De Rooij, J.J. (1979): Speech punctuation. An acoustic and perceptual study of some aspects of speech prosody in Dutch, Dissertation, Utrecht.

COMMENTS FROM THE PANELISTS

[Since it is impossible to reproduce here the slides shown by several of the discussants, those parts of their presentations that refer to slides have been edited to make them reasonably comprehensible without visual aids.]

R. Bannert reiterated his conviction that the domain of quantity patterns in Standard Swedish and in a number of other languages is the stressed vowel and the following consonant, and questioned the claim that the syllable boundary falls in the middle of a long consonant. He also presented additional evidence concerning the effect of sentence accent on the durational structure of words like stöka and stöcka. Sentence accent lengthens not only the durations of the segments which make up the sequences, but it lengthens all segments of the word in focus, including the second, unstressed vowel of the test word. The segments /s/, /t/ and /a/ have the same duration in both types of test word. The clear difference between the two minimally contrastive words is in the VC sequences of complementary length. The significance of the VC sequences has also been confirmed by perceptual experiments.

D. H. Klatt formulated some general questions that relate to the problem discussed in his paper: 1) what are the phenomena to be described in a particular language, 2) how do all the rules interact, 3) what is an appropriate underlying representation for an utterance in a particular language, if one wants to predict durations or do a complete synthesis by rule? In a linguistics framework, one would like to start with an as abstract--but psychologically real--representation as possible. As regards the rhythm component, it is true that the paper makes the impression that no attempt has been made to account for it; but there are some rules that make the segmental patterns tend to be isochronous, such as cluster shortening rules and polysyllabic shortening rules within words (but not within feet). These two rules, and perhaps some interactions of other rules, bring about a tendency toward isochrony.

B. Granström pointed out that the primary aim of their paper was not to evaluate Klatt's rule system, but to look into what things are important in rule systems in general, and how naturalness of a rhythmic structure is related to intelligibility. Isochrony in perception is obviously there, or the observation would not have been made in the first place; the question is how important it is in production. It might be that it is not even desirable to have isochrony in production. Parallel studies of rhythm in music indicate that music generated by computer with perfect isochrony is often very dull. Another reason why we believe isochrony is not necessary in the description of durational structure is that it turned out that the rule system is actually very good: in the evaluation process, the utterances generated by the rule system were evaluated as being more natural than the actual productions by Dennis Klatt! And measurements show that the output of the rule system was more isochronous than the actual productions. We believe therefore that an isochrony component is not needed, at least not for the generation of the types of isolated sentences produced in our experiment.

G. D. Allen asked how one should handle short and long quantity in intrinsic timing models. According to the extrinsic view, the motor plan includes temporal features which are used by an extrinsic controller (a "speech clock"), which somehow signals the motor system when to begin and end a specified activity. In the intrinsic view, however, the temporal properties of the act are never specified as such but rather are the result of other, not specifically temporal properties of the act. As an example, consider long versus short vowels. An extrinsic timing model would deliver the command to produce the segment (e.g. /a/) along with a "start" command and a durational feature, which would be used by the clock to generate a "stop" command. An intrinsic timing model, on the other hand, would select either the short or long /a/, which must be represented as distinct acts within the motor repertoire, and that short or long /a/ would then be produced as an integrated part of the overall syllable, word, and/or phrase. Its resulting duration would be a complex function of the several interacting levels of structure and behavior which all together define the act.

Asking how one might test for the existence of intrinsic versus extrinsic timing, Allen reviewed an experiment by Laver (cf. J. Laver's comment below) as an example of a potentially useful experimental paradigm.

S. G. Nooteboom presented some data showing that the perceived boundary between short and long vowels shifts in accordance with speech production regularities. The listener has at his disposal a very detailed knowledge of the temporal regularities of speech: he knows how speech should sound in his language. It is more difficult to know how the listener uses this knowledge, and even more difficult to know how it is stored. In the paper, Nooteboom had made a proposition that all this knowledge is stored as a set of rules in the brain, and that the listener rapidly calculates the expected durations of both short and long vowels, places his criterion in the middle between these two, and thus adjusts his judgment according to context. He considers this now to be a very unlikely procedure, mainly because it must be time-consuming to do so much calculation, and also because he does not believe that all these higher-order effects are going straight back to the level of phoneme decision. There is another way of accounting for the same data, in accordance with some psychological models of word recognition.

