SYMPOSIUM NO. 5: TEMPORAL RELATIONS WITHIN SPEECH UNITS

(see vol. II, p. 241-311)

Moderator: Ilse Lehiste

Panelists: George D. Allen, Robert Bannert, Christopher J. Darwin,
Hiroya Fujisaki, Björn Granström, Dennis H. Klatt, and
Sieb G. Nooteboom

Chairperson: Claes-Christian Elert

ILSE LEHISTE's INTRODUCTION

The title of the symposium leaves open the question of the type and size of the speech units. The contributors to the symposium have indeed chosen to address themselves to units of quite different types and sizes. Likewise, they have approached the problems connected with the temporal structure of speech units both from the perspective of speech production and from that of speech perception. The contributions include highly theoretical papers, papers presenting detailed results of experiments, and papers falling between these two poles. Some systematization appears to be in order. I would like to present herewith a framework within which I believe the issues can be profitably formulated for the discussions which I hope will follow.

The framework involves three dimensions. One of them concerns the relationship between timing control in production and the role of timing in perception. The second dimension deals with the direction of determination in the temporal organization of spoken language: specifically, with the question whether the timing of an utterance is determined by its syntax, or whether there exist rhythmic principles in production and perception that are at least partly independent of syntax. The third dimension follows directly from the previous two and relates to the type and size of speech units. What is the nature of those units, and are they to be established on the basis of a morphosyntactic analysis of the sentence, or on some kinds of independent phonetic criteria?

Clearly both production and perception are involved in oral communication by spoken language, and it would seem unnecessary to elaborate the point. However, I have had occasion to argue--against considerable weight of opinion--that durational differences in production, be they ever so significant statistically, cannot play a linguistically significant role if they are so small as to be below the perceptual threshold. It would be wise, I think, to remind oneself periodically of "the evident fact that we speak in order to be heard in order to be understood" (Jakobson et al. 1952). I hope, therefore, that in our discussion of temporal relations within speech units, models of production and models of perception will be related to each other.

The second and third questions concern the direction of determination: does phonology follow syntax, or are we dealing with interacting, but parallel hierarchies? Some researchers have developed programs for generating the temporal structure of a sentence on the basis of segments and syntactic structure, without paying any attention to rhythm. This is, I believe, due to a particular theoretical orientation. Generative phonology operates with segmental features; even suprasegmental features are attached to segments. And in a generative grammar, phonetic output is the last step in the generation of a sentence. An independent rhythm component simply has no place in the theory. For these scholars, then, the speech units are segments, phrases, clauses, and sentences. (And it is quite interesting to see them struggle with units not foreseen in the theory, like syllables and phonetic words.) Researchers who are not fully committed to this theoretical viewpoint operate with certain other units, such as speech measures or metric feet. Again, the reality of both kinds of units can be studied from the point of view of production as well as from that of perception.

Practically all the issues I have outlined are treated in the papers contributed to this symposium. Production is the main concern of the papers of Allen, Bannert, Klatt, and Öhman et al.; perception is the focus in the papers of Carlson et al., Donovan and Darwin, Fujisaki and Higuchi, Huggins, and Nooteboom.

In my brief summary of the papers, I shall address some specific questions to the authors, and raise some general questions that I hope will be discussed at the end of the presentations.

Among the papers dealing with production, Bannert considers the relationship between the durations of vowels and consonants in stressed syllables of disyllabic words in Central Swedish--words of the types stöka (V:C) vs. stöcka (VC:). When sentence accent is added to these words, both segments are lengthened, but by unequal amounts. The increase is largest for the long segment of each type of sequence, i.e. the long vowel in stöka and the long consonant in stöcka. Bannert finds that the temporal structure of quantity is best described by using the concept of vowel-to-sequence ratio, V/(V + C), and he proposes that the VC sequences be viewed as units of production and perception.

I have a comment and a question. The comment relates to the observation that lengthening affects the long segment of the VC sequence. It might be useful to recall here that already Trubetzkoy defined the difference between long and short phonemes in terms of stretchability: tokens of long phonemes are stretchable, while short ones are not. Knowing that it is the long element that is stretchable, one could have predicted Bannert's result: that the addition of sentence accent to quantity increases the temporal distance between the two word types.

