Language and Speech
1977, Vol. 20

ON PUSHING THE VOICE-ONSET-TIME (VOT) BOUNDARY ABOUT*

L. LISKER**, A. M. LIBERMAN†‡, D. M. ERICKSON,
D. DECHOVITZ†, and R. MANDLER†
Haskins Laboratories

There is voluminous evidence that homorganic stop consonants are distinguishable on the basis of voice onset time relative to their supraglottal articulation. For initial stops a convenient acoustic reference point is the onset of the release burst, and VOT has been defined as the interval between this point and onset of glottal signal. VOT boundary values between voiced and voiceless initial stops of English have been established by spectrographic measurements of naturally produced isolated words and by perception testing of synthesized CV syllables. The close match between the two kinds of boundary values suggests that fairly natural values were chosen for the invariant features of the synthetic speech patterns tested. It is known, however, that certain of these affect voicing perception. New data from synthesis experiments show that VOT boundaries shift with changes in transition duration, and that it is the first formant and *not* higher ones which are responsible.

There is a good deal of evidence that homorganic stops are distinguishable on the basis of voice onset time (VOT) relative to their supraglottal articulation (Fant, 1960, p. 255; Lisker and Abramson, 1964). For initial stops a convenient acoustic reference point is the onset of the release burst, and VOT has been defined as the interval between this point and the onset of glottal signal. VOT boundary values between English /b,d,g/ and /p,t,k/ have been determined by spectrographic measurements of naturally produced isolated words and by perception testing of synthesized CV syllables. The close match between the two kinds of boundary values encourages the belief that fairly natural values were chosen for features of the synthetic speech patterns that were not systematically varied in the perceptual experiments. It has been known for some time, however, that certain of these features *do* affect the perception of stop voicing. Thus Cooper, Delattre, Liberman, Borst, and Gerstman (1952) pointed to a rising first formant as a cue to voicing, and the same group (Liberman, Delattre, and Cooper, 1958) later singled out first-formant "cutback" as a formidable cue to the English voiceless stops. More recently Fujimura (1971) and Haggard, Ambler, and Callow

* *This research was supported by a grant to Haskins Laboratories from the National Institute of Child Health and Human Development.*

** *Also The University of Pennsylvania.*

† *Also The University of Connecticut.*

‡ *Also Yale University.*

(1970) have shown that fundamental frequency can also serve as a cue to the stop voicing contrast. Most recently Stevens and Klatt (1974) have emphasized the role of transition duration, showing that with greater durations there is an increase in the VOT value at the boundary between synthetic /da/ and /ta/ syllables. From all these studies it is clear that listeners do not attend exclusively to VOT in judging synthetic stop-vowel patterns, where by "VOT" is meant the duration of the interval between burst onset and the point in time where simultaneously the periodic signal source is switched on and the first formant shifts from zero to full amplitude. Stevens and Klatt have argued that listeners, at least a significant proportion of them, respond in categorical manner to the presence *v.* absence of rapid frequency shifts in the formants, particularly the first formant, following the onset of voicing. This not only accounts for their data, but also, as they point out, serves to explain why VOT boundaries vary with place of stop closure, since it has been observed that burst and transition durations also vary with place in natural speech. The Stevens-Klatt theory emphasizes the fact that there is another temporal landmark, aside from burst and voicing onset, that may have perceptual importance for stop voicing, namely, the point where formant frequencies achieve values appropriate to the following vowel. It might be the case, they seem to be saying, that the choice of the burst as the reference point for measuring voice-onset timing is more a matter of visual convenience for the spectrogram reader than of selecting the most useful landmark for the human auditor. At least two questions may be raised:

(1) Is the Stevens-Klatt hypothesis the only one suggested by their data?

(2) Is their proposed new measure of voice-onset timing any more nearly sufficient than VOT as a basis for categorizing the stops?

To help answer these questions, let us look at some new data which show, to begin with, that the Stevens-Klatt finding is in fact replicable.

In Fig. 1 we have the responses of 20 phonetically naive young talkers of American English asked to label as either /da/ or /ta/ a set of appropriately designed synthetic speech patterns. The variables are VOT and transition duration. VOT was varied in 10 msec. steps from 5 to 65 msec. delay in onset of pulsing and first formant relative to the burst. In Test I six transitions ranging from 20 to 85 msec. in duration were used; in Test II durations varied from 40 to 115 msec., this last value being the largest for which acceptable /da/ and /ta/ syllables could be heard. Just as Stevens and Klatt found, the 50% crossover points along the VOT dimension move to higher values with increasing transition duration. The crossover for the shortest transition tested differs from the one for the longest by somewhat more than 25 msec. This shift is just about twice as large as the shift which Stevens and Klatt reported. Of course the 25 msec. shift shown here is occasioned, as we see, by a change of 95 msec. in transition duration; in the Stevens-Klatt experiment the transitions were varied by only 30 msec. Insofar as they are comparable, our data show the closest possible agreement with theirs. However, on the basis of our data it is as easy to emphasize the *stability* of the VOT boundary in the face of an extreme change in transition duration as it is to point out the undeniable fact that it is not absolutely immutable.