H. Fujisaki stated that the motivation for his contribution to this symposium was to provide some quantitative means and frameworks for discussing temporal relations within speech units. The successive units of connected speech manifest themselves not as discrete, separable acoustic events, but rather as overlapping and mutually interfering events. Thus, for example, in discussing the issue of isochrony, one cannot claim that a certain point represents the timing of a speech unit just by looking at the speech signal waveform or its spectrogram. In order to decide whether isochrony is a characteristic of speech production or of speech perception, experimental techniques are needed that allow one to infer the timing of the production of segments as well as the timing of their perception. In his paper, Fujisaki showed quantitative techniques to determine these timing relationships. Thus his contribution was concerned not only with perception, but also with production. The material was deliberately restricted to disyllabic two-mora words of Japanese, since they can be regarded as the smallest examples of connected speech. The materials were further restricted to disyllabic words consisting only of vowels (which are quite common in Japanese), since the articulatory transition from a vowel to the following vowel can be most clearly observed and analyzed from the trajectory of formant frequencies.

Presenting several slides to illustrate the points made in the paper, Fujisaki pointed out a rather wide range of distribution of the onset of articulatory transition among utterances with different combination and order of vowels. At the same time, a strong negative correlation was found between the onset time of such a transition and the rate of transition. In other words, slower transitions were almost always initiated earlier, while faster transitions were almost always initiated later. The onset was distributed over the range from 90 msec to 150 msec within a total utterance duration of approximately 300 msec, which is at least several times larger than the DL for the perception of temporal differences at these durations.

The determination of perceptual timing is based on listening experiments using the same speech material, but by truncating the waveform at various points and presenting only the initial portions as stimuli. The time instant corresponding to 50% judgments was defined as "the perceptual onset" of the second vowel (syllable). The perception of the second vowel starts not at the onset of the formant transition, but at some point where more than 60-70% of the total formant transition has been traversed. The perceptual onsets of the second vowel in various disyllables are concentrated within a very narrow range (about ± one DL) centered around the midpoint of the utterance. Thus the initial and the final vowels are almost always perceived as being of equal duration within a vowel disyllable. The results indicate that the isochrony in this case is neither a mere illusion nor a perceptual distortion of the acoustic reality, but the timing of perception actually occurs isochronously. These findings may be interpreted in the light of a model for the control of speech timing (cf. Figure 7, p. 281 of Volume II). One may safely assume that the articulatory control under ordinary utterance conditions is open-loop control. The findings of this research support the hypothesis that motor commands are programmed in such a way that the perceptual durations of the two vowels within a disyllable are perceived as equal, at least as far as Japanese vowel disyllables are concerned.

In reply to Lehiste's questions, Fujisaki remarked that the work is presently being extended into two directions. One is the case of sequences of three or more vowels which are also quite common in Japanese. Preliminary results indicate that the same conclusion holds for these polysyllabic words. The other direction for future study is to include CV-syllables. It is necessary, however, either to establish an analysis technique whereby one can infer from the speech signal the exact timing of consonantal articulation, not just its acoustic consequences, or to rely on physiological observation to determine the timing of speech production and compare it with the timing of speech perception.

C. J. Darwin recalled the purpose of the reported experiment: to distinguish perceptually between two models for the production of speech durations. According to one model, each phoneme has a sort of "platonic" duration which is shortened as a function of syntactic influences; according to the other, there is an underlying rhythmic structure which is perturbed on the basis of the incompressibility of the elements that one is trying to fit into it. The prediction from this theory is that we are aware of the underlying regular rhythmic foot rather than its surface manifestation.

Darwin also presented additional data which supported the claims made in the paper--that people perceive rhythm to be more isochronous than it really is, and also that this does not apply to non-speech. Additional work has been done at Sussex addressing the question whether syntactic boundaries are signalled just by phrase-final lengthening or by lengthening the whole foot in which the boundary occurs. The results show that the latter is the case.

