(OSU-CISRC-TR-72-6)

WORKING PAPERS IN LINGUISTICS NO. 12

by

Ilse Lehiste, David Meltzer,
Linda Shockey and Richard Gregorski

Work performed under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43210
June 1972

Foreword

The Computer and Information Science Research Center of The Ohio State University is an inter-disciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report presents research accomplished in cooperation with the Department of Linguistics.

The work of the Center is largely supported by government contracts and grants. The preparation of four of the papers contained in this report was partly supported by the Office of Science Information Service, National Science Foundation under Grant No. GN-534.1.

Ilse Lehiste
Chairman
Linguistics

Marshall C. Yovits
Director
CIS Research Center

Introduction

This issue of Working Papers in Linguistics is devoted to work in the area of experimental linguistics. The five papers included in the current volume were supported, in part, by the National Science Foundation under Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University. They are presented here together as a progress report of research completed in this area since the publication of Working Papers in Linguistics No. 9 (July 1971). It is expected that all papers will be eventually published through normal channels.

The first three of the papers deal more or less explicitly with problems of the temporal organization of speech. The first paper is mainly a survey paper; however, the last third of the paper reports the results of an experiment concerning the perception of syntactic units. The second paper deals with durational patterns characterizing the production of monosyllabic English words whose syllable nuclei consist of vowels and resonants. It is argued that vowels and resonants fuse into syllable nuclei that function as a whole with regard to certain timing rules. The third paper presents some durational data gained from the analysis of repeated productions of Estonian words with contrastive quantity. It should be added at this point that I have some reservations concerning the normalization procedure which I used in the second half of the paper; the statistical interpretation of the results should therefore be viewed with some caution.

The fourth paper concerns the perception of synthetic vowels produced on our Glace-Holmes synthesizer. The last paper in the series deals with the perception of place of articulation cues; theoretical implications of the results of this paper are treated in the first article in this volume. Appendices and figures for a paper can be found at the end of the paper.

I. Lehiste

TABLE OF CONTENTS

The Units of Speech Perception*

Ilse Lehiste

1. Introduction.

Speech perception is a vast topic that might be approached in several different ways. Much interesting work has been done recently with regard to models of speech perception. There is continuing interest in the question of categorical perception and the differences in perception depending on whether or not a listener is responding in the speech mode; related questions involve the role of lateralization in speech processing, and the relationship between speech perception and short-term memory. I have decided to limit the topic to a survey of recent work concerning the units of speech perception. It will occasionally be necessary to relate these units to units of production; likewise, it will be impossible to refrain completely from discussing certain speech perception models. However, I shall not attempt exhaustive coverage of these latter topics; in fact, it will not be possible to achieve exhaustive coverage even of the more limited subject. However, I hope to touch upon some of the more interesting theories and experimental findings at the several levels at which perception units may be established. I shall proceed from the smallest to the largest, starting with the perception of sub-phonemic phonetic differences and concluding with clause- and sentence-level units and their relationship to syntax.

2. The minimal units of speech perception.

2.1. Listening in the speech mode.

One of the problems in trying to establish what constitutes the minimal unit of speech perception is drawing a boundary between the perception of signals in a psycho-acoustic experiment (auditory processing) and the perception of signals in a speech mode (phonetic processing). It is well known that an identical physical stimulus may be perceived in two different ways, depending on the psychological setting. For example, the F_2 transitions of a synthetic CV syllable may sound as chirps of a bird or as glides in pitch, when presented out of context; provided with a following synthetic vowel, they signal the point of articulation of the consonant preceding the vowel (Liberman, 1970). The question is now whether listeners are capable of distinguishing subphonemic

phonetic detail while listening in a speech mode.

One of the characteristics of listening in a speech mode is the so-called categorical perception of phonemes. This means that a listener's ability to discriminate variations in the acoustic cue is much better at the boundary of phone classes than within the phone class (Liberman, Harris, Hoffman, and Griffith (1957); Liberman, Harris, Kinney, and Lane (1961); Stevens, Liberman, Öhman, and Studdert-Kennedy (1969)). Presented with a set of simulated CV- syllables in which F2 transitions are separated by the same frequency intervals, the listener groups the transitions according to the number of distinctive points of articulation employed in his language; within the range, adjacent sounds are classified as 'same', and crossing from one range to another, adjacent sounds are classified as 'different'.

There are some problems with categorical perception. In early experiments, it appeared to work well for consonants, but poorly for vowels. Categorical perception appeared to be associated with a discontinuity in articulation; in the case of vowels, there is no such articulatory discontinuity, which might explain a lack of categorical perception in vowels.

The problem has been recently re-considered by Chistovich and Kozhevnikov (1969-1970). It had been shown earlier (Fry, Abramson, Eimas and Liberman (1962); Stevens, Liberman, Öhman, and Studdert-Kennedy (1969)) that listeners are capable of distinguishing among a large number of stimuli (synthetic vowels) which are classified by them in the same phonemic category. This result could be interpreted in two ways. One interpretation is that phonetic images of vowels form a continuum; in hearing a vowel, the listener 'locates' the stimulus on the continuum by reference to certain articulatory target positions kept in memory. The other interpretation is that a listener is capable of remembering, for a certain time, not only the phoneme which has been selected on the basis of the heard stimulus, but also some spectral characteristics of the sound. If the two stimuli which are being compared prove to be different phonemes, subphonemic spectral information is discarded (Chistovich, Fant, de Serpa-Leitão, and Tjernlund (1966); Chistovich, Fant, and de Serpa-Leitão (1966); Fujisaki and Kawashima (1968)).

2.2. The subphonemic level.

The experiments discussed by Chistovich and Kozhevnikov showed that in certain cases, man is capable of perceiving subphonemic phonetic differences even while listening in a speech mode. This suggests that minimal units of perception may be found at a subphonemic level. A proposal to that extent has been recently made by Wickelgren (1969a, 1969b), who submits 'context-sensitive allophones' as candidates for the role of minimal perceptual units.

Wickelgren claims that sounds are determined by context in such a way that, for example, a /p/ preceded by /a/ and followed

by /i/ is uniquely determined as the kind of allophone that follows /a/ and precedes /i/, and such an allophone of /p/ is different from one that is both preceded and followed by /a/.

There are several problems connected with this model, some of which came up in connection with a recent study by Lehiste and Shockey (1971). In this paper, we explored the perceptual significance of transitional cues in one or the other of the vowels of a VCV sequence that are due to the influence of the transconsonantal vowel. Öhman (1966) had shown that the transitions from the first vowel in a VCV sequence to the intervocalic consonant depend on the quality of the second vowel. Likewise, there are differences in the transitions from the same consonant to the same second vowel that depend on the quality of the first vowel. In our study, we used taped VCV sequences (where V = /i æ a u/ and C = /p t k/) in which either the first or the second vowel was removed by cutting the tape during the voiceless plosive gap. Although the transitional cues were present, and were of the same kind and order of magnitude as those observed by Öhman, the listeners were unable to recover the missing vowels from these modified transitional cues.

According to Wickelgren's model, the context to which allophones are sensitive consists of one preceding and one following sound; thus a following /i/ in an /api/ sequence will not exert any influence on /a/, although it will influence the realization of /p/. The results of the experiment just reported might be considered supportive of Wickelgren's claim; although influence from the second vowel was physically present during the first vowel, that influence was perceptually insignificant. It would seem then that perceptually, the context to which allophones are sensitive is indeed limited to one preceding and one following sound.

There is another possible interpretation: the transitions both to and from the intervocalic consonant are part of the consonant; thus it cannot be claimed at all that V2 has affected V1, even though the transitions from V1 to C have been modified.

The first interpretation is supported by the vowel data, but contradicted by certain consonant data obtained in the same experiment (Lehiste and Shockey (1971)). Perceptually, the influence of the transconsonantal vowel was insufficient to recover the missing vowel; thus allophones seem not to be sensitive to non-contiguous context. However, the first vowel in a V1CV2 sequence is coded, according to Wickelgren's model, as $\#V_c$, the c being the same for different V1's regardless of the quality, or even the presence, of V2. In other words, to take a concrete example, the first /a/'s in /api/, /apa/ and /ap#/ should all be identically coded as $\#a_p$. It seems reasonable to assume that if the context-sensitive allophone is the minimal unit of perception, the context to which the allophone is sensitive should be perceptible. Thus the /p/ should be equally perceptible, i.e. equally recoverable, under all three conditions described above. Our experiments in consonant identification show extensive differences in identifiability between consonants that appear in final position as a

result of elimination of the second vowel on the one hand, or as a result of having been produced by the speaker as unreleased final consonants, on the other. Although the modifications of transitions to an intervocalic consonant due to the quality of a following vowel were not sufficient to recover that vowel, they did have an effect on the identification of the consonant when the second vowel was removed.

The stimuli used in the final consonant identification experiment should have been identical: the left-hand context of the intervocalic consonants and the unreleased final consonants was the same, and the right-hand context was effectively removed by elimination of the releases. If identification was based only on left-hand context, we would have obtained identical scores. Since the scores were considerably different, perception must have been influenced by the anticipatory effect of the right-hand context, manifested within the segment preceding the consonant.

As a digression, I would like to remark that the claim that sounds are not sensitive to noncontiguous context cannot be upheld anyway in the light of historical sound changes. There are numerous processes which affect sounds, e.g. vowels, across intervening consonants and vice versa. For example, in the so-called palatal umlaut that has occurred in Germanic languages, there must have been a stage at which the /a/ of, say, /api/ was clearly distinct from the /a/ of /apa/. Whether the intervocalic consonants were involved or not is a moot question; it is difficult to prove or disprove whether in the Germanic languages the intervocalic consonant was first palatalized and then lost its palatalization after transmitting it to the preceding vowel. There exist instances, however, in which a consonant that is otherwise susceptible to palatalization was not palatalized by a following high vowel under umlaut conditions.

Let us now return to the second possible interpretation: that the transitions are not part of the vowel at all, but part of the consonant. Then the vowel would consist only of the steady state. In principle, if a context-sensitive allophone is the basic unit of perception, the context to which it is sensitive should play a part in perception. In other words, if the transitions are part of the consonant, it should be possible to recover both the preceding and the following consonant in a C1VC2 sequence, given only the steady state of the vowel. We have not run such an experiment, but the recoverability of C1 and C2, in the correct order, from the steady state of the vowel seems implausible considering what is known of the effect of preceding and following consonants on vowel targets. For example, both a preceding and a following /r/ will lower the third formant of an interconsonantal vowel; but given only the steady state, it will not be possible to discover whether the lowering was due to left-hand or right-hand context.

Wickelgren's hypothesis thus seems to be in need of modification. It is clear that the effects of coarticulation reach beyond

contiguous sounds. On the other hand, the context is not always perceptually recoverable. It may be that the 'context-sensitive' allophones fit a production model better than a perception model. The physical modifications are undoubtedly there, but if the context of a context-sensitive allophone is not perceptible, it seems unjustified to assume that context-sensitive allophones are the basic units of perception.

Considering allophones as minimal units of speech perception is one way to approach a level of perception lower than the phoneme. Another is to consider phonemes as "bundles" of distinctive features, and to investigate perception at the feature level. There is no question but that certain features can be perceptually isolated from the "bundles" in which they appear; e.g., voicing can be extracted from the other characteristics of a voiced consonant. The fact that features can be responded to apart from the phonemes to which they belong supports the notion that the brain is capable of parallel processing of incoming information (Miller and Nicely (1955)).

Parallel processing has been discussed in detail in several recent publications (Chistovich and Kozhevnikov (1969-1970); Bondarko, Zagorujko, Kozhevnikov, Molchanov, and Chistovich (1968) (translated by I.L. (1970)); Liberman (1970)). In essence, it means that the same physical signal (e.g. a frequency change in the second formant) carries more than one kind of information (e.g. the phonetic value of a vowel and the point of articulation of an adjacent consonant). A corollary assumption is that it is difficult, if not impossible, to draw precise boundaries between acoustic segments in such a way that the first acoustic segment would contain no information regarding the perception of the second segment, and vice versa.

It will turn out that the first characteristic of parallel processing encourages us to seek the minimal units of speech perception at a level lower (in a certain sense) than traditional allophones, while the second characteristic leads to the conclusion that the smallest units of perception must be located at a higher level--the level of something like a syllable. Let us consider both propositions in somewhat greater detail, and relate them to the role of phoneme-sized units in speech perception.

But first of all I should remark that an assumption of parallel processing would partly save Wickelgren's 'context-sensitive allophones' as minimal units in speech perception: in effect, the perception process could operate with information contained in several time segments, and the problem of non-contiguous influence could be ignored. On the other hand, allophones would lose their unit-like character: their features, perceived separately and in parallel, would not necessarily be coterminous, and instead of phone-like units (which one assumes 'context-sensitive allophones' to be) we would be dealing with something like 'long components' (cf. Lehiste (1967-1970), discussing Harris (1944)).

The question of the perception of sub-phonemic phonetic detail leads back to the question of categorical perception. To the extent that listeners are capable of distinguishing between stimuli falling within the same phonemic category, we are dealing with the perception of sub-phonemic phonetic detail. Reference was made above to the work of Chistovich et al. (1966a, 1966b) which showed that listeners were able to make finer distinctions in vowels than those prescribed by their phonemic system. For evidence of sub-phonemic perception of a suprasegmental feature--duration--I should like to quote Lisker and Abramson (1971). In their experiments with the duration of voice onset time, one of the authors serving as listener distinguished five clear labeling categories, while the phonemic system of English would provide only two.

The differential perception of duration leads to the question of the perception of temporal segments in speech. Several phoneticians have expressed doubt concerning the possibility of perceptual segmentation of speech into units whose duration can be objectively established. It is, of course, known that acoustical signals are largely continuous; nevertheless, they also exhibit some drastic and abrupt changes. The continuous nature of the clues signalling the point of articulation has been used to argue that the minimal unit of perception is a unit of the order of a syllable (for a recent summary, cf. Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967)). On the other hand, continuous speech signals are perceived in ordinary listening as if they consisted of a sequence of discrete units (phonemes). The question is whether the boundaries of these units--or a modified version thereof--can in some way be associated with characteristics of the acoustic patterns. The basic question is thus whether it is possible to segment speech in a perceptually meaningful way.

The obvious place to begin is to consider signals that differ only in the duration of a segment, in such a manner that the differences in duration are not associated with any qualitative differences. The voice-onset-time experiments provide one such condition; they have shown both a possibility of categorical perception (which would serve as evidence for the phonemic level) as well as subphonemic perception (providing evidence for the ability of the ear to analyze duration in a phonetic rather than categorical manner). Further evidence is provided by languages with distinctive quantity.

It is a linguistic fact that in some languages the length of a vowel or consonant may have distinctive function. Experiments with synthetic speech (Lehiste (1970b)) show that listeners agree in a very high degree in assigning linguistic labels to stimuli that differ only in the duration of a vowel or consonant. This implies that listeners are able not only to compare the duration of two stimuli (such as the duration of a voiceless plosive gap), but also to match the stimuli with some kind of 'durational image', an abstract durational pattern characterizing a particular

word type. If a difference in duration of 10 milliseconds can switch 42% of the listeners from one category of linguistic response to another, the difference must be perceptually significant. Obviously it is impossible to tell, during the voiceless plosive gap itself, whether the plosive is qualitatively shorter or longer; the listeners must be comparing durations, which means that they must be using some fixed point of reference. I submit that at least in languages with distinctive quantity, abrupt changes in the manner of articulation serve as reference points with regard to timing judgments.

This is fully in accord with the notion that speech is processed in parallel: whatever the process by which the duration of one segment is compared with that of another (or with a stored 'durational image'), it can very well take place at the same time as the cues for point of articulation are processed which are extracted from the same acoustic signal (e.g. the same vocalic sound). In fact, all suprasegmental information must be processed in a similar way. For example, the presence of voicing serves to establish the voicedness of a vocalic sound at the same time as a possible fundamental frequency change taking place during the voiced segment may signal a distinctive lexical tone. I have discussed the perception of suprasegmentals in detail elsewhere (Lehiste (1967-70); Lehiste (1970a), and shall not elaborate any further on this topic within the present context.

There is additional, somewhat circumstantial, evidence of the importance of the manner of articulation in speech perception. In a study of the perceptual parameters of consonant sounds, Sharf (1971) established seven-point scales for duration, loudness, frequency, sharpness, and contact. Substantial numbers of significant differences were obtained only for duration comparisons based on manner of articulation (and for contact comparisons based on place of articulation; but since the contact parameter was specifically chosen to provide an indication of how well subjects related sounds to place of articulation, the latter finding appears unsurprising). In an earlier study, Denes (1963) showed that manner of articulation carries by far the greatest functional load in the English sound system, and suggested that the acoustic correlates of manner might be used for segmentation in automatic speech recognition systems.

Perception of duration thus appears associated with the perception of manner of articulation. Both represent perception of phonetic detail which may or may not be distinctive. The perception of such phonetic detail serves to substantiate the claim that the minimal elements of speech perception must be located at the subphonemic level, which may thus be considered as established.

2.3. The phonemic level.

The question is now whether the unit next in size is a phoneme-like unit or a syllable. The evidence for the psychological and

perceptual reality of phoneme-like units has been summarized by Chistovich and Kozhevnikov (1970). Savin and Bever (1970) have argued for the "non-perceptual reality" of the phoneme. Let us review the arguments of Chistovich and Kozhevnikov first.

Much of the evidence for phoneme-like perceptual units comes from studies of categorical perception (cf. above). To the extent that the categorical perception idea is valid, the psychological reality of phonemes as perceptual units must be accepted. There is a connection between categorical perception and the motor theory of speech perception; both seem to apply better to consonants than to vowels (or to other signals of a continuous nature) (Liberman (1957); Stevens (1960); Liberman, Cooper, Harris, and MacNeilage (1962); Lane (1965); Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967); Studdert-Kennedy, Liberman, Harris, and Cooper (1970)). Chistovich and Kozhevnikov (1969-70) have shown, first, that vowels are also perceptible in a categorical fashion. Since the articulatory process involved is continuous rather than discontinuous, this would argue against the motor theory. Second, they suggested that the number of categories in vowel perception may be larger than the number of traditional phonemes in the language; and further, that a listener is capable of remembering for a certain time not only a phoneme, but what they call 'timbre description'--subphonemic phonetic detail, which makes it possible to make distinctions within a category. The authors call their perceptual categories 'psychological phonemes'. It has been shown, for example, that Russian subjects classify [i̯] and [i] as different psychological phonemes, although they are never encountered in the same environment and thus may be considered as constituting allophones of a single phoneme. Vowels between hard and soft consonants were classified by Russian subjects as belonging to different sound types, although they would again constitute positionally conditioned allophones according to classical phonemic theory.

Savin and Bever (1970) studied the order in which listeners make decisions at the phonemic and syllabic levels in the course of speech perception. Their method was to ask a listener to monitor a sequence of nonsense syllables for the presence of a certain linguistic unit, either a phoneme or a syllable, and to respond (by releasing a telegraph key) as quickly as possible when he had heard it. The target was a complete syllable (e.g. "bæb", "sæb") or a phoneme from that syllable: the syllable-initial consonant phoneme for some subjects (e.g. /b/ or /s/) and the medial vowel phoneme for other subjects (e.g. /æ/). Subjects responded more slowly to phoneme targets than to syllable targets (by 40 msec for /s-/, 70 msec for /b-/ and 250 msec for medial /æ/). Savin and Bever interpret these results as supportive of the view that phonemes are identified only after some larger linguistic sequence (e.g. syllables or words) of which they are parts. The reality of the phoneme, the authors say, is demonstrated independently of speech perception

and production by the natural presence of alphabets, rhymes, spoonerisms, and interphonemic contextual constraints.

These results do not disprove the existence of a phonemic level of perception, and therefore the title of the paper by Savin and Bever ("The nonperceptual reality of the phoneme") appears somewhat misleading. Before the general conclusion is accepted, one would like to see what the reaction times to final consonants are, i.e. whether subjects would respond more slowly to a final /-b/ than to the syllable /sæb/. While not directly comparable to the reaction time experiments carried through by Fry (1970, to be discussed below), the results of Savin and Bever are sufficiently different from those of Fry to suggest additional studies.

It seems that a level of perception at which phoneme-like units are responded to should be recognized; it remains to relate it to the other levels of perception for which evidence has likewise been provided by studies of speech perception.

3. Higher-level units of perception.

3.1. Unitary perception of sequences of segments.

The parallel processing of speech signals is compatible with the suggestion that the minimal unit of perception must be of the order of a syllable (Savin and Bever (1970); Liberman, Cooper, Shankweiler, and Studdert-Kennedy (1967)). There is a good bit of evidence that the ear is particularly well suited to the perception of changes in acoustic parameters rather than their steady states (Abbs and Sussman (1971)). Without going into details, let me just recall the experience of most researchers who have synthesized isolated vowels: produced on a monotone, the vowels frequently seem to occupy a borderline between speech-like and nonspeech-like stimuli, while the imposition of a fundamental frequency glide shifts the listener clearly into the speech mode. It is also well known that the majority of point of articulation cues of consonants are manifested in adjacent vowels. It seems thus reasonable to look for higher-level units of perception beginning with sequences of two speech sounds. The first major problem involves the perception of sequential order.

Wickelgren's idea of context-sensitive coding could certainly explain the correct perception of sequential order; but the notion of parallel processing, which seemed essential for upholding that theory, appears to be incompatible with the decoding of order from simultaneously received feature cues. The perception of temporal order is a vast topic, deserving a review on its own; I shall restrict myself in this survey to a few recent experiments which shed some new light on the problem.

The mechanisms employed in the perception of consonant clusters have been investigated in a series of experiments by Bond (1971) and Day (1970a, 1970b).

The study by Bond (1971) deals explicitly with the perceptually unitary nature of consonant clusters. Bond studied 15 pairs of English words which differed from each other only in the order of obstruents in the cluster. The pairs /ps-sp/, /ts-st/ and /ks-sk/ were all represented five times (some examples: task-tax, lisp-lips, coast-coats). The words were produced by a male native speaker of English; randomized listening tests were constructed, in which the signal was degraded by addition of white noise. 19 subjects took the listening test, writing down what they heard. Five of the subjects took the test a second time, producing a spoken response (a repetition) to each stimulus. These subjects' responses were analyzed for reaction time in addition to being scored for correctness. It was found that reaction time was consistently faster for correct than for incorrect responses; but the pattern of confusions for written responses and spoken responses was essentially the same. It was further found that reversal errors were the most common errors. Bond argues from this that minimal perceptual units must be larger than the phoneme. If consonant clusters were perceived phoneme by phoneme, there is no reason for the listener to reverse the order. To be sure, the listener may occasionally be forgetful; but there is no reason to suppose that he would be more likely to forget the order of the consonants than to forget one of the consonants. Since reversal errors were much more common than substitution errors, some special perceptual mechanisms must be postulated for the perception of consonant clusters. Bond's findings thus confirm a suggestion made by Neisser (1967), according to which a listener gradually learns to distinguish a cluster like /ts/ from a cluster like /st/, rather than perceiving a sequence of /t/ followed by /s/, or /s/ followed by /t/. Clusters of this type thus seem to constitute a perceptual unit.

