IN QUALIFIED DEFENSE OF VOT*

LEIGH LISKER
Haskins Laboratories
and
University of Pennsylvania

The VOT measure has been said to provide the single most nearly adequate physical basis for separating homorganic stop categories across a variety of languages, granted that other features may also be involved. That transition duration affects perceived voicing of synthesized initial stops of one specific language, English, has suggested the hypothesis by Stevens and Klatt (1974) that a detector responsive to rapid formant-frequency shifts after voice onset better explains the child's acquisition of the contrast than does some mechanism which responds to VOT directly. If such a detector is part of our biological equipment, then it seems remarkably underutilized in language, for the hypothesis asserts that basic to voicing perception is whether laryngeal signal is or is not present during the interval in which the stop-vowel transition occurs. In effect, the "archetypical" voiceless stop is aspirated. Not only do many languages not possess voiceless aspirates, but even in English aspiration is severely restricted. Of course the VOT measure has its limitations – it is inapplicable to prepausal stops. However, there are much more serious difficulties with the posited detector, since even for the English initial stops there is evidence that the presence of a voiced first-formant transition is not required for the perception of /bdg/, nor does the absence of such a transition necessarily yield /ptk/, provided appropriate VOT values are provided.

BACKGROUND

To judge from much current discussion of the phonetic feature of stop consonant voicing, it appears that voice onset time (VOT) is often taken quite simply as the duration of the time interval between onset of the release explosion, or "burst," and the onset of glottal pulsing, i.e. as a purely acoustic measure considered quite apart from the laryngeal and supraglottal events whose temporal relation it reflects. In some experimental work, in fact, stimuli for listener evaluation have been constructed in which this purely acoustic feature is incorporated in a way that no human vocal tract could begin to match (Winitz, LaRiviere, and Herriman, 1975). In such stimuli the temporal relation between burst and pulsing onset cannot be useful as an index of speech activity, whatever the relevance of the responses it elicits for our understanding of auditory processes and the relation between more and less speechlike signals. Bursts and pulsing onsets are readily discoverable, given favorable recording conditions, in the acoustic representations of stop-vowel sequences, and *here* VOT as an acoustic measure is one of several conceivable

* *Much of the experimental data presented here was reported at the 8th International Congress of Phonetic Sciences, Leeds, England, in August 1975. The support of the National Institute of Child Health and Human Development (Grant HD-01994 to Haskins Laboratories) is gratefully acknowledged.*

measures by which to express the temporal relation between the underlying laryngeal and supraglottal manoeuvers. For initial prevocalic stops the choice of VOT as the acoustic index of this relation is the only practical one; burst onset is not the only possible market of the stop articulation, but it is the only one that can be easily located by eye in either the spectrogram or oscillogram of the speech signal. By contrast, any acoustic measure of laryngeal timing requiring that we fix the time at which a stop articulation is complete would be more difficult to perform, even if we could agree on how to locate the reference point in time; it is only in synthetic speech patterns that the onsets and terminations of transitions can be reliably described.

Measurements of extensive samples of speech have suggested that VOT provides the single most nearly adequate physical basis for separating the homorganic stop categories of English and a number of other languages (Lisker and Abramson, 1964). At the same time it appears that VOT values for a given stop category are not entirely independent of certain other factors as well. For English, the language on which we are best informed, knowing the VOT value for an initial stop is not quite enough to assign the stop to one or the other of the two stop sets /bdg/ and /ptk/; it may be necessary to have information as to place of closure, degree of stress on the syllable, and some other factors as well (Lisker and Abramson, 1964, 1967). If we can justifiably assume that these stops are phonetically unambiguous, it would seem that the listener must attend to some other feature or features in addition to, or perhaps other than, the absolute duration from burst to voicing onset (or from voicing onset to burst). Of course it is more often the case that the absolute VOT value *does* suffice for the correct classification of a stop, so that we might believe VOT to be the paramount, and possibly the sole cue for the listener. We cannot decide whether this is true by observing natural speech, however, since (1) the interval between burst and pulsing onset is not the only feature by which stops of contrasting categories differ, and (2) we cannot be sure that an acoustically salient property is correspondingly significant for the listener. Because the several acoustic consequences of a change in the timing of vocal fold adduction and stop release do not vary in a mutually independent manner in natural speech, we must turn to speech synthesis, gaining thereby the advantage of a more tractable "speech" source at the cost of an acceptable loss in naturalness.

