LANGUAGE AND SPEECH, Vol. 23, Part 3, 1980

ON DEFINING THE VOWEL DURATION THAT CUES VOICING IN FINAL POSITION*

LAWRENCE J. RAPHAEL**
Herbert H. Lehman College, C.U.N.Y.
MICHAEL F. DORMAN
Arizona State University
and
ALVIN M. LIBERMAN†
Haskins Laboratories

Most investigations into the perceptual relevance of vowel duration have employed patterns of synthetic speech in which only the steady-state portions of syllables have been used as the variable. The experiments reported here were designed to discover if CV transitions are also included by the listener in determining the effective duration of the "vowel." More specifically, they measured the contribution of syllable-initial formant transitions to that "vowel" duration on which the voicing of a stop in syllable-final position can be made to depend. It is clear from the results that the syllable-initial transitions do contribute to that duration; indeed, they appear to contribute to the same degree as an equal duration of steady-state. Apparently, the effective duration of a vowel extends over all parts of the acoustic signal that may be said to have been influenced by it, including especially the transitions that reflect the consequences of the coarticulation of vowel and consonant.

INTRODUCTION

There is, in general, no acoustic criterion that will directly divide the stream of speech into segments corresponding in size to consonants and vowels. That is so because the processes of articulation, and especially of coarticulation, cause the phonetic information to be overlapped as it is encoded in the acoustic signal. As a consequence, the segments of the signal, however defined, cannot be matched straightforwardly to the segments of the phonetic message. Thus, a segment of sound will usually convey information simultaneously about several segments of the message, as, for example, when consonant and vowel are both encoded in the formant transitions (Lisker, 1957, 1974). Conversely, a segment of the message will often be spread through several segments of the signal, as,

* *This research was supported by a Grant [HD01994] from the National Institute of Child Health and Human Development to Haskins Laboratories. We wish to thank Suzi Pollack and Anthony Levas for their assistance in collecting and tabulating portions of the data.*

** *Also Haskins Laboratories, New Haven, Connecticut.*

† *Also Yale University and University of Connecticut.*

for example, when the closing and opening gestures appropriate for the stop consonant in a syllable like /spa/ have perceptible effects on all three of the segments into which the sound can be most clearly divided – namely, the band-limited noise associated with the fricative, the period of silence following the noise, and the (formant) structure of the periodic sound following the silence.

From both practical and theoretical points of view, the mismatch between phonetic and acoustic segments is of considerable importance. As a practical matter, it accounts, on the one hand, for our inability to synthesize speech by concatenating pre-recorded phones; on the other, it has presented a formidable obstacle to those who try to make automatic speech recognition a reality. As for theory, it raises the obvious, and obviously thorny, questions associated with the fact that human perceivers can normally cope with the mismatch and recover segments of the message.

A further consequence of the mismatch in segmentation ought to be seen wherever duration is a cue for the perception of segmental structure. Consider, then, the role of duration in the contrast between voiced and voiceless stops in syllable-final position. In that connection, it has long been known that a difference in duration characterizes the distinction between /bæd/ or /æd/ on the one hand, and /bæt/ or /æt/ on the other. The difference has been observed in real-speech utterances, and the importance of that difference for perception has been established in experiments on synthetic speech. But where, exactly, does the difference in duration lie? Customarily, investigators have put it on the "vowel." In the case of measurements that have been made on real-speech utterances, however, it is hard to know just what "vowel" duration means, since, as we have pointed out, there is often no unambiguous acoustic sign that marks a boundary between the vowel and the consonants that may, at the level of phonetic structure, be contiguous to it. As House (1961) has pointed out, the changes in source excitation that have been used as the principal clues to segmentation, do not invariably coincide with the articulatory boundaries of vowels. The procedural difficulties to be encountered in delimiting vowels in the acoustic signal have been discussed in some detail by Peterson and Lehiste (1960). The acoustic analyses in their study and in the one by House generally included (voiced) formant transitions in durational measurements of vowels. The perceptual studies using synthetic speech, however, did not investigate whether, or to what extent, transitions contribute to the effective vowel duration that cues the voicing class of following consonants. Thus, in the early work by Sholes (1954, 1956, 1959a, 1959b), the synthetic patterns were VC syllables in which the experimental variable was simply the duration of a steady-state portion of the pattern. Later, when the research was extended to syllables of the CVC type (Denes, 1955; Raphael, 1972; Raphael, Dorman, Freeman and Tobin, 1975), it was, again, only a steady-state portion of the pattern that was varied. From procedures of that kind we have learned that duration does, in fact, affect the perception of a syllable-final stop as voiced or voiceless. But we have no experimental answer to the question: what counts in effective vowel duration, especially in the case of CVC syllables?

