, Wroclaw
JAORSZ, Mrs. A.
Samodzielna Pracownia Foniatrii, Independent Laboratory of Phoniatrics, ul. Chalubinskiego 2, Wroclaw
JASSEM, Prof. Dr. W.
Polish Academy of Sciences, Acoustic Phonetics Research Unit, Noskowskiego 10, 61-704 Poznan
MAJEWSKI, W.
Technical University of Wroclaw, Wybrzeze Wyspiańskiego 27, 50-370 Wroclaw
OZGA, Miss J.A.
Jagiellonian University of Krakow, Institute of English, UJ, al. Nickiewicze 9/11, 31-120 Krakow
ROPA, Dr. A.
Uniwersytet Jagoellonski, ul. Golebia 24, 31-007 Kracow
SOBKOWIAK, W.
Adam Mickiewicz University, Institute of English, Marchlewskiego 124/126, 61-874 Poznań
WEGNER, Ms. Dr. A.A.
Adam Mickiewicz University, Institute of English Philology, Marchlewskiego 124/126, Poznań 61-874
WEISS, Dr. Med. Mrs. B.
Samodzielna Pracownia Foniatrii. Medical Academy, ul Chalubinskiego 2, Wroclaw
ZALEWSKI, J.M.
Technical University of Wroclaw, Ul. Prusa 55, Wroclaw

Portugal

CRUZ-FERREIRA, Mrs. M.
Lisbon University, Dept. of Anglo-American Studies, Faculty of Arts, Lisboa-Codex

Saudi Arabia

BAKALLA, Dr. M.H.
King Saud University, Riyadh

Singapore

BROWN, Dr. R.S.
National University of Singapore, Department of English Language and Literature, 0511 Singapore
NIHALANI, Dr. P.
National University of Singapore, Department of English, Singapore 0511

South Africa

UYS, Drs. J.Z.
University of Fort Hare, Department of Afrikaans-Nederlands, Private Bag X1314, Alice, 5700 Ciskei

Spain

IRUN, Miss M.
Universitat de Barcelona, Laboratori de Fonètica Experimental, Estudi General de Lleida, Cavallers, 15, Lleida/Catalonia
JULIA, J.
Universitat de Barcelona, Laboratori de Fonètica Experimental, Estudi General de Lleida, Cavallers 15 (El Roser), Lleida/Catalonia
MONROY, R.
University of Murcia, Faculty of Letters, Dept. de Inglés, St. Cristo 1, Murcia
MOT, Dr. B.L.
Universidad de Barcelona, Departemento de Inglés, Facultad de Filologia, Plaza Universidad, Barcelona 7
SALESA, E.
Instituto Auditivo Espanôl SA, Pau Clarés 98, Barcelona 10
SORRIBES, J.
Universitat de Barcelona, Laboratori de Fonètica Experimental, Estudi General de Lleida, Cavallers 15, Lleida/Catalonia
WORSFOLD, B.J.
Universitat de Barcelona, Laboratori de Fonètica Experimental, Estudi General de Lleida, Carrer Cavallers 15, Lleida/Catalonia

Sweden

ALME, Miss A.M.
University of Stockholm, Institute of Linguistics, 106 91 Stockholm
BANNERT, Dr. R.
Lund University, Institute of Linguistics, Helgonabacken 12, 223 62 Lund
BREDVAK-JENSEN, Mrs. A.C.
Lund University, Institute of Linguistics, Department of phonetics, Helgonabacken 12, 223 62 Lund
BRUCE, Dr. G.
Department of Linguistics and Phonetics, Helgonabacken 12, 223 62 Lund
CARSLON, R.
Department of speech communication and music acoustics, Fack, 100 44 Stockholm
ELERT, Prof. C.C.
University of Umeå, Department of Linguistics, 901 87 Umeå
ERICSSON, M.D., G.B.E.
University Hospital, Department of Phoniatrics, 581 85 Linköping
FANT, Prof. G.
Royal Institute of Technology, Department of Speech Communication and Music Acoustics, 100 44 Stockholm
GALYAS, Mrs. K.
Royal Institute of Technology, Department of speech communication, Stockholm 100 44
GÅRDING, Ms. Prof. E.
Lund University, Department of Linguistics and Phonetics, Helgonabacken 12, 223 62 Lund
GRANSTRÖM, B.
Department of speech communication and music acoustics, KTH, Fack, 100 44 Stockholm
HOLMGREN, Ms. K.
Stockholm University, Institute of Linguistics, 106 91 Stockholm
HOUSE, B.A., D.
Lund University, Department of Linguistics and Phonetics, Helgonabacken 12, 223 62 Lund
HUNNICUTT, Mrs. S.
Department of Speech Communication and Music Acoustics, KTH, Fack, 100 44 Stockholm