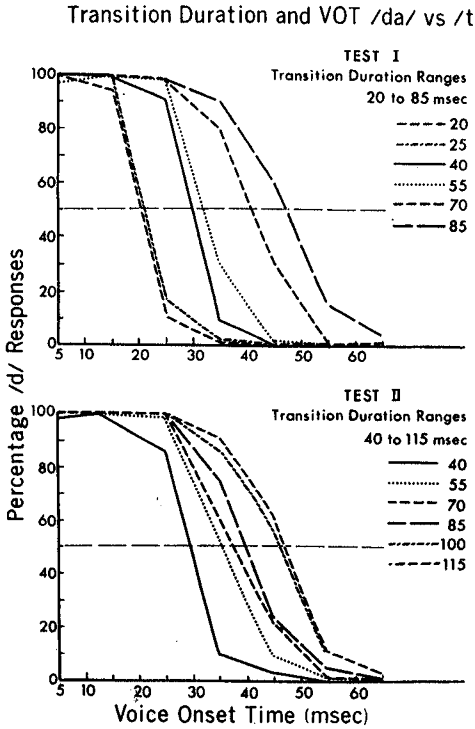

Fig. 1. Each of 20 phonetically naive American listeners gave four labelling responses in a forced-choice (/da/ or /ta/) test of a set of 56 acoustically different stimuli (8 transition durations × 7 values of VOT). Stimuli were presented in two parts (Test I and Test II) to reduce subject fatigue.

Now let us ask again whether these data, and the Stevens-Klatt data as well, point unequivocally to the duration of the voiced rather than the unvoiced transition as the feature that determines listeners' labelling judgments. In Fig. 2 we have represented schematically the first-formant trajectories of our test stimuli. For each transition duration, at VOT = +5 msec., the F_1 frequency rises linearly from an onset of 154 Hz to a steady-state value of 769 Hz. Since in general the F_1 intensity is zero until the periodic source in our synthesizer is turned on, for VOT values greater than +5 msec. the actual onset frequency of F_1 depends directly on VOT. Thus for a transition of 20 msec. duration the F_1 onset frequency at the VOT crossover value is about 620 Hz. In the display the F_1 trace is a solid line to the right of the VOT crossover. To the left of that value the dashed line indicates the absence of acoustic energy at the frequency of the first formant, while the higher formants (not shown) are excited by the random noise source of the synthesizer. VOT measurements include the 5-msec.

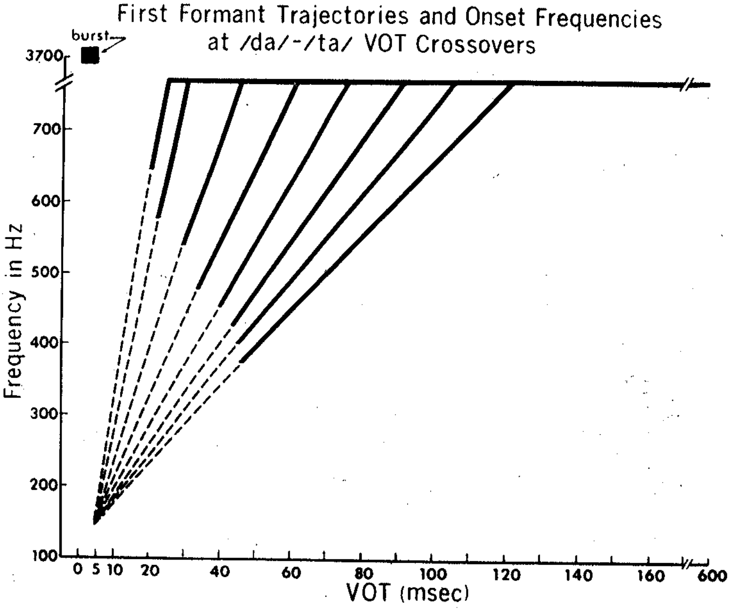

Fig. 2. First-formant trajectories for each of the transition durations used to generate the data of Fig. 1 are represented by a family of straight lines with origin at VOT = 5 msec. and F_1 frequency = 154 Hz, and terminating at F_1 steady state frequency 769 Hz appropriate for vowel /a/. At the 50% VOT crossover value for each transition duration F_1 switches from zero amplitude (dashed line) to full amplitude (solid line). The burst simulating the acoustic effect of consonantal release occupies the interval from 0 to 5 msec. along the abscissa and has a centre frequency near 3700 Hz.

burst which precedes the onset of amplitude in the formants. We see that with increasing transition duration not only is there a rightward shift in VOT crossover, but that there are also changes in F_1 onset frequency and in the duration of the transition following onset of the periodic excitation. These relations are more directly seen in Fig. 3.