Day (1970a) studied phonemic fusion in dichotic listening, in which listeners received two speech stimuli at the same time with various relative onset times. The stimuli differed in their initial consonants (e.g. /bæŋkət/ and /læŋkət/). On some trials, either /bæŋkət/ or /læŋkət/ led by 25, 50, 75, or 100 msec; on other trials, both stimuli began at the same time. Subjects reported hearing /blæŋkət/ regardless of which consonant led. When specifically asked to judge the temporal order of the initial phonemes, most subjects reported hearing /b/ first, no matter whether /b/ or /l/ actually led. Day concludes that instead of processing temporal order in an accurate fashion, subjects responded to the stimuli according to the constraints imposed by the phonological system of English. In English, stop + liquid clusters are permissible in initial position, but liquid + stop clusters do not occur. The responses thus clearly imply the presence of a linguistic level of processing.

A similar study was carried out with reversible clusters (Day (1970b)). Since there are no reversible clusters in English in initial position, a final cluster was selected. The stimuli

were /tæs/ and /tæk/, whose fusion would yield acceptable English words in either order, viz. /tæsk/ and /tæks/. All trials were dichotic pairs, consisting of /tæs/ to one ear and /tæk/ to the other ear. The onsets of the syllables were aligned over a wide range of values: stimuli either started at the same time, or one or the other stimulus led in steps of 5 msec to a 100 msec lead.

In contrast with the nonreversible case, temporal order judgment was very good when the cluster could occur in either order in the language. One of the temporal orders (/ks/) was somewhat more preferred. Day suggests that this may be due to the fact that the acoustic shapes of stop consonants undergo greater changes as a function of context than do fricatives; thus the acoustic shape of /k/ in /tæk/ may be more important than that of the /s/, to the extent of biasing the perceived order of the two phonemes. (I would suggest that segmental duration may have played a perhaps decisive part. The stimuli were synthesized with equal duration given to /æ/ in both /tæk/ and /tæs/. In actual speech, /æ/ would be longer before a fricative; thus listeners may have been biased toward a /tæks/ response by the relative shortness of the /æ/).

In a further experiment, subjects were asked to decide which ear led, rather than which phoneme. Performance on the ear task was much better: subjects were highly accurate, even though they were language-bound on the phoneme task.

The difference between the results obtained with nonreversible and reversible clusters is explained by Day as follows. Two general levels of processing are postulated: a linguistic level and a nonlinguistic level. Both operate in normal listening situations, but the linguistic level appears to be prepotent: it can effect selective loss of information obtained from the nonlinguistic level. Correct temporal order may be represented in the system at some point in time, but later stages of processing mold this information to conform to the linguistic structure of the language. Hence nonlinguistic information, concerning acoustic shape and temporal order, may be lost or ignored. Day suggests that temporal order information is lost only after it enters higher stages of linguistic processing.

3.2. Primary processing and linguistics processing.

Day called the two levels of speech processing which her experiments had isolated linguistic and non-linguistic. It appears, however, that both levels have to be further subdivided. Even at the non-linguistic level, there is a difference in perception depending on whether one is listening in the "speech mode". Evidence for this is available from many sources, among which are laterality studies (Studdert-Kennedy and Shankweiler (1970); Day and Cutting (1970)). I would like to call the processing of an auditory signal in the speech mode "phonetic processing". Attempts to separate auditory and phonetic modes of

processing have been recently discussed by Fujisaki and Kawashima (1969) and by Pisoni (1971). The linguistic level suggested by Day could perhaps be called the phonological level of speech processing. At this level, information available to the listener about the phonological structure of the language (e.g. information concerning permissible sequences) is interposed between primary recognition and perceptual decision. The experiments of Chistovich et al. (1966a, 1966b) regarding the mimicking and perception of vowels show the possibility of separating the phonetic and phonological levels of perception, as do the experiments in the perception of reversible and non-reversible clusters by Day.

There are higher levels within the linguistic level of processing, and some attempts have been made recently to explore them experimentally. A very intriguing set of experiments by Fry (1970) deals with reaction time to monomorphemic and bimorphemic words that are identical as to their phonemic composition. Fry used the minimal pair lacks/lax, serving both as speaker and listener. Responding 100 times to the randomized stimuli, he made only 2 wrong responses to 50 occurrences of lax, and likewise only two errors in responding to lacks--a result surprising to Fry, who had not expected a subject to be able to respond consistently to the difference between the two items. The mean reaction times were 557 msec for lax and 518 msec for lacks, a difference that just misses significance at the .05 level of probability. Fry considers it worth noting that the direction of the difference points to a longer reaction time to the monomorphemic word.

Fry also tested the reaction time to longer sequences differing in the presence and absence of a word boundary. The items were the two sentences It's a sign of temporizing and It's a sign of temper rising, which are segmentally identical in Fry's pronunciation. There were six errors in the perception of 50 presentations of temporizing and 3 in the case of 50 presentations of temper rising. Mean reaction times (measured from the beginning of the syllable /tem/ in each case) were 711 msec for temporizing and 858 msec for temper rising, a difference which was significant below the .01 level of probability. The item containing the word boundary thus took significantly longer to produce a response, although the difference in duration between the two items was negligible (30 msec in a total of 1430 msec).

Fry's starting assumption had been that processing time increases with the complexity of the task. The results of the experiment with sentences support this view; the two sentences differ in their syntactic structure, and it is quite probable that the syntactic level of processing was involved in addition to primary processing. However, the results of the lax - lacks experiment seem to imply that a monomorphemic word presents a more complex task than a bimorphemic one. This appears counter-intuitive; and there might be alternative explanations to Fry's

findings. If the results should be substantiated by further experiments, it might be assumed that a bimorphemic word contains more information than a monomorphemic word and therefore can be processed faster. If additional data should show that the effect observed by Fry may have been due to chance, it might be concluded that there exists no separate morphemic level of linguistic processing.

Such experiments were in fact carried through by Bond (1971). Bond used ten minimal pairs, each pair consisting of one monomorphemic and one bimorphemic word of the same phonemic shape. Each pair of words composed a sub-list, within which the two words were recorded in random order, each word being produced ten times. Care was taken to insure that the speaker intended the 'right' word every time. 29 listeners took the test, which consisted of 200 stimuli. Reaction times and correct scores were obtained by techniques similar to those used by Fry.

The overall scores indicated that subjects were not able to identify the words correctly at levels significantly above chance. The mean scores ranged from 45.1% for lax - lacks to 55.4% for lapse - laps. When the responses of the subjects to each production were analyzed, however, it was found that subjects were very consistent in their responses to some of the test items. Significant scores (at the .02 level) were obtained for three items in the 20 productions of members of the pair bard - barred (15.4%, 84.6% and 15.4% correct), and one item each in the pairs wade - weighed (100% correct), lax - lacks (18.2% correct), baste - based (85.7% correct) and mist - missed (100% correct). As the scores show, while the subjects could be highly consistent in agreeing on a particular response, they did not necessarily identify the word correctly; the identification scores for utterances on which the subjects agreed on one response were still at chance level (57% correct).

There was no significant systematic difference in reaction time between correct and incorrect responses. There was, however, some tendency for reaction time to be shorter to the bimorphemic word, as Fry had discovered; the differences were not statistically significant.

This cannot be considered supportive of Fry's findings, because reaction time differences become meaningful only if the subjects can identify the words correctly, which was not the case with Bond's subjects. Bond explains the high degree of agreement shown by the subjects in response to some of the stimuli as follows. Faced with the task of the experiment, listeners develop a strategy for making use of fine phonetic detail (duration, spectral characteristics of /s/ etc.). In this manner they arrive at some consistent labelings. But since the identifications based on this strategy are equally likely to be correct or incorrect, the strategy cannot be considered to be part of ordinary speech perception.

Within the framework developed in this paper, I would propose that we are dealing with phonetic processing rather than linguistic

processing. The perception of fine phonetic detail is certainly documented by Bond's results, but this information plays no part in establishing a possible morphological level within linguistic processing.

While the morpheme level evidently has to be rejected as a level of processing within the level of linguistic processing, it might be inquired whether a word constitutes a perceptual unit at some level. Fry's reaction time experiments provide some evidence that the word is certainly not the minimum unit of perception. In testing reaction times to 18 contrasts like bid-big, or begin-began, Fry found that in only three cases did the mean reaction time exceed the total duration of the stimulus. In most cases, subjects had no difficulty whatever in responding before a word or syllable were complete. The processing mechanism was evidently capable of dealing with segments smaller than the whole syllable or word.

Whether the word constitutes a perceptual unit does not emerge from Fry's experiment with sentences containing the items temporizing - temper rising, since in examples of this kind it is impossible to separate lexical differences from syntactic ones. However, certain techniques have been developed within the past ten years for studying the perception of syntactic units, and the rest of the paper will deal with perception at this level.

3.3. Perception of syntactic units.

To a large extent, recent studies of sentence-level perceptual units go back to a seminal paper by Ladefoged and Broadbent (1970). In the research on which the paper is based, Ladefoged and Broadbent presented a series of tape-recorded sentences to various groups of listeners. During each sentence, a short extraneous sound (a "click") was present on the recording, and listeners had to indicate the exact point in the sentence at which the click occurred. Errors were large compared to the duration of a single speech sound; Ladefoged and Broadbent concluded that the basic unit of perception is larger than a phoneme, and that the listener does not deal with each sound separately but rather with a group of sounds. Subjective location of clicks, as reported by the subjects, differed from their objective location according to a regular pattern; Ladefoged and Broadbent argue that the points toward which the clicks were displaced constituted boundaries of perceptual units.

Fodor and Bever (1965) used the same technique to investigate the hypothesis that the primary units of speech perception correspond to the constituents of which a sentence is composed, i.e. the more abstract segments revealed by a constituent analysis of the sentence provided by the grammar of the language. Fodor and Bever found that clicks were attracted toward the nearest major syntactic boundaries in sentential material. The number of correct responses was significantly higher in the case of clicks located objectively at major boundaries than in the case of

clicks located within constituents. Fodor and Bever consider these results supportive of the view that the segments marked by formal constituent structure analysis do in fact function as perceptual units, and that the click displacement is an effect which insures the integrity of these units: the units resist click intrusion.

In a subsequent study, Garrett, Bever and Fodor (1965) attempted to determine whether the earlier results should be interpreted as reflections of the assignment of constituent structure during the processing of sentences, or were rather effects of correlated acoustic variables (such as pause and intonation) which tend to mark constituent boundaries in spoken language. They constructed and recorded pairs of sentences for which some string of lexical items was common to each member of a pair. The common portions of each pair were made acoustically identical by cross-splicing, i.e. by splicing a recorded version of a portion of one member of the pair to the opposite member of the pair. When a spliced version is paired with a copy of the original recording,

(Example: A. (In her hope of marrying) (Anna was surely impractical)
B. (Your hope of marrying Anna) (was surely impractical).)

there are two sentences in which part of the acoustic material is identical, but for which the constituent boundaries are different. The results showed that exactly the same acoustic signal was responded to differently in every case, and the differences were uniformly as predicted by the intended variation in the constituent structure.

Bever, Lackner and Stolz (1969) further tested the hypothesis that the perceptual segmentation of speech depends on transitional probabilities. The fact that clicks are subjectively located at boundaries between clauses might be a reflection of the low transitional probability between clauses rather than a demonstration that syntactic structure is actively used to organize speech processing. In this experiment, subjects were asked to indicate the subjective location of clicks placed in sentences which differed in terms of transitional probabilities between clauses. It was found that high-probability sequences within clauses attract clicks, while low-probability sequences do not. The authors interpret these results as indicative that transitional probability has different effects within and between clauses and thus is not a general mechanism for the active segmentation of speech.

In another set of experiments, Bever, Lackner and Kirk (1969) found that within-clause phrase structure boundaries do not significantly affect the segmentation of spoken sentences, and that divisions between underlying structure sentences determine segmentation even in the absence of corresponding

clause division in the surface phrase structure.

In most of these studies, subjects were ostensibly involved in only one task, namely click localization; but in fact they were performing a far more complex assignment. They had to listen to a sentence, pay attention to the click, remember the sentence, write it down, remember the click location, and mark that on the written version of the sentence. The sentences were usually quite long; it seems obvious that we are dealing here with a complex interaction of perception and memory. Techniques used up to this point did not attempt to separate the effects of memory and perception.

Abrams and Bever (1969) attempted to minimize the effects of memory by giving the subjects a different task: pressing a key in response to a click. In a second presentation of the test sentences, subjects had to write the sentences and locate the click as before. Reaction times were thus obtained in addition to click localization data.

The results turned out somewhat ambiguous. Abrams and Bever had expected that clicks objectively occurring in clause breaks should receive faster reaction times than clicks in any other location. This turned out not to be so. There was also no systematic interaction between reaction time and subjective click location. Reaction time to clicks before clause breaks was affected by clause length and by familiarity with the sentence more than the reaction time to clicks after clause breaks. According to Abrams and Bever, this indicates that syntactic structure does systematically modify attention during speech perception. In sentences, the clause is a natural unit for internal perceptual analysis. During clauses one listens to the speech and nonspeech stimuli; at the end of clauses one encodes perceptually what was just heard. Accordingly, a click at the end of a clause is responded to relatively slowly, since it coincides with the point of internal perceptual analysis of the preceding sentence. At the beginning of a clause, a click is reacted to quickly because it conflicts with relatively little internal perceptual processing.

Abrams and Bever suggest further that the attentional system tapped by the reaction-time measure is distinct from the behavioral process which produces the systematic errors in click location. Immediate reaction time interacts with the process of developing the internal perceptual organization of speech. Listeners first organize the speech into major segments, then they relate the speech and click temporally. It is this latter process that maintains the integrity of the speech units as revealed in the location of clicks.

In another study, Bever, Kirk and Lackner (1969) tried to avoid conscious participation of the listeners altogether by measuring their galvanic skin response to shocks. In this experiment, subjects heard sentences in one ear, during which a brief shock was administered before, in or after the division between two clauses. The galvanic skin response to shocks

objectively at the end of a clause was larger than the response to shocks at the beginning of a clause. Bever, Kirk and Lackner view this as confirmation of the hypothesis that the syntactic structure of a sentence can influence systematically the change in skin resistance in response to a mild shock presented during the sentence.

An independent effect was that galvanic skin response to shocks at the end of a clause decreased as a function of clause length; responses to shock at the beginning of a clause were relatively unaffected by the length of the preceding clause. According to the authors, this supports the claim that listeners respond to the syntactic structure of speech as they hear it.

Fodor and Garrett (1971) revised the earlier view that click location is affected only by major constituent boundaries. Under appropriate conditions (when a listener is given more than the usual amount of time to consider a sentence), minor boundaries were found to affect click location. Fodor and Garrett suggest that assignment of minor constituent boundaries is a relatively late operation in the processing of sentences. If the listener has a chance for developing a more fine-grained analysis of the sentence containing a click, effects of minor constituent boundaries on click location are increased.

The series of studies just reviewed thus presents the following claims: listeners use grammar actively to impose syntactic structure on the speech stimulus as they hear it. Listeners respond in terms of the underlying structure of the sentence rather than its surface structure. Acoustic cues alone do not determine the boundaries of perceptual units.

Certain of these findings have been challenged in several recent studies. Abrams and Bever (1969) had found that subjects did not react faster to clicks placed in major constituent breaks than to clicks within the constituents. Holmes and Forster (1970) found exactly the opposite: reaction times to clicks at the major syntactic break of the sentence were faster than reaction times to clicks not at a break. This confirmed their hypothesis that processing load is a function of the surface structure of sentences, and that it decreases at major constituent boundaries.

The second result of the study by Holmes and Forster is likewise in direct contrast to the findings reported by Adams and Bever: reaction times were slower when the click was in the first rather than in the second half of the utterance. Holmes and Forster interpret this result likewise in terms of differential processing loads. It is obvious that these results place in question the conclusions drawn by Abrams and Bever from their data.

Chapin, Smith and Abrahamson (1972, in press) produce a detailed critique of Bever, Lackner and Kirk (1969) who had claimed that underlying structure sentences are the primary units of immediate speech processing. Chapin, Smith and Abrahamson found that clicks were attracted to major surface constituent boundaries, even when these did not coincide with the boundaries

of underlying structure clauses. Another finding was that clicks are attracted to preceding constituent boundaries. This suggests an overriding perceptual strategy in speech processing: the listeners attempt to close constituents of the highest possible level at the earliest possible point.

Bond (1971) studied both click localization and reaction time, testing the hypothesis that subjects segment an incoming sentence on the basis of stress and intonation patterns. Reaction time is then predicted to be shorter to clicks between phonological phrases, and longer to clicks within phonological phrases; it is also expected to be different to clicks located in stressed syllables, as compared to clicks placed in unstressed syllables.

When reaction time to clicks in stressed and unstressed syllables was compared, it was found that reaction time was significantly faster to the click located in an unstressed element, either in the consonant preceding the unstressed vowel or in the unstressed vowel itself. Subjects were much more accurate in locating a click when it occurred in a stressed vowel than when it occurred in a consonant or in an unstressed vowel (correct scores 46% vs. 12%). Clicks were thus much less likely to be 'attracted away' from stressed vowels than from unstressed vowels; the error responses, however, were in the direction toward major boundaries.

Reaction time was also examined on the basis of an 'intonation phrase', i.e. any phrase that was demarcated by a clear intonation curve. Reaction time was found to be progressively slower as the click occurred further into the intonation phrase; thus there is a correlation between reaction time and the position of the click within an intonation phrase.

Bond suggests that in sentence perception, the listeners segment the sentence into phrases defined on the basis of stress and intonation; they then process the sentence further, to arrive at a syntactic analysis. Reaction time is apparently sensitive to initial segmentation, while click localization is sensitive to the final analysis.

3.4. The Role of Stress in the Perception of Sentence-Level Units.

Bond's study did not attempt to separate the parts played by stress and intonation. I conducted an experiment, described below, to investigate further the role of stress in click localization.

The purpose of this experiment was to explore the role played by suprasegmental features, especially stress, in the analysis of an incoming sentence. If the assumption is true that linguistic processing presupposes phonetic processing, it stands to reason that stress and intonation are not ignored by a listener in the perception of a sentence. This, as may be recalled, has been more or less generally assumed since the 1965 paper by Garrett, Bever and Fodor (cf. above).

It was decided to place clicks in identical positions within a sentence, varying the stress in such a manner that the words

within which clicks occurred would appear both with and without stress, all other factors being equal. If listeners react differently to clicks placed in the same position under different stress conditions, the role of suprasegmental factors in perceptual processing will be confirmed.

In order to control stress and click placement precisely, the experiment was carried through with synthetic speech. The stimuli were produced at the Bell Telephone Laboratories using the following technique. A normal utterance was analyzed by a formant-tracking program (Olive (1971)). The automatically tracked formants and fundamental frequency were later modified by hand; changes in time, formant structure, and fundamental frequency were produced by a suitable computer program. The program allows the researcher to specify the frequencies of the three formants, the fundamental frequency, and the overall amplitude at each 10 msec sampling period. Specific changes that were made will be described below. The re-synthesis was produced by a digital hardware synthesizer (Rabiner et al. (1971)). The entire process was controlled by a Honeywell DDP 224 computer (Denes (1970)).

The experimental technique used in the experiment differs from earlier methods in several ways. In most previous experiments, clicks had been recorded on the second channel of a two-track tape recorder, and the stimuli had been presented to listeners dichotically through headphones. Dichotic presentation introduced into the experimental situation a whole array of complicating factors, including competition between speech and nonspeech in relation to hemispheric specialization (Day and Cutting (1970)), and the problem of right- or left-handedness of the subjects. To avoid these probably unnecessary complications, the stimuli were recorded on full-track tape, with clicks introduced synthetically within the recording, and were presented to listeners over a good-quality loudspeaker in a sound-treated environment. It should be recalled that Ladefoged and Broadbent had likewise used a loudspeaker in their original experiment reported in 1960.

In most earlier experiments, listeners were required to write down the sentence that had been presented, and to indicate the position of the click on their own transcription. As was mentioned above, this technique introduces a memory component into the picture whose magnitude is difficult to estimate. It has been known for some time that the human short-term memory has a capacity of something like seven syllables (Miller (1956)). Memory units have been studied intensively by Johnson (1970), who found the 'chunks' of recall to be approximately the same size. In a recent paper, Gamlin (1971) has shown that subjects matched for intelligence may differ in their short-term memory capacity, and that low short-term memory subjects structure sentences differently than high short-term memory subjects. Gamlin suggests that low short-term memory subjects may be forced by their memory limitations to process sentences into smaller syntactic units.

Clearly most of the test sentences used in earlier click experiments have been long enough to overtax the short-term memory; thus it is entirely possible that the results confuse the sentence processing strategies with memory strategies.

The way chosen to eliminate the memorization problem was to use only one sentence with which the listeners became familiar during the introduction to the test, and to provide the subjects with written versions of the sentence. This represents again a return to the Ladefoged-Broadbent (1960) technique. In that study, subjects were presented both with unknown sentences over headphones, and with sentences that were written out and read out before the stimuli that contained the clicks were played over a loudspeaker. Ladefoged and Broadbent found that prior knowledge of the content of the sentence did not affect accuracy.

The sentence chosen for the experiment was one used by Bever, Lackner and Kirk (1969) in the experiment which provided the basis for their claim that the underlying structures of sentences are the primary units of immediate speech processing. The sentence, together with the phrase structure assumed by Bever et al., is as follows:

If (you (did ((call up) Bill))) (I (thank you (for (your trouble))))

Bever et al. placed clicks in the major clause break, in the middle of each of the two words immediately preceding the major break, and in the middle of each of the two words immediately following the major break. Separate results are not reported for this sentence, but one may assume the general conclusions to be applicable, i.e. that the boundary after Bill attracted clicks, while boundaries within the two clauses had no consistent effect on the subjective location of clicks.

The sentence was synthesized by the procedure described above. The sentence was produced by a male speaker with no special emphasis on any word and without any pauses. After resynthesis, the pitch of the sentence was changed to monotone at 100 Hz. Stress was then simulated on each of the four words did, Bill, I, and thank. This was done by time expansion and by introducing a pitch inflection on the appropriate word. The values of the parameters are specified by the program at 10 msec intervals. In time expansion, the number of sampling intervals is specified to which a given word is to be expanded, and the program interpolates the values of the parameters proportionately. The expansion factors had been obtained previously by comparing the durations of stressed and unstressed versions of the test words in different productions of the sentence; they were 25/33 for did, 32/58 for Bill, 16/34 for I and 31/42 for thank.