We consider it unlikely that only steady-state formant duration is significant in this regard, both because of the segmentation practices employed in the acoustic studies and because of the general difficulty in delimiting the domains of consonant and vowel, even

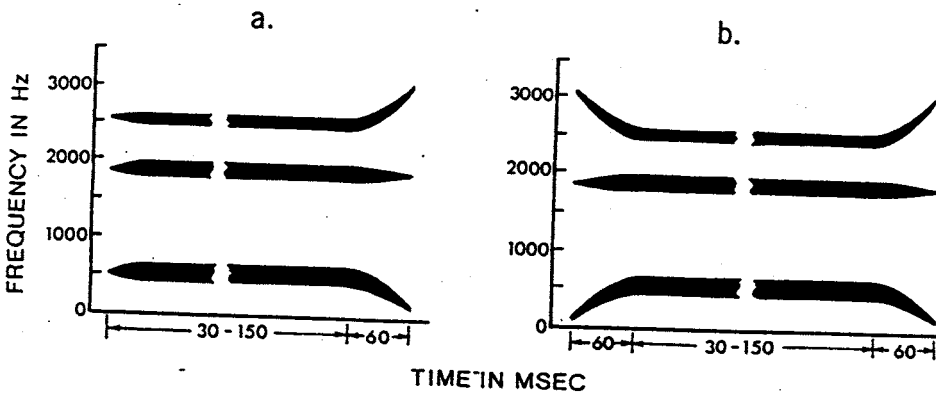

Fig. 1. Stimulus patterns used in Experiment 1. Schematic spectrograms of patterns sufficient to cue the perception of (a) /ɛt/ and /ɛd/ and (b) /dɛt/ and /dɛd/.

in synthetic patterns that contain formants in steady state. We should expect, rather, that the effective duration would be some larger part of the acoustic signal, more specifically some part large enough to include a significant fraction of the initial consonant-vowel transitions. The aim of the experiments to be reported here is to test that expectation.

Experiment 1

The purpose of the first experiment was to determine whether the presence of syllable-initial formant transitions appropriate for a stop consonant would alter the steady-state vowel duration necessary for listeners to identify a syllable-final stop consonant as voiced or voiceless.

Method

Two series of synthetic speech stimuli were generated using the Haskins Laboratories parallel-resonance synthesizer. In one series, all of the stimuli were vowel + consonant (VC) syllables consisting of the steady-state vowel /ɛ/ followed by 60-msec formant transitions appropriate for /d/ (see Fig. 1a). The stimuli varied in the duration of their steady-state portions from 30 to 150 msec in 10-msec steps. Informal pretests, using laboratory workers as listeners, established that this range of steady-state vowel durations was sufficient to cue perception of the syllable-final stop as /t/ at the short end of the range and as /d/ at the long end.

In the other series, the stimuli were consonant + vowel + consonant (CVC) syllables

which differed from the VC stimuli only in the addition of formant transitions to the beginning of each stimulus. The transitions, sufficient to cue the stop consonant /d/, were 60 msec in duration (see Fig. 1b).

The stimuli in both VC and CVC series were synthesized with a constant fundamental frequency of 120 Hz. The amplitude envelopes at the onset of the VCs and CVCs were equated for rise time.

Four tokens of each stimulus in both series were recorded. These were then randomized with a four-second interstimulus interval into a single test sequence. That sequence was reproduced for listeners in a large, sound-attenuated room via a Revox 1240 tape recorder and an AR-4X loudspeaker.

The 11 subjects who participated in this experiment were undergraduate volunteers who had not previously participated in experiments on speech perception. They were told that they would hear approximations to the syllables /ɛd/, /ɛt/, /dɛd/ and /dɛt/ and were asked to indicate, on a printed response sheet, the identity of the final stop in each syllable. The subjects heard 20 representative stimuli, randomly ordered, in a practice session before the start of the experiment.

