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

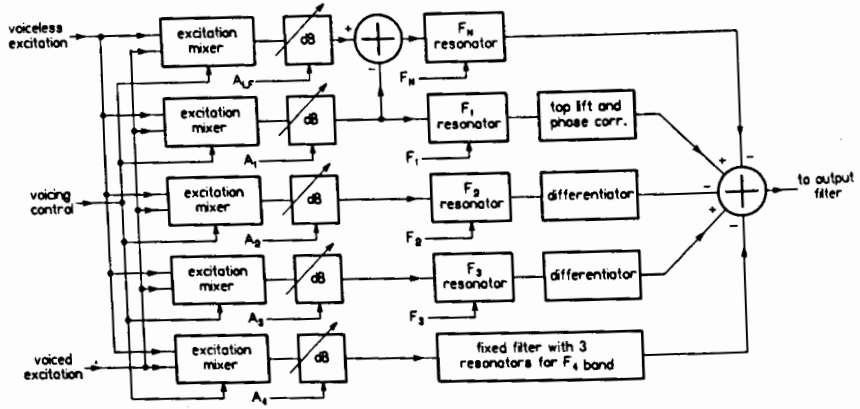

Figure 1. Arrangement of resonators in parallel-formant synthesizer. After Holmes, 1983.

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

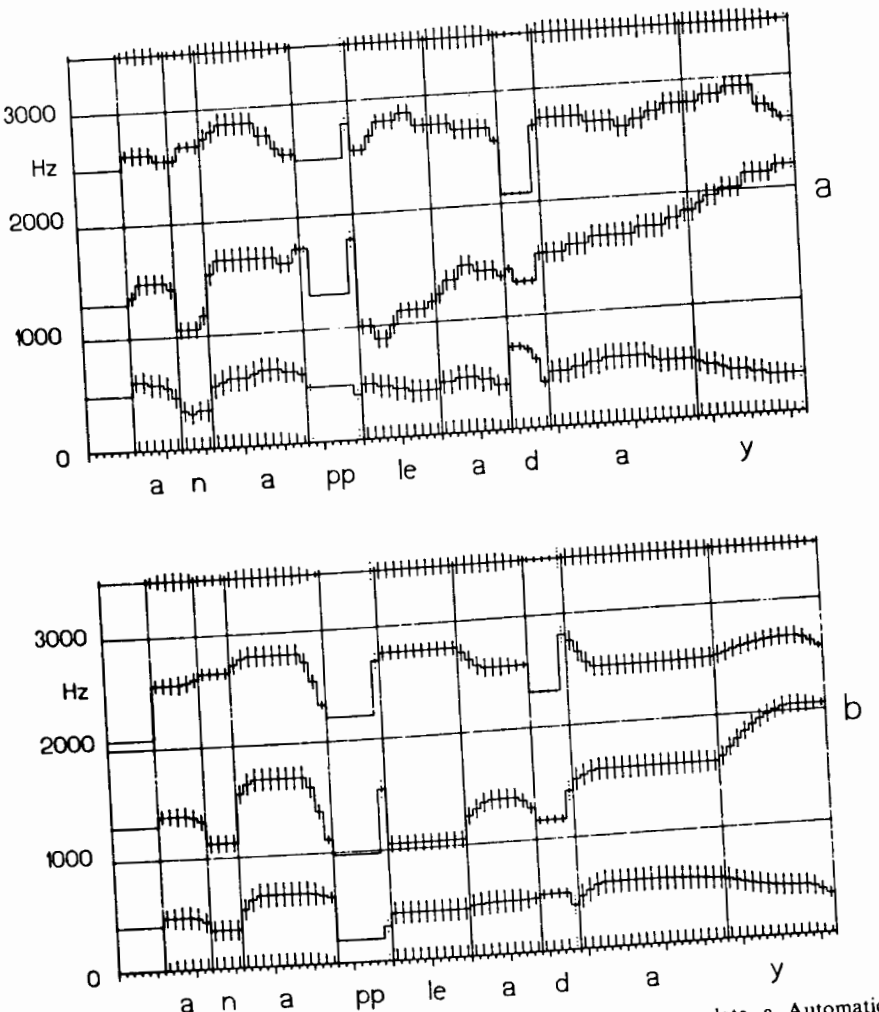

Figure 2. Pseudo-spectrographic reresentation of formant parameter data. a. Automatically derived from natural