JANSON, Dr. T.
University of Stockholm, Department of Linguistics, 106 91 Stockholm
KARLSSON, Mrs. I.A.
Department of Speech Communication and Music Acoustics, KTH, Fack, 100 44 Stockholm
KITZING, Mrs. K. Ph.D.
Institute of Phonetics, Lund
LACERDA, F.
Stockholm University, Institute of Linguistics, 106 91 Stockholm
LINDBLOM, B.
Stockholm University, Department of Linguistics, 106 91 Stockholm
LINDGREN, M.A., R.A.
University of Stockholm., Institute of Linguistics, Department of Phonetics, 106 91 Stockholm
LÖFQVIST, Dr. A.
Lund University, Department of Phonetics, Helgonabacken 12, 223 62 Lund
MAGNUSSEN, Ph. D., E.I.K.
Lund University, Institute of Linguistics, Helgonabacken 12, 223 62 Lund
NETTELBLADT, Mrs. U.K.
Lund University, Department of Linguistics and Phonetics, Helgonabacken 12, 223 62 Lund
NORD, L.
Department of speech communication and music acoustics, KTH, Fack, 100 44 Stockholm
OSTER, Mrs. A.M.
Department of speech communication and music acoustics, KTH, Fack, 100 44 Stockholm
PROHOVNIK, Mrs. B.
Med. rehab. Clinic, Västra Klinikerna, Jönköping 55185
RISBERG, Dr. A.
Royal Institute of technology KTH Fack, Department of speech communication, Stockholm 100 44
SCHULMAN, R.
Universitet Stockholm, Department of Phonetics, 106 91 Stockholm
SPENS, K.E.
Swedish Institute for the Handicapped, Box 303, 161 26 Bromma
STÅLHAMMER, J.U.J.
Department of Speech Communication and Music Acoustics, KTH, 100 44 Stockholm
STRANGERT, Ms. E.
Umeå University, Department of Phonetics, 901 87 Umeå
SUNDBERG, Prof. J.E.F.
Royal Institute of Technology (KTH), Dept. of Speech Communication and Music Acoustics, 100 44 Stockholm
SVANTESSON, J.O.
University of Lund, Avd. F. Kinesiska, Tunavägen 39 G, 223 63 Lund
TOUATI, P.
Lund University, Department of Linguistics and Phonetics, Helgonabacken 12, 223 62 Lund
WINGSTEDT, Mrs. M.
University of Stockholm. Department of Linguistics, 106 91 Stockholm

Switzerland

COLLOVA, P.
Phonetisches Laboratorium der Universität, Freiestr. 36, 8032 Zürich
SPORRI-BUTLER, Dr. Phil., Mrs. S.
Universität Zürich, Phonetisches Laboratorium, Freiestrasse 36, 8032 Zürich
VATER, Mrs. Dr. S.
Centre de poétique et de phonétique expérimentales, Université 1, 12211 Genève 4

Tanzania

TEMU, C.W.
Institute of Kiswahili Research, P.O. Box 35110, Daresalaam

Tunesia

GHAZALI, S.
Universite de Tunis, Institut Bourguiba des langues Vivantes, 47 Avenue de la Liberté, Tunis 1004

Turkey

SELEN, Prof. Dr. N.
Ankara Universitesi, D.T.C. Fakültesi Türkei, Institut für Deutsche Sprache und Literatur, Ankara