Experiments in the perception of synthetic speech patterns have yielded results fairly consistent with findings for natural speech. For the patterns most often used in these experiments it can be asserted that, as with natural speech, listeners' labelings are largely determined by the duration of the interval between a burst and onset of a periodic excitation (Abramson and Lisker, 1965, 1970). But it must be borne in mind that, following Liberman, Delattre, and Cooper (1958), these stimuli not only simulate a variable VOT, but incorporate a covarying "cutback" of the first formant as well. Thanks to this F1 cutback, members of the stimulus set differ in their F1 onset frequencies and in the durations and frequency ranges of their first formants. Since from Liberman, Delattre, and Cooper (1958) it appears that pulsing onset and F1 cutback are both needed to effect a shift from /bdg/ to /ptk/, it makes sense to think of a VOT *dimension* as articulatory rather than purely acoustic in nature (Lisker and Abramson, 1965, 1971). Only in this sense can one speak of a VOT *continuum*, a term quite inapplicable, in any purely acoustic sense, to most of the stimuli used in experiments in the perception of stop voicing. Thus not only the timing of burst and pulsing onset, but F1 cutback, fundamental-frequency contour (Haggard, Ambler, and Callow, 1970; Fujimura, 1971), and any other acoustic consequences of a variation in the timing of laryngeal action, are all, in this view, components of the VOT dimension (Abramson, 1977; Summerfield and Haggard, 1974). Not to be excluded from this set is burst intensity, which has classically but incorrectly been ascribed to an independent dimension of articulatory force (Heffner, 1950).

We must, in short, distinguish between VOT, an acoustic measure of signals originating in the human vocal tract, and another measure, which we might call POT, to describe the timing of onset of a pulse train quite independently of how the signal being measured was generated. Since most experiments in stop voicing have involved stimuli in which the variables *as described* are at least two: POT and F1 cutback, we cannot say very much about the significance of POT as a cue. If VOT is retained as the referent to the temporal interval between burst and the simultaneous onsets of pulsing and first-formant transition (and perhaps it is best to retain it in this usage), it is not the name of an acoustic dimension; its rationale derives from certain assumptions, perhaps oversimplified, as to what the acoustic output of a vocal tract executing a fixed articulatory gesture should be, given a larynx capable of shifting from an open voiceless state to a closed vibratory one at any time during the supraglottal articulation. A simple dimension is thus to be found at the articulatory level.

VOT AND F1 TRANSITION

Aside from its role as a convenient measure of the degree of synchrony between laryngeal and supraglottal events during stop-vowel production, it is of interest to assess the perceptual significance of VOT in its narrowly acoustic sense, and to compare the effect of varying VOT with that of other acoustic features associated with stop voicing. One such that has been the focus of attention recently is the duration of the first formant transition following the onset of pulsing (Stevens and Klatt, 1974). This feature is of particular interest because it significantly affects voicing judgments and yet is perhaps more nearly independent of the larynx than other features affecting stop voicing perception. Stevens and Klatt used synthetic /da ta/ patterns of varying transition durations to find that the boundaries for the different transitions were more directly related to the duration of the voiced F1 transition than to VOT. For longer transitions a greater delay in VOT was needed to elicit /ta/ responses. Stevens and Klatt concluded that /ptk/ require a lag in voicing onset sufficiently long to effectively eliminate any F1 energy during the period of transition to the following vowel.

The Stevens and Klatt (1974) experiment has been repeated, and their findings replicated (Lisker, Liberman, Erickson, Dechovitz, and Mandler, 1977). Using a wider range of transition durations, Lisker *et al.* found VOT boundary values for English that ranged from about +20 msec. for the shortest duration of transition (20 msec.) to almost +50