Results

The results of varying the steady-state vowel duration in the VC and CVC series are shown in Fig. 2a. It is apparent that this variation was sufficient to cue the distinction between voiced and voiceless stops: at short durations the stimuli in both series were heard as voiceless, while at the longest duration they were heard as voiced. This outcome is in general agreement with the earlier findings of Raphael (1972) and Denes (1955).

Of greater interest, from our point of view, is the fact that the presence of initial formant transitions reduced the steady-state vowel duration necessary for listeners to identify the final stop as voiced. We can see that the CVC function is displaced, relative to the VC function, toward smaller values of steady-state vowel duration. Taking the 50% crossovers of /d/ and /t/ responses as an approximate measure of the magnitude of the separation of the functions, we note that listeners required about 55 msec less of steady-state vowel duration to hear the final consonant as voiced (i.e., as /d/) in the CVC stimulus series. This difference between the 50% crossovers indicates that the initial formant transitions not only supplied information about the initial stop, but also contributed by their duration to the perception of the voicing class of the stop at the end of the syllable.

We should also note that the 55 msec difference between the 50% crossover points approximates the duration of the syllable-initial formant transitions of the CVC series (60 msec). If, as we have assumed, the difference between the crossover points reasonably reflects the overall separation of the CVC and VC functions, then it would appear that the initial formant transitions are about as effective as the steady-state vowel in determining the voicing class of the syllable-final stop. That being so, the labeling functions for the two stimulus series should be similar if we plot them as a function of syllable duration. We have done this in Fig. 2b. There we see that although the functions are quite similar, the data points for the CVC stimuli generally lie above those for the VC stimuli. Statistical analysis indicates that the difference between the percentage of /t/ responses in the CVC

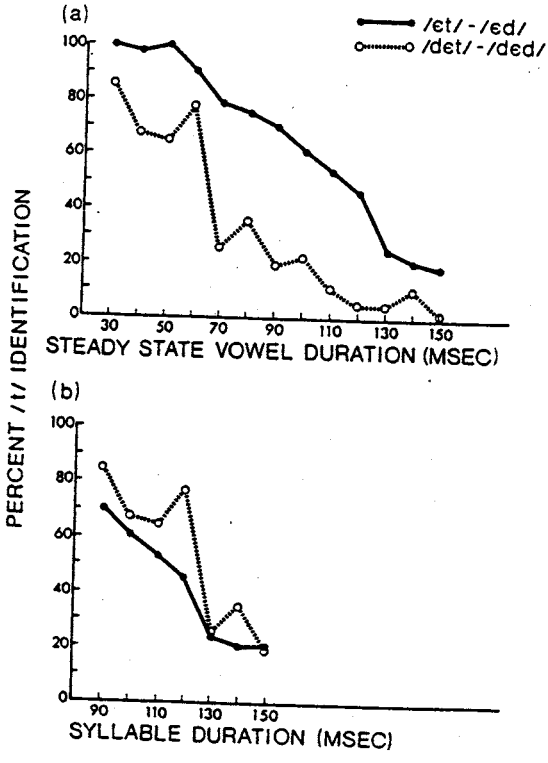

Fig. 2. (a) Percent /t/ identification as a function of steady-state vowel duration for VC and CVC stimuli. (b) Percent /t/ identification as a function of syllable duration for VC and CVC stimuli.

condition (52) and in the VC condition (42) is significant ($T = 3$; $p < 0.005$). We conclude, therefore, that the duration of the initial transitions is not quite as effective as the duration of the steady-state vowel in cueing the voicing contrast of stops in final position.

EXPERIMENT 2

We designed the second experiment to test the generality of the results of Experiment 1 in two ways. First, we wished to determine if /d/-transitions of durations both longer and shorter than 60 msec would contribute as much as 60-msec /d/-transitions to the perception of the voicing class of final consonants. Second, we wanted to discover if, and to what extent, the transitions associated with a class of consonants other than stops contributed by their duration to the perception of the voicing class of final consonants.

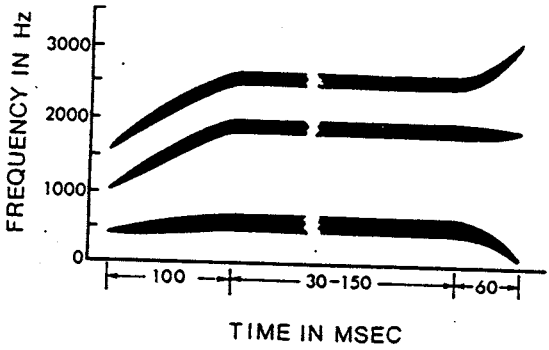

Fig. 3. Schematic spectrogram patterns sufficient to cue the perception of /rɛt/ and /rɛd/.