United Kingdom

ABBERTON, Mrs. Dr. E.R.M.
University College London, Department of Phonetics and Linguistics, Gower St., London WC1E 6BT
AL-BAMERNI, A.
University of Oxford, Phonetics Laboratory, 41 Wellington Square, Oxford 2FJ
AWBERRY, Dr. G.M.
Welsh Folk Museum, St. Fagan's, Cardiff, Wales
BAILEY, Dr. P.J.
University of York, Department of Psychology, York YO1 5DD
BERESFORD, R.
University of Newcastle-Upon-Tyne, Dept, of Speech, Joseph Cowen House, Newcastle-Upon-Tyne
BLADON, R.A.W.
University of Oxford, Phonetics Laboratory, 41 Wellington Square, Oxford, OX1 2JF
BUTCHER, Dr. A.
University of Reading, Dept. of Linguistic Science, Whiteknights, Reading RG6 2AA
CARTER, Mrs. P.
Cambridge University Press, The Edinburgh Building, Shafterbury Road, Cambridge CB2 2RU
CLARK, C.J.
University of Reading, Dept. of Linguistic Science, Reading RG6 2AA
CLEMENT, R.D.
University of Edinburgh, Linguistic Survey of Scotland, 27 George Square, EH8 9LG Edinburgh
COWLISHAW, Mrs. K.C.
IBM United Kingdom Science Center Athelstan HS, St. Clement SDtreet, 5023 9DR Winchester
CRUTTENDEN, Dr. A.
University of Manchester, Department of Linguistics, Manchester M13 9PL
CUTLER, Mrs. Dr. A.
MRC Applied Psychology Unit, 15 Chaucer Rd, Cambridge C62 2EF
DARWIN, Dr. C.J.
University of Sussex, Laboratory of Experimental Psychology, Brighton BN1 9Q4
FLETCHER, Miss J.M.
The University of Reading, Department of Linguistic Science, Reading RG6 2AH

FOURCIN, Prof. A.J.
University College London, Department of Phonetics and Linguistics, Wolfson House, 4 Stephenson Way, London NW1 2HE
HARDCASTLE, Dr. W.J.
University of Reading, Dept. Linguistic Science, Whiteknight, Reading Berks
HARRINGTON, J.M.
University of Cambridge, Linguistics Department, Sidgwick Avenue, CB3 9DA Cambridge
HAZAN, Miss V.
University College London, Department of Phonetics and Linguistics, Wolfson House, 4 Stephenson Way, London NW1 2HE
HIGGS, Mrs. J.
University of Edinburg, Department of Linguistics, A. Ferguson Building, George Square, EH8 9LL Edinburg
HOLMES, Dr. J.N.
Joint Speech Research Unit, Princess Elizabeth Way, Cheltenham, Gloucestershire, GL 52 5AJ
HOUSE, Ms. J.E.
University College London, Department of Phonetics, Gower Street, London WC1E 6BT
HOWELL, Dr. P.
University College London, Department of Psychology, Gower Street, London WC1E 6BT
HUTCHESON, Mrs. S.
University of Edinburgh, Department of Linguistics, Adam Ferguson Building, George Square, Edinburgh EH8 9LL
KEMP, J.A.
University, Department of Linguistics, Adam Ferguson Building, George Square, Edinburgh EH8 9LL
LAVER, Dr. J.D.M.H.
University of Edinburgh, Department of Linguistics, Adam Ferguson Building, George Square, Edinburgh EH8 9LL
LISTERRI BOIX, J.
University of Salford, Modern Languages Department, Salford M5 4WT
MACCURTAIN, Ms. Dr. F.
Middlesex Hospital Medical School, Ferens Institute Otolaryngology, Mortimer Street, London W1.
MACINNES, Dr. J.
University of Edinburgh, School of Scottish Studies, 27 George Square, EH8 9LL Edinburgh
MARTIN, Ms. Dr. M.
University of Oxford, Department of Experimental Psychology, South Parks Road, Oxford OX1 3HG
MOORE, Dr. R.K.
Royal Signals and Radar Establishment L120 (s), RSE, Sint Andrews Road, Malvern, Worcs WR14 3PS
MORGAN, Mrs. R.A.
University of Reading, Department of Linguistics Science, Reading RG6 2AA
PETRIE, Ms. H.
University College London, Gower Street, WC1E 6BT London
POINTON, G.E.
Britsh Broadcasting Corporation, Room 2079, Broadcasting House, W1A 1AA London
RAMSARAN, Drs. S.
University College London, Dept. of Phonetics and Linguistics, Gower Street, London WC1 6BT
ROSEN, S.
University College London, Department of Phonetics and Linguistics, 4 Stephenson Way, London NW1 2HE