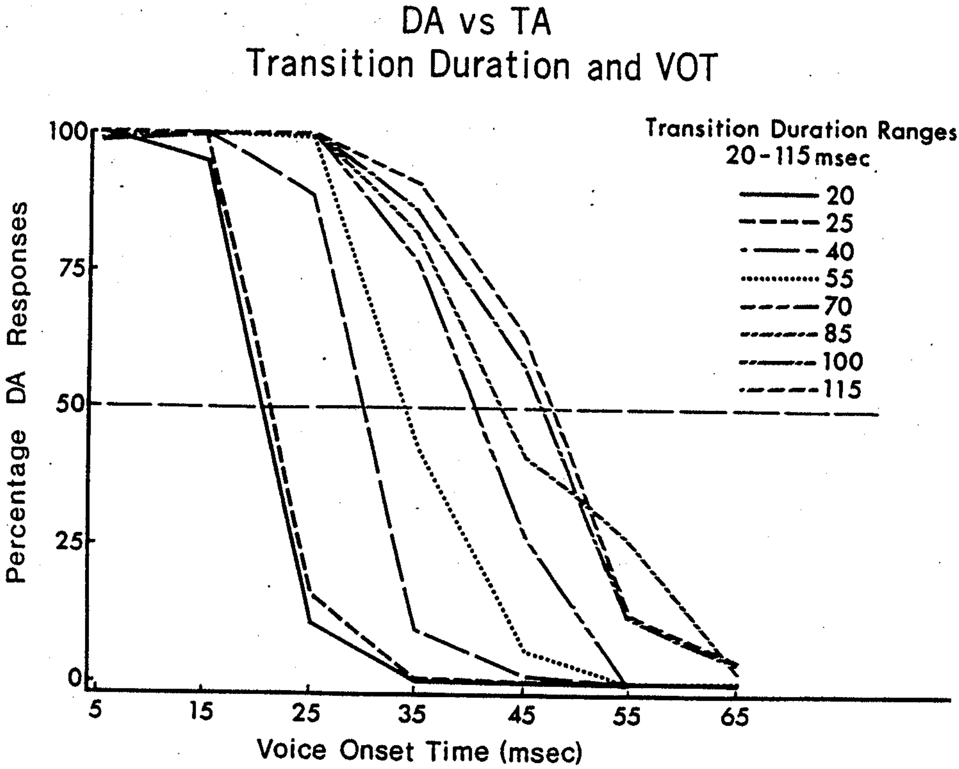

Fig. 1. Response data in a forced-choice labeling test of /da/–/ta/ patterns synthesized on the Haskins Laboratories parallel formant synthesizer. Each data point on the curves representing patterns with transition durations of 40, 55, 70, and 85 msec. represents 8 judgments by each of 20 Ss, On the other curves each point derives from 4 judgments by the same 20 Ss.

msec. for the longest (115 msec.). Their data (Fig. 1) support the view that the absolute duration of the interval between burst and the onset of pulsing and first-formant energy cannot be held solely accountable for listeners' response behavior.

The Place Effect on Stop Voicing

The Stevens and Klatt (1974) discovery that transition duration affects the perception of stop voicing led them to suggest an explanation for observed VOT differences in stops of different places of occlusion (cf. Peterson and Lehiste, 1960; Lisker and Abramson, 1964), – namely, if the duration of the interval from stop to following vowel position