The fundamental frequency contour applied to the test word started at 100 Hz, rose to a peak of 111 Hz, and dropped back to 100 Hz. The peak of the contour was placed at the point of occurrence of the fundamental frequency peak in a normal stressed production.

Clicks were produced by setting formant frequencies to 1 for one sampling period and introducing random noise through the formants at an intensity equal to that of the strongest vowel. The duration of the clicks was 10 msec. Clicks were placed before, within and after each of the four words; the clicks within words were located at the pitch peak. With the method of time expansion used in the study, the clicks remained in precisely the same position relative to the word under both stress conditions. A table of click placements and stress conditions is given below.

TABLE 1
SURVEY OF CLICK PLACEMENT

Stressed word	Test word	Click placement relative to test word		
		Before	Within	After
Did	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x
Bill	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x
I	did	x	x	x
	Bill	x	x	x
	I	x	x	x
	thank		x	x
thank	did	x	x	x
	Bill	x	x	x
	I		x	x
	thank		x	x

Two comments should be added. In order to simulate stress on I, a glottal stop (with a duration of 17 sampling periods, i.e. 170 msec) was inserted before I. In the sentence in which I carried simulated stress, two click placements were used for the sequence Bill, I: a click was placed in the last frame of Bill, immediately preceding the glottal stop, and in the first frame of I, immediately following the glottal stop. In other instances, only one click placement was used between words. This is true also of sequences of Bill, I (i.e. the major clause break) in all other cases in which I was not stressed, including those in which Bill carried simulated stress.

The first part of the listening test was designed to check the effectiveness of the stress simulation. A set of ten randomized sentences was prepared, containing two productions

each of the test sentence produced on a monotone (and without time expansion, i.e., without stress simulation), and two sentences each with stress placed respectively on did, Bill, I, and thank. (The sentences contained no clicks.) The listeners were asked to underline the stressed word. The results are presented in the following table.

TABLE 2
SUBJECTIVE PLACEMENT OF STRESS, DEPENDING ON STRESS SIMULATION
Scores in per cent

	If	you	did	call	up	Bill,	I	thank	you	for	your	trouble.
Monotone	4	4	12	16	16	28		10				10
Stress on did		4	90	2				4				
Stress on Bill			6	2	2	84		6				
Stress on I			2	2		4	92					
Stress on thank	2		2			2		94				

As may be seen from the table, the syllables on which simulated stress was placed were overwhelmingly accepted as being stressed. The neutral sentence provided two surprises. I had expected the word did to be judged as stressed, since it is lexically marked as emphatic; however, there was a wide scatter of responses, and the word judged relatively most frequently as stressed was the word Bill. It will be reported later that this word behaved in an unexpected way in other respects too. Whether its position before the clause break is in any way connected with this behavior has to remain a matter of conjecture; further experimentation is clearly needed to solve the problem.

After the first part of the test, some examples of sentences containing clicks were played to the listeners, and instructions were given to draw a slash line through that part of the sentence that contained the click. Subjects were informed that clicks may occur between words or within a word. Sample sentences with slashes were provided on the handout. The subjects then proceeded to the main part of the test, which contained the 41 stimuli in two different randomizations (for a total of 82 stimuli), balanced in such a way that each stimulus occurred once during the first half and once during the second half of the test. The whole test took approximately twenty minutes to complete. The test was

administered singly or in small groups to 25 listeners, mainly graduate students and staff members of the Department of Linguistics of the Ohio State University. The results consist of 50 judgments per stimulus, for a total of 4100 judgments. The results of the listening tests will be presented with reference to Tables 3, 4, and 5. The question of correct identification will be discussed first.

The evidence for the listener's analysis of the sentence in terms of underlying structure units had been largely derived from subjective localization of clicks at major syntactic boundaries. Specifically, it had been claimed that clicks objectively at such boundaries were correctly located more frequently than clicks placed elsewhere, and that clicks placed elsewhere had a strong tendency to migrate toward the major syntactic boundaries. This experiment contained sentences in which clicks were placed at various boundaries, including the major clause boundary. The per cent correct identification of click location at various boundaries was as follows:

```
If (you   ( did (  ( call up )    Bill )))    ( I   (thank
      24.0      40.0         51.5         27.6   41.5     16.5

    you (for (your trouble ))))
```

The total number of clicks correctly identified between Bill and I was 69 out of a possible 250 (5 sentences), or 27.6%. The total number of clicks objectively placed in the boundary, but subjectively shifted elsewhere, was 181, or 72.4%. Most of these clicks were attracted into the following word, i.e. into I. When I was unstressed, it attracted 37 clicks away from the boundary (from 150 possibilities, 3 sentences), and when it was stressed, 71 (from 100 possibilities, 2 sentences). As far as attracting clicks objectively located elsewhere, there were 150 such cases out of a possible 1800 (36 sentences), which amounts to 8.3%.

It must be concluded that the results of this experiment do not support the claim that the major syntactic boundary attracts clicks.

Table 3 presents the average correct scores for the subjective location of clicks objectively placed in stressed and unstressed productions of the words did, Bill, I, and thank. The unstressed scores combine stresses on the three other words; e.g. unstressed did combines scores for instances in which stress was simulated on Bill, I, and thank. A study of the scores reveals a number of regularities. There is a common pattern for the words did, I, and thank, while Bill shows a highly divergent pattern. Table 4 gives the average scores of the three words with similar behavior.

TABLE 3
CORRECT SCORES (PER CENT)

Word	Objective click placement		
	Before	Within	After
did, stressed	16.0	68.0	62.0
did, unstressed	26.7	56.0	32.7
Significance of difference*	> .10	> .10	< .01
Bill, stressed	56.0	38.0	24.0
Bill, unstressed	50.0	64.7	32.7
Significance of difference	> .10	< .01	> .10
I, stressed	16.0	40.0	66.0
I, unstressed	34.7	24.0	33.3
Significance of difference	< .05	< .10	< .001
thank, stressed	28.0	78.0	22.0
thank, unstressed	46.0	43.3	14.7
Significance of difference	< .05	< .001	> .10

*See Spiegel (1961, p. 171).

TABLE 4
CORRECT SCORES FOR DID, I, AND THANK (IN PER CENT)

Word	Objective click placement		
	Before	Within	After
Stressed	20	62	50
Unstressed	35.8	41.0	26.9
Significance of difference	< .10	< .05	< .001

In unstressed versions of did, I, and thank, clicks placed before the word tended to be identified more correctly than clicks placed in analogous position in stressed words. The difference is significant at the .10 level. Clicks within and after stressed words were identified more accurately than within and after unstressed words. This, too, is a significant difference, with the significance increasing from the .05 level

for position within the test word to the .01 level for position after the test word. The word Bill, however, shows the opposite result. In the case of Bill, the relationships between the scores are reversed, although only the difference between the scores for position within stressed and unstressed versions of Bill reaches significance (at the .01 level).

The various kinds of subjective shifts are shown in Table 5.

TABLE 5
CLICK PLACEMENT AND CLICK LOCATION IN STRESSED AND UNSTRESSED WORDS
(per cent)

Objective click placement	Subjective click location				
	Within preceding word	Before test word	Within test word	After test word	Within following word
Before did, stressed	8.0	16.0	46.0	8.0	2.0
Within did, stressed		10.0	68.0	16.0	6.0
After did, stressed		2.0	20.0	62.0	14.0
Before did, unstressed	11.3	26.7	50.7	2.0	0.7
Within did, unstressed	2.0	14.0	56.0	16.7	6.0
After did, unstressed	0.7	6.7	20.0	32.7	16.7
Before Bill, stressed	10.0	56.0	26.0	4.0	
Within Bill, stressed			38.0	48.0	6.0
After Bill, stressed		2.0	8.0	24.0	16.0
Before Bill, unstressed	9.3	50.0	26.0	8.7	0.7
Within Bill, unstressed	0.7	2.7	64.7	25.3	4.0
After Bill, unstressed			7.3	35.3	45.3
Before I, stressed	2.0	16.0	64.0	8.0	8.0
Within I, stressed		2.0	40.0	44.0	12.0
After I, stressed		2.0	20.0	66.0	12.0
Before I, unstressed	10.0	34.7	24.7	6.7	6.7
Within I, unstressed	2.0	17.3	24.0	22.0	28.0
After I, unstressed	1.3	9.3	3.3	33.3	50.7
Before thank, stressed	2.0	28.0	64.0	2.0	
Within thank, stressed	2.0	4.0	78.0		
After thank, stressed		2.0	36.0	22.0	20.0
Before thank, unstressed	9.3	46.0	33.3		
Within thank, unstressed	6.6	29.3	43.3	4.0	
After thank, unstressed	3.3	18.0	53.3	14.7	4.7

Study of this table explains why clicks preceding stressed words received low correct scores: there is an overwhelming tendency for such clicks to be subjectively located within the stressed word. To put it differently, stress attracts the click from the preceding boundary into the stressed word. For did, correct identification of a click before the test word was 16%, compared to subjective shifts in 46% of the cases; for I, the 16% correct location of the click occurring at the boundary contrasts with a 64% shift into the stressed word, and for thank, 28% correct contrasts with a 64% shift. The subjective shift in the case of I is particularly noteworthy, since it involves a shift away from the major syntactic boundary, which supposedly attracts clicks and certainly should resist their being attracted away. Table 6 shows the level of significance of differences in scores due to some of the shifts.

TABLE 6
DEGREE OF SIGNIFICANCE OF SUBJECTIVE SHIFTS

Objective Click Placement	Subjective shift (by one-half step) to	
	Within test word	Within following word
Before did, stressed	< .01	
After did, stressed		< .001
Before did, unstressed	< .01	
After did, unstressed		> .20
Before Bill, stressed	< .001	
After Bill, stressed		> .20
Before Bill, unstressed	< .01	
After Bill, unstressed		> .20
Before I, stressed	< .001	
After I, stressed		< .001
Before I, unstressed	> .20	
After I, unstressed		< .05
Before thank, stressed	< .001	
After thank, stressed		> .20
Before thank, unstressed	< .20	
After thank, unstressed		> .20

Table 6 requires some interpretation. It is obvious that the shifts from before a stressed word into the stressed word are highly significant. In some instances, shifts from after the

test word to the following word are also significant; but failure to shift is equally important. This is not shown directly on this table, but can be realized by comparing Table 6 with Table 5. For example, the probability that a click objectively placed after stressed did would be attracted into the following word is exceedingly small; the reason is the high accuracy of click location in that position in general, and the fact that no stressed word ever followed did. It is the stressed words that attract preceding clicks; there was no comparable systematic tendency for clicks to be subjectively shifted from a preceding boundary to the middle of an unstressed word.

As regards the word Bill, the degree of significance shows the failure to shift in both cases in which the click was placed before the word.

Clicks objectively placed within a stressed word receive high correct scores and show little tendency to shift away. This tendency is greater in unstressed words. The direction of these shifts is not systematic in any way.

Clicks placed after stressed words are highly identifiable. If they migrate, it is toward the following word. The tendency to shift into the following word is much more pronounced in the case of clicks placed after unstressed words. After Bill and I, in particular, the click was subjectively shifted to the following word more frequently than it was correctly located. Interestingly, this is the only instance in which unstressed Bill shares the behavior of other unstressed words; in all other respects, it seems as if stress and lack of stress were reversed in the case of Bill. The reason why clicks are not shifted to the following word after unstressed did and thank is most probably the lack of stress on the words immediately following the click.

Except for the matter just described, no particular regularities seemed to be associated with the position of the word relative to the beginning or end of the sentence. The behavior of clicks associated with Bill remains a problem calling for further study.

The results of the experiment demonstrate that stress does indeed have an effect on the subjective location of clicks. Without trying to read too much into the outcome of the limited experiment, I feel justified in saying that click localization is more sensitive to surface phenomena than as been previously assumed. The underlying structure of the sentence remained the same during the experiment; if the listeners somehow proceed directly to the analysis of underlying structures, clicks should have been treated similarly in the same words, regardless of their stressed or unstressed realization. Since there were significant differences, one may conclude that click localization is not exclusively dependent on the underlying syntactic structure of the sentence.

4. Summary and Conclusion.

In this paper, I have attempted to establish the units of perception and the levels at which perception operates. Evidence has been adduced for two basic steps in perception: primary processing and linguistic processing. Primary processing consists of auditory processing and phonetic processing, which constitutes listening in a speech mode. There are several levels within the linguistic level, of which the phonological and syntactic level are considered better documented than a possible morphological level. Linguistic processing presupposes primary processing. Auditory processing must logically precede other levels of processing; phonetic processing is considered as presupposed by the other levels, but the possibility is admitted that phonetic and linguistic processing may proceed concurrently. The units at the various levels may differ in size, and there is extensive interaction between them, as there is, for example, between the phonetic and phonological levels on the one hand and the syntactic level on the other hand. Processing at the syntactic level presupposes analysis at the phonetic level, which seems to be largely suprasegmental. Parallel processing is accepted as part of the model, and a strict separation of levels is considered unwarranted.

Footnote

*I am grateful to the College of Humanities of The Ohio State University for releasing me from teaching duties during the autumn quarter of 1971, while this paper was being written. I wish also to express my appreciation to Dr. P. B. Denes and Dr. J. P. Olive of the Bell Telephone Laboratories for their help with the experimental part of this paper, to Dr. A. W. F. Huggins (of M.I.T.) and Dr. T. Smith (of the University of California, San Diego) for their challenges and suggestions, and to my research assistants Linda R. Shockey and Richard P. Gregorski for their help in administering the listening test. This paper was presented at the April 1972 Vancouver symposium on "Speech Production--Speech Perception: Their Relationship to Cortical Functioning".

References

Abbs, J. H., and H. M. Sussman (1971) "Neurophysiological feature detectors and speech perception: a discussion of theoretical implications." JSHR 14.23-36.

Abrams, Kenneth, and Thomas G. Bever (1969) "Syntactic structure modifies attention during speech perception and recognition." Quarterly Journal of Experimental Psychology 21.280-290.

Bever, T., R. Kirk, and J. Lackner (1969) "An autonomic reflection of syntactic structure." Neuropsychologia 7.23-28.

Bever, T. G., J. R. Lackner, and R. Kirk (1969) "The underlying structures of sentences are the primary units of immediate speech processing." Perception and Psychophysics 5.225-234.

Bever, T. G., J. Lackner, and W. Stolz (1969) "Transitional probability is not a general mechanism for the segmentation of speech." Journal of Experimental Psychology 79.387-394.

Bond, Z. S. (1971) "Units in speech perception." Working Papers in Linguistics No. 9, viii-112. Computer and Information Science Research Center Technical Report Series, OSU-CISRC-TR-71-8. The Ohio State University, Columbus, Ohio.

Bondarko, L. V., N. G. Zagorujko, V. A. Kozhevnikov, A. P. Molchanov, and L. A. Chistovich (1968) "A model of speech perception by humans." Academy of Sciences of the U.S.S.R., Siberian Section: Nauka, Novosibirsk. Translated by I. Lehiste, Working Papers in Linguistics No. 6, Ohio State University, Columbus (1970) 88-132.

Chapin, Paul G., Timothy S. Smith, and Adele A. Abrahamson (1972, in press) "Two factors in perceptual segmentation of speech." Journal of Verbal Learning and Verbal Behavior.

Chistovich, L., G. Fant, A. de Serpa-Leitão, and P. Tjernlund (1966a) "Mimicking of synthetic vowels." Speech Transmission Laboratory Quarterly Progress and Status Report No. 2.1-18.

Chistovich, L., G. Fant, and A. de Serpa-Leitão (1966b) "Mimicking and perception of synthetic vowels, Part II." Speech Transmission Laboratory Quarterly Progress and Status Report 3.1-3.

Chistovich, L. A., and V. A. Kozhevnikov (1969) "Perception of Speech." in Voprosy teorii i metodov issledovanija vosprijatija recevyx signalov, Leningrad; Translated as L. A. Chistovich et al. "Theory and methods of research on perception of speech signals." JPRS 50423 (1970).

Day, Ruth S. (1970a) "Temporal order judgments in speech: are individuals language-bound or stimulus-bound?" (Paper presented at the 9th Annual Meeting of the Psychonomic Society, St. Louis, November, 1969). Haskins Laboratories Status Report SR-21/22, 71-87.

Day, Ruth S. (1970b) "Temporal order perception of a reversible phoneme cluster." Paper presented at the 79th meeting of the Acoustical Society of America, Atlantic City, 21-24 April.

Day, Ruth S., and James E. Cutting (1970) "Perceptual competition between speech and nonspeech." Paper presented at the 80th meeting of the Acoustical Society of America, Houston, 3-6 November.

Denes, Peter B. (1963) "On the statistics of spoken English." JASA 35.892-904.

Denes, Peter B. (1970) "On-line computers for speech research." Transactions of the IEEE on Audio- and Electroacoustics, December, Vol. AU-18, No. 4, 418-425.

Fodor, J. A., and T. G. Bever (1965) "The psychological reality of linguistic segments." Journal of Verbal Learning and Verbal Behavior 4.414-420. Also in: L. A. Jakobovits and M. S. Miron (eds.), Readings in the Psychology of Language. Englewood Cliffs, N.J.: Prentice-Hall, Inc. (1964) 325-332.

Fodor, J. A., and M. F. Garrett (1971) "A consolidation effect in sentence perception." M.I.T. Research Laboratory of Electronics Quarterly Progress Report No. 100, January 15, 182-185.

Fry, D. B. (1970) "Reaction time experiments in the study of speech processing." Nouvelles Perspectives en phonétique, Institut de Phonétique, Université Libre de Bruxelles: Conferences et Travaux, Vol. 1, 15-35.

Fry, D. B., A. S. Abramson, P. D. Eimas, and A. M. Liberman (1962) "The identification and discrimination of synthetic vowels." Language and Speech 5.171-189.

Fujisaki, H., and T. Kawashima (1968) "The influence of various factors on the identification and discrimination of synthetic speech sounds." Reports of the 6th International Congress on Acoustics, Tokyo, No. 2.B-95-98.

Fujisaki, H., and T. Kawashima (1969) "On the modes and mechanisms of perception of speech sounds." Paper presented at the 78th meeting of the Acoustical Society of America, San Diego, November 4.

Gamlin, Peter J. (1971) "Sentence processing as a function of syntax, short term memory capacity, the meaningfulness of the stimulus and age." Language and Speech 14.115-134.

Garrett, M., T. Bever, and J. Fodor (1965) "The active use of grammar in speech perception." Perception and Psychophysics 1.30-32.

Harris, Z. (1944) "Simultaneous components in phonology." Language 20.181-205.

Holmes, V., and K. Forster (1970) "Detection of extraneous signals during sentence recognition." Perception and Psychophysics 7.5.297-301.

Johnson, Neal F. (1970) "The role of chunking and organization in the process of recall." Psychology of Learning and Motivation, Vol. 4, Academic Press, Inc.: New York, 171-247.

Ladefoged, P., and D. E. Broadbent (1960) "Perception of sequence in auditory events." Quarterly Journal of Experimental Psychology 12.162-170.

Lane, H. (1965) "The motor theory of speech perception: a critical review." Psychological Review 2.275-309.

Lehiste, Ilse (1967) "Suprasegmental features, segmental features, and long components." Actes du Xe congrès international des linguistes, Bucarest, 1967: Editions de l'academic de la Republique socialiste de Roumanie, Bucarest, Vol. IV.1-7 (1970).

Lehiste, Ilse (1970a) Suprasegmentals.. M.I.T. Press:Cambridge.

Lehiste, Ilse (1970b) "Experiments with synthetic speech concerning quantity in Estonian." Proceedings of the 3rd International Congress of Fenno-Ugricists, Tallinn (in press).

Lehiste, Ilse, and L. Shockey (1971) "The perception of coarticulation." Two papers presented at the 82nd meeting of the Acoustical Society of America, Denver, October 20.

Liberman, Alvin M. (1957) "Some results of research on speech perception." JASA 29.117-123.

Liberman, Alvin M. (1970) "The grammars of speech and language." Cognitive Psychology 1.301-323.

Liberman, A. M., F. S. Cooper, K. S. Harris, and P. F. MacNeilage (1962) "A motor theory of speech perception." Proceedings of Speech Communication Seminar, Stockholm, Session D-3, 1-10.

Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy (1967) "Perception of the speech code." Psychological Review 74.431-461.

Liberman, A. M., K. S. Harris, N. Hoffman, and B. Griffith (1957) "The discrimination of speech sounds within and across phoneme boundaries." Journal of Experimental Psychology 54.358-368.

Liberman, A. M., K. S. Harris, J. Kinney, and H. Lane (1961) "The discrimination of relative onset time of the components of certain speech and nonspeech patterns." Journal of Experimental Psychology 61.379-388.

Lisker, L., and A. S. Abramson (1971) "Distinctive features and laryngeal control." Language 47.767-785.

Miller, G. A. (1956) "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Psychological Review 63.81-97.

Miller, G. A. and P. E. Nicely (1955) "An analysis of perceptual confusions among some English consonants." JASA 27.338-352.

Neisser, Ulric (1967) Cognitive Psychology. New York: Appleton-Century-Crofts.

Öhman, S. E. G. (1966) "Coarticulation in VCV utterances." JASA 39.151-168.

Olive, J. P. (1971) "Automatic formant tracking by a Newton-Raphson technique." JASA 50.661-670.

Pisoni, David B. (1971) "Very brief short-term memory in speech perception." Paper presented at the 82nd meeting of the Acoustical Society of America, Denver, October 19.

Rabiner, L. R., et al. (1971) "Digital formant synthesis." Paper 23C8, Proceedings of the 7tn International Congress on Acoustics, Budapest, Vol. 3.157-158.

Savin, H. B., and T. G. Bever (1970) "The nonperceptual reality of the phoneme." Journal of Verbal Learning and Verbal Behavior 9.295-302.

Sharf, Donald J. (1971) "Perceptual parameters of consonant sounds." Language and Speech 14.169-177.

Spiegel, Murray R. (1961) Theory and Problems of Statistics. McGraw-Hill: New York, 171.

Stevens, K. N., A. M. Liberman, S. E. G. Öhman, and M. Studdert-Kennedy (1969) "Cross-language study of vowel perception." Language and Speech 12.1-23.

Studdert-Kennedy, Michael, and Donald Shankweiler (1970) "Hemispheric specialization for speech perception." JASA 48.579-594.

Studdert-Kennedy, Michael, A. M. Liberman, K. S. Harris, and F. D. Cooper (1970) "Motor theory of speech perception: A reply to Lane's critical review." Psychological Review 77.234-249.

Wickelgren, Wayne A. (1969a) "Context-sensitive coding, associative memory, and serial order in (speech) behavior." Psychological Review 76.1.1-15.

Wickelgren, Wayne A. (1969b) "Context-sensitive coding in speech recognition, articulation and development." In K. N. Leibovic, ed., Information Processing in the Nervous System. Springer: New York-Heidelberg-Berlin, 85-95.