We decided to use a resonant consonant, /r/, for this purpose. We chose /r/ for two reasons: (1) It is closer to /d/ in place of articulation than are the other resonant consonants (/w/ and /j/); and (2) it provides an alternative and more natural test of the effect of long transitions than /d/, which is not normally produced with transitions exceeding 60 msec.

Method

Five series of stimuli were synthesized. Two of the series were those used in Experiment 1: the VC series and the CVC series with initial /d/-transitions of 60 msec duration. To these we added two new CVC series: one in which the initial /d/-transitions were 30 msec in duration, and another in which they were 90 msec in duration. We should note that the stimuli with the 90-msec transitions sounded acceptable to both experienced and naïve listeners, despite their unnaturally great duration (see Lisker, Liberman, Erickson, Dechovitz and Mandler, 1977). We created a fifth, /r/-initiated, CVC stimulus series (see Fig. 3) in which the /r/ was cued by transitions 100 msec in duration.

In the first part of Experiment 2, four tokens of each stimulus in the VC and the three CVC series initiated by /d/ were randomized into a single test sequence. In the second part of the experiment, four tokens of each stimulus in the VC and /rVC/ series were randomized into one test sequence. In both sequences the interstimulus interval was 4 seconds. Fifteen undergraduate volunteers who had not served previously in experiments on speech perception were the subjects. The listening conditions and the response task were identical to those of Experiment 1.

Results

The results of the first part of Experiment 2 are shown in Fig. 4a. We note first that the results of Experiment 1 have been replicated. In both experiments the /d/–/t/

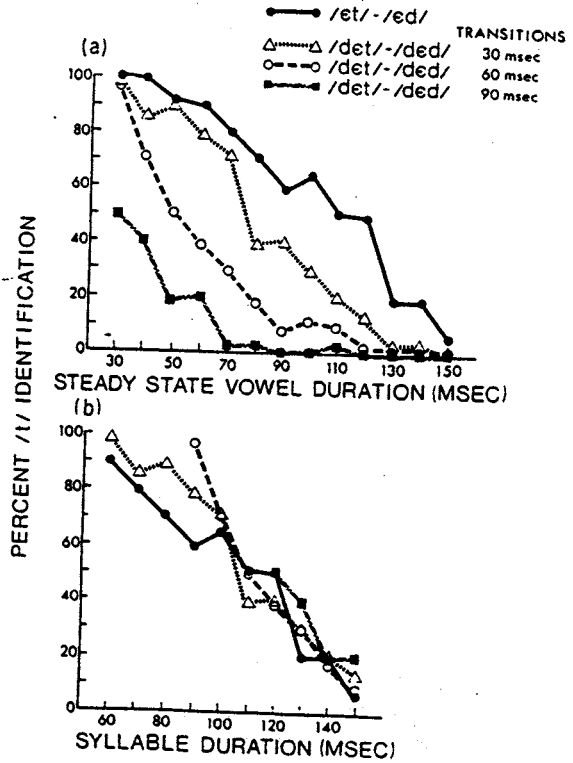

Fig. 4. (a) Percent /t/ identification as a function of steady-state vowel duration for VC stimuli and for CVC stimuli with 30-, 60- and 90-msec initial transitions. (b) Percent /t/ identification as a function of syllable duration for VC stimuli and for CVC stimuli with 30-, 60- and 90-msec initial transitions.

boundaries (50% crossovers) for the CVCs with 60-msec transitions fell at a shorter steady-state vowel duration than did the boundaries for the VC stimuli. Further, the boundaries for the VC series in both experiments are virtually identical at 120 msec of steady-state vowel duration, while the boundaries for the 60-msec CVC series differ by no more than 15 msec of steady-vowel duration: 65 msec in Experiment 1, 50 msec in Experiment 2.