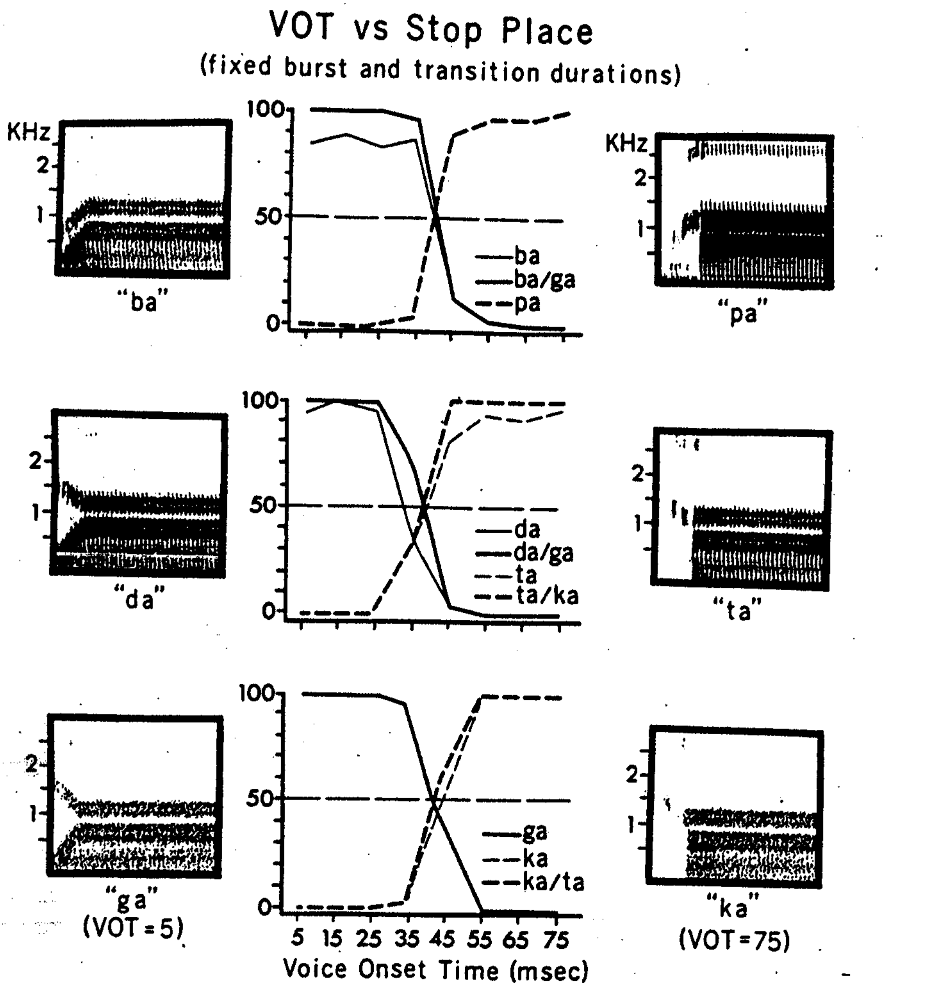

Fig. 2. Labeling response data from 8 Ss. (4 trials) required to classify perceived initial stop in each of 24 stimuli (8 VOT values, 3 transition configurations) with respect to voicing state and place of articulation. Each stimulus included a 5-msec. burst and a transition (end of burst to steady-state "vowel") of 50 msec.

differs for stops of different place (Lehiste and Peterson, 1961), then if stop voicing depends on the time interval between onset of the first formant and the termination of the transition, we should expect to find the progessively larger VOT boundary values as we go from labials to apicals to dorsals. Transition durations reported by Lehiste and Peterson (1961) are 67, 79, and 88 msec. for /p/, /t/, and /k/ respectively, while the same

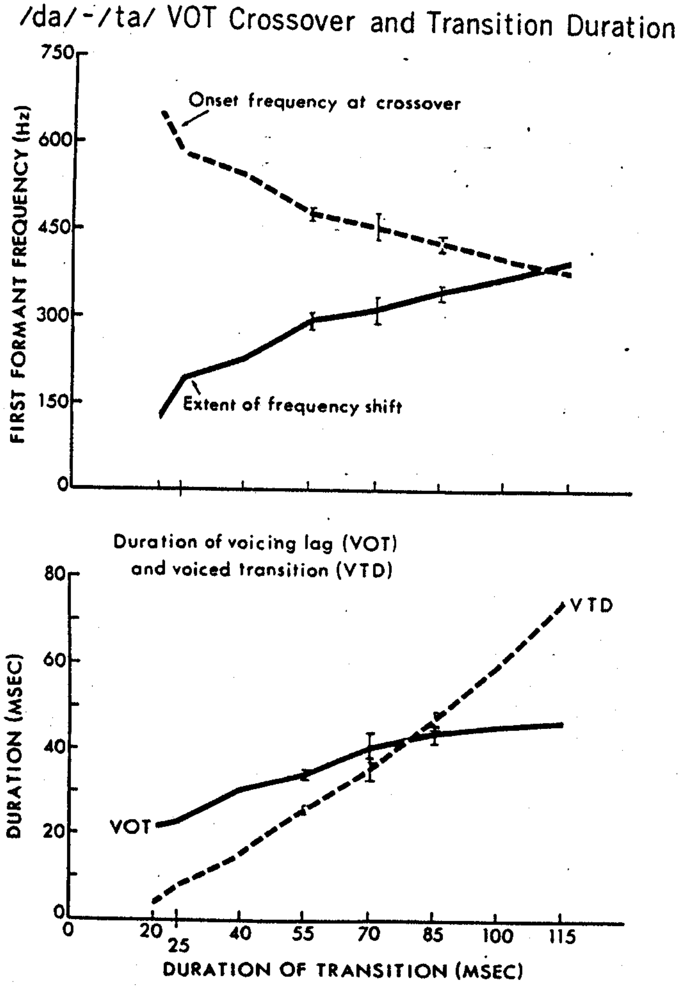

Fig. 3. The four curves are derived from the same data that are represented in Fig. 1.

authors report VOT values of 58, 69, and 75 msec. for the same stops respectively (Peterson and Lehiste, 1960). The mean F1 residue, if F1 "turns on" with the onset of voicing, has therefore a duration, of about 10 msec., that is constant for the three places of articulation; according to Stevens and Klatt such a transition duration is not readily detected.

