LANGUAGE AND SPEECH, Vol. 21, Part 4, 1978

VOWEL DURATION CHANGE AND ITS UNDERLYING PHYSIOLOGICAL MECHANISMS[*]

KATHERINE S. HARRIS
The City University of New York
and
Haskins Laboratories

Two explanations have been proposed for the relationship between vowel-target formant frequency and articulatory stress. The first, the "extra energy" hypothesis, suggests that stressing is accompanied by larger signals to the articulators, so that stressed syllables are longer and have more extreme formant values. The second, the "undershoot" hypothesis, suggests that the signals sent to the articulators are of constant magnitude, but that changes in timing result in differences in formant frequency. This view leads to a prediction that the relationship between target formant frequency and duration is fixed, whatever the cause of the duration variation. Acoustic and electromyographic measures were made of productions of nonsense syllables with varying stress and speaking rate, by three speakers. Results fail to support the undershoot hypothesis, since syllable duration and vowel target frequency are independent. While speaking rate variations are accomplished in a different manner by the three speakers, the "extra energy" model for stressing seems to be supported.

INTRODUCTION

The central problem of speech production research has been to explain allophonic variation, that is, the effect of various factors on the articulatory manifestation of a given phone. Traditionally, it has been common to separate context effects into two classes – the so-called "coarticulation" and "timing" effects – however the theoretical problems are the same for both classes. It is recognized also that, in general principle, some allophonic variations are to be considered either idiosyncratic or language-specific, while others may be considered to arise from general properties of the motor organization of the articulatory system. It is with this last, poorly defined, class that most studies have been concerned.

One of the best studied allophonic effects is the effect of syllable stress on vowel production. It has been shown that stressed vowels are commonly both more intense and longer than their unstressed counterparts. In addition, stressed and unstressed vowels

[*] *Earlier versions of this paper were presented at the 50th Meeting of the American Speech and Hearing Association, Washington, D.C., November 1975, and at the 94th Meeting of the Acoustical Society of America, Miami, Florida, December, 1977. The research was supported by NIH grants DE-01774 and NS-13617 to Haskins Laboratories. The techniques for making the acoustic measurements were developed by Thomas Baer; the difficult task of analyzing the data was accomplished by Betty Tuller.*

differ in vowel quality, in that unstressed vowels tend to be neutralized (Lindblom, 1963). The vowel-quality difference is sufficiently substantial that it has been shown to cue the perception of stress in disyllables (Fry, 1958). This sort of quality difference has been shown for a number of languages (Lehiste, 1970).

Two models have been proposed for the effect of stress on vowel colour. The first might be called an "extra energy" model. While the details of the model are not worked out at a physiological level, the general idea is that extra energy is applied to the stressed vowel, with the result that it lasts longer, and the signals to the articulators are a little larger, so that the vowel is further from a neutral vocal tract position (Öhman, 1967; Jones, 1940).

A second model might be called the "undershoot model" (Lindblom, 1963). In Lindblom's model, the difference between target formant values for the vowels of differing duration is a consequence of the change in duration itself. A vowel is specified in the nervous system by a set of signals. When these signals are sent to the articulators, they result in a given vocal-tract shape, the target, unless the signals for a subsequent phone arrive at the articulators too soon, so that the target is not attained because the path of articulatory movement is deflected towards the new target.

These two theories make different predictions about events at the acoustic and physiological levels. For the "undershoot" theory, at the acoustic level, the effect of any change of context on the relationship between duration and formant target frequency is the same, so that the effects of changing stress and speaking rate, for example, are identical. A given decrease in duration will be accompanied by a fixed undershoot. At the control signal level, the size of the signals to the articulators should be constant, as stress or speaking rate is varied. For the "extra energy" hypothesis, vowel duration and target frequency are separable, and stressed syllables should show more extreme formant values than unstressed syllables.

The experiment to be described was developed to test these theories, and beyond that, to gain insight into the speech timing mechanism, by studying the effect of stress and speaking rate on simple nonsense syllables.

METHODS

The speakers were three adults, two females (KSH and FBB) and one male (LJR), all native speakers of American English, and personnel of Haskins Laboratories.