The data also reveal that in the process of determining the voicing class of syllable-final stops, listeners are sensitive to the absolute durations of initial formant transitions. As the transition duration increases, from one stimulus series to the next, the /d/–/t/ phoneme boundary, expressed as steady-state vowel duration in Fig. 4a, decreases. Thus the phoneme boundaries for the 30-, 60- and 90-msec transition series are located, res-

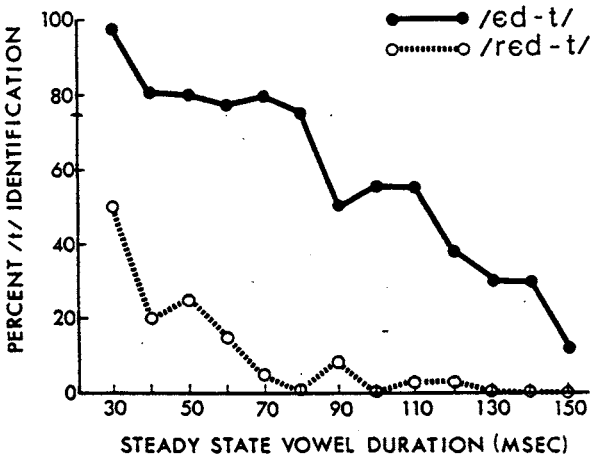

Fig. 5. Percent /t/ identification as a function of steady-state vowel duration for VC stimuli and CVC stimuli initiated by /r/.

pectively, at 77, 50 and 30 msec of steady-state vowel duration.[1]

We note that in this experiment, as in Experiment 1, the transitions are almost as effective as the steady-state vowel in cueing the voicing class of the syllable-final stop. In Fig. 4b we have re-plotted the identification functions of Fig. 4a as functions of syllable duration. As in the analogous figure from Experiment 1 (i.e., Fig. 2b) we see that the functions for all the stimulus series are quite similar, though not identical. For the data points shown in Fig. 4b the differences in the percentages of /t/ responses among the various stimulus series were small but significant: 57% for the 30-msec transitions v. 51% for the VC stimuli ($T = 6, p < 0.005$); 48% for the 60-msec transitions v. 39% for the VC stimuli ($T = 5, p < 0.005$). There were not enough data points for the 90-msec transition stimuli for a valid comparison to be made or statistically tested.

We should point out another similarity between the results of this experiment and Experiment 1: We can see in Fig. 4a that the magnitudes of the differences between the 50% crossover points of the identification functions are similar to the differences between the durations of the transitions of the various stimulus series. We find a 27-msec difference between the 30- and 60-msec transition series, and a 20-msec difference between the 60- and 90-msec transition series. Further, the difference between the VC and 90-msec CVC series is exactly equal to duration of the transitions of the CVCs: 90 msec. We interpret these data as a further indication that the transitions are almost (and in one instance –

[1] *We call the reader's attention to the fact that our subjects, as a group, never heard the final consonant in the 90-msec transition series as /t/. Thus, the 50% point for this function is not literally a crossover in /d/–/t/ judgments, and our use of the term "phoneme boundary" is a convenience, rather than a description.*

VC v. 90-msec CVC – equally) as effective as the steady-state vowel in cueing the voicing class of the syllable-final stop.

Let us turn now to the comparison of the data for the /r/-initiated and VC syllables (Fig. 5). We shall again use the 50% points of the functions as an approximate measure of their separation. The 50% points[2] fell at 30 msec of steady-state vowel duration for the /r/-initiated series and at 113 msec of steady-state vowel duration for the VC series. We note that the 83-msec difference between these points is only slightly less than the duration of the 100-msec transitions in the /rVC/ stimuli. Moreover, this 83-msec difference is quite similar to the 90-msec difference found in the first part of this experiment between the 50% points of the VC stimuli and the CVC stimuli initiated by the 90-msec /d/-transitions. On the basis of this evidence we suggest that the transitions of syllable-initial resonant consonants, like those of stop consonants, are almost fully incorporated into the durational estimates that listeners make in determining the voicing class of syllable-final stops.

DISCUSSION

The results of the experiments reported here are consistent with those of earlier studies (Denes, 1955; Raphael, 1972; Raphael *et al.*, 1975) in showing that vowel duration is a cue to voicing of syllable-final consonants. What is new in our results is that, when the syllable begins with a stop consonant, the duration that is effective in determining syllable-final voicing includes almost all of the initial formant transitions appropriate to that syllable-initial stop consonant. There are two somewhat different ways to interpret that finding. One is to suppose that the formant transitions are not part of the vowel – that is, that they do not simultaneously convey information about both the consonant and the vowel – in which case we should conclude that it is not vowel duration that determines voicing in final consonants, but rather the sum of the durations of the several segments. The other is simply to assume that the initial formant transitions do contain information about the vowel, and that perceived vowel duration takes that into account.