Manner of Articulation, Parallel Processing, and the Perception of Duration*

Ilse Lehiste

A number of problems are connected with the study of the temporal aspects of speech production and perception. An important one is the problem of segmentation. Any study involving the measurement of duration presupposes the establishment of boundaries. The researchers in the field are by no means unanimous regarding the boundaries of speech sounds. It has to be decided at which level the boundaries are to be located--the articulatory or the acoustic level--and whether these boundaries have any perceptual reality.

Some time ago (Peterson and Lehiste (1960)) I established some practical guidelines for segmenting an utterance on the basis of the acoustic characteristics of the sound wave. Naeser (1969) has recently elaborated these rules, and they have been used by several investigators of speech sound duration. Underlying these rules was a basic assumption which I would now like to make explicit: the production and perception of timing patterns takes place with reference to major changes in manner of articulation.

Before presenting some new evidence that, im my opinion, supports this hypothesis, I should review some of the arguments against the possibility of segmentation.

One of the arguments is connected with the continuous nature of the speech wave, and the fact that there is no one-to-one correspondence between acoustic segments and linguistic segments. Fant has devoted a considerable amount of attention to this problem (Fant (1962)). He has made the observation that although the speech wave is basically continuous, spectrographic pictures of speech often display quite distinct boundaries between successive parts along the time axis. These boundaries are, according to Fant, related to switching events in the speech production mechanism, such as a shift in the primary sound source (e.g. from voice to noise), or the opening and closing off of a passage within the vocal cavities. Boundaries between sound segments are due to the beginning or end of at least one of simultaneously present sound features. But sound segment boundaries are not to be confused with phoneme boundaries. Several adjacent sounds of connected speech may carry information on several adjacent phonemes. A typical example would be the influence exerted by a consonant on a following vowel.

The notion that the same sound segment may carry information on several adjacent phonemes is intimately connected with the hypothesis of parallel processing. In essence, parallel processing

means that the same physical signal may carry more than one kind of information, which in the process of speech perception may be extracted simultaneously, as if over separate channels. Parallel processing has been discussed extensively in various recent publications (Neisser (1967); Chistovich and Kozhevnikov (1969-1970); Liberman (1970)). Given the continuous nature of the acoustic signal and the fact that perceptual cues may overlap in time, it is quite understandable that some linguists have claimed that it is not possible to establish the duration of segments in any perceptually meaningful way.

Granted that the acoustic signal is continuous and that speech processing may take in parallel fashion, I still believe that the picture is unduly complicated by not making a distinction between manner of articulation and point of articulation characteristics. It is the point of articulation cues that are continuous; they may be spread out over several adjacent segments, while the manner of articulation cues provide the abrupt changes that are seen in the visual display of an acoustic waveform. While several adjacent segments may carry information on the point of articulation, the manner of articulation can usually be determined from an examination of the pertinent segments themselves. Duration of segments relates to the manner of articulation rather than to the point of articulation, and timing information is extracted primarily from manner of articulation cues.

This is, of course, a hypothesis that should be supported by experimental evidence. I shall try to present some toward the end of this paper. But first let me bring up some further considerations that lead me to believe in the possibility of making rather precise and perceptually meaningful measurements of the duration of various sound segments.

It is an established fact that differences in duration are perceptible. In principle, the ear is capable of distinguishing between durations, be it the duration of a continuous signal (like gated white noise) or the duration of a silent interval embedded in a continuous signal. If it is claimed that listeners can distinguish the durations of non-linguistic stimuli, but cannot perceive differences in the duration of linguistic stimuli, it is assumed that speech sounds have some characteristics that make their boundaries perceptually blurred. Being a native speaker of a quantity language, in which differences in duration carry high linguistic significance, I find this notion intuitively quite unacceptable. If durational differences can serve as part of the linguistic signaling system, listeners must be able to compare the durations of speech sounds (or whatever unit possesses contrastive duration). And if they compare durations, they must know at which moment a given sound begins and ends. In other words, there must exist unambiguous boundaries to which the listener may refer in comparing either two durations heard in succession, or a perceived duration with a stored "durational image". I propose that major changes in manner of articulation

constitute the acoustic and perceptual correlates of such boundaries.

It has been suggested that differences in duration are perceived as qualitative rather than quantitative differences. While sounds differing in duration may also differ in quality in a number of cases, it is not necessarily true in other instances. As a limiting case, I would like to quote my own experiments with a set of synthetic Estonian words, in which certain stimuli differed solely in the duration of the intervocalic plosive gap (Lehiste (1970a)). The duration of the rest of the word, including the duration of the transitions to and from the consonant, was held constant. The listeners had no difficulty in assigning different linguistic labels to words falling into different intervocalic plosive duration categories. It should be obvious that there was no qualitative difference in the silence corresponding to the duration of the plosive; thus the judgments must have been based on the perception of differences in the duration of the plosive gap.

I find no contradiction between the claim that listeners can compare the durations of segments and that they perceive speech by a kind of parallel processing of the incoming signal. The timing information is simply another feature which is extracted at the same time as the information concerning the segmental nature of the incoming speech waeve.

There is additional, somewhat circumstantial evidence of the importance of the manner of articulation in speech perception. In a study of the perceptual parameters of consonant sounds, Sharf (1971) established seven-point scales for duration, loudness, frequency, sharpness, and contact. Substantial numbers of significant differences were obtained only for duration comparisons based on manner of articulation. In an earlier study, Denes (1963) showed that manner of articulation carries by far the greatest functional load in the English sound system, and suggested that the acoustic correlates of manner of articulation might be used for segmentation in automatic speech recognition systems.

One way to test the hypothesis that the timing of speech sounds takes place with reference to major changes in manner of articulation would be to find some instances in which the application of a timing rule depends on such differences. The study I want to report about during the rest of this paper deals with a limited attempt to find such a situation.

I investigated the duration of segments in monosyllabic English words beginning and ending in an obstruent consonant and containing syllable nuclei consisting of long and short vowels, preceded and/or followed by resonants. The boundaries between obstruents and both vowels and resonants are clearly manifested and should be easily detectable, whereas the boundaries between vowels and resonants are relatively less well defined and do not correspond to what I would call major changes in manner of

articulation. If the hypothesis is true that timing takes place with reference to major changes in manner of articulation, the span between the release of the initial obstruent and the onset of the final obstruent should function as a unit of timing, regardless of the position or even the presence of the resonant; the resonants should fuse with the vowels into syllable nuclei functioning as a whole with regard to some timing rules.

The structure of the test words may be symbolized as C_1 R_1 V R_2 C_2 (consonant-resonant-vowel-resonant-consonant). The first consonant was either a voiced or voiceless plosive. In the latter case, aspiration was present. The duration of initial consonants was not measurable, especially in the case of voiceless plosives; thus only the duration of aspiration will be presented in the following tables. The first resonant could be present or absent. If present, it was either /r/ or /l/. The vowel was always present; it could be either long or short. For purposes of this study, all vowels were considered long with the exception of [I ɛ ʌ U]. The second resonant could be present or absent; if present, it was either /m/, /n/, /r/, or /l/. The final consonant was a voiced or voiceless obstruent (in most cases, plosive). The tables report the duration of the closure part of the final consonant; release and aspiration (if present) are not included in the tables. For purposes of processing, the sounds were coded in the following manner:

Aspiration	Resonant	Vowel	Resonant	Consonant
1 = +	2 = /r/	1 = long	4 = /m/	1 = voiced
0 = -	1 = /l/	0 = short	3 = /n/	0 = voiceless
	0 = -		2 = /r/	
			1 = /l/	
			0 = -	

For example, the code 0001l refers to a word beginning with a voiced initial consonant, containing no first resonant, a short vowel, /l/ as second resonant, and ending in a voiced final consonant. An example would be the word build. The code 000R1 refers to all words of this type in which a resonant was present in the slot indicated by R.

There were 156 test words of this general structure. Each word was produced five times by three native speakers of English; thus the data consist of 15 x 156 = 2350 productions. A list of the test words, together with their codes, is presented at the end of this paper in Appendix A.

The tapes were processed by means of a Frøkjær-Jensen Pitch Meter and Intensity Meter and displayed on an Elema-Schönander Mingograph, operated at a speed of 10 cm/sec. The boundaries of segments were established mainly on the basis of duplex oscillograms, using principles summarized by Naeser (1969). Durations of all segments were measured, and average durations were computed for

all segments in all word types. These average durations are presented at the end of the paper in Appendix B.

The results of the study will be discussed with reference to three summary tables and five figures. Table 1 presents average durations of segments in words with syllable nuclei consisting of vowel + resonant.

TABLE 1
AVERAGE DURATIONS OF SEGMENTS IN WORDS WITH SYLLABLE NUCLEI CONSISTING OF VOWEL + RESONANT

Word type	N	Asp.	R	V	R	C	SN	SN+C
000R1	105			229.1	174.4	54.8	403.5	458.3
000R0	105			163.4	103.6	106.7	267.0	373.7
001R1	150			290.0	141.1	95.6	431.1	526.7
001R0	150			196.9	89.0	112.9	285.9	398.8
100R1	75	78.7		212.1	156.9	55.4	447.7	503.1
100R0	75	72.2		143.7	92.7	104.6	308.6	413.2
101R1	120	86.6		284.4	134.4	61.9	505.4	567.3
101R0	135	83.6		192.4	79.7	102.2	355.7	457.9
0R1R1	15		90.7	347.0	100.4	38.7	538.1	576.8
0R1R0	15		75.3	253.5	72.4	78.9	401.2	480.1

The word type is given in Column 1. Five pairs of word types are presented, differing in the voicing of the final obstruent consonant. The first pair consists of words beginning with a voiced plosive, followed by a short vowel and a resonant. The second pair is similar, except that the vowels are long. The third pair consists of words beginning with a voiceless plosive, followed by a short vowel and a resonant. The fourth pair contains a long vowel. The fifth pair finally consists of words in which a voiced initial plosive was followed by a sequence of resonant, long vowel and resonant, followed by a voiced and voiceless plosive. (Only one example of each type was available--bland and blank--and therefore the averages have to be interpreted with caution.)

The second column contains the number of productions used for averaging. Since there were three speakers, each producing the word five times, the number of different words may be obtained by dividing N by 15.

The third column contains the duration of aspiration, which was present in words beginning with a voiceless plosive. (All durations are in milliseconds). The fourth column shows the average duration of the prevocalic resonant, where present. The fifth

column contains the average duration of the vowel. The next column shows the average duration of the postvocalic resonant. This is followed in the next column by the average duration of the final consonant. The duration is that of the hold of the consonant and does not include release and/or aspiration. The following column gives the duration of the span from the release of the first obstruent to the onset of the second. The last column gives the sum of the syllable nucleus (consisting of vowel and one or two resonants) and the final consonant. Figure 1 presents the same information graphically.

It is a well known rule in English that vowels are shortened before a voiceless final consonant and lengthened before a voiced final consonant. In an earlier study (Peterson and Lehiste (1960)), we had established the ratio between vowel durations before voiceless and voiced final consonants as 0.66, i.e. approximately 2/3. The present set of data shows that both the vowel and the postvocalic resonant are subject to either shortening or lengthening, depending on the voicing of the final obstruent.

When all parts of the syllable nucleus from the release of the initial plosive to the onset of the postvocalic resonant were combined, the ratio between their average durations before a voiceless and a voiced plosive was 0.73. The duration ratio for postvocalic resonants was 0.62. The ratio of the durations of the whole span from the release of the initial plosive to the closure of the final plosive was 0.69.

In a recent study devoted to vowel length variation as a function of the voicing of the consonant environment, Chen (1970) included 96 word tokens containing vowel + resonant sequences. He obtained comparable ratios: 0.73 for the vowel, 0.60 for the resonant, and 0.66 for the whole vowel + resonant sequence. This, as may be remembered, is identical with the ratio obtained for vowels by Peterson and Lehiste (1960), and very close to the 0.69 ratio obtained in the present study.

I believe it to be obvious that with regard to the timing rule in question, the sequence vowel + resonant functions indeed as a unitary syllable nucleus, albeit a segmentally complex one. The timing of the sequence appears to proceed indeed from the release of the initial obstruent to the formation of the closure of the final obstruent, which constitute major changes in the manner of articulation.

The question whether sequences of resonant + vowel function in the same manner turned out to be somewhat more complicated. Table 2 presents four sets of words in which resonants, if present, preceded and/or followed vocalic syllable nuclei.

TABLE 2
AVERAGE DURATIONS OF SEGMENTS, SYLLABLE NUCLEI AND WORDS
IN VARIOUS WORD TYPES INVOLVING RESONANTS

Word type	N	Asp.	R	V	R	C	SN	SN+C
10100	135	75.1		234.9		135.0	310.0	445.0
101R0	135	83.6		192.4	79.7	102.2	355.7	457.9
1R100	165	95.5	42.9	211.4		125.0	349.8	474.8
1R1R0	30	105.3	20.8	223.5	65.7	88.6	415.3	503.9
00101	45			428.1		76.4	428.1	504.5
001R1	150			290.0	141.1	95.6	431.1	526.7
0R101	60		83.2	394.6		68.1	477.8	545.9
0R1R1	15		90.7	347.0	100.4	38.7	538.1	576.8
10000	45	71.0		180.3		130.2	251.3	381.5
100R0	75	72.2		143.7	92.7	104.6	308.6	413.2
1R000	75	94.0	45.8	172.0		120.2	311.8	432.0
1R0R0	15	89.2	42.5	149.7	79.6	75.4	361.0	436.4
00001	45			292.5		84.9	292.5	377.4
000R1	105			229.1	174.4	54.8	403.5	458.3
0R001	60		83.6	258.4		85.5	342.0	427.5
0R0R1	15		90.6	238.1	144.9	49.7	473.6	523.3

The first set of four consists of words containing long vowels and beginning and ending in a voiceless plosive. The second set is similar, except the words began and ended in voiced plosives. The third set is analogous to the first, except for the vowel being short; the fourth set is in the same way analogous to the second. Sets one and two are shown on Figure 2; Figure 3 presents comparable material for sets three and four.

The differences in the average duration of syllable nuclei (including vowels and resonants) range from 3.0 msec (for 00101 - 001R1) to 181.1 (for 00001 - 0R0R1). In trying to assess the relative significance of the differences, it appears resonable to ask first whether the differences are perceptible. Just noticeable differences (jnd's, or difference limens - DLs) in duration have been studied by several investigators (summarized in Lehiste (1970b)). Table 3 gives the differences between the average durations for all pairs within each set of word types presented in Table 2.

TABLE 3
DIFFERENCE IN THE AVERAGE DURATIONS OF SYLLABLE NUCLEI
INVOLVING RESONANTS BEFORE AND AFTER THE VOWEL

Word type	N	Average duration of SN, in msec	Difference in SN, in msec	Nearest absolute DL, in msec
10100	135	310.0		
101R0	135	355.7	45.7	48.0 (Stott, 1935)
10100	135	310.0		
1R100	165	349.8	39.8	48.0 (Stott, 1935)
10100	135	310.0		
1R1R0	30	415.3	105.3	48.0 (Stott, 1935)
101R0	135	355.7		
1R100	165	349.8	5.9	48.0 (Stott, 1935)
00101	45	428.1		
001R1	150	431.1	3.0	48.0 (Stott, 1935)
00101	45	428.1		
0R101	60	477.8	49.7	68.64 (Henry, 1948)
00101	45	428.1		
0R1R1	15	538.1	110.0	69.0 (Stott, 1935)
001R1	150	431.1		
0R101	60	477.8	46.7	68.64 (Henry, 1948)
10000	45	251.3		
100R0	75	308.6	57.3	47.64 (Henry, 1948)
10000	45	251.3		
1R000	75	311.8	60.5	47.64 (Henry, 1948)
10000	45	251.3		
1R0R0	15	361.0	109.7	48.0 (Stott, 1935)
100R0	75	308.6		
1R000	75	311.8	3.2	47.64 (Henry, 1948)
00001	45	292.5		
000R1	105	403.5	111.0	48.0 (Stott, 1935)
00001	45	292.5		
0R001	60	342.0	49.5	47.64 (Henry, 1948)
00001	45	292.5		
0R0R1	15	473.6	181.1	48.0 (Stott, 1935)
000R1	105	403.5		
0R001	60	342.0	61.5	48.0 (Stott, 1935)

The last column contains the absolute DL, in msec, established for reference durations that are closest to the duration of the syllable nuclei under consideration. Table 4 summarizes the pertinent data for durational difference limens. The information is presented graphically in Figures 4 and 5.

TABLE 4

Reference duration (msec)	Δ T/T	Absolute DL (msec)
200	.142	28.4 (Stott, 1935)
277	.172	47.64 (Henry, 1948)
400	.120	48.0 (Stott, 1935)
480	.143	68.64 (Henry, 1948)
600	.115	69.0 (Stott, 1935)

Looking at the first set, we see one difference that is clearly nonperceptible; another that is considerably above threshold and should be perceptible; and two that hover around the difference limen. Similar observations may be made with regard to the other sets.

Some generalizations may be drawn from comparing all four sets. The picture seems a little more systematic with long vowels than with short vowels. Here all differences are below or near the threshold with the exception of that between a vowel occurring alone and a vowel flanked on both sides by a resonant. In words with short vowels and voiceless initial and final plosives, the relative shortness of the vowel raises these differences slightly above threshold in those word pairs in which a vowel occurring alone is compared with vowel preceded and/or followed by a resonant. However, the ordering of the resonant before or after the vowel does not affect the timing in any significant way, as had also been the case with long vowels.

In words with short vowels and voiced plosives, there are two pairs whose differences are clearly above threshold, and two that are close to threshold value.

If the sets are combined according to the voicing or voicelessness of the plosives (ignoring the intrinsic differences in vowel duration), the differences drop below threshold except for the word types containing two resonants. These are longer than the other words by approximately the average duration of one resonant.

Table 2 reveals a number of other interesting facts about

the temporal structure of the test words which are, however, not directly relevant to the question under consideration. For example, the duration of final consonants stands in a compensatory relationship to the duration of syllable nuclei, so that the differences between the average durations of words are usually smaller than those between syllable nuclei. For many word types, these differences are likewise below the perceptual threshold.

Consideration of words with resonants preceding and following vowels thus adds some further support to the hypothesis that the timing patterns are related to major changes in the manner of articulation. Roughly speaking, long vowels seem to fuse into a timing unit with either a preceding or a following resonant; with short vowels the evidence is less clear, but at least with voiced initial and final consonants, the vowel and a preceding resonant seem to have the same average duration as the vowel by itself. The ordering of the vowel - resonant sequence is irrelevant for overall duration. Number of segments begins to play a part when more than one resonant is involved; these cases thus provide the limit to which the argument can be carried. It is possible that some of the exceptions to the general pattern are due to the accidents of test word selection; a larger corpus, with a better balanced set of test words, might yield a clearer picture.

Footnote

*The research on which the paper is based was supported in part by the National Science Foundation through Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University.

References

Chen, Matthew (1970) "Vowel length variation as a function of the voicing of the consonant environment." Phonetica 22. 129-159.

Chistovich, L. A. and V. A. Kozhevnikov (1969) "Perception of speech." in: Voprosy teorii i metodov issledovanija vosprijatija recevyx signalov, Leningrad; translated as L. A. Chistovich et al. Theory and methods of research on perception of speech signals, JPRS 50423, (1970).

Denes, Peter B. (1963) "On the statistics of spoken English." JASA 53.892-904.

Fant, C. Gunnar M. (1962) "Descriptive analysis of the acoustic aspects of speech." Logos 5.3-17.

Lehiste, Ilse (1970) "Experiments with synthetic speech concerning quantity in Estonian." Proceedings of the 3rd International Congress of Fenno-Ugricists, Tallinn (in press).

Lehiste, Ilse (1970) Suprasegmentals, M.I.T. Press, Cambridge.

Liberman, A. M. (1970) "The grammars of speech and language." Cognitive Psychology 1.301-323.

Naeser, Margaret (1969) "Criteria for the segmentation of vowels on duplex oscillograms." Technical Report No. 124, Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, Madison.

Neisser, Ulric (1967) Cognitive Psychology, Appleton-Century-Crofts, New York.

Peterson, Gordon E. and I. Lehiste (1960) "Duration of syllable nuclei in English." JASA 32.693-703.

Sharf, Donald J. (1971) "Perceptual parameters of consonant sounds." Language and Speech 14.169-177.

APPENDIX A

List of test words used in the study, arranged according to their structural code.

1.	felt	00010	41.	daunt	00130
2.	guilt	00010	42.	bank	00130
3.	built	00010	43.	faint	00130
4.	dug	00001	44.	burnt	00130
5.	dead	00001	45.	feigned	00131
6.	bed	00001	46.	joined	00131
7.	guild	00011	47.	dawned	00131
8.	build	00011	48.	mound	00131
9.	felled	00011	49.	band	00131
10.	sent	00030	50.	found	00131
11.	dint	00030	51.	burned	00131
12.	bent	00030	52.	bled	01001
13.	shunt	00030	53.	blend	01031
14.	bent	00030	54.	gloat	01100
15.	bend	00031	55.	bleat	01100
16.	send	00031	56.	black	01100
17.	bend	00031	57.	blurt	01100
18.	dinned	00031	58.	blade	01101
19.	shunned	00031	59.	blurred	01101
20.	beat	00100	60.	glues	01101
21.	doubt	00100	61.	blank	01130
22.	beach	00100	62.	bland	01131
23.	back	00100	63.	bread	02001
24.	goat	00100	64.	dread	02001
25.	gape	00100	65.	drug	02001
26.	died	00101	66.	grape	02100
27.	bayed	00101	67.	drought	02100
28.	bird	00101	68.	breach	02100
29.	goos	00101	69.	dried	02101
30.	molt	00110	70.	brayed	02101
31.	bolt	00110	71.	puck	10000
32.	bold	00111	72.	tuck	10000
33.	mold	00111	73.	kick	10000
34.	sort	00120	74.	pug	10001
35.	mart	00120	75.	ted	10001
36.	marred	00121	76.	tilt	10010
37.	sword	00121	77.	cult	10010
38.	fount	00130	78.	tilled	10011
39.	mount	00130	79.	culled	10011
40.	joint	00130	80.	tent	10030

81.	tint	10030
82.	tent	10030
83.	tend	10031
84.	tend	10031
85.	tinned	10031
86.	cap	10100
87.	peach	10100
88.	pot	10100
89.	coke	10100
90.	cape	10100
91.	peat	10100
92.	tout	10100
93.	tight	10100
94.	kirk	10100
95.	tooth	10100
96.	pod	10101
97.	cause	10101
98.	cowed	10101
99.	pan	10101
100.	paid	10101
101.	tied	10101
102.	goos	10101
103.	poise	10101
104.	colt	10110
105.	cold	10111
106.	cart	10120
107.	tart	10120
108.	court	10120
109.	card	10121
110.	tarred	10121
111.	cord	10121
112.	cant	10130
113.	pint	1013
114.	paint	10130
115.	pant	10130
116.	canned	10131
117.	panned	10131
118.	pined	10131
119.	pained	10131
120.	camp	10140
121.	click	11000
122.	pluck	11000
123.	plug	11001
124.	cloak	11100
125.	clap	11100
126.	pleat	11100
127.	plot	11100
128.	clerk	11100
129.	cloud	11101
130.	plod	11101
131.	played	11101
132.	plan	11101
133.	claws	11101
134.	clues	11101
135.	ploys	11101
136.	plant	11130
137.	planned	11131
138.	clamp	11140
139.	truck	12000
140.	crick	12000
141.	tread	12001
142.	trent	12030
143.	trend	12031
144.	preach	12100
145.	crepe	12100
146.	crap	12100
147.	trout	12100
148.	croak	12100
149.	trite	12100
150.	truth	12100
151.	tried	12101
152.	crowd	12101
153.	craws	12101
154.	prayed	12101
155.	crews	12101
156.	cramp	12140

APPENDIX B

Average durations, in milliseconds, of segments occurring in the test words, each produced 5 times by three speakers. N = number of words of a given type.