speech. b. Generated by rule using tables adjusted to suit utterance. After Bridle and Ralls, 1983.

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

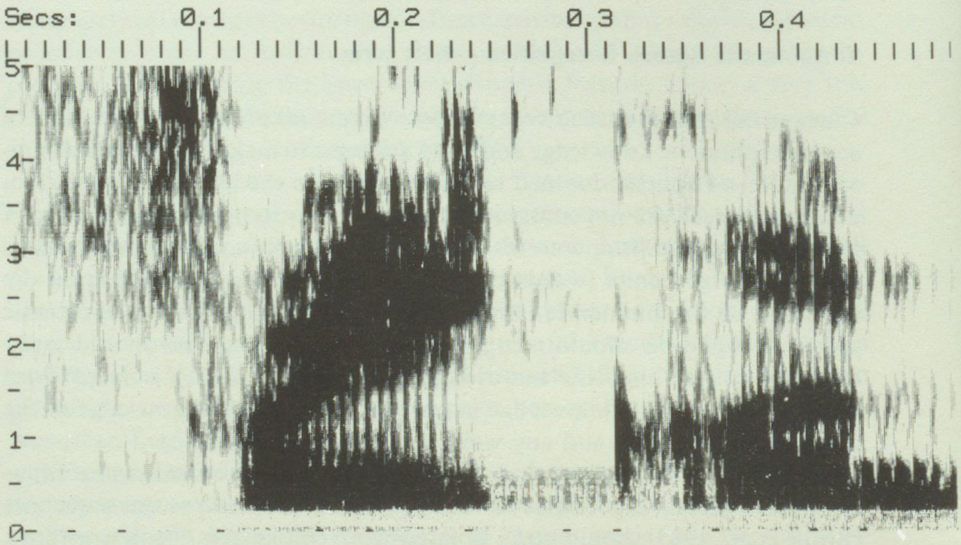

Figure 3a. Conventional display of the first part of the word "frequency" spoken by a male talker.

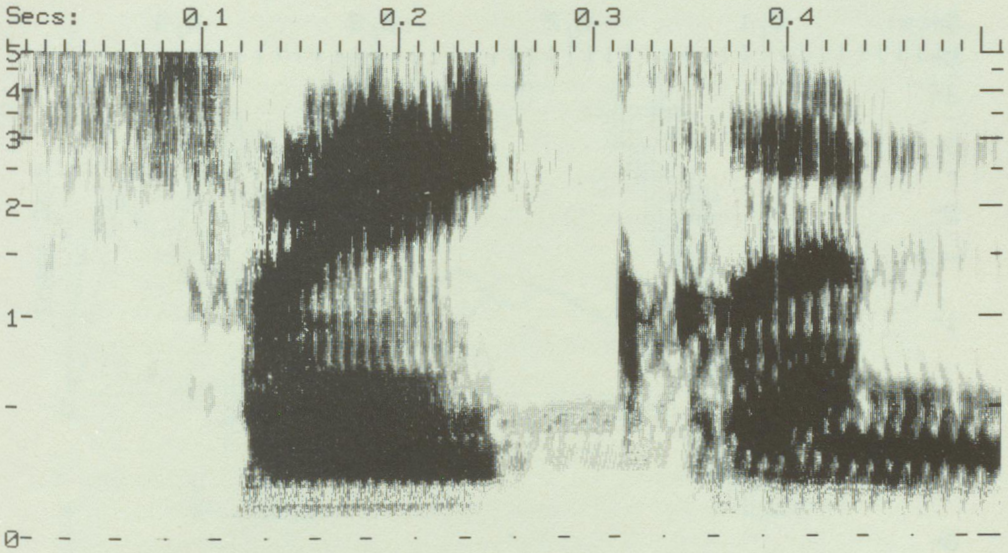

Figure 3b. Spectrogram of Fig. 3a modified to a Bark scale. After Darwin.