The speech sample was the four-syllable nonsense words /əpipipə/ and /əpipibə/, with stress placed on either the second or third syllable. Subjects read semi-random lists of these four "words" at two self-selected speaking rates, "slow" and "fast." Although 25 repetitions were produced of each utterance, later processing failures reduced the lists to 24 repetitions for LJR and 20 and 21, respectively, for KSH and FBB. Acoustic recordings were made, as well as electromyographic recordings from the genioglossus muscle. Since the genioglossus bunches the main body of the tongue, and brings it forward (Raphael and Bell-Berti, 1975; Smith, 1970), we might expect greater activity from the muscle as fronting increases. In order to ensure at least one successful recording, two electrodes

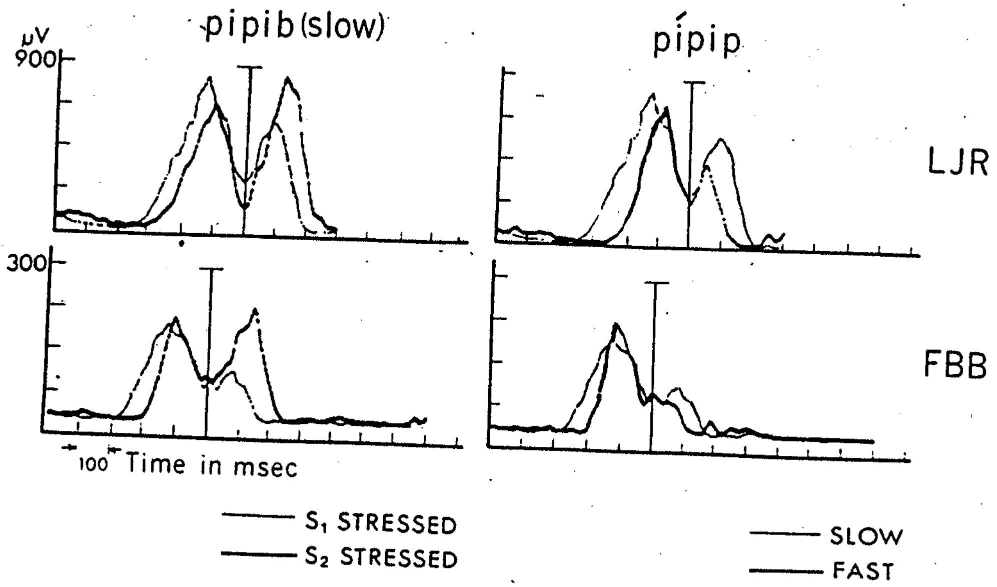

Fig. 1. Averaged EMG activity for the genioglossus muscle for two speakers. Stress contrasts are shown at the left; speaking rate contrasts are shown at the right.

were inserted into the muscle for KSH and FBB, and three for LJR. Since all insertions were successful, we selected those recordings which appeared, on preliminary inspection, to be most stable, for further analysis. The electrode preparation and insertion technique has been reported in detail elsewhere (Hirose, 1971).

All acoustic measurements were made on an interactive computer system at Haskins Laboratories. Duration measures were made on the waveform; the duration for each syllable represents the duration of closure, burst, aspiration and voicing for the central two syllables in the nonsense word, indicated as the first and second syllables below. Measurements of F_2 and F_3 peak frequency were made after spectrographic transformation, although as measures of F_2 frequency were so closely parallel to those of F_3 frequency, they will not be reported. Since low-frequency room noise was recorded during the experiment, reliable measurement of F_1 frequency was not feasible.

The EMG signals were rectified, filtered and averaged using the Haskins Laboratories system, as previously described (Kewley-Port, 1973). Peak EMG activity was measured for each syllable, as a crude indication of overall muscle activity. Some typical averaged interference patterns are shown in Fig. 1.