Word type	N	Asp.	R	V	R	C	SN	SN + C
00001	3			292.5		84.9	292.5	377.4
00010	2			147.1	102.9	125.0	250.0	375.0
00011	2			227.1	181.2	65.5	408.3	473.8
00030	5			179.7	104.2	88.4	283.9	372.3
00031	5			231.1	167.5	44.1	398.6	442.7
00100	6			270.6		152.3		422.9
00101	3			428.1		76.4		504.5
00110	2			190.7	97.5	127.7	288.2	415.9
00111	2			298.6	148.1	62.0	446.7	508.7
00120	1			161.1	85.4	122.5	246.5	369.0
00121	1			208.1	150.1	174.2	358.2	532.4
00130	7			238.8	84.1	88.5	322.9	411.4
00131	7			363.4	125.1	50.6	488.5	539.1
01001	1		89.0	256.1		82.7	345.1	427.8
01031	1		90.6	238.1	144.9	49.7	473.6	523.3
01100	4		89.2	225.4		131.1	314.6	445.7
01101	2		91.1	376.6		70.0	467.7	537.7
01130	1		75.3	253.5	72.4	78.9	401.2	480.1
01131	1		90.7	347.0	100.4	38.7	538.1	576.8
02001	3		78.2	260.6		88.4	338.8	427.2
02100	3		72.8	243.2		158.3	316.0	474.3
02101	2		75.3	412.5		66.3	487.8	554.1
10000	3	71.0		180.3		130.2	251.3	381.5
10001	2	71.6		260.6		95.4	332.2	427.6
10010	2	77.9		114.8	90.8	116.3	283.5	399.8
10011	2	87.6		205.2	153.5	66.3	446.3	512.6
10030	3	66.4		172.6	94.6	92.9	333.6	426.5
10031	3	69.7		219.0	160.3	44.6	449.0	493.6
10100	9	75.1		234.9		135.0	310.0	445.0
10101	7	77.6		373.9		77.2	451.5	528.7
10110	1	83.5		154.2	88.1	125.0	325.8	450.8
10111	1	86.5		252.2	146.6	62.5	485.3	547.8
10120	3	90.9		165.1	73.6	118.9	329.6	348.5
10121	3	97.1		260.9	126.6	70.7	484.6	555.3

Word type	N	Asp.	R	V	R	C	SN	SN+C
10130	4	76.8		218.6	84.8	83.6	380.2	463.8
10131	4	76.2		340.0	130.1	52.6	546.3	598.9
10140	1	83.1		231.9	72.1	81.3	387.1	468.4
11000	2	91.1	47.8	156.1		129.5	295.0	424.5
11001	1	100.7	47.5	244.4		90.3	392.6	482.9
11100	5	94.5	47.3	206.0		116.2	347.8	464.0
11101	5	95.1	51.9	412.0		82.9	559.0	641.9
11130	1	97.9	46.7	248.6	85.3	69.3	478.5	547.8
11131	1	92.9	38.9	322.8	132.3	46.5	586.9	633.4
11140	1	117.1		211.4	64.0	94.4	392.5	486.9
12000	3	96.8	43.9	187.9		110.9	328.6	439.5
12030	1	89.2	42.5	149.7	79.6	75.4	361.0	436.4
12031	1	103.3	50.3	208.9	162.3	47.1	524.8	571.9
12100	6	96.5	38.4	216.8		133.9	351.7	485.6
12101	4	95.6	47.7	375.4		72.2	518.7	590.9
12140	1	93.5	41.6	235.5	67.5	82.8	438.1	520.9

Fig. 1

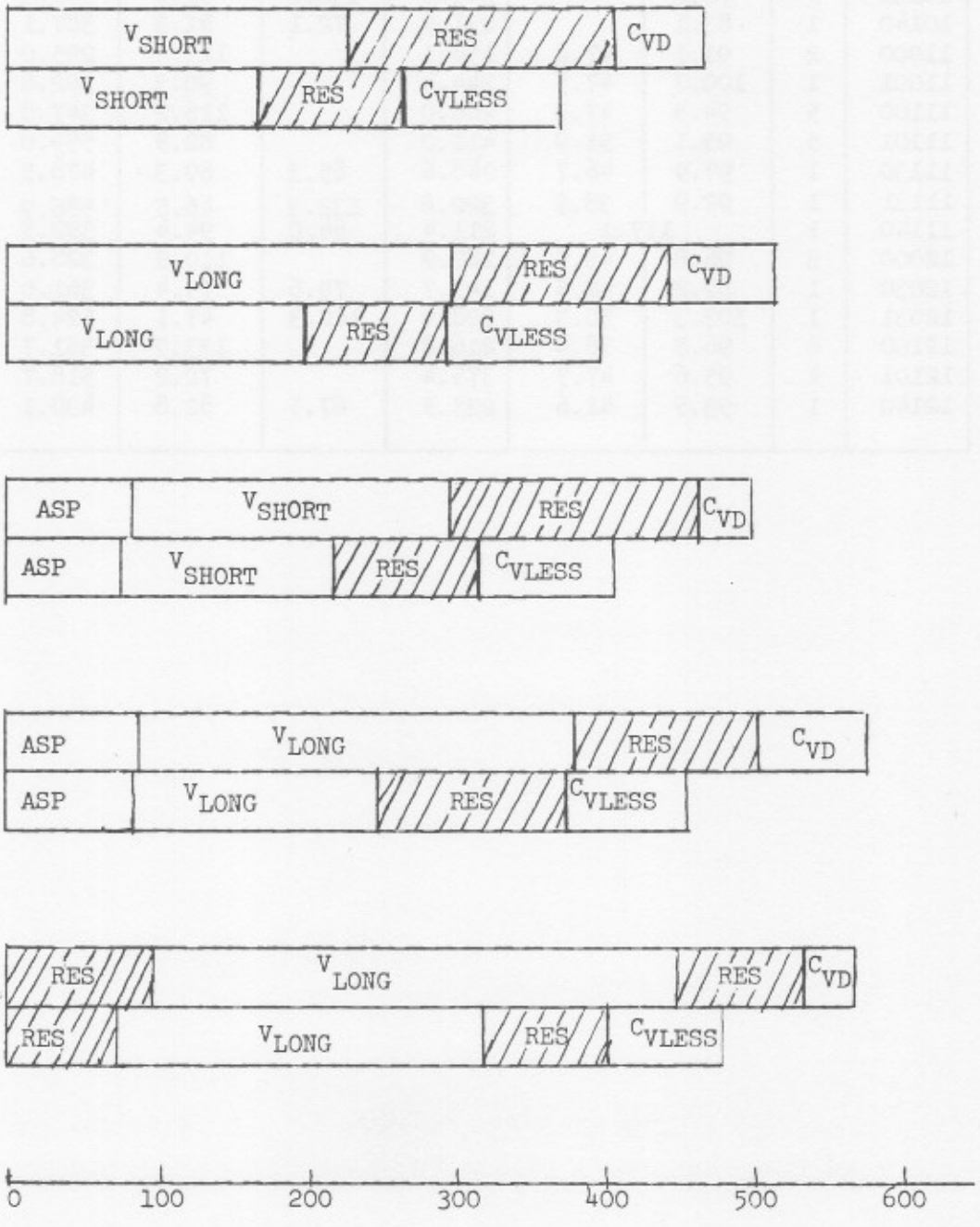

V SHORT
RES
C VD
V SHORT
RES
C VLESS
V LONG
RES
C VD
V LONG
RES
C VLESS
ASP
V SHORT
RES
C VD
ASP
V SHORT
RES
C VLESS
ASP
V LONG
RES
C VD
ASP
V LONG
RES
C VLESS
RES
V LONG
RES
C VD
RES
V LONG
RES
C VLESS
0
100
200
300
400
500
600
AVERAGE DURATION (msec)

Fig. 2

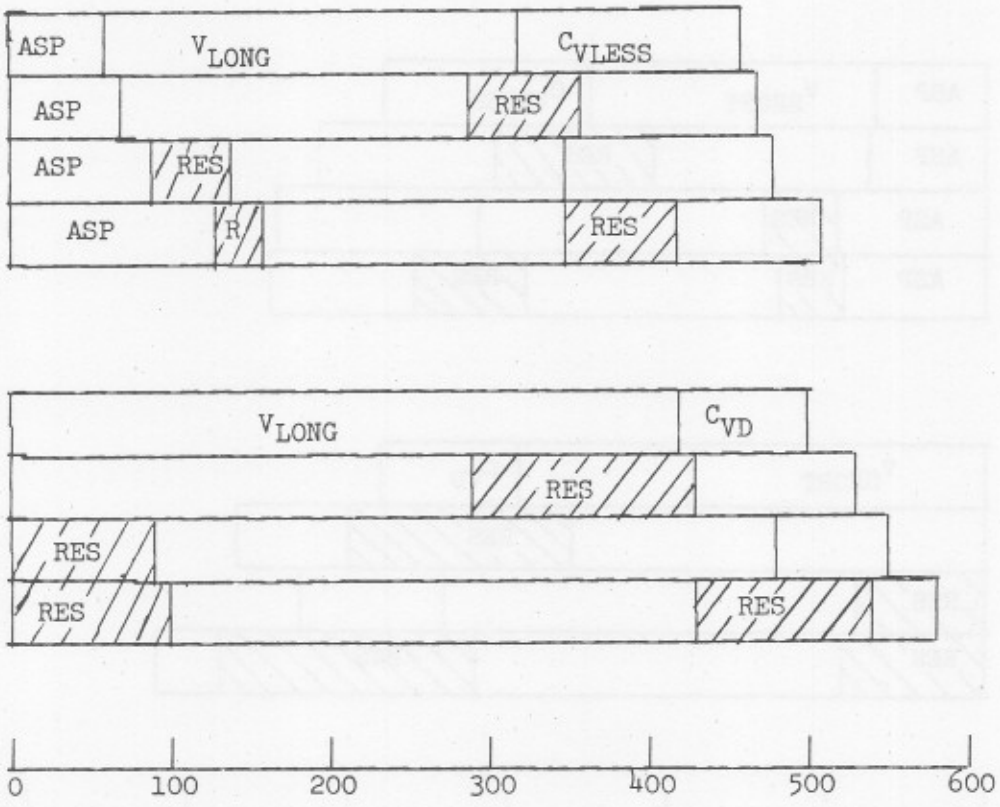

ASP
V_{LONG}
C_{VLESS}
ASP
RES
ASP
RES
ASP
R
RES
V_{LONG}
C_{VD}
RES
RES
RES
0
100
200
300
400
500
600
AVERAGE DURATION (msec)

Fig. 3

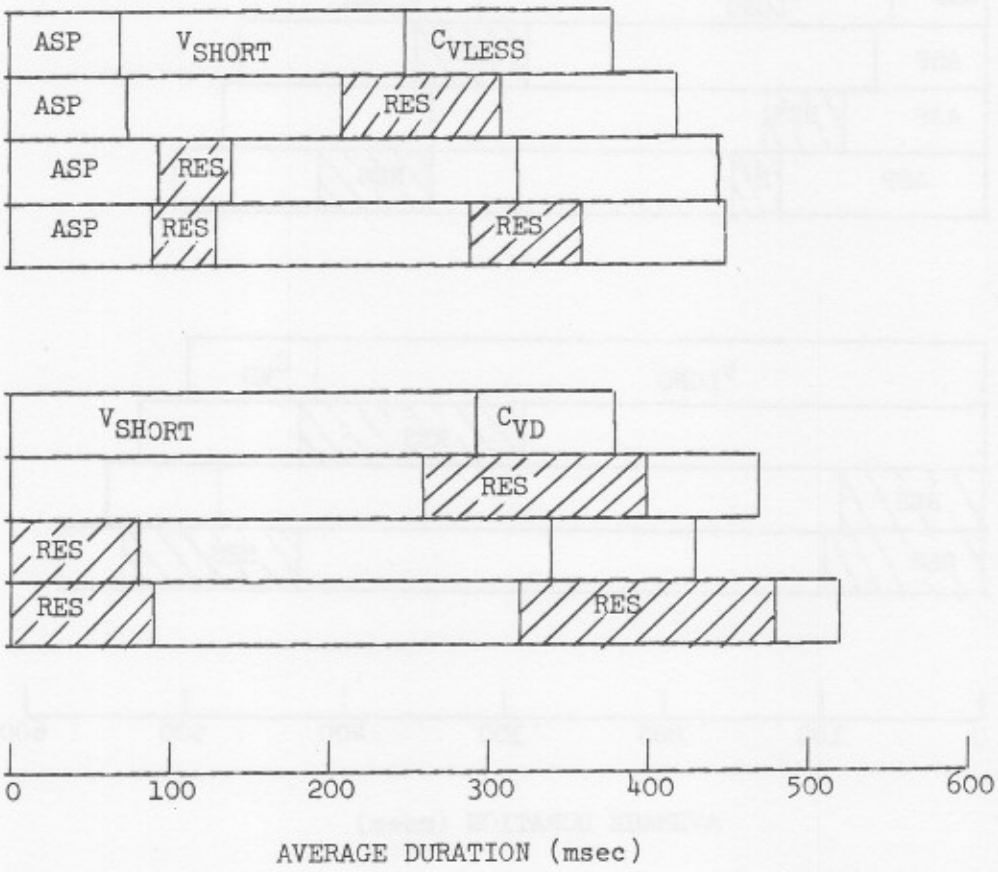

ASP
V_{SHORT}
C_{VLESS}
ASP
RES
ASP
RES
ASP
RES
RES
V_{SHORT}
C_{VD}
RES
RES
RES
RES
0
100
200
300
400
500
600
AVERAGE DURATION (msec)

Fig. 4

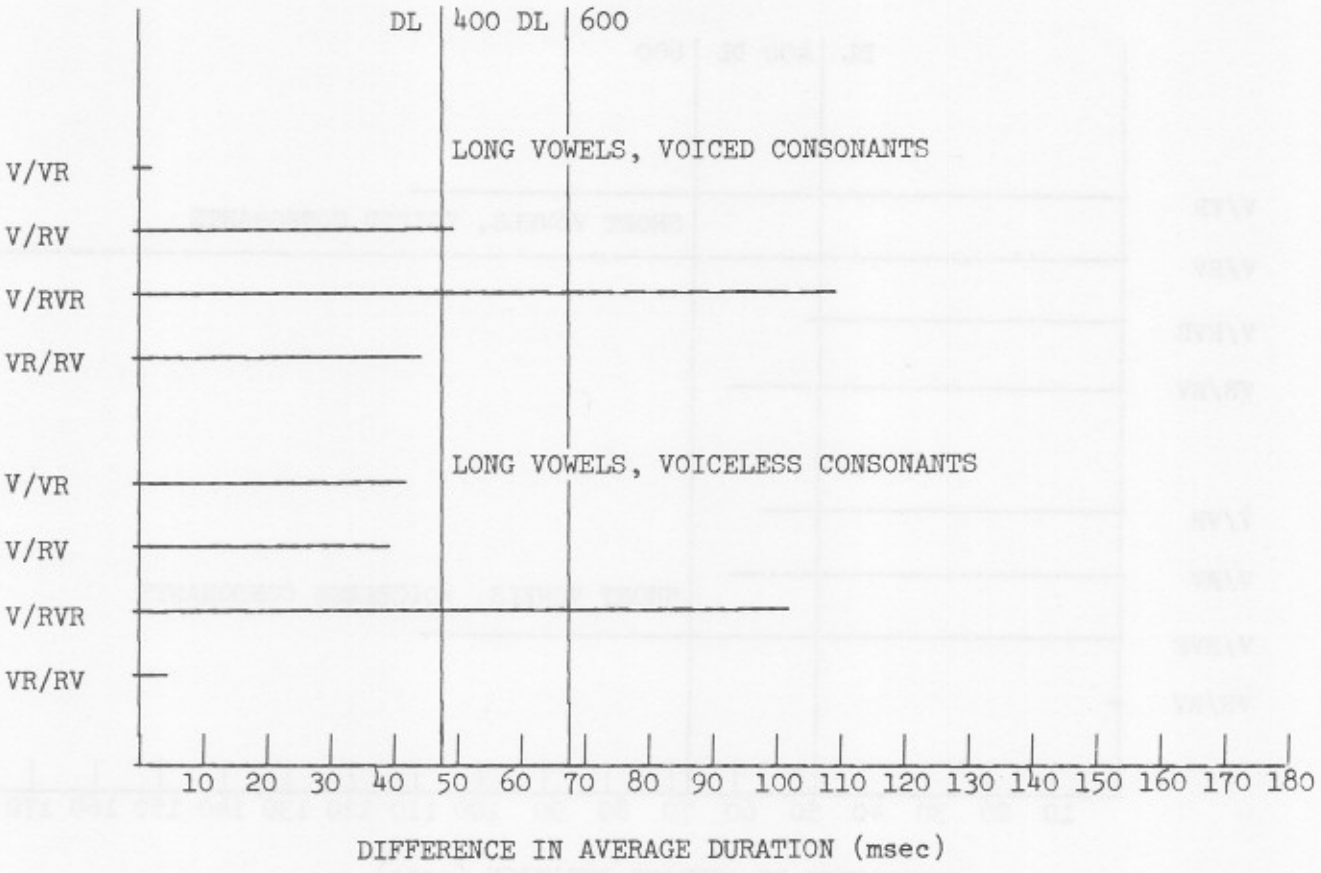
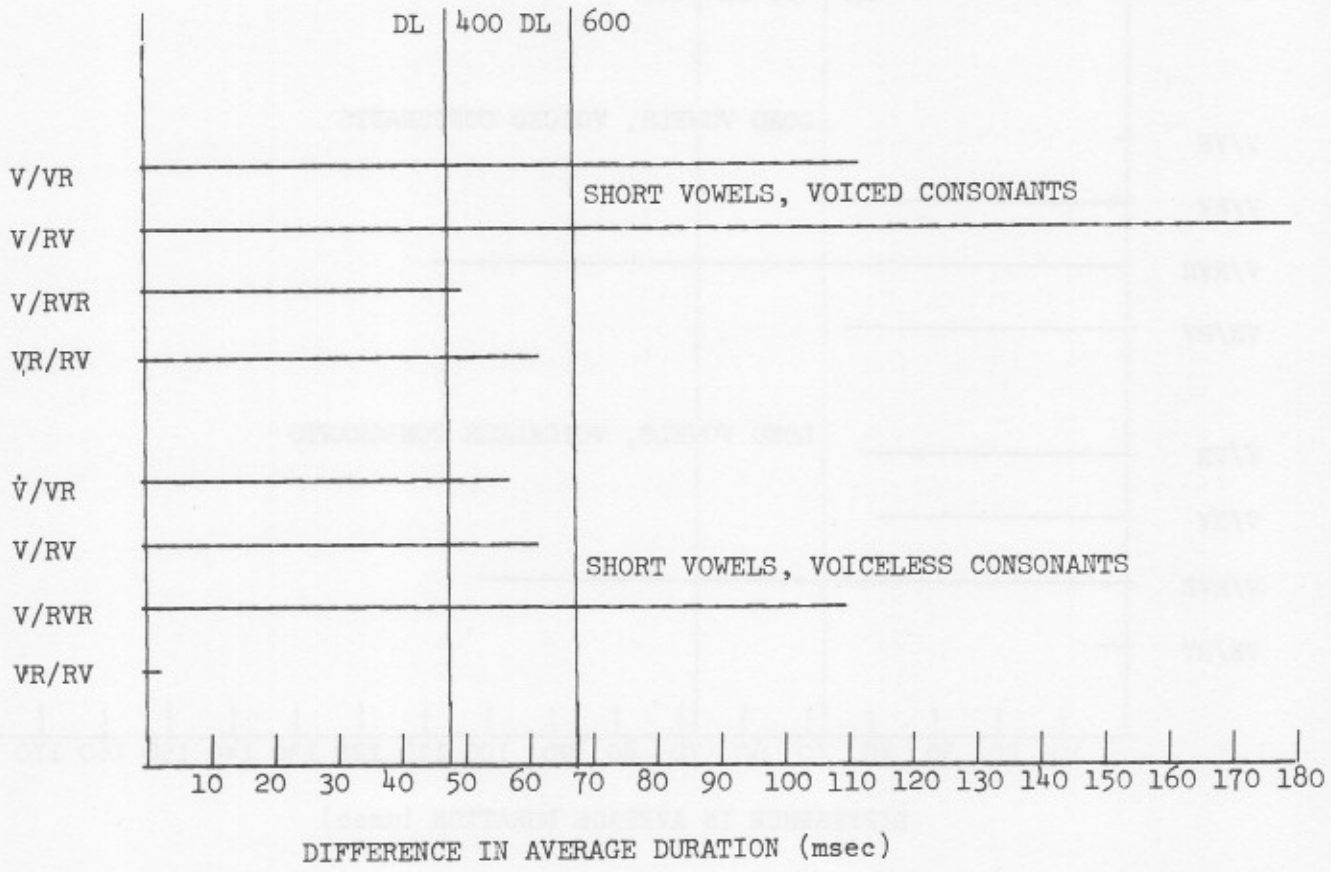

DL
400 DL
600
LONG VOWELS, VOICED CONSONANTS
V/VR
V/RV
V/RVR
VR/RV
LONG VOWELS, VOICELESS CONSONANTS
V/VR
V/RV
V/RVR
VR/RV
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
DIFFERENCE IN AVERAGE DURATION (msec)

Fig. 5


DL
400 DL
600
V/VR
V/RV
V/RVR
VR/RV
SHORT VOWELS, VOICED CONSONANTS
V̇/VR
V/RV
V/RVR
VR/RV
SHORT VOWELS, VOICELESS CONSONANTS
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180
DIFFERENCE IN AVERAGE DURATION (msec)

Temporal Compensation in a Quantity Language*

Ilse Lehiste

This paper explores the temporal relationships between the segments in three sets of Estonian words. While a considerable amount of interest has been shown previously in the duration of sounds with contrastive function, the temporal structure of the whole word of which the contrastive sounds constitute a part has received relatively little attention.[1] It is the thesis of this paper that the word is programmed as a whole, and that significant relationships exist among all segments that constitute a word, although not all segments participate in segmental quantity oppositions.[2]

The material analyzed for the study consists of three minimal triples: vaga 'pious' (nom. sg.), vaka 'bushel' (gen. sg.), vakka 'bushel' (part. sg.); sada '100' (nom. sg.), saada 'send' (2. sg. imperative), saada 'send' (-da-inf.); saag 'saw' (nom. sg.), saak 'prey' (nom. sg.), sakk 'sawtooth' (nom. sg.). According to traditional analyses, the intervocalic consonant /k/ is in short, long and overlong quantities in the first set (short, long and overlong will be referred to as quantities 1, 2 and 3); in the second set, the contrastive sound is the vowel of the first syllable, which appears in quantities 1, 2, and 3; and in the third set, /a/ is in quantity 3 in the first two words and in quantity 1 in the third, while final /k/ is in quantity 1 in the first word and in quantity 3 in the second and third.[3] These words were recorded by two speakers, each of whom repeated each test word between 100 and 110 times in sequence. The recordings were made in Tallinn in the autumn of 1970.[4] The tapes were processed through a Frøkjær-Jensen trans-pitchmeter and intensity meter; the curves were displayed by means of an Elema-Schönander Mingograf (at a speed of 10 cm/second). Measurements of duration were made from mingograph traces; the results were analyzed statistically by means of an IBM 360 computer.[5] To normalize for variations in tempo, the average durations of all words were computed, and a subset of 50 utterances whose durations were closest to the mean duration was extracted for each word. Further computations were performed on these subsets. This procedure of tempo normalization is essentially the same as that employed by Ohala and Kozhevnikov and Chistovich in previous temporal studies.[6]

One form of temporal compensation within Estonian words has been frequently referred to in previous descriptions. This is the compensation in the duration of the vowel of the second syllable, which adjusts itself inversely to the duration of the

first syllable, so that a first syllable in quantity 1 is followed by a so-called half-long vowel in the second syllable, and first syllables in quantity 2 and 3 are followed by successively shorter second syllable vowels. Evidence for this type of compensation, which is part of the phonological structure of Estonian words, is given in Table 1 and in Figures 1, 2, and 3.