DISCUSSION

I. Lehiste recalled the results of some of her earlier experiments which had shown that speakers can use several strategies to signal syntactic boundaries. The strategies have a common result, namely lengthening the foot containing the boundary. These experiments had not tested the relative importance of the different strategies, e.g. of phrase-final lengthening, as boundary cues. Lehiste challenged Klatt and Granström to respond to Darwin. In the discussion which followed, it emerged that even though lengthening of the foot is of primary importance, it does matter what part of the interstress level is lengthened: listeners feel uncomfortable if the lengthening is limited to the part that follows the syntactic boundary. It appears that both phrase-final lengthening and lengthening of the foot are necessary for listeners to identify the position of a syntactic boundary.

G. D. Allen commented that it is perhaps wrong to call isochrony in English "largely perceptual" (as had been done by Lehiste), since speech is already temporally highly structured in production. He also questioned those of Darwin's results that showed that non-speech was not perceived as more isochronous than the stimuli really were. This finding appears to be at variance with previous research on time perception, and Allen therefore asked (1) was there in fact a trend in the right direction which was smaller than the one for speech and not statistically significant, and (2) what would be the effect on the nonspeech temporal interval perceptions of filling the intervals with various sounds, as the intervals of speech are filled?

C. J. Darwin responded saying that one of the nonspeech results did depart significantly from actual durations, but it went in the other direction--it was perceived as significantly less isochronous. Darwin agreed with the need to perform experiments with different kinds of nonspeech controls with filled intervals. He would also like to perform similar experiments with music.

I. Lehiste expressed the hope that temporal patterning in other languages besides English and the Scandinavian languages might be considered during the discussion, and urged the discussants to remain conscious of the general theme of the symposium: what are the units within which temporal structures are manifested, how does sentence rhythm relate to the durations of these smaller units, and how does sentence rhythm relate to nonphonological aspects of language--e.g. to syntax.

B. Granström found that perhaps too much attention had been given to isochrony in the discussion, and presented some data that showed that a word can be a very important unit for temporal programming.

P. L. Divenyi, referring to his 1977 dissertation, stated that he had found context effects in rhythmic perception in music. If there is no isochrony in the microscopic sense, there could be in the macroscopic sense, even for nonspeech. Rate is a variable that can affect rhythmic perception. Isochrony is an inherent property of the production system; one could relate isochrony found in perception to production by simply postulating certain listening habits. Thus he does not see any contradiction between productive isochrony and perceptual patterns found in perceptual experiments.

L. Lisker suggested an experiment: to assign segment durations by a random process (in synthesis), and find out what loss in intelligibility and naturalness there would be.

R. Gsell discussed temporal relations in Thai, a quantity and tone language. Stress has a leveling effect on quantity contrasts. Temporal constraints and perceptual limitations produce for the listener neutralization of contour tones in shortened and unstressed syllables.

E. Selkirk took issue with the moderator's characterization of generative phonology as a theory which is in principle unable to countenance such notions as syllable, timing, and rhythm. The notion of the phonological representation within the theory was one of a purely linear kind which saw it as a sequence of segments and boundary elements. In recent years, though, workers who see themselves as operating within the context of generative phonology have been rediscovering that this conception of phonological representation has to be radically revised, allowing for far richer hierarchically arranged suprasegmental structures.

Some workers, Selkirk included, have been arguing for a rather different conception than that in the Sound Pattern of English of Chomsky and Halle, of the relation between phonology and syntax in a generative grammar. In this conception, syntax is seen as bearing on phonology only insofar as phonological units, like syllables or intonational phrases, may have specific syntactic domains over which they are defined, but phonological and phonetic processes are seen as functioning only in terms of these phonological hierarchical structures. It is a claim of this theory that something like final lengthening has its domain defined in terms of phonological units (such as intonational phrase and perhaps others); it would not be immediately sensitive to syntactic structure. What is predicted here is that there would be a systematic convergence of various types of phonological phenomena; the unit at the end of which one finds lengthening would be the same one with which, for example, an intonation contour would be associated, or it may also be the domain of rules of segmental phonology.