RESULTS

Results of the acoustic measurements are shown in Fig. 2. Each panel shows the eight data points for F_3 peak frequency as a function of duration, for /p/ or /b/ syllables, for a single speaker. Each point is labelled as to whether it was produced in a first or second

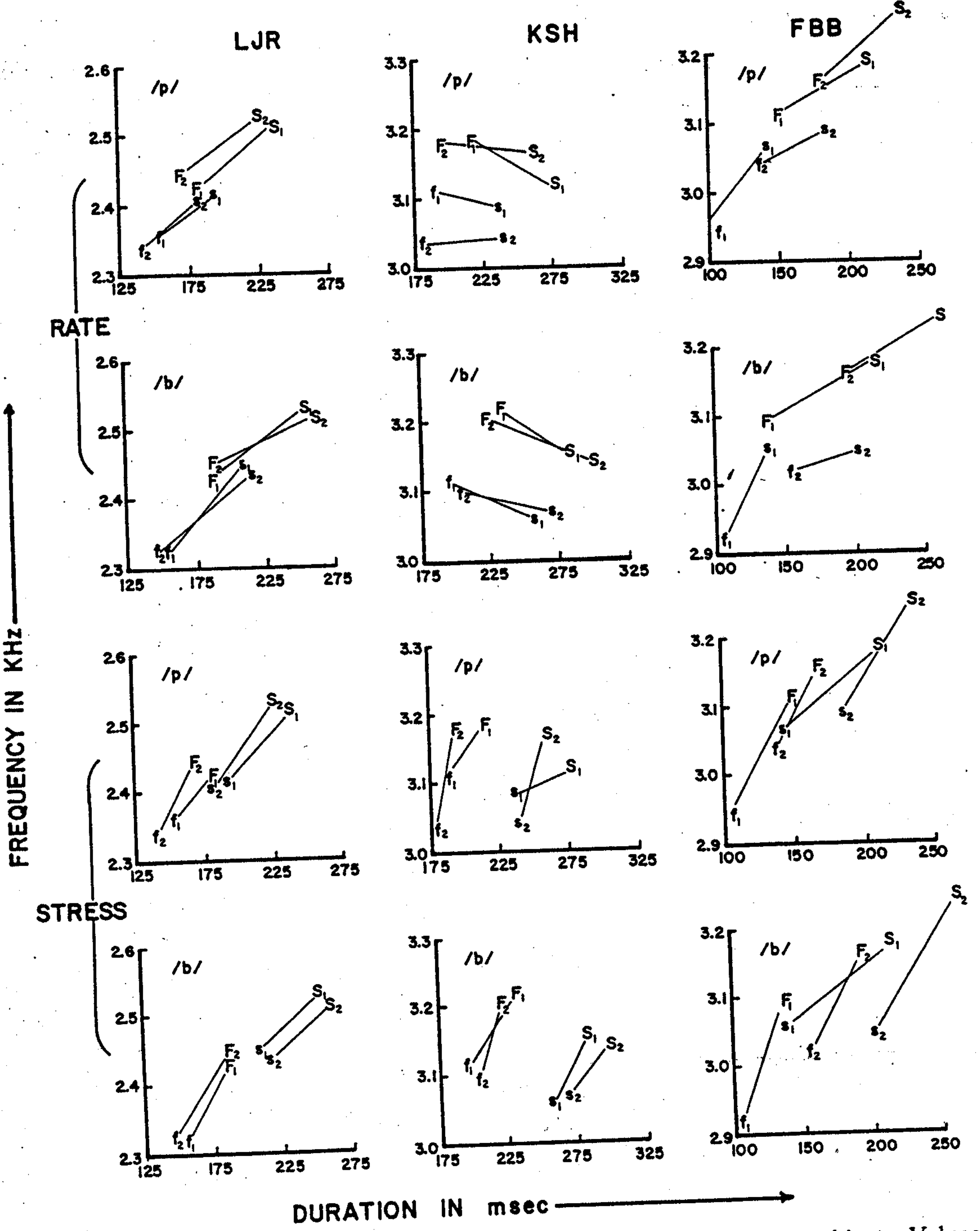

Fig. 2. Peak F_3 frequency plotted against syllable duration, for three subjects. Values for utterances whose final consonant is /p/ are plotted in the first and third rows; those for utterances whose final consonant is /b/ are plotted in the second and fourth rows. Points representing minimal contrasts in speaking rate are connected in the upper six panels; points representing minimal stress contrasts are connected in the lower six panels. Stressed syllables are indicated with upper case letters; unstressed syllables are indicated with lower case letters. Values for first syllables are subscripted with "1"; those for second syllables, with "2." Values for fast speaking rate are indicated with "F"; those for slow speaking rate are indicated with "S."