If VOT for English /ptk/ is tied to transition durations, then synthetic speech patterns having burst and transition durations that do not vary with upper formant-transition

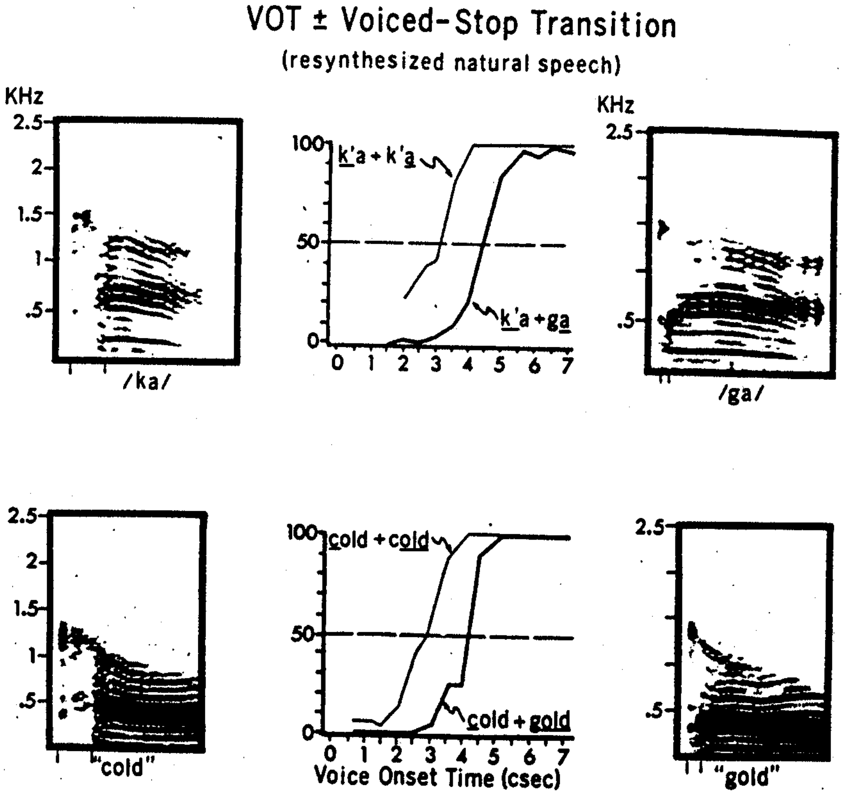

Fig. 4. The spectrograms are of naturally produced syllables, the nonsense pair /ka/ and /ga/ and the minimal word pair *cold-gold.* Test stimuli were prepared by combining the /k/ burst and varying durations of the following aspiration with the voiced portions of both members of each syllable pair. Editing was performed electronically with the Haskins Laboratories pulse-code-modulation (PCM) system. Twelve Ss gave ten responses each to each of 23 stimuli derived from /ga,ka/ and 28 stimuli from *gold-cold.*

patterns (the place cues) should yield VOT boundaries that are the same for the different places. In fact, if stops of the three places are simulated by such patterns, the VOT crossover values fail to show any regular increase as place of closure shifts from front to back in the mouth. For the particular stimuli used in such an experiment, with 5-msec. bursts and transitions of 50 msec., the crossover values at all three places showed no significant deviation from +40 msec. along the VOT dimension (Fig. 2). The boundary value between /bdg/ and /ptk/ might equally well be specified as being about 15 msec. before the end of the transitions.

VOT V. F1 TRANSITION

Despite the persuasiveness of the data on the effect of varying first-formant duration, it is not at all certain that they support the Stevens and Klatt (1974) hypothesis that there exists an F1 transition detector whose response to a frequency shift following voice onset is a more reliable basis for a voiced stop judgment than the VOT measure. The display in Fig. 3 represents the same data as those of Fig. 1, and the lower panel provides a comparison of the VOT boundary, as a function of transition duration, with the duration of the voiced F1 transition. It would appear from this that the voiced-transition-duration measure is rather more susceptible to variation with changing transition duration than is the VOT measure.