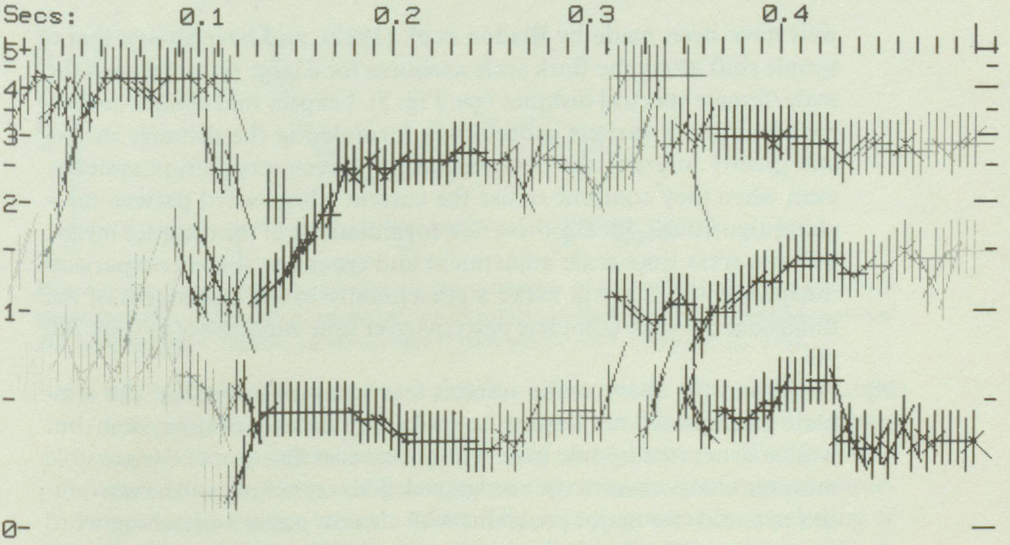

Figure 4a. Shows the effect of convolving the spectral representation of Fig. 3b with a series of masks, designed to detect formant movements of different slopes. After Darwin.

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-

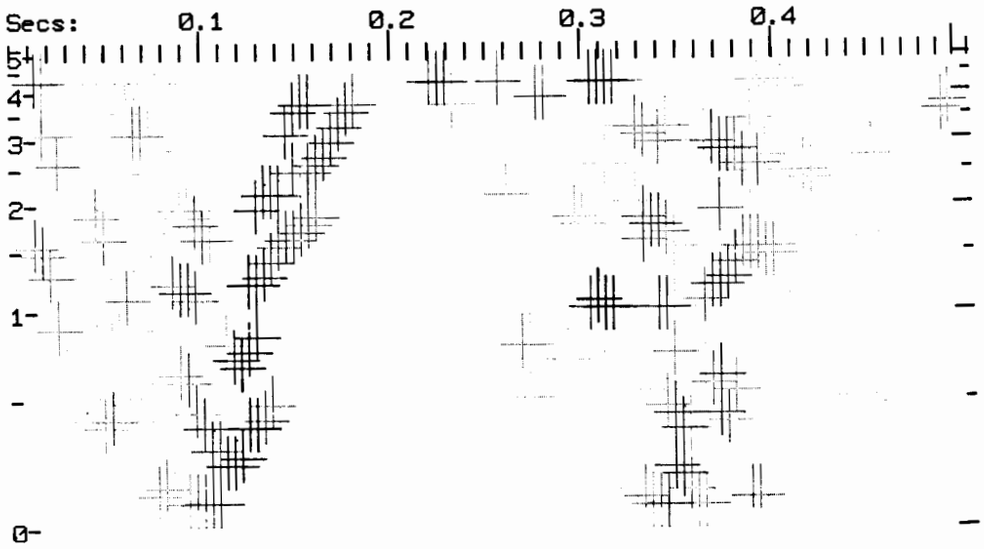

Figure 4b. Shows the effect of plotting the first difference of the signal derived using a Gaussian mask, to highlight the onsets of energy. After Darwin.