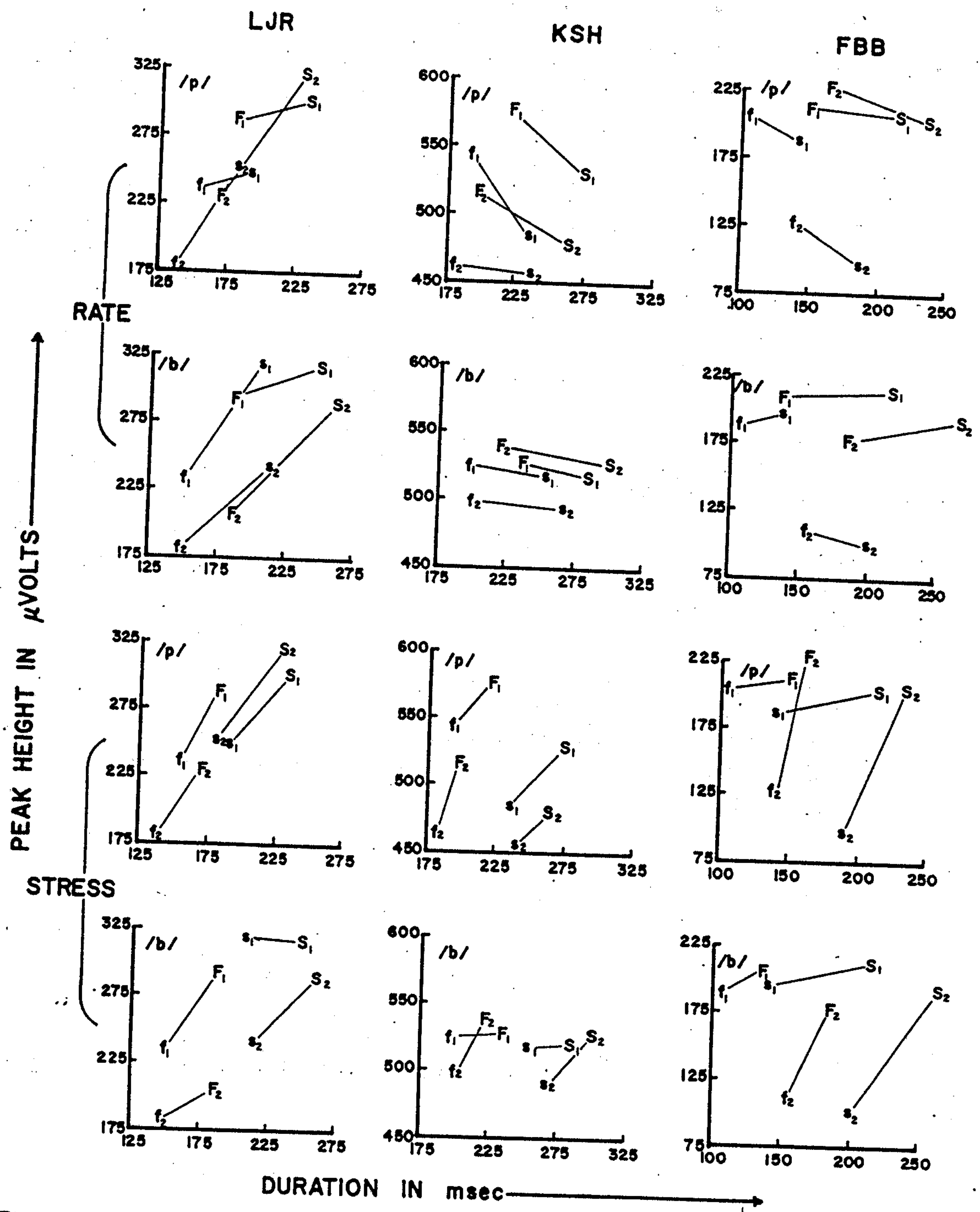

Fig. 3. Peak EMG activity (in microvolts) plotted against syllable duration, for three subjects. Values for utterances whose final consonant is /p/ are plotted in the first and third rows; those for utterances whose final consonant is /b/ are plotted in the second and fourth rows. Points representing minimal contrasts in speaking rate are connected in the upper six panels; points representing minimal stress contrasts are connected in the lower six panels. Stressed syllables are indicated with upper case letters; unstressed syllables are indicated with lower case letters. Values for first syllables are subscripted with "1"; those for second syllables, with "2." Values for fast speaking rate are indicated with "F"; those for slow speaking rate are indicated with "S."

syllable, as to the stressed or unstressed character of the syllable, and the speaking rate condition. In the top set of six panels, lines are drawn between syllables which show a minimal contrast in speaking rate. If the "undershoot" hypothesis were supported, we would expect all lines to slope upward and to the right; that is, "slow" syllables would be longer in duration, indicating that the subject was following instructions, and that the fast syllables would have lower F_3 frequency values, indicating undershoot at fast speaking rates. In fact, the proposed pattern is followed, for two subjects, LJR and FBB, but is not for the third, KSH.

In the lower set of six panels, lines are drawn between points which show a minimal stress contrast. For contrasts of stress, lines connecting corresponding syllable pairs slope upward and to the right for all three subjects. Furthermore, in those cases where there is an overlap of duration values from different syllables, there is a tendency for stressed syllables to lie at higher values of F_3 peak frequency than unstressed syllables.

There are two minor variables which might have shown some systematic pattern. Since the nonsense words were produced in citation form, somewhat longer durations might be expected for the second syllable; furthermore, longer durations should be expected for those syllables terminating in /b/. A comparison of the twelve relevant cases shows that while all /b/ S_2 syllables are longer than S_2 /p/ syllables, F_3 frequency values are higher for /b/ for only five of the twelve comparisons, indicating no systematic relationship between the two variables. In short, speaking-rate variation seems to be quite different in the effects on vowel target than stress variations; and individuals differ in how they accomplish speaking-rate variation. Furthermore, the relatively small variations in duration which accompany the effect of terminal voicing and phonetic environment do not seem to be accompanied by systematic changes in F_3 frequency.