SCULLY, Ms. C.
University of Leeds, Department of Linguistics and Phonetics, Leeds LS2 9JT
SHIRT, Mrs. M.A.
Leeds University, Dept. of Linguistic and Phonetics, The University, Leeds LS2 9JT
SUMMERFIELD, Dr. Q.
Medical Research Council Institute of Hearing Research, Science Road, University Park, Nottingham NG7 2RD
ULDALL, Mrs. E.
Edinburgh University, Linguistics Dept., Adam Ferguson Building, George Square, Edinburgh 8
WATERSON, Mrs. N.
University of London, Department of Phonetics and Linguistics, School of Oriental and African Studies, Malet Street, London WC1E 7HP
WELLS, Dr. J.C.
University College London, Department of Phonetics and Linguistics, Gower Street, London WC1E 6BT

USA

ABRAMSON, Prof. A.S.
University of Connecticut, Department of Linguistics, U-145, Storrs, CT 06268
AKERS, G.A.
Kurzweill Speech Systems Inc., 57 Wells Avenue, Newton, MA 02159
AL-ANI, Prof. S.H.
Indiana University, Near Eastern Language and Cultures, Goodbody Hall 204, Bloomington 47405
ALFONSO, P.J.
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut
ALLEN, Prof. J.
Massachusetts Institute of Technology, R.L.E. - 36-413, 50 Vassar Street, Cambridge, MA 02139
BALTAXE, Ph.D. Ms. C.A.M.
University of California, Neuropsychiatric Institute, School of Medicine, Department Psychiatry, 760 Westwood Plaza, Los Angeles 90024
BARBER, Mrs. S.
Speech Plus, Inc., 461 N. Bernardo Av., CA Mountain View
BELL-BERTI, Mrs. F.
St. John's University, Jamaica, New York 11439
BEST, Mrs. C.T.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511
Also: Columbia University, Teachers College, New York, 10027
BICKLEY, Ms. C.A.
Massachusetts Institute of Technology, Research Laboratory of Electronics, 36-521, Cambridge, MA 02139
BOND, Z.S.
Ohio University, Lindley Hall 203, Athens, OH 45701
BOYCE, Ms. S.E.
Haskins Laboraties, Yale University, 270 Crown Street, New Haven, CT 06511
CHILDERS, Ph.D., Prof. D.G.
University of Florida, Department of Electrical Engineering, Gainsville, Florida 32611
CLERMONT, F.
Voice Control Systems, Center for Speech Research and Education, 5951 Encina Rd, Suite 127, Goleta, CA 93117

COPPIETERS, R.
Pomona College, ML Dept., Mason Hall, Claremont, CA 91711
DART, Ms. S.
University of California, Los Angeles, Linguistic Dept. U.C.L.A., Los Angeles CA
DISNER, Ms. Dr. S.F.
Mattel Electronics, 5150 Rosecrans Avenue (Mail stop 729-12), Hawthorne, California
DIVENYI, P.L.
Veterans Administration Medical Center, Martinez, California, 94553 California
DORSEY, E.
Speech Plus, Inc., 461 N. Bernardo Av., Mountain View, CA 94043
DOUGHERTY, E.S.J.
Georgetown University, Jesvit Community, Washington DC 20057
ESPY, Mrs. C.Y.
Massachusetts Institute of Technology, Research Laboratory of Electronics, MIT 36, Cambridge, MA 02139
ESTILL, J.
Haskins Laboratory, New Haven, NY
FLANAGAN, Dr. J.L.
Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974
FLEGE, J.E.
University Alabama in Birmingham, Department of Biocommunication, University Station, Birmingham, AL 35294
FOKES, Ms. J.
Ohio University, Lindley Hall 203, Athens, Ohio 45701
FROMKIN, Mrs. Prof. V.A.
University of California, Graduate Division, UCLA, CA 90024, Los Angeles
FUJIMURA, O.
Bell Laboratories, 600 Mountain Avenue, Room 2D-545, Murray Hill, NJ 07974
GALTON, Prof. Dr. H.
The University of Kansas, 2807 Ousdahl Road, Lawrence, KS 66044
GENTIL, Ms. M.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06511
GOLDSTEIN, L.M.
Yale University and Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
HARRIS, Mrs. K.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
HAWKINS, Mrs. Dr. S.
Massachusetts Institute of Technology, 36-511, Research Laboratory of Electronics, Cambridge, MA 02139
HERBERT, Prof. R.K.
State University of New York, Department of Linguistics, Binghamton, NY 13901
HICKS, Dr. J.W., Jr.
University of Florida, Institute for Advanced Study of the Communication Processes, 63-ASB, Gainesville, FL 32611
HIERONYMUS, Dr. J.L.
National Bureau of Standards, Institute of Computer Sciences & Technology, 225/A216, Washington, DC 20234
HOLLIEN, Prof. Dr. H.
University of Florida, ASB-50, IASCP, Gainesville, FL 32611
HOLMBERG, Mrs. E.B.
Boston University, Department Communication Disorders, 48, Cummington Street, Boston, MA 02215
HURTIG, R.R.
The University of Iowa, 125 D Wendell Johnson Speech and Hearing Center, Iowa City, Iowa 52242