In the upper panel of Fig. 3 we see how F_1 onset frequency, or alternatively the extent of F_1 shift following voice onset, varies at the VOT boundary with changing transition duration. Given the limitations of the experiment, the two curves of course say exactly the same thing, and pending further work one cannot say which measure is more relevant perceptually, or indeed how much meaning either of them has independently of the purely temporal measures of voicing. Perhaps we might note for now that as the transition duration is increased, given a fixed VOT value, there are changes in the onset frequency and extent of frequency movement of F_1, and that one or the other of these changes increases the impression of voicing; consequently the

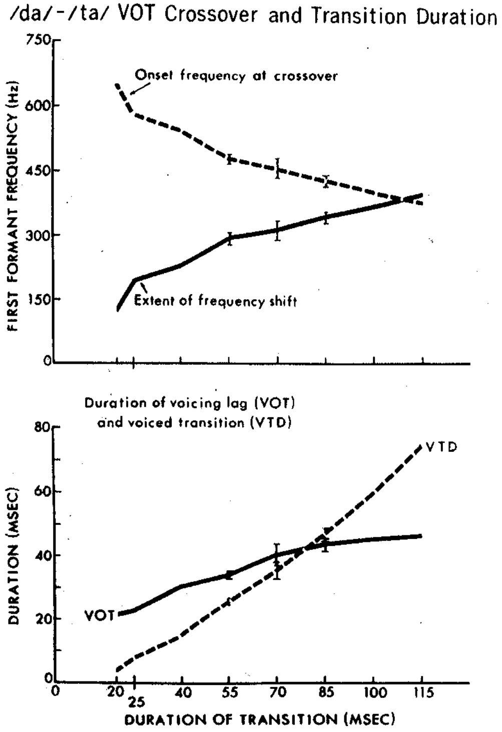

Fig. 3. The four curves in this figure represent the same data as Fig. 1. In the upper panel the curve marked "onset frequency" shows F_1 frequency at VOT crossover as falling with increasing transition duration. For transition durations of 55, 70, and 85 msec. the endpoints of the short vertical lines represent the somewhat different F_1 onsets obtained by Test I and II. The second curve in the upper panel represents extent of F_1 frequency shift from onset of full amplitude at the VOT boundary to achievement of /a/ steady-state frequency of 769 Hz. The short vertical lines again indicate differences between Test I and II results.

The curves of the lower panel represent crossover values of VOT and duration of buzz-excited F_1 as functions of transition duration. Vertical lines indicate differences between Test I and II results.

onset of glottal pulsing must be delayed to achieve a stimulus which is ambiguous as between /da/ and /ta/.

The lower display of Fig. 3 suggests an answer to the question of whether the measure of voice-onset-time proposed by Stevens and Klatt provides a more stable index of stop voicing than does VOT. Their measure, here labelled "VTD" for "voiced transition duration", ought to yield a curve of smaller slope than the standard

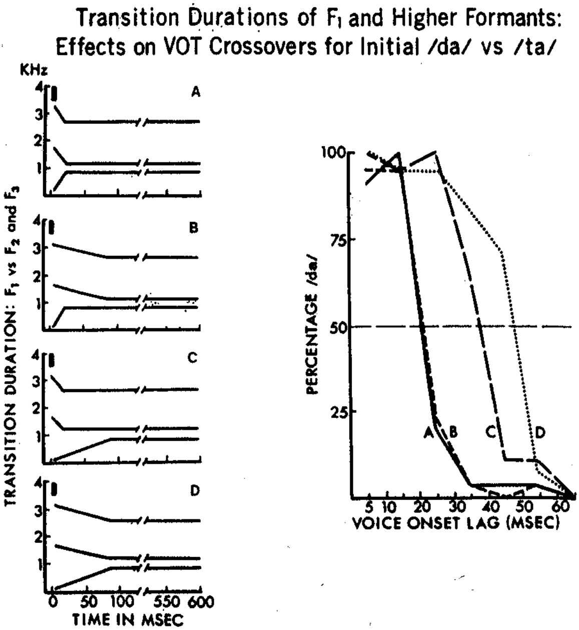

Fig. 4. Each data point represents 144 labelling responses (12 Ss × 12 trials) to each of 28 stimuli (4 transition configurations × 7 values of VOT). The four possible combinations of F_1 transition duration (15 and 80 msec.) and the same two values for F_2 + F_3 are shown in the schematic spectrograms on the left; labelling responses for each transition type are represented on the right.