Figure 1 shows the average durations of segments in productions of vaga-vaka-vakka by the two speakers. As may be seen, there is some adjustment in the duration of the vowel of the second syllable, which partly compensates for the increasing duration of the intervocalic consonant. A similar observation may be made with respect to the set sada-saada (2) - saada (3), displayed on Figure 2. Here the duration of the vowel of the second syllable is inversely correlated with the vowel of the first syllable, whose duration is contrastive.[7]

Figure 3 shows the set saag-saak-sakk. Here the compensation is between the vowel and the final consonant, both of which are contrastive on the segmental level. It is interesting that the total durations of the words saag and saak are practically identical: the compensation is complete, and the two monosyllabic words really differ only in the distribution of duration among the two contrastive segments. The third member of the set, sakk, contains what is commonly analyzed as a quantity 1 vowel and a quantity 3 final consonant. In terms of measurable duration, this adds up to somewhat less than the 3 + 1 sequence in saag, where a vowel in quantity 3 is followed by a consonant in quantity 1. As far as the word saak is concerned, the assignment of the vowel and the final consonant to phonemic quantities remains ambiguous on phonetic grounds. For both speakers, the duration of /a/ in saak is longer than that of /a/ in saada (2), but shorter than /a/ in saada (3); the duration of /k/ in saak is likewise between the durations of /k/ in vaka and vakka. The pertinent data are given in Table 1.

The temporal compensation with which we are primarily concerned in the current paper is of a different kind. It is manifested not in the pattern itself, but in its realization. We hypothesize that there exists a temporal program for the production of an utterance. At a certain level in the process of the production of the utterance, the sequence of articulatory gestures is programmed, and the utterance is assigned an overall basic duration. If this is true, then repeated productions of the same utterance will aim at a duration close to the average for a series of productions. In order that this may be accomplished, temporal adjustment will take place between successive segments during a single production: if one of the segments is produced with a duration that is longer than its own average, another segment within the same utterance will be relatively shorter than its respective average, so that the duration of the word as a whole will remain more or less constant, i.e. vary as little as possible from the average duration programmed for the word. Each segment will, of course, have some variability, which may be

statistically expressed in terms of variance. If the segments were independent of each other, their variances would be additive, and the variance of the whole word would be the sum of the variances of the segments. If, however, there is temporal compensation among the segments constituting the word, the variance of the word should be less than the sum of the variances of the segments.

Table 2 contains the mean durations of each test word, the sum of variances of the segments, and the variance of the word taken as a whole. Figure 4 presents the same data graphically for the vaga-vaka-vakka set. As is obvious from the table and the figure, temporal compensation is indeed present in all test words, and in general the hypothesis appears to be validated. The study was continued to establish the statistical significance of correlations between all subsets of segments in each test word. A summary of the results is presented in Table 3 and in Figures 5-7.

Figure 5 shows the correlation coefficients (Pearson correlations) for all segments contained within the words vaga, vaka and vakka produced by the two speakers. Specifically, these correlations show the relationship of the first three segments to the fourth. (Correlations between various other combinations of segments are given in Table 3.) As may be seen, the degree of negative correlation is extremely high. The two vertical lines on the figure represent r values that show significance at the .005 and .0005 level respectively; the actually obtained correlations are significant at an even higher level.

Figure 6 presents correlation coefficients for the words sada, saada (2), and saada (3). In this case, the displayed negative correlations were found to obtain between the two syllables--segments 1 and 2 on the one hand, and 3 and 4 on the other hand. As before, the correlations are highly significant.

Figure 7 presents similar data for the monosyllabic words saag, saak and sakk. Here the first consonant and vowel have been correlated with the final consonant. Again, the degree of negative correlation is highly significant.

Not all combinations of segments yielded equally high negative correlations. In most cases, correlations involving the initial consonant and other parts of the word were either significant at a lower level or not significant at all. This may reflect the fact that the duration of the initial consonant is non-contrastive at the segmental level. However, combinations that involved all segments yielded significant negative correlations in all cases.

The hypothesis presented at the beginning of this paper was that words are programmed as units, and that significant relationships exist among all segments that constitute a word. The results of the study have clarified these relationships: the durations of segments constituting a word are negatively correlated, and the level of significance of these negative correlations is much too high to be attributed to chance. Since the timing patterns extend

over the whole word, it may be concluded that words do indeed constitute units of programming. Further research is needed to establish to what an extent these patterns are modified when the word becomes part of a higher-level unit such as a phrase or sentence.

Footnotes

*This research was supported in part by PHS Research Grant No. 1 R03 MH18122-01 from the National Institute of Mental Health, and in part by Grant No. 534.1 from the National Science Foundation to the Computer and Information Science Research Center, The Ohio State University.

[1]The problem is surveyed, and literature cited, in Ilse Lehiste (1970) Suprasegmentals. Cambridge: M.I.T. Press.

[2]For a discussion of the problem, cf. Ilse Lehiste (1971) "Temporal organization of spoken language," in Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen. Edited by L. L. Hammerich, Roman Jakobson, and Eberhard Zwirner. Copenhagen: Akademisk Forlag, 159-169.

[3]The words are given in standard spelling. The letter g stands for a voiceless lenis plosive, which is the realization of /k/ in quantity 1. Traditional spelling does not distinguish between long and overlong vowels, both of which are written with two vowel letters.

[4]I would like to thank my informants for the generous contribution of their time, and the researchers at the Institute for Language and Literature and the Laboratory of Experimental Phonetics of the Academy of Sciences of the Estonian S.S.R. for their cooperation and assistance in making the recordings.

[5]The analysis techniques are described in detail in L. Shockey, R. Gregorski, and I. Lehiste (1971) "Word unit temporal compensation"." Ohio State University Working Papers in Linguistics No. 9.

[6]John Ohala (1970) Aspects of the Control and Production of Speech. UCLA Working Papers in Phonetics No. 15, Los Angeles; V. A. Kozhevnikov and L. A. Chistovich (1965) Speech:Articulation and Perception. Translated by J.P.R.S., Washington, D.C., No. JPRS 30543. Moscow-Leningrad.

[7]There is some controversy over the question whether a first syllable in quantity 2 is followed by a half-long vowel or not. In the present set of data, one of the speakers had successively shorter second-syllable vowels in vaga-vaka-vakka, the other in sada-saada-saada.

Table 1

Mean durations (in milliseconds) of segments in nine test words produced by two speakers. N = 50.

Word and speaker	C_1	V_1	C_2	V_2
ÕP vaga	71.34	120.74	98.68	257.26
ÕP vaka	57.24	103.54	188.82	223.84
ÕP vakka	57.86	105.54	397.26	187.68
EJ vaga	51.40	128.22	71.94	257.24
EJ vaka	45.06	94.54	204.20	210.28
EJ vakka	50.18	92.42	376.32	203.02
ÕP sada	131.30	128.38	73.56	251.98
ÕP saada (2)	136.72	255.88	76.96	190.68
ÕP saada (3)	139.30	454.56	100.80	185.40
EJ sada	130.96	101.62	53.84	232.16
EJ saada (2)	117.88	191.46	72.36	196.44
EJ saada (3)	122.94	275.96	78.10	165.76
ÕP saag	141.14	486.20	85.96	
ÕP saak	136.78	316.46	271.00	
ÕP sakk	119.86	115.98	316.70	
EJ saag	158.16	419.82	118.40	
EJ saak	131.72	222.64	351.14	
EJ sakk	143.68	104.82	350.80	

Table 2

Mean durations (in milliseconds) and variances of nine test words produced by two speakers. N = 50.

Word and speaker	Mean duration	Sum of variances of segments	Variance of word
ÕP vaga	548.02	527.33	118.38
ÕP vaka	573.44	582.72	84.06
ÕP vakka	748.34	1345.85	487.63
EJ vaga	508.80	796.56	199.75
EJ vaka	554.08	598.35	132.31
EJ vakka	721.94	1138.20	193.13
ÕP sada	585.22	452.00	238.56
ÕP saada (2)	660.24	896.20	161.56
ÕP saada (3)	880.06	1257.32	305.63
EJ sada	518.58	610.43	203.75
EJ saada (2)	578.14	390.80	70.88
EJ saada (3)	642.76	751.75	213.50
ÕP saag	713.30	601.10	297.19
ÕP saak	724.24	592.22	173.06
ÕP sakk	552.54	621.51	202.13
EJ saag	696.38	1753.70	655.06
EJ saak	705.50	1196.46	264.25
EJ sakk	599.30	1005.59	330.63

Table 3

Correlation coefficients between various combinations of segments in productions of nine test words by two speakers.

$N = 50$; $r = \frac{1}{N} \sum \left(\frac{X - \bar{X}}{\sigma X} \right) \left(\frac{Y - \bar{Y}}{\sigma Y} \right)$. Significance of r at .235 - .95, at .279 - .99, at .361 - .995, and at .451 - .9995.

Word	Segments involved in the correlation	Correlation coefficient	
		Speaker ÕP	Speaker EJ
vaga	1, 2	-0.178	-0.468
	2, 3	-0.142	-0.367
	3, 4	-0.353	-0.386
	1, 2, 3, 4	-0.738	-0.660
vaka	1, 2	-0.310	-0.148
	2, 3	-0.046	-0.333
	3, 4	-0.652	-0.523
	1, 2, 3, 4	-0.770	-0.730
vakka	1, 2	0.058	-0.035
	2, 3	-0.186	-0.219
	3, 4	-0.509	-0.717
	1, 2, 3, 4	-0.556	-0.776
sada	1, 2	-0.236	-0.170
	2, 3	-0.404	-0.428
	3, 4	0.166	-0.391
	1, 2, 3, 4	-0.530	-0.543
saada (2)	1, 2	0.121	-0.381
	2, 3	-0.328	-0.314
	3, 4	0.104	-0.116
	1, 2, 3, 4	-0.834	-0.765

Word	Segments involved in the correlation	Correlation coefficient	
		Speaker ÕP	Speaker EJ
saada (3)	1, 2	-0.467	-0.288
	2, 3	-0.382	-0.609
	3, 4	-0.139	0.028
	1, 2, 3, 4	-0.667	-0.668
saag	1, 2	-0.155	-0.191
	2, 3	-0.604	-0.529
	1, 2, 3	-0.637	-0.590
saak	1, 2	-0.257	-0.312
	2, 3	-0.475	-0.205
	1, 2, 3	-0.668	-0.768
sakk	1, 2,	-0.100	0.112
	2, 3	-0.406	-0.326
	1, 2, 3	-0.708	-0.685

Figure 1. Average durations of segments in the three words vaga, vaka and vakka, produced by two informants.

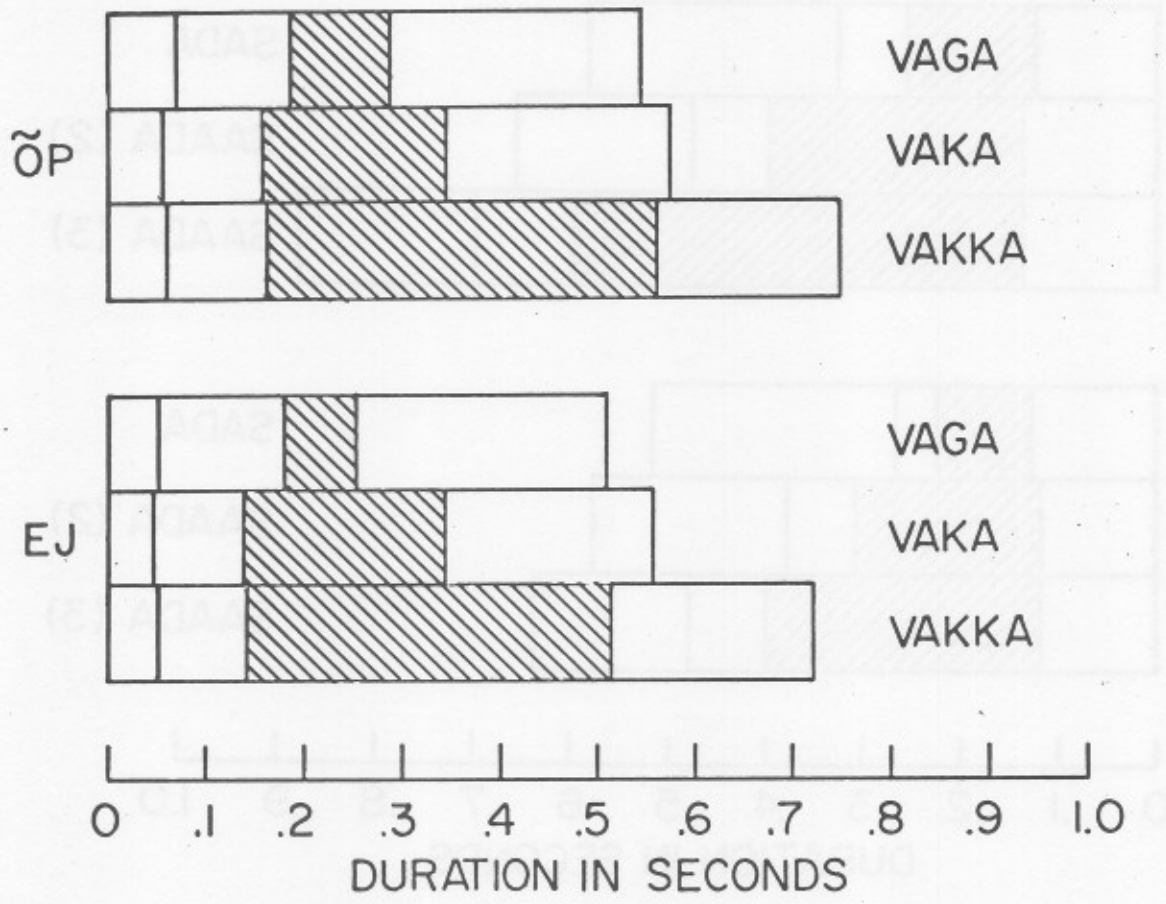


Figure 2. Average durations of segments in the three words sada, saada (2) and saada (3), produced by two informants.

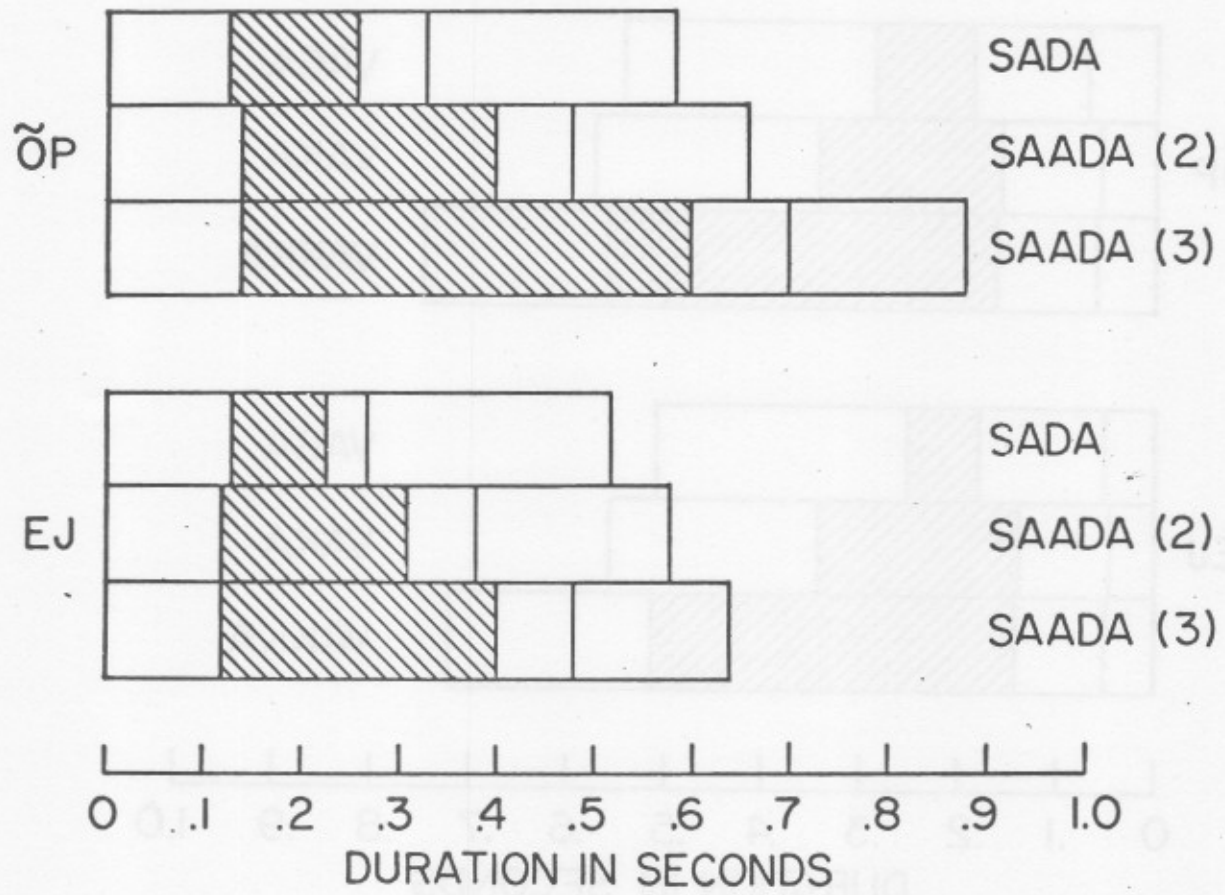
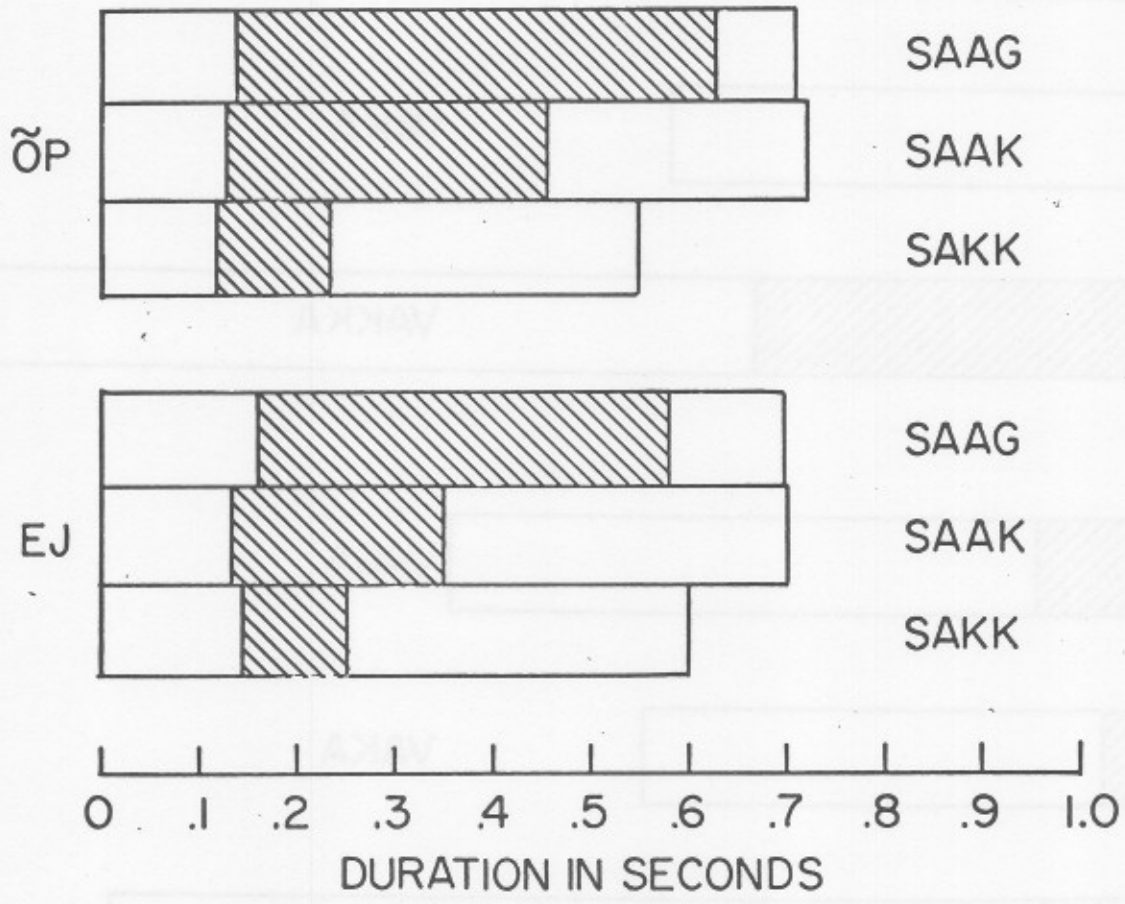


Figure 3. Average durations of segments in the three words saag, saak and sakk, produced by two informants.