In both the Stevens and Klatt (1974) and Lisker *et al.* (1977) stimuli the variation in VOT and transition duration brought with it a variation, not only in the duration of F1, but in its onset frequency and amplitude of frequency shift (Fig. 3, upper panel). Either or both of these features might well make a significant contribution to stop voicing perception. Experimental data do in fact exist which indicate that it is not so much F1 movement as onset frequency that cues the stop voicing distinction (Lisker, 1975; Summerfield and Haggard, 1977).

The notion of an F1 transition detector implies that the presence of an F1 transition is a stronger cue to voicing than a short voicing lag, and that the absence of a voiced F1 transition is of greater importance perceptually than a long lag in voicing onset. In Fig. 4 data are shown which represent the labeling responses to stimuli derived by the editing of natural speech samples. They indicate that the presence of F1 transitions produced by voiced velar articulations (which ought to produce long transitions) is not incompatible with /k/ responses, provided the interval between burst onset and onset of voicing is long enough. Moreover, the absence of a /g/ transition (or the presence of no more F1 transition than that provided by a /k/ articulation following voice onset) is not incompatible with /g/ responses, this time provided the voicing lag is of brief duration.

CONCLUSION

The hypothesis that stop voicing perception depends entirely on the absolute duration of the interval separating burst and the onset of voicing and first formant is rejected. It is probable that no single acoustic parameter can be found to account for listeners' categorization of initial stops into the /bdg/ and /ptk/ sets. If, however, we ask whether VOT or voiced-transition duration is the more effective cue, then the answer seems clear. Since the presence of a voiced F1 transition is neither necessary nor sufficient to elicit voiced stop judgments, while the absence of the same feature is neither necessary nor sufficient to induce voiceless stop judgments, it is difficult to accept the hypothesis of an F1 transition detector that operates more reliably than even the simplest conceivable device responding to the absolute duration of the interval from burst to onset of voicing and first formant. This by no means rules out entirely a role for voiced F1 transition, – first, on the general ground that it is probably impossible to demonstrate that *any* acoustic feature plays absolutely no role in speech perception, and second and more compellingly, it accounts so neatly for the relation between VOT and transition duration as these vary with place of stop occlusion.

REFERENCES

ABRAMSON, A.S. (1977). Laryngeal timing in consonant distinctions. *Phonetica*, **34**, 295-303.

ABRAMSON, A.S. and LISKER, L. (1965). Voice onset time in stop consonants: acoustic analysis and synthesis. In D.E. Commins (ed.), *Proc. 5th Int. Congr. on Acoustics* (Liège), vol. 1A, paper A51.

ABRAMSON, A.S. and LISKER, L. (1970). Discriminability along the voicing continuum: cross-language tests. *Proc. 6th Int. Congr. Phonetic Sci., Prague, 1967* (Prague), 569-73.

FUJIMURA, O. (1971). Remarks on stop consonants: synthesis experiments and acoustic cues. In L.L. Hammerich, R. Jakobson, and E. Zwirner (eds.), *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen* (Copenhagen), 221-32.

HAGGARD, M., AMBLER, S. and CALLOW, M. (1970). Pitch as a voicing cue. *J. acoust. Soc. Amer.*, **47**, 613-17.

HEFFNER, R.-M.S. (1950). *General Phonetics* (Madison, Wisc.).

LEHISTE, I. and PETERSON, G.E. (1961). Transitions, glides, and diphthongs. *J. acoust. Soc. Amer.*, **33**, 268-77.

LIBERMAN, A.M., DELATTRE, P.C. and COOPER, F.S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, **1**, 153-67.

LISKER, L. (1975). Is it VOT or a first-formant detector? *J. acoust. Soc. Amer.*, **57**, 1547-51.

LISKER, L. and ABRAMSON, A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, **20**, 384-422.

LISKER, L. and ABRAMSON, A.S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech*, **10**, 1-28.

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**, 209-16.

PETERSON, G.E. and LEHISTE, I. (1960). Duration of syllable nuclei in English. *J. acoust. Soc. Amer.*, **32**, 693-703.

STEVENS, K.N. and KLATT, D.H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *J. acoust. Soc. Amer.*, **55**, 653-59.

SUMMERFIELD, A.Q. and HAGGARD, M.(1974). Perceptual processing of multiple cues and contexts. *J. Phonetics*, **2**, 279-94.

SUMMERFIELD, A.Q. and HAGGARD, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *J. acoust. Soc. Amer.*, **62**, 435-48.

WINITZ, H., LaRIVIERE, C. and HERRIMAN, E. (1975). Variations in VOT for English initial stops. *J. Phonetics*, **3**, 41-52.