That we incline to the latter view is plain from what we said in the Introduction. Now we would add that there is evidence that supports it quite directly. Most direct, perhaps, is that provided by Mermelstein, Liberman and Fowler (1977), who asked listeners to adjust the duration of a steady-state vowel until it sounded equal to the same vowel when it was preceded by formant transitions appropriate to a stop. They found that, for the same perceived duration, the steady-state vowel had to be significantly longer than the steady-state portion of the vowel in the stop-vowel syllable. Other, only slightly less direct, support comes from experiments on vowels when the perceived color is sensitive

[2] *We must note again that, as in the case of the 90-msec CVC stimulus series, the function for the /r/-initiated stimuli does not provide us with a literal crossover point, since the listeners' /t/ responses never rose above the 50% level.*

to duration. In one experiment, Verbrugge and Isenberg (1978) synthesized /bVb/ and /V/ syllables in which the vowel was perceived either as /ɛ/ or /æ/ depending on the duration of the steady-state. Listeners reported hearing /æ/ more often in the /bVb/ stimuli than in the /V/ stimuli. That is, when a /bVb/ stimulus was equal in duration to a /V/ stimulus, listeners were more likely to judge the vowel in the /bVb/ stimulus to be longer in duration than the vowel in the /V/ stimulus. Thus, the listeners included not only the initial and final formant transitions in the duration that was effective for vowel color, but rather those transitions and more. In a second experiment, Verbrugge and Isenberg varied transition duration and steady-state duration independently. When they held overall syllable duration constant, they found that /æ/ judgments increased as transition duration increased and steady-state vowel duration decreased. In contrast, isolated steady-state vowels matched to the durations of the steady-state portions of the /bVb/ stimuli elicited fewer /æ/ judgments as their duration decreased. Thus, it appears again that formant transitions contribute to those estimates of vowel duration that are important for perception of vowel color; indeed, it appears again that the transitions contributed more than their actual duration.

It is in regard to this last point – the amount of duration contributed by the transitions – that our study and the others do not quite agree. In our study, the contribution of the transitions to effective duration was about 90% of the transition, in the Mermelstein *et al.* study it was roughly 50%, and, as we have just noted, in the Verbrugge and Isenberg studies it was more than 100%. There were several differences among the stimulus patterns used in these experiments – open v. closed syllables, duration and rate of the transitions, amplitude envelopes, and others – any one of which might have caused the difference. Further research can surely uncover those causes.

What is clear in all the studies, however, is that the effective duration of the vowel may be assumed to include significant parts of the abutting formant transitions. We suppose that this is because those transitions contain information about the vowel (as well as the consonant). Our hypothesis, then, is that the effective (and perceived) duration of the vowel is taken as that span of the signal that contains information about the vowel. Presumably, this will include any portion that shows the acoustic results of coarticulation between consonants and vowels. To enlighten us further on that matter, we look forward to experiments similar to those we have reported here in which we prefix other types of segments and, correspondingly, other types of acoustic cues. We should also look to experiments of the type recently reported by Strange, Jenkins, and Edman (1977, 1978). They had listeners identify vowels in CVC syllables from which various acoustic portions – vocalic centers, transitions – had been excised. The finding was that listeners identified the vowels very well when the vocalic centers were missing, better indeed than when the vocalic centers were present and the transitions were gone. That result shows that vowel information is present, in the transitions, which is surely relevant to our concern here; but also relevant is the technique, for it can be used to delimit the span of the signal over which the vowel information is to be found.

REFERENCES

DENES, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 27, 761-764.

HOUSE, A.S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, 33, 1174-1178.

LISKER, L. (1957). Linguistic segments, acoustic segments, and synthetic speech. *Language*, 33, 370-374.

LISKER, L. (1974). On time and timing in speech. In T.A. Sebeok, A.S. Abramson, D. Hymes, H. Rubin, E. Stankiewicz and B. Spolsky (eds.), *Current Trends in Linguistics*, Vol. 12 (The Hague), pp. 2387-2418.