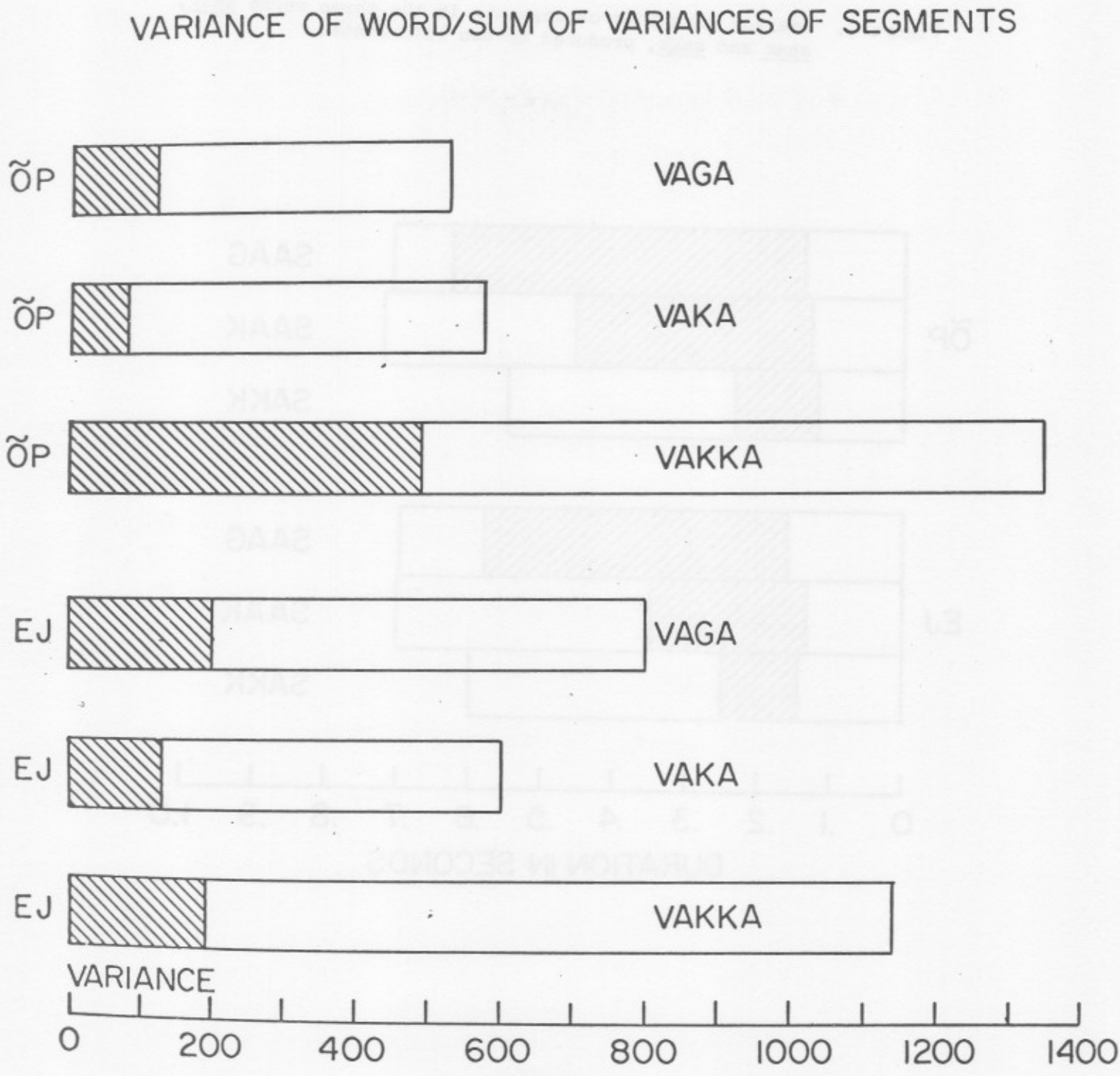


Figure 4. Variance of the word (diagonally hatched) and sum of variances of segments in productions of the words vaga, vaka and vakkā by two speakers. Variance of word is superimposed on the sum of variances of segments.

Figure 5. Correlation coefficients (r) between the first three segments and the fourth segment contained in the words vaga, vaka and vakka produced by two speakers.

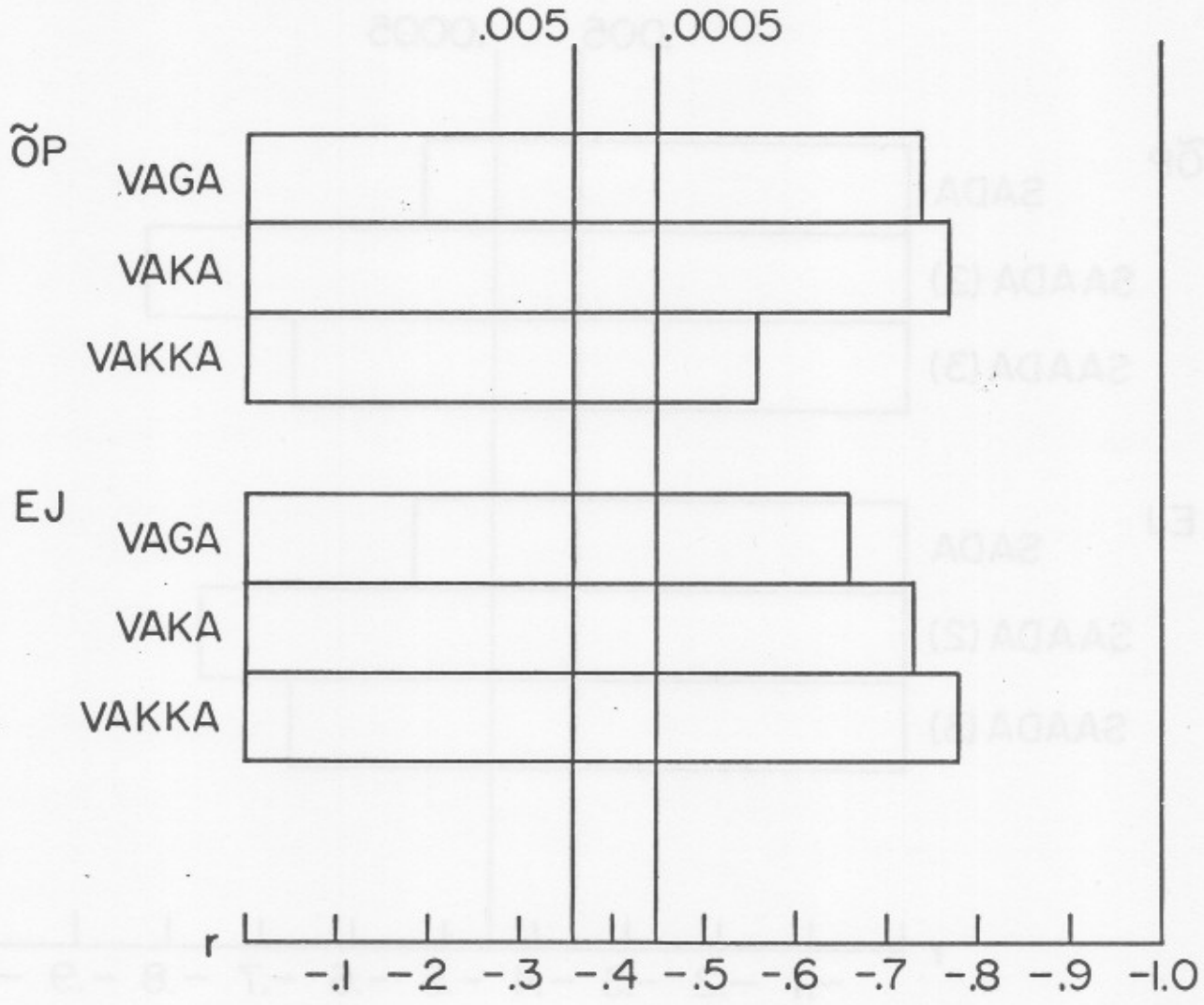


Figure 6. Correlation coefficients (r) between the first two and last two segments contained within the words sada, saada (2) and saada (3) produced by two speakers.

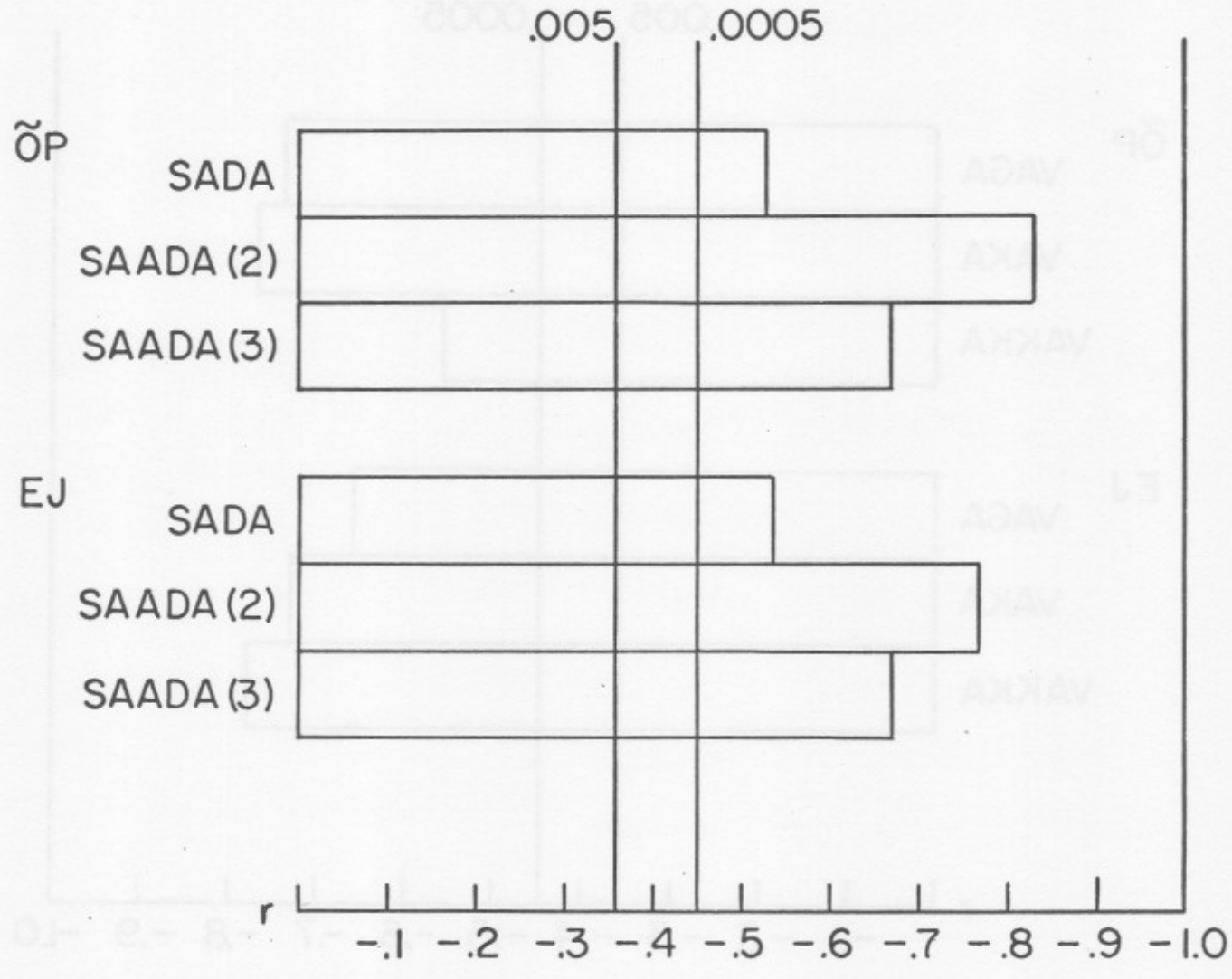
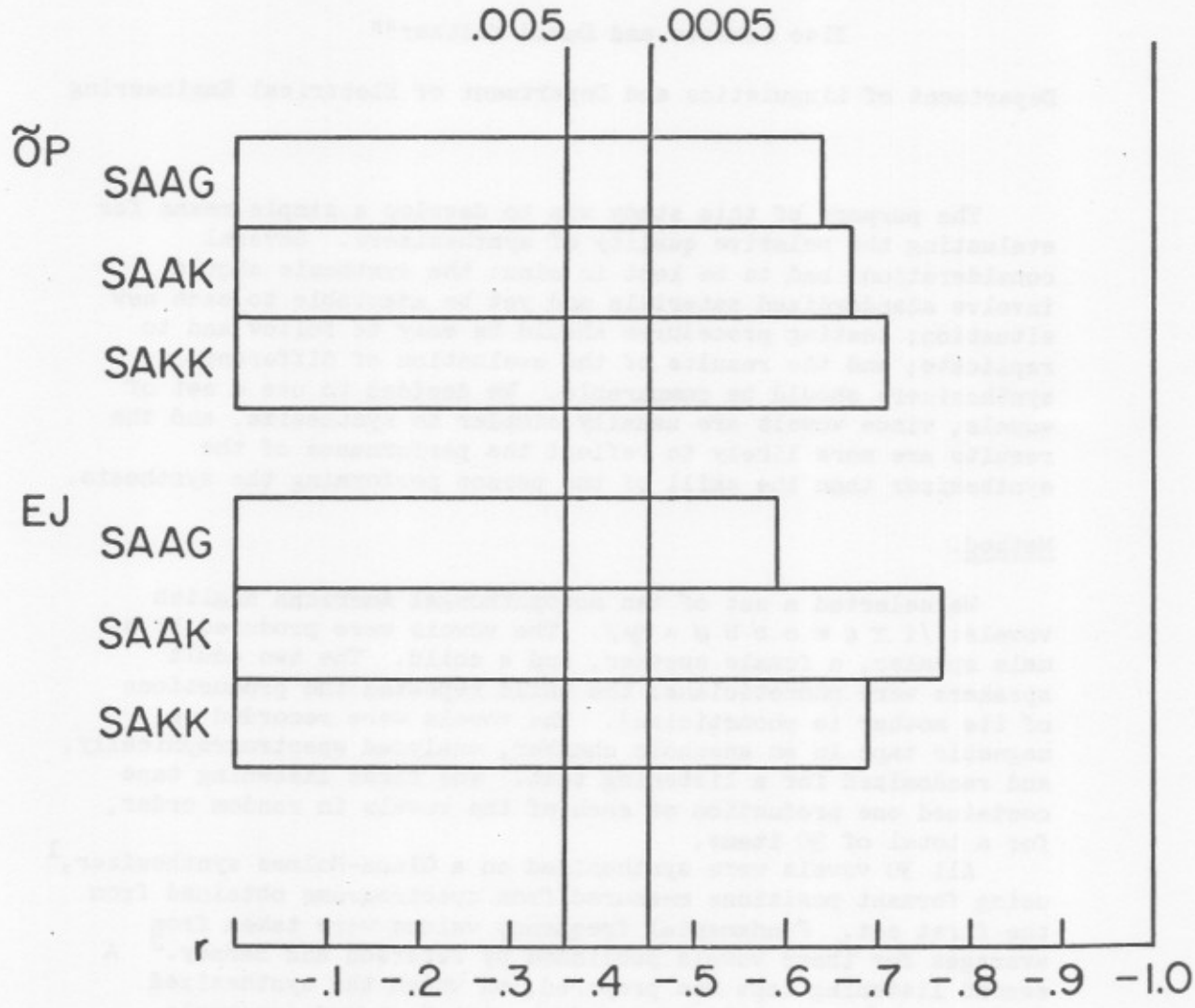


Figure 7. Correlation coefficients (r) between the first two segments and the third segment contained within the words saag, saak and sakk produced by two speakers.



Vowel and Speaker Identification in Natural and Synthetic Speech*

Ilse Lehiste and David Meltzer**

Department of Linguistics and Department of Electrical Engineering

The purpose of this study was to develop a simple means for evaluating the relative quality of synthesizers. Several considerations had to be kept in mind: the synthesis should involve standardized materials and yet be adaptable to each new situation; testing procedures should be easy to follow and to replicate; and the results of the evaluation of different synthesizers should be comparable. We decided to use a set of vowels, since vowels are usually simpler to synthesize, and the results are more likely to reflect the performance of the synthesizer than the skill of the person performing the synthesis.

Method.

We selected a set of ten monophthongal American English vowels: /i ɪ ɛ æ ɑ ɔ U u ʌ ɝ/. The vowels were produced by a male speaker, a female speaker, and a child. The two adult speakers were phoneticians; the child repeated the productions of its mother (a phonetician). The vowels were recorded on magnetic tape in an anechoic chamber, analyzed spectrographically, and randomized for a listening test. The first listening tape contained one production of each of the vowels in random order, for a total of 30 items.

All 30 vowels were synthesized on a Glace-Holmes synthesizer,[1] using formant positions measured from spectrograms obtained from the first set. Fundamental frequency values were taken from averages for these vowels published by Peterson and Barney.[2] A second listening tape was prepared, on which the synthesized 30 vowels appeared in random order, with 5-second intervals.

A third set of vowels was generated on the basis of average formant and fundamental frequency values published by Peterson and Barney. In this set, formant values for men, women and children were combined with the respective fundamental frequencies, resulting in 9 different combinations for each of the ten vowels. Three of the nine sets are directly comparable to the materials contained on the first two tapes; six represent combinations which do not occur in normal speech. These combinations were synthesized to gain some information about the relative importance of formant structure and fundamental frequency in the identification

of speakers and vowels. The third set of vowels, 90 items in all, was randomized and re-recorded in the same manner as the second set.

Six listening tapes were prepared, containing all 150 stimuli. On each tape, the order of the three sets of vowels was varied, so that the effects of order of presentation of normal vowels and different synthetic vowels would be equalized. The tapes were presented to 10 listeners each, for a total of 60 listeners. The listeners had had approximately three months' training in (English) phonetics and were familiar with the phonetic symbols. The task of each listener was to identify both the vowel and the speaker by placing the proper phonetic symbol in one of three columns, thus assigning the vowel to a male speaker, female speaker, or a child.

Results.

The results of the listening tests are presented in Tables 1-7. Table 1 presents the results of vowel identification for normal productions. Table 2 gives comparable data for the set of vowels synthesized on the basis of measurements made from the first set.

Table 3 presents comparable data for the set of vowels synthesized on the basis of the Peterson-Barney averages. This table contains only normal combinations, i.e. male formants and fundamental frequency, female formants and fundamental frequency, and child's formants and fundamental frequency. It is thus directly comparable to Tables 1 and 2. Table 4 summarizes the vowel and speaker identification data for these three sets of vowels: normal productions, synthesis from measured values (attempting to recreate the first set synthetically), and synthesis from average values.

Tables 5, 6 and 7 present data for the set of 90 vowels synthesized on the basis of averages. The tables contain information obtained for all nine possible combinations of formant and fundamental frequencies. Table 5 presents speaker identification scores. Table 6 was generated by averaging Table 5 results across fundamental frequency changes and across formant structure changes. Table 7 gives vowel identification scores.

Discussion.

1. Speaker Identification.

A first observation is that male speakers are identified more easily than women and children, who are frequently confused with each other. This would seem to be a trivial observation; it is interesting, however, that the confusions are much greater in synthesized sets than in the normal set. Evidently the normal productions contain some additional information which is used by listeners in making the decision, and which is not reproduced on

the Glace-Holmes synthesizer.

Tables 5 and 6 show that formant structure is a relatively more important cue in speaker identification than fundamental frequency. For example, vowels produced with male formants, but female fundamental frequency, were assigned to a male speaker in 80.8% of instances, while vowels synthesized with female formants, but with male fundamental frequency, were assigned to a male speaker in only 18.6% of the cases.

2. Vowel Identification.

First of all, it is obvious that children's vowels are relatively difficult to identify. In the case of the first two sets (Tables 1 and 2), one might attribute this to the fact that the child whose recording of the vowels was used in this test may not have succeeded in pronouncing the vowels correctly. But a comparison with synthesis from the Peterson-Barney averages (Table 3) shows that this is not so: here, too, the score for children's vowels was the lowest, and the reason must be sought elsewhere. A simple answer might be provided by observing that children's formants are usually not well defined, since the high fundamental frequency of a child's voice would furnish only one or two harmonics per formant. If this is the true reason, the identifiability of a child's vowels should increase when a man's fundamental frequency is used. Table 7 shows that this is not the case: children's formants, with a male fundamental frequency, resulted in an average vowel identification score of 43.9%, compared to 67.9% for children's formants combined with children's fundamental frequency.

It is noticeable also that synthesis from averages produced relatively higher vowel identification scores than synthesis from measurements of the normal set. A possible reason is that Peterson and Barney used for their averages only vowels that had been correctly identified by a panel of listeners, and discarded those that were not unanimously accepted. Thus the Peterson-Barney averages represent some kind of idealized vowels--not what an average speaker would produce, but what an average listener would accept.

Vowels obviously differ a great deal in their relative identifiability. In normal productions, the vowels /U/ and /ʌ/ had the lowest scores. Surprisingly, /ʌ/ had a relatively high score in the synthetic set based on measurements; in this set, the lowest scores were obtained for /ɛ/ and /U/. For the set of vowels synthesized from averages, the lowest scores were associated with /ʌ/ and /U/, as had been the case with the normal set. High front vowels and /ɝ/ had consistently high identification scores.

A surprising result was the low identification score of /i/ in the set synthesized from measurements (Table 2). We hypothesize that this might be due to the fact that the fourth

formant was not used in the synthesis; however, /i/ synthesized on the basis of averages received a high score, even though F_4 was not used either. The relatively high score for the child may be explained by the fact that a modification was introduced into the synthesizer to obtain the characteristic high third formant for the child's /i/, which would otherwise have been out of range of the Glace-Holmes synthesizer.

An analysis of the substitutions made by the listeners would add some interesting information, but would contribute little to the primary aim of the study: establishing an evaluation measure for synthesizers.

We propose to use the difference between normal scores and scores obtained with synthetic vowels as an evaluation measure. The use of the Peterson-Barney data will provide a fixed reference. For the current state of our Glace-Holmes synthesizer, we have to evaluate its performance as approximately 25% below normal speech. This is based on a comparison of overall scores. The overall vowel identification score for the normal set (all three speakers combined) was 79.46%; the overall speaker identification score (all ten vowels combined) was 90.03%. The corresponding scores for the set synthesized from measured spectrograms were 50.87% and 69.73% respectively. The differences between the scores obtained for the normal set and the synthesized set were -28.59% for vowel identification and -20.30% for speaker identification, giving an approximate degradation of the signal of 25%. Compared with the synthesis from averages, the performance of the Glace-Holmes synthesizer is much better: the difference for vowel identifications between the normal set and the synthesis from averages was -4.23%, and for speaker identification, -15.76%, for an average degradation of 10%.