Fig. 3 shows peak EMG activity as a function of acoustic duration, in an analogous form of presentation to that of Fig. 1. It is clear that the second Lindblom hypothesis is not supported; peak activity varies substantially for stress, for most comparisons, for all three subjects; for speaking rate, there are no consistent effects on peak activity.

As to the effects of the minor variables of syllable position and terminal consonant, out of twelve possible comparisons, peak activity for /b/-syllables is greater than that for /p/-syllables in six. Thus, the overall conclusions from a study of peak EMG activity is that signals to the articulators are not a constant size; stressed syllables show greater activity than do unstressed syllables. Other variables do not show consistent effects.

DISCUSSION

The results of the experiment lead to the general conclusion that the mechanisms for the control of speaking rate and duration in vowel articulation are disjunct, a conclusion which finds some support in the literature.

As to the acoustic result, the experiment closest to that reported above is that of Gay (1978). He concluded that speaking rate and lexical stress are controlled by different mechanisms, on the basis of results which show variations of vowel-target frequency as a function of stress, but no variation as a function of speaking rate. Our results are identical to his for stress, but at variance with respect to speaking rate.

There are some studies of articulator movement under conditions of varying stress and speaking rate which, although no acoustic measures are reported, are consonant with those reported here. Kent and Netsell (1971), in a cineradiographic study, examined articulator position as a function of stress. Most of their observations show that under conditions of contrastive stress, the articulators move further from neutral position, and there is, in addition, some evidence for increased articulatory velocity with increased stress.

Using a technique identical to that of Kent and Netsell, Kuehn and Moll (1976) examined articulator displacement for vowels under conditions of variable speaking rate. They found that, as speaking rate increased, some speakers decreased articulator target position, while others increased articulator velocity to reach the target in a reduced amount of time.