IZDEBSKI, Ph.D., K.
University of California San Francisco, 533 Parnassus, R 494-U, San Francisco, CA 94143
JONGMAN, Drs. A.
Brown University, Department of Linguistics, Box E, Providence 02912
KEATING, Ms. P.A.
UCLA, Department of Linguistics, Los Angeles, CA 90024
KLATT, D.
Massachusetts Institute of Technology, Room 36-523 MIT, Cambridge, MA 02139
LADEFOGED, Mrs. J.
Linguistics Department, UCLA, Los Angeles, CA 90024
LADEFOGED, Prof. P.
Linguistics Department, UCLA, Los Angeles, CA 90024
LEHISTE, Mrs. Prof. I.
Ohio State University, Department of Linguistics, 1841 Millikin Road, Columbus, OH 43210
LIBERMAN, Prof. A.
University of Minnosota, Department of German, 9 Pleasant St. S.E., Minneapolis, MN 55455
LIBERMAN, M.
Bell Laboratories, 600 Mountain Avenue, Room 2D-551, Murray Hill, NJ 07974
LINDAU-WEBB, Ms. M.E.
UCLA, Department of Linguistics, Los Angeles, CA 90024
LISKER, Prof. L.
University of Pennsylvania, Department Linguistics, 613 Williams Hall, Philadelphia, PA
MADDIESON, I.
UCLA Phonetics Laboratory, Linguistics Department, 405 Hilgard Av., Los Angeles, CA 90024
McCUTCHEON, Prof. Ph.D.
University of Alabama in Birmingham, Department of Biomedical Engineering, 224 CUD, University Station, Birmingham, AL 35294
McGARR, Mrs. Dr. N.S.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
MINOW, M.
Digital Equipment Corporation, 146 Main Street, ML3-3/U8, Maynard, MA 01754
MOULTON, Prof. W.G.
Princeton University, 68 Lovers Lane, Princeton, NJ 08540
OHALA, M.
San Jose State University, Linguistics Program, San Jose, CA 95192
OHALA, Prof. J.J.
University of California, Berkeley, Department of Linguistics, Berkeley, CA 94720
PERKELL, J.S.
Massachusetts Institute of Technology, Research Laboratory of Electronics, Room 36-543, Cambridge, MA 02139
PICKETT, Prof. J.M.
Gallaudet College, SCRL/MTB Kendall Green, Washington, DC 20002
PIERREHUMBERT, Mrs. J.
Bell Laboratories, 600 Mountain Avenue, Room 2D-444, Murray Hill, NJ 07974
PRECODA, Ms. K.E.
University of California, Department of Linguistics, 2113 Campbell Hall, Los Angeles, Calif. 90024
PRICE, Mrs. Dr. P.J.
Massachusetts Institute of Technology, 36-511, Research Laboratory of Electronics, Cambridge, MA 02139
RAPHAEL, Prof. L.T.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510

RECASENS, D.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
REPP, B.H.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
SARA, S.I.
Georgetown University, School of Languages & Linguistics, Washington, DC 20057
SHADLE, Miss C.H.
Massachusetts Institute of Technology, Speech Communication Group, Research Laboratory of Electronics, 36-521, Cambridge, MA 02139
SHATTUCK-HUFNAGEL, Mrs. Dr. S.
Massachusetts Institute of Technology, Speech Communication Group, Research Laboratory of Electronics, 36-511, Cambridge, MA 02139
SKOUSEN, R.
Brigham Young University, Department of English, 3167 JKHB, Provo, Utah 84602
SOLOMON, J.R.
California State University, Hayward, Speech Pathology and Audiology Program, Hayward, CA 94542
STEVENS, K.N.
Massachusetts Institute of Technology, MIT, Room 46-511, Cambridge, MA 02139
STUDDERT-KENNEDY, M.
City University of New York, Haskins Laboratory, 270 Crown Street, New Haven, CT 06510
TEITELBAUM, Ph.D. M.
B. Laurel Court, West Hempstead, NY 11552
TITZE, Dr. I.R.
University of Iowa, 127 A.S.H.C., Iowa City, Iowa 52242
VELLEMAN, Mrs. Dr. S.L.
University of Texas, Linguistics Department, Austin, TX 78712
VEROFF, Ph.D. Ms. A.E.
Casa Colina Hospital, 255 East Bonita Avenue, Pomona, CA 91767
WALZ, J.C.
University of Georgia, Department of Romance Languages, 109 Moore College, Athens, GA 30602
WEISS, Prof. R.
Western Washington University, Department of Foreign Languages, Bellingham, Washington 98225
WHALEN, D.H.
Haskins Laboratories, 270 Crown Street, New Haven, CT 06510
WOLF, J.J.
Bolt Beranek and Newman Inc., 10 Moulton Street, Cambridge, MA 02238
ZUE, V.W.
Massachusetts Institute of Technology, Bldg. 36-541, 77 Massachusetts Avenue, Cambridge, MA 02139