LISKER, L., LIBERMAN, A.M., ERICKSON, D.M., DECHOVITZ, D. and MANDLER, R. (1977). On pushing the voice-onset-time (VOT) boundary about. *Language and Speech*, 20, 209-216.

MERMELSTEIN, P., LIBERMAN, A.M. and FOWLER, A. (1977). Perceptual assessment of vowel duration in consonantal context and its application to vowel identification. *Journal of the Acoustical Society of America*, 62, S101 (Abstract).

PETERSON, G. and LEHISTE, I. (1960). Duration of syllabic nuclei in English. *Journal of the Acoustical Society of America*, 32, 693-703.

RAPHAEL, L.J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in English. *Journal of the Acoustical Society of America*, 51, 1296-1303.

RAPHAEL, L.J., DORMAN, M.F., FREEMAN, F. and TOBIN, C. (1975). Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies. *Journal of Speech and Hearing Research*, 18, 389-400.

SHOLES, G. (1954). Terminal stops and nasals. *Haskins Laboratories Quarterly Progress Reports*, 13, Appendix 2.

SHOLES, G. (1956). Stop consonants in final position. *Haskins Laboratories Quarterly Progress Reports*, 21, Appendix 5.

SHOLES, G. (1959a). Acoustic cues for final /s/ and /z/. *Haskins Laboratories Quarterly Progress Reports*, 31, Appendix 3.

SHOLES, G. (1959b). Synthesis of final /z/ without voicing. *Journal of the Acoustical Society of America*, 31, 1568 (Abstract).

STRANGE, W., JENKINS, J.J. and EDMAN, T.R. (1977). Identification of vowels in "vowel-less" syllables. *Journal of the Acoustical Society of America*, 61, S39 (Abstract).

STRANGE, W., JENKINS, J.J. and EDMAN, T.R. (1978). Dynamic information specifies vowel identity. *Journal of the Acoustical Society of America*, 63, S5 (Abstract).

VERBRUGGE, R.R. and ISENBERG, D. (1978). Syllable timing and vowel perception. *Journal of the Acoustical Society of America*, 63, S4 (Abstract).

ADDRESSES OF CONTRIBUTORS

BERKOVITS, Dr. Rochele, Dept. of Linguistics, Tel Aviv University, Ramat Aviv, Israel.
COLLINS, Dr. Patrick J., John Jay College of Criminal Justice, City University of New York, 444 West 56th Street, New York, NY 10019, U.S.A.
DORMAN, Prof. Michael, Dept. of Speech and Hearing Science, Arizona State University, Tempe, AZ 85281, U.S.A.
JAEGER, Dr. Jeri J., Department of Linguistics, S.G.S., Australian National University, Box 4, Canberra, A.C.T. 2600, Australia.
LIBERMAN, Prof. Alvin M., Haskins Laboratories, 270 Crown Street, New Haven, CT 06510, U.S.A.
RAPHAEL, Prof. Lawrence J., Herbert H. Lehman College, C.U.N.Y., Bedford Park Boulevard, W. Bronx, NY 10468, U.S.A.
RIETVELD, Dr. A.C.M., Institut voor Fonetiek, Katholieke Universiteit Nijmegen, Erasmuslaan 40, Nijmegen, The Netherlands.
ROMAINE, Dr. Suzanne, Dept. of Linguistics, University of Birmingham, P.O. Box 363, Birmingham B15 2TT, U.K.
WOLFF, Dr. J. Gerard, Dept. of Psychology, The University of Dundee, Dundee DD1 4HN, Scotland.

PUBLICATIONS RECEIVED

Behavioral Science, **25** (1980), 2, 3.
Cahiers Roumains d'Etudes Littéraires, **2** (1979), 6.
International Journal of the Sociology of Languages, **23** (1980).
Journal of Experimental Psychology: Animal Behavior Processes, **6** (1980), 2.
Journal of Experimental Psychology: Human Learning and Memory, **6** (1980), 2.
Lenguaje y Ciencias, **20** (1980), 1.
Leuvense Bijdragen, **68** (1979), 4.
Revue Roumaine de Linguistique, **16** (1979), 2.
Zeitschrfit für Psychologie, **187** (1979), 3, 4.
Zurnal Vyššej Dejatel mosti Imeni – I.P. Pavlov (A.K. Nauk SSSR), **29** (1979), 6 and **30** (1980), 1.