Finally, in a study of single motor units, in the anterior belly of the digastric muscle, Sussman and MacNeilage (1978) observed a pattern of recruitment and discharge reorganization in emphatic stress, characterized by, among other things, changes in motor-unit discharge rate and recruitment of additional motor units. They conclude that their findings support the Öhman (1967) notion of "an instantaneous addition of a quantum of physiological energy underlying stressed productions."

It can be concluded, then, that speakers articulate using independent controls of the duration and magnitude of movement, in generating allophonic variations in varying stress and speaking rate. There is some acoustic evidence (Nord, 1974) that the mechanism associated with terminal lengthening may lead to allophones with still a different relationship between duration and target frequency. Terminal lengthening may be particularly important in conveying syntactic information about sentence structure to the listener (Klatt, 1976; Cooper, 1976).

There is a possible reason, from a perceptual point of view, why these independent controls are necessary. Klatt (1976) has pointed out that, for a listener faced with the acoustic representation of a sentence, there is information encoded in duration about both the suprasegmental and segmental structure of the sentence. He suggested that if the listener already understood the sentence, he could interpret the durational variations, but he must use the durational information to decode the message. However, if durational variation is accomplished so that there are different relations among duration, target formant frequencies and transition velocity, for, for example, stressing and clause terminal lengthening, some of the ambiguities in the message may be resolvable.

It has recently been shown by Strange, Verbrugge, Shankweiler and Edman (1976) that vowel identification is aided by consonant context; presumably, as we understand movement dynamics better, we will better understand the way in which the listener makes use of the acoustic counterpart of the articulatory act.

REFERENCES

COOPER, W.E. (1976). Syntactic control of timing in speech production. *J. Phonetics*, **4**, 151-71.

FRY, D.B. (1958). Experiments in the perception of stress. *Language and Speech*, **1**, 126-52.

GAY, T. (1978). Effect of speaking rate on vowel formant movements. *J. acoust. Soc. Amer.*, **63**, 223-30.

HIROSE, H. (1971). Electromyography of the articulatory muscles: current instrumentation and technique. *Haskins Laboratories Status Report on Speech Research*, SR-25/26, 73-86.

JONES, D. (1940). *An Outline of English Phonetics*, 6th ed. (New York).

KENT, R. and NETSELL, R. (1971). Effects of stress contrasts on certain articulatory parameters. *Phonetica*, **24**, 23-44.

KEWLEY-PORT, D. (1973). Computer processing of EMG signals at Haskins Laboratories. *Haskins Laboratories Status Report on Speech Research*, SR-33, 173-83.

KLATT, D. (1976). The linguistic uses of segment duration in English: acoustic and perceptual evidence. *J. acoust. Soc. Amer.*, **59**, 1208-21.

KUEHN, D.P. and MOLL, K.L. (1976). A cineradiographic study of VC and CV articulatory velocities. *J. Phonetics*, **4**, 303-20.

LEHISTE, I. (1970). *Suprasegmentals* (Cambridge, Mass.).

LINDBLOM, B.E.F. (1963). Spectrosgraphic study of vowel reduction. *J. acoust. Soc. Amer.*, **35**, 1773-81.

NORD, L. (1974). Vowel reduction – centralization or contextual assimilation? Preprints of the *Speech Communication Seminar, Stockholm, August 1-3, 1974*, V2, 149-54.

ÖHMAN, S.E.G. (1967). Word and sentence intonation: a quantitative model. *Speech Transmission Laboratory QPSR*, 2-3, Royal Inst. Tech., 20-54.

RAPHAEL, L.J. and BELL-BERTI, F. (1975). Tongue musculature and the feature of tension in English vowels. *Phonetica*, **32**, 61-73.

SMITH, T. (1970). *A Phonetic Study of the Function of the Extrinsic Tongue Muscles*. Ph.D. dissertation, U.C.L.A.

STRANGE, W., VERBRUGGE, R.R., SHANKWEILER, D. and EDMAN, T. (1976). Consonant environment specifies vowel identity. *J. acoust. Soc. Amer.*, **60**, 213-44.

SUSSMAN, H. and MacNEILAGE, P.F. (1978). Motor unit correlates of stress: preliminary observations. *J. acoust. Soc. Amer.*, **64**, 338-40.