USSR

LEKOMTSEVA, Dr. M.I.
Institute of Slavistik & Balkanistik, Acad. Sciences, Leningradskij pr. 7, 125040 Moscow
LOMTADZE, A.E.
Tbilisi State Pedagocical Institute of Foreign Languages, No 45 I. Chavchavadze Ave., Georgia, USSR, Tbilisi 380062
METLYUK, Ms. A.
Institute of Foreign Languages, Department of Experimental Phonetics, Zacharova 21, Minsk

SHEVCHENKO, Mrs. T.
Moscow Foreign Languages Institute, 38, Metrostroevskaya, Moscow 119034
STEPANOVICIUS, Prof. A.
Vilnius University, Department of English Philology, Universite to 3, 232734 Vilnius
VENDE, K.
Tallinn Teacher Training Institute, Narva mnt. 25, Tallinn 200101
ZHGHENTI, I.
The Tbilisi State Institute of Foreign Languages, Department of Phonetics, Chavchavadze Ave., 45, Georgian SSR, Tbilisi 380062

Venezuela

CHELA-FLORES, Prof. G.
Universidad del Zulia, Maracaibo, Venezuala

Yugoslavia

DJOKIC, D.
Kolarac Peoples University, 5, Studentski TRG, 11000 Beograd

Index of Contributors

Note: Italic numbers refer to Plenary Sessions (Invited Lectures) and Semi-Plenary Sessions (Symposia)

Preface

As organisers of the 10th International Congress of Phonetic Sciences we saw it as one of our major aims to make sure that the Congress Proceedings would appear in the year of the actual Congress. Through a number of unforeseen circumstances, of which a month long postal strike was merely one, we have not been able to achieve this aim. Nevertheless we are happy to announce in this note the completion of the Proceedings.

In view of the fact that we have sought to give as wide a coverage of the scientific topics of the congress as possible, apart from publishing the full reports of (semi)plenary sessions, and Chairmen's reviews of all Symposia, we have accommodated a large number of section papers as well. This decision has led to the present bulky volume, containing 788 pages of scientific papers.

In order to stay within reasonable bounds we have reluctantly felt forced to restrict the coverage of ceremonial sessions and social events to a bare minimum.

Suffice it to say that the opening statement on behalf of the Permanent Council was given by Prof. A.S. Abramson, who also delivered a moving obituary to the deceased President Prof. D.B. Fry.

Prof. P. Ladefoged of U.C.L.A. was elected as the new President on the final day of the Congress.

During the Opening Ceremony further addresses were given by the Director-General of the Dutch Ministry of Education and Sciences, dr. E. van Spiegel, with a speech on 'Science Policy in the Netherlands' and by the Rector Magnificus of Utrecht University, Prof. O.J. de Jong. Receptions were given by the University of Utrecht, the City of Utrecht and the Province of Utrecht.

For further information about the programme readers are referred to the official Congress Programme.

Special mention should be made of two addresses, by Prof. V.A. Fromkin of U.C.L.A. and Prof. S.G. Nooteboom, Leyden University/I.P.O. Eindhoven, who each gave congress members a bird's eye review of the Congress. We regret that, mainly for reasons of space, we have been unable to publish these remarks which, valuable though they were, after all were meant to be delivered orally.

We are grateful to all authors who have on the whole reacted promptly to our requests for modifications, for abiding by our deadlines and for trusting us to carry out the final editorial work. We regret that in a few cases we had to resort to reprinting the Abstract version of papers.