Lengthening or the realization of intonational contours and so on are not conceived as individually and separately sensitive to units of syntactic structure.

H. Fujisaki, responding to comments by P. Divenyi and L. Lisker, agreed that we need to look at both microscopic and macroscopic levels of timing. There should be a hierarchy of levels in which speech timing is programmed and maintained. For instance, the problem of compensation between the duration of a consonant and the following vowel is a matter of timing within a syllable, but the compensation between the duration of a vowel in a CV syllable and the following consonant of the next syllable is a matter of interaction between sub-syllabic units across syllable boundaries. Fujisaki had looked at vowel disyllables in order to investigate the relationship between durations of the two syllables without having to consider the problem of consonant-vowel compensation.

J. Laver reviewed his "motoric balance point" experiment mentioned by Allen in connection with two opposing views of the nature of the control of temporal relations. The argument is between the extrinsic view of temporal control, where a "speech clock" acts as an external, overlaid control device, versus the intrinsic view, where temporal relations are the direct product of characteristics of segmental representations themselves. Laver singled out one finding in his experiment which tends to support one of these views. When his subjects were faced with the need to produce forms which had a quantity difference as well as a quality difference between them, such as PEEP and PIP, then the link between quantity and quality was very labile in their productions, and very easily perturbed. There were many errors made, where the right quality but the wrong quantity was produced. So there were examples of PIP with a long vowel duration and of PEEP with a short vowel duration, where both nevertheless showed appropriate articulatory quality. This tends to support the extrinsic view, where duration is at least to some extent the product of specific neuromuscular programming separate from programming for articulatory spatial targets as such.

N. Thorsen addressed a question to Nooteboom, who, with his last slide, had appealed to the audience to have the courage to assume that word identification precedes phoneme recognition. Thorsen asked how Nooteboom would account for the perception of slips of the tongue, which are generally perceived as such, i.e., as slips or mistakes, while at the same time the word is being identified correctly.

K. L. Pike, in his comments, made the point that in English, both isochronic and non-isochronic timing are essential. Under certain circumstances, we must not have isochronic stress groups; under other instances we must indeed have them. This is connected with the fact that in his normal use of English there are some items which one might call "double stresses". These are, in general, related to certain kinds of syntactic groups. There is also a kind of a semantic component which often goes with these double stresses. It is a unitizing effect, tying the items together in some kind of a single concept to be viewed as a unit rather than as components loosely strung together. We must not be so inflexible that we assume that we must have either isochronic stress groups or else we must have largely non-isochronic stress groups. In Pike's analysis of the material one must leave room for both in English. This, in its turn, forces another conclusion: we cannot assume that there is a single rigid set of rules mapping directly, and in only one manner, material from the grammatical hierarchy on to the phonological one; nor of semantically oriented units from a referential hierarchy on to the grammatical or phonological one. We need three hierarchies, always interacting one with another, but never the one totally determining the other. Our rule systems, therefore, cannot be inflexibly from grammar to semantics and phonology; nor from semantics to grammar and then phonology. Rather we must have some interdependence in which the purpose of the speaker is distributed in ways which are vastly more complex than a one-way rule system can tell us.

S. Nooteboom, responding to Thorsen, disclaimed having ever implied that listeners cannot extract phonemes from the acoustic signal. In the normal recognition of known errorless words--which is usually very fast indeed--it is not necessary to assume that phonemes are mediating in perception. Hearing unknown words, or words containing detected mispronunciations, listeners must have been listening in a "phoneme mode".

L. Nakatani questioned the existence of isochrony in production. Even though in comparing black dog with blackish dog there seems to be isochrony, this can very easily be explained by the fact that the first syllable in a bisyllabic word becomes shortened relative to the same syllable in monosyllabic words. There is another factor operating here--resyllabification. In blackish, the /k/ is aspirated, indicating that the /k/ now belongs to the second syllable. So one cannot compare black and blackish, for the syllables are different. If one controls for this by using reiterant speech, some kind of compensation can indeed be found; but if one controls for that and looks at the effect due to the insertion of an unstressed syllable in medial position, one does not find any compensation. Similarly, if one inserts an unstressed syllable at the beginning of the second word, there is no compensation. There is a very linear relationship between the number of intervening unstressed syllables and the interval between stressed syllables. This is consistent with data collected by Wayne Lea.