The question concerns Bannert's proposal that VC sequences be viewed as units of production and perception. I would like to know how such units relate to already well established units such as syllables. Presumably the syllable boundary falls before the single intervocalic consonant in words like stöka and within the long intervocalic consonant in words like stöcka. I find it difficult to conceptualize the psychological reality of the VC sequence as distinct from segments on the one hand and syllables on the other. It seems to consist of non-comparable parts of the two syllables. Where would these VC sequences fit in a hierarchy of units of production? And what is the evidence for the claim that they also constitute units of perception?

The paper by Klatt presents a detailed scheme for the synthesis by rule of segmental durations in English sentences. It is an almost pure example of that approach that starts from an abstract linguistic description and ends up as a sequence of segments whose durations are conditioned by other segments and by syntactic constraints. The paper does not address itself to the question of overall speech rhythm. A companion paper by Carlson, Granström and Klatt is devoted to testing the output of Klatt's synthesis algorithm. Among the interesting results are the observations that certain aspects of the durational pattern are of greater perceptual importance than others. Vowel duration is more important than consonant duration; the durations between stressed vowel onsets seem to constitute a particularly important aspect of sentence structure. Now it is known that English is a stress-timed language; there exists an extensive literature dealing with isochrony in English, and some of the arguments in favor of the existence of isochrony are quite persuasive. I would like to address a question to the three authors of the two papers, concerning the role of rhythm in the production and perception of English sentences. Would it not be advisable to include a rhythm component in the synthesis scheme?

The papers by Öhman et al. and by Allen concern themselves with production models in general. Öhman's et al. paper argues for a gesture theory of speech production. The authors claim that "the linguistically functional, intended acoustic effects are not, in general, required to have any particular duration; ...acoustic segments with quasi-stationary qualities will arise not as a final end of the phonetic action but as a secondary consequence of the effort to reach a certain final end (the simultaneous sounding of the effects in question)". Öhman and co-authors maintain that the phonological contrast between Swedish words like vila and villa can be eliminated using this analysis. Namely, the stress effect, which takes relatively long to produce, is coarticulated with the vowel /i/ in vila--thus making the quickly producible /i/ long, while the stress is coarticulated with the sequence /i . l/ in villa, thus making the /l/ long.

I would like to ask the authors--if they were here--how they would handle contrasts between long and short vowels in unstressed position--contrasts which are found in a large number of languages, e.g. in Czech and Hungarian.

Allen's paper draws a useful distinction between descriptive models and theoretical models of speech timing, and makes the intriguing prediction that theoretical models may be about to undergo substantial modification, primarily due to the emergence of an "action theory" of speech production. According to that theory, neural activity is hierarchically organized into successively higher levels of coordination, until the highest level of all can only be described in terms of the overall goal of the action. The models of "intrinsic timing" which Allen describes seem to operate at levels higher than a segment; I would like to ask Allen, too, how the segmental short-long opposition can be handled within these theories. It would have been quite interesting to hear some discussion about the almost diametrically opposed approaches taken in the papers by Allen and Öhman et al. Öhman, as you may recall, states that manifested segmental durations are generally secondary consequences of the effort to produce simultaneous acoustic effects. Thus there appears to be no room for temporal programming as such. The models Allen refers to claim that intrinsic timing is an inherent property of the speech act. Can these two views be reconciled, or will one of them be proved wrong?

Among the papers devoted primarily to perception, Nooteboom's presents a decision strategy for the disambiguation of vowel length in Dutch. The strategy presupposes knowledge on the part of the listeners of temporal regularities of speech, and the ability to shift an internal criterion--the boundary between long and short vowels--depending on the speech context. For example, the listener is assumed to know that vowels followed by pause are generally longer than vowels followed by a consonant; that vowels are longer when that consonant is a fricative than when the consonant is a plosive; that vowels are shorter with increasing number of unstressed syllables following the syllable containing the stressed vowels, etc. Nooteboom hypothesizes that listeners do indeed possess this knowledge and shift the perceptual boundary between long and short vowels according to speech context. The data presented by Nooteboom are quite impressive; it seems to me, however, that there is something artificial in the described situation. When the listeners adjust the criterion depending on the speech context, they are in fact perceiving the total speech act, not just the vowels. Otherwise there would be no need to perform the adjustment. The environment is just as much part of the percept as the vowel. From my experience with English, I would predict that the durations of vowels and postvocalic consonants stand in a compensatory relationship, and that both are related to the overall duration of the word. Even though the strategy Nooteboom proposes is quite complex, I submit that it is actually an oversimplification.