We wish to thank all those readers who were kind enough to help us in selecting section papers for possible inclusion in the Proceedings well in advance of their actual appearance. In alphabetical order: Ir. G. Bloothooft, Prof. Dr. R. Collier, Prof. P.H. Damsté, Dr. Tj. de Graaf, Dr. D. Günzburger, Mr. J. 't Hart, Dr. V. van Heuven, Dr. F.J. Koopmans-van Beinum, Dr. S. Marcus, Prof. Dr. S.G. Nooteboom, Prof. L.C.W. Pols, Drs. T. Ringeling, Dr. M.E.H. Schouten, Mr. I.H. Slis, Ir. L. Willems, Dr. W. Zonneveld.

We gratefully acknowledge the unfailing zest and accuracy of our secretary and typist who had to retype all MS's due to a new printing technique, Mrs. A.M. van der Linden and Mr. J.A.M. Renkers.

Finally we wish to thank Mr. H. laPorte of Foris Publications, who provided the Congress with a flying start in turning out the Abstracts Volume well in time before the Congress.

M.P.R. Van den Broecke
A. Cohen
Editors

Acknowledgement

The Editors gratefully acknowledge financial support from the following sources which made the publications of these Proceedings possible:
The Foundation for Linguistic Research, funded by The Netherlands Organization for the Advancement of Pure Research (Z.W.O.).
The Executive Council of Utrecht University.

Contents

Section 4: Automatic Speech Recognition

Section 5: Physiology and Acoustics of Speech Production

Section 6: Physiological Manifestations of Speech Sounds

Section 7: Psychoacoustics of Speech

Section 8: Perception of Phonemes

Section 9: Word Perception

Section 10: Stress and Accent

Section 11: Temporal Organisation of Speech

Section 12: Communicative Functions of Prosody

Section 13: Pitch and Intonation

Section 14: Speech Acquisition

Section 15: Sociophonetics

Section 16: Phonetics and Phonology

Report of the meetings of the Permanent Council held during the Tenth International Congress of Phonetic Sciences in Utrecht, 1-6 August 1983.

1. After commemorating the death of Professor Dennis Fry, the President of the Council, the death of Professor Roman Jakobsen, honorary President and founder of the present Council and the death of Prof. Romportl, vice-President, the Council elected Professor Peter Ladefoged as President.
2. Professor Eli Fischer-Jørgensen was elected as vice-president.
3. The Council agreed to request the collaboration of the Soviet Academy of Sciences in the arrangement of the Eleventh International Congress of Phonetic Sciences to be held under the auspices of the Academy of Sciences of the Estonian Soviet Republic in Tallin, Estonia, USSR in 1987.
4. Professors P. Janota, (Czechoslovakia), V. Fromkin (USA), J. Laver (UK), I. Lehiste (USA) M. Remmel (USSR), K. Kohler (F.R. Germany), M. Rossi (France) and M. Sawashima (Japan) were invited to become members of the Council.
5. Professors A. Martinet (France) and M. Onishi (Japan) were elevated to the dignity of Honorary Members of the Council.

International Permanent Council
President: P. Ladefoged, USA
Vice-President: E. Fischer-Jørgensen, Denmark
General Secretary: R. Gsell, France

Honorary Members:

V.A. Artemov, U.S.S.R.
A. de Lacerda, Portugal
A. Martinet, France,
P. Moore, U.S.A.
I. Ochiai, Japan
M. Onishi, Japan
W. Pee, Belgium
A. Rosetti, Rumania
G. Straka, France
E. Zwirner, F.R. Germany

Members:

A.S. Abramson, USA
H. Bolla, Hungary
R. Charbonneau, Canada
L.A. Chistovich, U.S.S.R.
A. Cohen, the Netherlands
G. Fant, Sweden
V. Fromkin, U.S.A.
H. Fujisaki, Japan
M. Halle, U.S.A.
P. Janota, Czsechoslovakia
W. Jassem, Poland
M. Kloster-Jensen, Norway
K. Kohler, F.R. Germany
J. Laver, UK
B. Lindblom, Sweden
B. Malmberg, Sweden
T.M. Nicolajeva, USSR
K.L. Pike, USA
M. Remmel, USSR
A. Rigault, Canada
M. Rossi, France
M. Sawashima, Japan
A. Sovijarvi, Finland
K.N. Stevens, USA
G. Straka, France
Wu, Zong-shi, Rep. China