Nakatani has also looked at duration patterns of words in different contexts. If there is a tendency toward isochrony, the durations of words should vary as a function of the context in which they occur. Looking at the same words in different positions in different sentences, Nakatani found that the duration patterns of words were extremely consistent, and concluded that there is no evidence for isochrony in production. Therefore it should be ascribed primarily to perception, and be based on the fact that content words and function words alternate, and that most bisyllabic words in English have the stress on the initial syllable.

I. Lehiste remarked that there are usually several principles operating at the same time, and they interact. Tendency toward isochrony is one of these principles, but there is certainly another one--the principle of maintaining the temporal integrity of the word, so that the duration of a monosyllabic word is roughly comparable to the duration of a disyllabic word. When these two principles interact, they will influence each other.

E. Uldall noted that we are devoting our attention almost entirely to "stress-timed" languages (though there have been references to Japanese). She expressed the wish to hear a lot more about the opposite case: for example, about French. Phoneticians very frequently refer to English as a classic case of stress-timing, and to French as a classic case of syllable-timing. Yet all the experimental evidence we have about English shows that the "rhythmic feet" are far from isochronous, and what Uldall has seen of French syllables makes her think that they are not isochronous either. So why do phoneticians go on saying what they do?

G. Fant stated that most of our data about durations have been obtained from speech waves--oscillograms and spectrograms. The question is, can we interpret this in terms of a production model to give a better perspective? The answer is affirmative. For instance, if we study vowels in sentence-final stressed position, we find that all the durations are the same, because what has determined the termination of the vowel is the phonatory gesture which is the same for all vowels and independent of the preceding consonant. On the other hand, if the vowel is followed by a consonant, the consonantal frame influences the vowel duration. This is the articulatory aspect. So the duration of a vowel can be set either by phonation or by articulation or, really, both. If a voiced stop comes after the vowel, then of course the vowel is terminated as the acoustical consequence of the constriction, but if it is an unvoiced plosive which comes after the vowel, then there is a separate neural command for the abduction of the vocal cords. That command is somewhat time-locked to articulation, but they are still separate events. This can be a fruitful way of scrutinizing the durational data.

S. M. Marcus gave a brief summary of his research concerning Perceptual Centres or P-centres, which involve rather more fine-grain aspects of speech timing than those determining the temporal structure, isochronous or otherwise, of continuous speech. In producing perceptually isochronous sequences of isolated monosyllables, perceptual regularity corresponded to no simple physical alignment. Subsequent experiments have shown the P-centre locations to be a function of the acoustic structure of the whole stimulus--for example extending the /t/ closure of "eight" shifts its P-centre. These results clearly demonstrate that before considering such questions as isochrony and "syllable-" or "stress-timing" in continuous speech, we need to be very clear what we are measuring the timing of. We must be wary of assuming that simple instrumental measurements, such as consonant and vowel onsets and durations, are related in other than a complex way to our perception. We should also be aware that much of the data which has been used to demonstrate either isochrony or lack of isochrony now needs to be carefully reexamined.

G. D. Allen urged the audience to view timing and rhythm as mental phenomena. Time as it is measured in spectrograms and oscillograms is but one correlate of timing and rhythm. These phenomena belong in the mind, several levels removed from the articulatory periphery.

I. Lehiste thanked the panelists, the very efficient chairman, and all contributors from the floor. She observed that many issues had remained unsolved--for example, the question whether isochrony in English is a property of production or perception. One underlying assumption, however, appears to have been generally accepted--namely that temporal organization operates within units that are larger than a single segment. The task still remains to establish these units for different languages. She concluded with the hope that this discussion has contributed some background that will be taken into account in future research directed toward the discovery of the temporal structure of language.