Footnotes

*This research was supported in part by the National Science Foundation under Grant GN-534.1 from the Office of Science Information Service to the Computer and Information Science Research Center, The Ohio State University. The paper was presented at the 82nd meeting of the Acoustical Society of America, Denver, Colorado, October 21, 1971. The authors are indebted to Dr. G. Powers of the Speech Department of The Ohio State University for his help in carrying out the listening tests.

**David Meltzer is currently with the I.B.M. Corporation, Poughkeepsie, New York.

[1]Glace, Donald A. A Parallel Resonance Synthesizer for Speech Research. Unpublished Manuscript.

[2]Peterson, Gordon E., and Harold L. Barney (1952) "Control Methods Used in a Study of the Vowels." JASA 24.175-184.

TABLE 1
VOWEL IDENTIFICATION: PHONATED VOWELS, NORMAL SPEAKERS
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (vowel & speaker)	Overall correct (MFC combined)
i	100	90	72	87.33	93.00
ɪ	96	74	87	85.67	93.67
ɛ	70	81	97	82.67	85.67
æ	96	77	90	87.67	73.67
a	94	57	25	58.67	73.67
ɔ	81	67	64	70.67	77.33
U	80	63	10	51.00	56.33
u	98	75	90	87.67	98.67
ʌ	72	54	0	42.00	48.33
ɝ	96	78	31	68.33	74.33
Average	88.3	71.6	56.6	72.17	79.46

TABLE 2
VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, BASED ON MEASUREMENTS OF PRODUCTIONS OF NORMAL SPEAKERS
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (Vowel & speaker)	Overall correct (MFC combined)
i	8	8	54	23.33	36.00
ɪ	12	21	14	15.67	19.67
ɛ	19	52	12	27.67	39.67
æ	79	65	70	71.33	93.00
a	46	45	42	44.33	68.00
ɔ	47	42	4	31.00	44.33
U	10	4	24	12.67	19.33
u	49	12	8	23.00	34.33
ʌ	73	30	44	49.00	72.33
ɝ	50	74	38	54.00	82.00
Average	39.3	35.3	31.0	35.2	50.87

TABLE 3
VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON & BARNEY (1952)
Scores given in per cent correct

Vowel	Male	Female	Child	Overall correct (Vowel & speaker)	Overall correct (MFC combined)
i	84	62	62	69.33	88.67
ɪ	76	47	70	64.33	77.00
ɛ	81	53	68	67.33	83.00
æ	79	60	72	70.33	89.67
ɑ	60	56	53	56.33	76.67
ɔ	67	42	20	43.00	59.67
ʊ	67	29	20	38.67	58.00
u	86	49	21	52.00	76.33
ʌ	37	38	29	34.67	47.00
ɝ	98	65	57	73.33	96.33
Average	73.5	50.1	47.2	56.93	75.23

TABLE 4
OVERALL SPEAKER AND VOWEL IDENTIFICATION
Scores given in per cent correct

Stimulus type	Speaker identification			Overall speaker identification score	Overall vowel identification score
	Male	Female	Child		
Normal speakers					
Male	99.2	0.4	0.4		
Female	2.2	81.0	16.8		
Child	0.0	10.1	89.9		
				90.03	79.46
Synthesis from measurements					
Male	96.2	3.0	0.8		
Female	9.8	62.2	28.0		
Child	5.2	44.0	50.8		
				69.73	50.87
Synthesis from averages					
Male	94.0	2.7	3.3		
Female	9.4	60.6	30.0		
Child	4.7	27.1	68.2		
				74.27	75.23

TABLE 5
SPEAKER IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON AND BARNEY (1952)
All vowels combined. Scores given in per cent correct.

Formants	Fundamental frequency	Identified as		
		Male	Female	Child
Male	Male	94.0	2.7	3.3
	Female	80.8	10.4	8.8
	Child	69.7	11.4	18.9
Female	Male	18.6	50.5	30.9
	Female	9.4	60.6	30.0
	Child	7.2	43.2	49.6
Child	Male	11.2	39.6	49.2
	Female	7.5	44.3	48.2
	Child	4.7	27.1	68.2
Average		33.68	32.20	34.12

TABLE 6
SPEAKER IDENTIFICATION, BASED ON A) FUNDAMENTAL FREQUENCY AND B) FORMANT STRUCTURE
All vowels combined. Scores given in per cent correct

		Identified as		
		Male	Female	Child
Fundamental frequency	Male	41.27	30.93	27.80
	Female	32.57	38.43	29.00
	Child	27.20	27.23	45.57
Formants	Male	81.50	8.17	10.33
	Female	11.73	51.43	36.84
	Child	7.80	37.00	55.20

TABLE 7

VOWEL IDENTIFICATION: SYNTHESIZED VOWELS, FORMANT STRUCTURE AND FUNDAMENTAL FREQUENCY BASED ON AVERAGES GIVEN BY PETERSON AND BARNEY (1952)

Scores given in per cent correct.

Formants	Fundamental frequency	i	ɪ	ɛ	æ	ɑ	ɔ	ʊ	u	ʌ	ɝ	Average
Male	Male	88	78	81	86	64	67	69	86	41	100	76.0
	Female	96	66	73	96	87	78	61	83	46	80	76.6
	Child	78	31	40	89	83	44	10	14	26	19	43.4
Female	Male	88	54	54	53	52	22	37	40	42	96	53.8
	Female	98	76	91	91	84	73	66	83	58	98	81.8
	Child	78	31	40	89	83	44	10	14	26	19	43.4
Child	Male	83	35	24	3	25	17	46	75	35	96	43.9
	Female	98	70	87	92	47	61	63	72	85	96	77.1
	Child	80	77	77	92	82	39	39	60	42	91	67.9
Average		87.44	57.56	63.0	76.78	67.44	49.44	44.56	58.56	44.56	77.22	62.66

On the Perception of Coarticulation Effects in English VCV Syllables

Ilse Lehiste and Linda Shockey

Abstract

Öhman's (1966) investigation of the acoustic correlates of coarticulation in VCV sequences indicates that terminal formant frequency transition values are strongly influenced by the nature of the transconsonantal vowel. This experiment was designed to explore the perceptual correlates of Öhman's spectrographic findings. It was discovered that when a VCV sequence (where C is a voiceless plosive) is cut in two during the period of consonantal closure, there are not enough remaining cues in either the resulting VC or CV sequences to allow for identification of the deleted segment or of its articulatory features. However, it appears that coarticulation effects may hinder recognition of non-final allophones placed artificially in final position: consonants in VC sequences spliced from original VCV utterances are more difficult to identify than unreleased final consonants of the same quality.

Introduction

This investigation was prompted by the observation that while a good deal is known about the acoustic manifestation of coarticulation (with respect to point of articulation), there seemed to exist no published data regarding the perceptibility of the effects of this kind of coarticulation. In 1966, Öhman published the results of an extensive study dealing with coarticulation in English VCV sequences. He found that formant transitions from the first vowel to the intervocalic consonant are strongly influenced by the phonetic quality of the vowel following the consonant. We decided to investigate whether the changes in formant transitions due to the anticipation of the following vowel are perceptually significant.

Method

A set of VCV utterances was constructed, in which the vowels were /i æ a u/ and the consonants /p t k/. The 4 x 3 x 4 = 48 utterances were recorded by one informant (a low-pitched female native speaker of English). In addition, 12 VC and CV syllables were recorded, in which the four vowels were followed and preceded

by the three consonants each. The recordings were made in an anechoic chamber, using high-quality equipment. The VCV syllables were cut in two parts, placing the cut in the voiceless plosive gap. Using splicing techniques, four randomized lists were constructed. The first consisted of syllables from which the consonant release and the second vowel were removed. Each stimulus appeared twice on the listening test, for a total of 96 items. The task of the listeners was to identify the missing final vowel. The purpose of the test was to determine whether the transitions from the initial vowel to the consonant carried enough information to make this possible.

The second listening test consisted of syllables from which the first vowel had been removed. There were 96 test items. The task of the listeners was to identify the missing initial vowel.

List three contained 12 syllables produced by removing the initial vowel and consonant transition from symmetrical VCV utterances. Each stimulus appeared twice, randomly mixed with 2 x 12 syllables consisting of the same consonants and vowels, produced as CV sequences. The task of the listeners was to identify the 48 initial consonants. List four was similar, except that the stimuli consisted of VC sequences and the listeners had to identify 48 final consonants.

The listening tests were administered to untrained listeners, who were mostly sophomore-level students at The Ohio State University. 23 listeners took the first test, 36 the second; test 3 was taken by 41 listeners, and test 4 by 50. The data thus consist of 2,208 responses to Test 1, 3,456 responses for Test 2, 1,968 responses for Test 3, and 2,400 responses for Test 4.

Results

The results of the first two listening tests were largely negative, even though the same kind and degree of coarticulation effects reported by Ohman were measured on spectrograms made from our test tape. The listeners were evidently not able to identify the missing vowel. They were told that it was one of the four vowels /i æ a u/, and the results show that they were assigning these four vowels in an essentially random manner.

Percentages of correct responses for Test 1 (identify missing final vowel) ranged from 16 to 30% over all possible VC-combinations. The average of correct responses was 24%. In Test 1, the vowel actually produced on the tape was chosen for an answer in 24.5% of the total responses.

In Test 2 (identify missing initial vowel), percentages correct ranged from 19 to 30% over all possible -CV combinations. The average correct was 24.4%. However, in Test 2 there was a strong tendency among subjects to indicate the missing vowel as being identical with the one following the consonant, i.e. the one plainly audible from the recording. Of the total responses, 45.9% were instances of choosing the vowel heard. Out of the total correct scores, nearly half (42.73%) were due to "correct"

identification of formerly symmetrical utterances, e.g. [apa], [iti]. Obviously then, the bias toward selecting the vowel heard is obscuring the number of correct responses. Whether this result is attributable to any sort of coarticulation phenomenon is unverifiable; it may simply be that the subjects were accustomed to identifying the vowel following the consonant after taking Test 1, but since 1) the subjects were given repeated instructions before taking Test 2, and 2) considerably more people took Test 2 than Test 1, this explanation seems unlikely. A satisfactory explanation does not appear to be possible at this time.

Except for the above, there seemed to be no significant trends in the incorrect responses for either test. Incorrect responses did not tend to fall into classes sharing some feature with the correct response, such as high/low or front/back.

The responses to Tests 3 and 4 show a clearer pattern, and will be discussed with reference to Tables 1 - 5.

Table 1 presents summary data about the identification of initial and final consonants.

TABLE 1

Identification of Initial Consonants (all vowels combined)				
	p	t	k	
#C	89.02%	90.54%	88.10%	89.22%
-C	88.71%	92.68%	86.28%	89.22%
Identification of Final Consonants (all vowels combined)				
Correct scores	p	t	k	Overall correct
C# (Released)	59.25%	90.25%	92.50%	80.67%
C- (Truncated)	54.75%	32.75%	29.50%	39.00%

A first observation is that in initial position, there is no difference between the correct identification scores of initial consonants produced as CV sequences and derived by tape-cutting from VCV sequences. The overall scores are identical and fairly high, 89.22% in both cases. The identification of final consonants is much less reliable. At 80.67%, the overall correct score for released final plosives approaches that of initial consonants; however, the three consonants differ in their relative identifiability, since /p/ has a significantly lower score than /t/ or

/k/. There was no such difference among the initial consonants. The identification of truncated final plosives has an overall score of 39.00%, with /p/ ranking higher than /t/ and /k/.

Table 2 presents a confusion matrix for released and truncated final plosives.

TABLE 2
IDENTIFICATION OF FINAL CONSONANTS (ALL VOWELS COMBINED)

Released	p	t	k
p	59.25	11.75	29.00
t	3.75	90.25	6.00
k	3.00	4.50	92.50
Truncated			
p	54.75	20.50	24.75
t	42.75	32.75	24.50
k	45.00	25.50	29.50

All vowels are combined in these results. It becomes obvious from this table that for released final plosives, the primary confusion was between final /p/ and /k/. As regards the truncated final plosives, the relatively high score of /p/ becomes less striking in view of the fact that /t/ and /k/ were both identified as /p/ far more frequently than they were correctly identified.

Table 3 gives an overview of the effect of different vowels on the identification of final consonants.

TABLE 3
CORRECT IDENTIFICATION OF FINAL CONSONANTS AFTER VARIOUS VOWELS
(ALL CONSONANTS COMBINED)

Preceding Vowel	Released	Truncated
i	87.67	43.0
æ	75.33	42.33
a	73.00	30.00
u	86.67	40.67
Overall correct	80.67	39.00

The highest scores were obtained for /i/ and /u/ for released consonants. In the truncated set, /a/ is associated with a significantly low score, while the other three vowels seem to have had no particular effect on the identifiability of the consonants.

Table 4 presents the data for final consonants arranged in the form of a complete confusion matrix.

TABLE 4
IDENTIFICATION OF FINAL CONSONANTS

Original stimulus (final vowel removed)	Perceived as			Original stimulus	Perceived as		
	p	t	k		p	t	k
ipi	59	23	18	ip	79	5	16
iti	31	42	27	it	2	92	6
iki	36	36	28	ik	6	2	92
æpæ	47	20	33	æp	41	15	44
ætæ	31	34	35	æt	3	93	4
ækæ	14	40	46	æk	4	4	92
apa	48	16	36	ap	45	24	31
ata	60	17	23	at	8	81	11
aka	66	9	25	ak	2	5	93
upu	65	23	12	up	72	3	25
utu	49	38	13	ut	2	95	3
uku	64	17	19	uk	0	7	93

Table 5 gives some results of a spectrographic analysis to which the 24 items of the final consonant test were submitted.

TABLE 5
F_2 - TRANSITIONS AND RELEASES OF FINAL PLOSIVES
FREQUENCIES IN HZ

Preceding Vowel	Final consonant								
	p- F_2	p^h F_2	Release	t- F_2	t^h F_2	Release	k- F_2	k^h F_2	Release
/i/	2800	2900	1600	3000	3000	2500 4500 5250	3000	3000	2500
/æ/	1900	1700	1500	2000	2000	2500 3500 4450	2250	2300	2300
/a/	1450	1500	1400	1500	1750	2500 3500 4250	1400	1450	1650
/u/	1000	950	1450	1250	1250	2500 4000	1050	1050	1500

The table contains terminal values of F_2 transitions toward the final consonant, and center frequencies of energy concentrations observed after the release of final consonants occurring in VC syllables produced as such.

Interpretation of the Data

Let us consider first the differences between the scores for released and truncated plosives. In the case of released /t/ and /k/, the scores are uniformly high. With /t/, the releases always had concentrations of energy at more than one frequency, which distinguishes /t/ releases from other releases following otherwise similar transitions. Compare, for example, the sequences /it/ and /ik/ (Table 5), where both the F_2 terminal frequency and the first energy concentration (and the only one for /k/ visible on the spectrogram) were at the same frequencies. On the other hand, the difference in the releases of /p/ and /k/ after /æ/ evidently was not strong enough to remove the confusion between released /æp/ and /æk/. The confusions between /p/ and /k/ after /a/ and /u/ seem obvious, when the terminal F_2 frequencies and energy concentrations in the release are compared.

A curious finding is the fact that releases did not improve the scores of /æp/ and /ap/ at all (Table 4). In fact, the release in /æp/ seems to have increased the tendency of listeners to identify this stimulus as /æk/. This is strange, since the release

of /æk/ has a high frequency concentration as compared to the release of /æp/. Evidently in this case, the contribution of the release toward differential identification was negligible.

Discussion

The purpose of this investigation was to study the effects of coarticulation on perception. The results turned out to be essentially negative. Whatever the effects of coarticulation in terms of their influence on formant transitions, these effects are not sufficient to have an influence on perception. Thus the anticipation of a following vowel may result in a modification of the transition from a preceding vowel to the intervocalic consonant; but this modification is apparently not sufficient to enable the listeners to identify the following vowel from stimuli from which the following vowel itself was deleted. Likewise, whatever the lingering effects of a preceding vowel on the intervocalic consonant, a deleted initial vowel cannot be identified by listeners on the basis of effects that may have been physically present in the transition from the intervocalic consonant to the second vowel.

There was also no difference between the identification scores of initial consonants produced as such and consonants that became initial after the first vowel was deleted from a VCV sequence.

Only final consonants produced some differences between allophones produced as final and allophones produced originally as medial. Here the allophones preceding final silence are clearly much more easily identified than allophones placed into final position by tape-cutting. It is not immediately obvious how much of that difference is due to the effects of coarticulation. Wang (1959), in an experiment which was in part similar to ours, studied the relative contributions of releases and formant transitions to the correct identification of final plosive consonants. He found little difference between identification scores of released final /p t k/ and final /p t k/ whose releases had been eliminated by tape-cutting. The former ranged between 90-98%, the latter between 73-85%. On the basis of these data, it would seem that the contribution of releases was approximately 15% and the contribution of transitions was approximately 85%. Our listeners achieved overall scores for released final /p t k/ of 80.67% and 39.00% for unreleased final plosives. If the contribution of the releases was approximately 15%, there is still a difference of 25% to be accounted for. We conjectured that the anticipation of another vowel may affect the characteristic transitions to pre-silence final consonants to such an extent that listeners make additional errors in identification.

In order to investigate this hypothesis, we conducted an auxiliary experiment. Using the same speaker, equipment and splicing techniques, we prepared a randomized listening test composed of 12 VCV sequences from which the consonant release and

second vowel were removed and 12 VC syllables in which the final consonant was unreleased, i.e. was produced by the speaker as an unreleased plosive. Twenty listeners were asked to identify the final consonants. These subjects were also given the original listening test for the identification of final consonants, as described above. The results are presented in Table 6.

TABLE 6
PER CENT CORRECT IDENTIFICATION OF FINAL CONSONANTS

Stimulus type	Per cent correct
Truncated VC- (re-test)	34.0
Released VC^{h} (re-test)	91.3
Truncated VC- (auxiliary test)	55.4
Unreleased VC$^{\neg}$ (auxiliary test)	73.7
Released VC^{h} (Wang, 1959)	95.3
Truncated VC- (Wang, 1959)	77.6

For these 20 subjects, identification of truncated plosives on the original test was about 5% lower than for the 50 sophomores, but identification of final released plosives was over 10% higher. The latter result may be attributed to the facts that, first, the re-tests were given to subjects singly or in groups of two, whereas the larger group was tested in a single session; thus the conditions for the re-test were more conducive to producing higher scores. Second, the subjects for the re-test were both more mature and more highly motivated than the 50 students. This, of course, does not explain the lower score in the identification of truncated plosives, but this difference is hardly significant.

On the auxiliary test, there was a higher correct score for the consonants whose releases were eliminated by tape cutting (55.4% as compared to 39.0%). A possible reason is the slightly slower rate of speech which the speaker chose for this recording session. Most significantly, the naturally unreleased consonants show a higher identification score (by nearly 20%) than the consonants placed in final position by tape-cutting. The 73.7% correct score falls within the lower range of Wang's results for final unreleased consonants.

It would thus appear that our hypothesis that coarticulation effects reduce intelligibility in the event that they are found in an environment where they do not occur naturally is supported by these additional data.

It is difficult to say what the actual physical cues or miscues were that caused the lowering of identification scores for plosives which had been placed in final position by the elimination of the second vowel from a VCV sequence. The only obvious case would be the sequence /aka/, in which the second formant transition to medial /k/ has a low terminal frequency, while the corresponding transition from the medial /k/ to the final /a/ has a high initial frequency (cf. also Green (1959), esp. pp. 50-52). It might be expected that the anticipation of the following high frequency would result in a raising of the terminal frequency of the transition from initial /a/ toward medial /k/. As Table 4 shows, no such raising occurred in the utterance produced by our informant.

References

Green, Peter S. (1959) Consonant-Vowel Transitions. Travaux de l'Institut de phonétique de Lund. Lund.

Öhman, S. (1966) "Coarticulation in VCV utterances." JASA 39.1.151-168.

Wang, William S-Y. (1959) "Transition and release as perceptual cues for final plosives." Journal of Speech and Hearing Research 2.66-73.

A Note on Temporal Compensation

Richard Gregorski and Linda Shockey

In "Word-Unit Temporal Compensation," (O. S. U. Working Papers in Linguistics No. 9, 1971) and "Implications of Temporal Compensation for Speech Production Models," (Proceedings of the VII International Congress of Phonetic Sciences. Mouton, in press) we presented some conclusions about the higher-level programming of speech based on negative correlation coefficients between speech units in words and short phrases. In the second of these papers, we expressed some reservations about the methodology we used. Since some interest has been shown in this method of investigating temporal programming, we feel we should discuss the problems that we have encountered.

First, the problem of speech rate. There has been, as far as we know, little research on what characteristics determine different rates of speech, if indeed there is a predominant strategy for changing tempo, and on whether the concept of speech rate should be viewed as absolute or relative; as a series of quantal steps or as a continuum. Even if we were able to distinguish speech rates accurately, we have very little information on how changes in rate affect correlations between segments in an utterance (this is discussed briefly in Kozhevnikov and Chistovich, Speech: Articulation and Perception. Moscow-Leningrad. Translated by J.P.R.S. No. 30, 543, pp. 99 ff.). We tried to solve this problem by 'normalization', which was our term for choosing for examination out of our total data set a subset the members of which were nearly identical in duration (as suggested by John Ohala in his dissertation, Aspects of the Control and Production of Speech, U.S.L.A. Working Papers in Phonetics 15, 1970. He was not working with correlation coefficients, however). Ohala has since pointed out (personal communication) that this procedure introduces negative correlations between elements. (When we tested our non-normalized corpus for negative correlations, we found few, but attributed it to rate mixing.)

Second, the problem of 'complementary halves'. If the normalization procedure mentioned above is applied to a set of utterances and any two mutually exclusive portions tested for correlation, the coefficient will always be very near 1, by definition. It was a mistake on our part to attach any further significance to this fact.

Third, 'correlations at a distance.' Similar correlation coefficients are found between, for example, segments A and B

(adjacent segments) and segments A and F (separated by several intervening elements). It seems that if one considers the A-B case significant, one must attribute a similar significance to the A-F case. But it is not clear to us what these correlations, taken at equal value, tell us about language programming.

Since we have not found solutions to these problems, we feel that it is too early to make any conclusions about temporal patterning of language based on the technique described in our 1971 papers.

Note

In "Word-Unit Temporal Compensation" (p. 153) we incorrectly attributed to John Ohala the notion that the mechanism for isochrony may be part of the linguistic competence of the speaker of English. The reference should have been to George Allen's "The Place of Rhythm in a Theory of Language," U.C.L.A. Working Papers in Phonetics No. 10, 1968.