Fujisaki and Higuchi present an analysis of the temporal organization of segmental features in Japanese disyllables consisting only of vowels, and find that although the onsets of the transition for the second vowel are distributed over a relatively wide range, a perceptual analysis of the onset of the second vowel shows relatively little temporal variation. It thus seems that the apparent diversity of the onset of transition in various disyllables is introduced for the purpose of maintaining the uniformity of perceived duration of segments. Fujisaki and Higuchi consider their results supportive of a model in which the motor commands and the articulatory/acoustic realizations of successive segments are programmed in such a way that the perceptual onsets of successive segments are isochronous.

I am quite impressed and convinced by these results and would really like to have more information. Japanese and English appear to have quite different temporal structures at the sentence level. How far does isochrony go in Japanese? Is the disyllabic sequence conceivably a basic unit of temporal programming--for example, if we have a word of four syllables, does it have the length of two disyllabic sequences? Is there any interaction between segments and syllables--for example, how would the inclusion of consonants in the disyllabic sequences influence their duration both in production and perception?

The paper by Huggins is mainly concerned with the intelligibility of temporally distorted speech. Huggins finds that a distorted timing pattern (which often characterizes the speech of the deaf) is a sufficient cause for catastrophic loss of intelligibility. While I have no argument with this particular claim, I would like to take issue with a statement concerning the relationship between pauses and other cues employed to indicate syntactic boundaries. Huggins states that boundaries that are marked by pauses need not be inferred from more subtle cues. In some recent work of mine on the perception of sentence boundaries, I found that listeners can completely ignore a fairly lengthy pause, if it is not preceded by a certain amount of preboundary lengthening and/or change in fundamental frequency. I wonder if Huggins would really persist in claiming that pause is a sufficient boundary signal?

The paper by Donovan and Darwin deals with the perceived rhythm of speech, with special consideration of the problem of isochrony. Their paper tests, among others, a hypothesis that I had formulated in 1973 and discussed in more detail in 1977. My observation was that listeners tend to hear utterances as more isochronous than they really are, and that listeners perform better in perceiving actual durational differences in non-speech as compared to speech. I concluded from this that isochrony is largely a perceptual phenomenon. Donovan and Darwin have confirmed

these results. They make two points in addition: first, that isochrony is a perceptual phenomenon which is not independent of intonation, and second, that it is a perceptual phenomenon confined to language, reflecting underlying processes in speech production. Donovan and Darwin question the value of seeking direct links between syntax and segmental durations rather than indirect ones by way of an overall rhythmic structure.

While I am in enthusiastic agreement with this particular conclusion, I would like to question the presumed role of intonation in establishing the rhythm of spoken language. There is recent evidence (De Rooij 1979) that intonation contributes very little, if at all, to the temporal structure of a sentence: perception of the temporal structure is not noticeably changed when the fundamental frequency is changed to a monotone. In some unpublished work I found that syntactically ambiguous sentences could not be disambiguated by manipulation of the fundamental frequency, whereas they could be successfully disambiguated by systematic changes in the time dimension. (This latter result has appeared in print: Lehiste, Olive and Streeter, 1976.) If Donovan and Darwin persist in their claim, I would like to hear stronger arguments than have been presented in their paper.

The discussion will be structured as follows. The authors will now have approximately five minutes each to make corrections and additions to their papers. Then we will have a panel discussion, lasting about 30 minutes, during which I hope the authors will respond to some of the questions I have brought up--as well as contribute questions of their own that we will all discuss. The last hour of the session will be devoted to a general discussion with participation from the floor. If there is time, I shall try to verbalize some of the final conclusions that emerge from the discussion.

References

Jakobson, R., C.G.M. Fant, and M. Halle (1952): *Preliminaries to speech analysis*, Cambridge, Mass.: MIT Press (tenth printing 1972).

Lehiste, I. (1973): "Rhythmic units and syntactic units in production and perception", *JASA* 54, 1228-1234.

Lehiste, I. (1977): "Isochrony reconsidered", *JPh* 5, 253-263.

Lehiste, I., J.P. Olive, and L.A. Streeter (1976): Role of duration in disambiguating syntactically ambiguous sentences", *JASA* 60, 1199-1202.