VOT measure, if in fact it is true that listeners pay more attention to the transition following voice onset than to the preceding voiceless interval. This is not the case here. We conclude, with Stevens and Klatt, that VOT is not alone sufficient to explain our listeners' behaviour, but that VTD, their proposed measure, is even less adequate, by itself, to account for that behaviour.

So far we have been talking about formant transitions as though only the first formant deserves attention in a discussion of stop voicing. To see whether this is justifiable we ran a second experiment in which, along with VOT, the transition duration of the first formant was varied independently of the two higher formants. VOT values ranged from +5 to +65 msec. in 10-msec. steps, with representative patterns as shown in Fig. 4. The various combinations of transitions tested were presented to twelve listeners, each of whom provided twelve independent labelling judgments of each of the 28 test stimuli. The transition duration values tested were 15 and 80 msec. It is apparent that with a short F_1 transition the effect of varying the duration of the higher formant transitions is nil. With the longer F_1 transition the higher formants have some effect, but that effect, as measured by VOT crossover shift, is considerably smaller

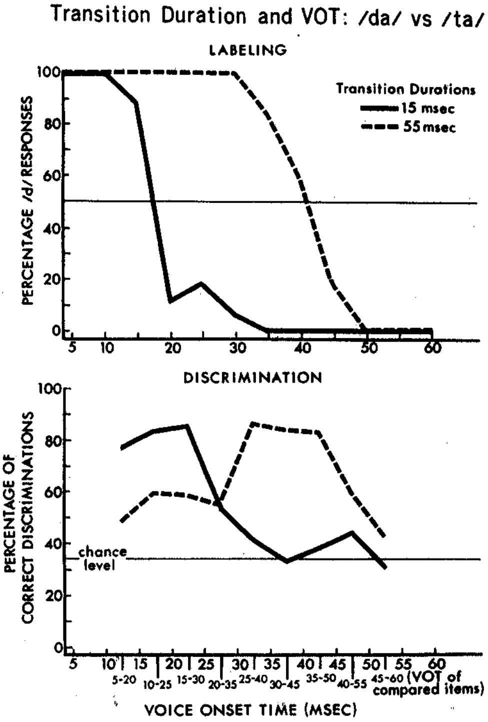

Fig. 5. The upper panel represents 12 responses by each of 12 subjects to 24 stimuli (2 transition durations × 12 values of VOT). The lower panel shows data from discrimination task in which stimuli of the same transition duration and VOT values differing by 15 msec. were presented as triads (two of them identical), with subject required to choose one of the triad as the "odd" member.

than the effect of a change in the first-formant transition. The conclusion to be drawn is, therefore, that the effect of transition duration on stop voicing perception is primarily ascribable to the first formant.

Finally, an experiment was run to determine whether the effect of transition duration on labelling behaviour was reflected in listeners' behaviour in a discrimination task. Two transition durations, with all three formants moving in synchrony, were tested: 15 and 55 msec. VOT was varied in 5 msec. steps from +5 to +60 msec. The upper panel of Fig. 5 indicates that the twelve listeners tested were consistent in their labellings with the behaviour of the twenty subjects whose responses were given in Figs. 1-3. The VOT crossovers are at about +18 and +41 msec., for transition durations of 15 and 55 msec. respectively. The closest corresponding values from the earlier test were at about +20 and +34 msec. for transition durations of 20 and 55 msec. The discrimination test used was of the "oddity" type, and the items in each comparison differed by

15 msec. in VOT. The functions obtained are shown in the lower panel of Fig. 5, from which it is clear that with the longer transitions the discrimination peak moves together with the category boundary along the VOT dimension. Moreover, the overall discrimination level is generally higher for the stimulus set with the longer transitions.

To conclude then, the VOT boundary is not fixed, varying directly with the transition duration. However, it is restricted in range, appearing to lie between limits at about 18 and 48 msec. following the burst onset. The duration of the voiced transition at the boundary also varies, over a range from near 0 to 75 msec., and our data fail to give any indication of a limiting value beyond which /p,t,k/ might not be heard. It is not entirely out of the question that for some listeners VTD is a more potent cue than VOT, but our data support the belief that for most it is the latter measure which better predicts the assignment of initial stops to the voiced and voiceless categories of English.

References

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. and Gerstman, L. J. (1952). Some experiments in the perception of synthetic speech sounds. *J. acoust. Soc. Amer.*, **24**, 597-608.

Fant, C. G. M. (1960). *Acoustic Theory of Speech Production* (The Hague).

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

Haggard, M. P., Ambler, S. and Callow, M. (1970). Pitch as a voicing cue. *J. acoust. Soc. Amer.*, **47**, 613-17.

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

Stevens, L. and Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *J. acoust. Soc. Amer.*, **55**, 653-9.