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.

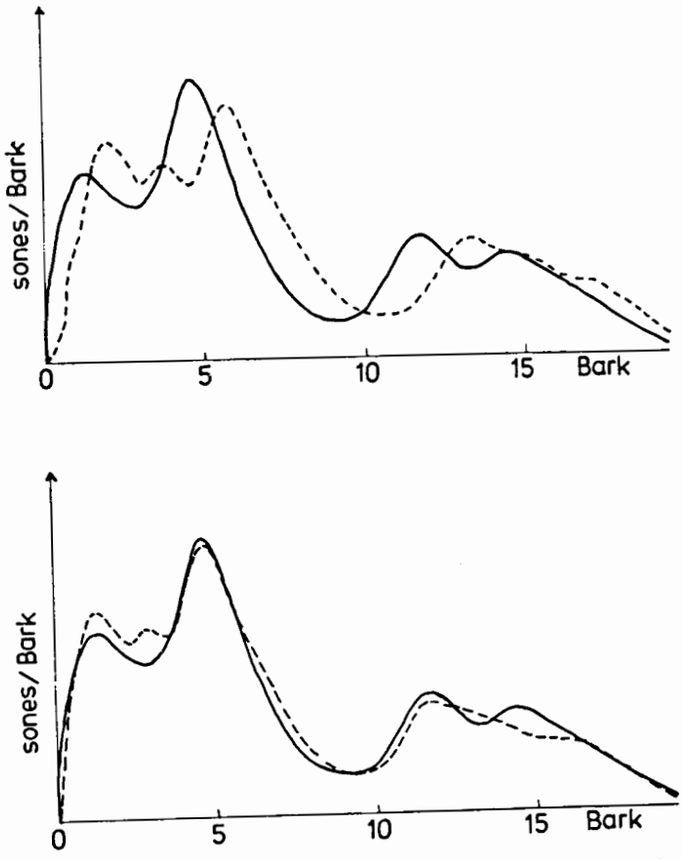

Figure 5. Psychophysically-based spectral representation of an English vowel, before (upper panel) and after (lower panel) normalization by a shift of 1 Bark. Solid curve: average vowel of five male speakers. Dashed curve: a single female vowel. After Bladon et al., 1982.

(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 if 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

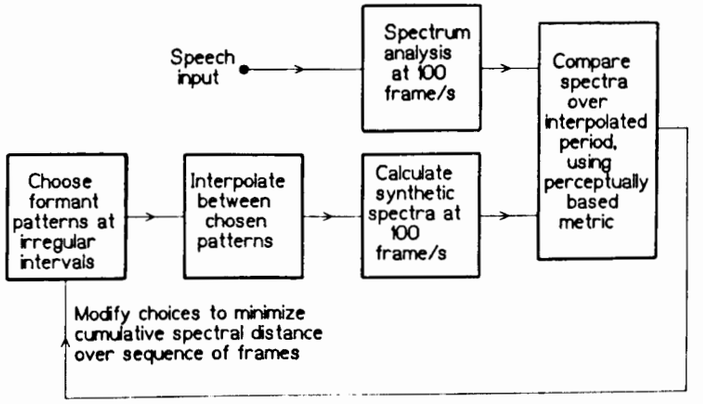

Figure 6. Analysis-by-synthesis system for formant coding. After Dupree.

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.