Is it VOT or a first-formant transition detector?*

Leigh Lisker

University of Pennsylvania, Philadelphia, Pennsylvania 19174
Haskins Laboratories, Yale University, New Haven, Connecticut 06520
(Received 26 December 1974; revised 12 March 1975)

Discussion of voicing as a distinctive property of English stop consonants in initial position has centered on the measure of "VOT," the time of onset of laryngeal signal relative to the noise pulse generated by the stop release, but it has been shown that listeners' selection of /b,d,g/ vs /p,t,k/ responses to synthetic stop+vowel stimuli is not determined entirely by VOT. Significant effects have been reported to depend on the behavior of the first formant (F_1) frequency immediately following voice onset, and on this basis it has been suggested that a "feature detector" responsive to a rapidly shifting F_1 better explains the infant's discrimination of the two stop categories than some mechanism which measures VOT directly. The relative importance of VOT as against the presence versus absence of F_1 frequency shift after voice onset is assayed in several synthesis experiments in which VOT and F_1 configurations are systematically varied. Labeling data obtained indicate that varying VOT regularly effects a significant change in listeners' judgments, and that varying F_1 has some effect too, but this latter variation is neither necessary nor sufficient generally to shift judgments decisively from one stop category to the other. The data further suggest that the presence of an F_1 rising transition after voice onset serves as a voiced-stop cue not so much because of its dynamic aspect, but simply because its onset frequency is low, i.e., at a value appropriate to a closed or almost closed state of the oral cavity.

Subject Classification: 70.30, 70.20, 70.70.

The large and still growing literature on the phonetic features that serve to distinguish linguistically distinct categories of homorganic stop consonants has been recently augmented by a short but significant contribution from Kenneth Stevens and Dennis Klatt.[1] The burden of their report is that perceptual importance attaches to the fact that for the voiceless aspirated stops of English the onset of voicing associated with a following stressed vowel occurs at about the time that the first formant has achieved the frequency appropriate to that vowel. Thus the so-called "VOT measure," i.e., the duration of the interval between onset of the burst resulting from stop release and the onset of glottal signal, has a value that is essentially equal to the duration of the oral opening gesture. By contrast, English /b, d, g/ are characterized by VOT values such that the formant transitions following release are excited by the glottal source over a significant portion of their total duration. On the basis of certain data from experiments in synthesis, Stevens and Klatt demonstrate that the boundary along the VOT dimension between /d/ and /t/ is not completely stable, but may vary considerably as a function of the rate and/or the duration of the transition. Of five subjects tested, one appeared to be responding primarily to VOT, while another's responses were rather to the interval between voice onset and the point in time at which the formant transition was completed. The other subjects were intermediate between these two, i.e., they seemed to use a mixture of these two strategies. On the basis of their findings Stevens and Klatt suggest that listeners generally have the ability to respond differentially to signals depending on whether or not they present a pulse-excited first formant of rapidly shifting frequency, and that *this* ability rather than one which "simply" measures VOT is what the language-acquiring infant relies on in the first steps toward a mastery of English phonology, that infant whom Eimas and his associates have shown to be able to distinguish a VOT of 20 msec from one of 40 msec quite like adult speakers of English.[2] The measure which Stevens and Klatt propose, a kind of complement to VOT, namely the transition duration *minus* VOT (we might call it simply "VTD" for "voiced transition duration"), has the merit, as they point out, that it is very probably more nearly independent of the place of stop articulation than is VOT, since it appears that VOT and burst and transition durations all increase from labial to alveolar to velar place of closure. This would seem to say that, inasmuch as speech production is for perception, the speaker of English controls the timing of voice onset, not with reference to the stop release, but rather in relation to the achievement of the steady-state vowel target formant frequencies. Of course, if there is no very significant variation in transition duration, at least for a given place of stop articulation before a given vowel, one might even suppose that the talker times the onset of voicing in relation to release, but that the listener attends to whether or not there is movement of the first formant after voicing begins.

A reading of the literature on the acoustic cues to stop voicing indicates that there should be nothing surprising about finding that the F_1 transition plays a role. A very early paper on speech synthesis reported that "the transitions of the first formant appear to contribute to voicing of the stop consonants" (Cooper *et al.*, 1952, p. 600).[3] Nor should it be thought at all extraordinary to find still other acoustic features, fundamental frequency (F_0) for example, that also control to some extent the phonetic classification of stop patterns as "voiced" or "voiceless." What would in fact be much more difficult to justify would be an assertion that any particular feature isolable in the acoustic signal plays absolutely no

role in a listener's phonetic categorizations. Some such features, we may feel sure, play a vanishingly small role (particularly if they appear to be independent of the larynx), but, given the experimental strategies used in discovering the acoustic cues, it is hard to imagine a feature not utterly imperceptible that could be shown to have no effect whatsoever on labeling behavior. A question that can reasonably be asked is: what is the relative importance of one feature compared with others? If it is claimed, for example, as Haggard *et al.* apparently do,[4] that F_0 has an importance of the order that may be claimed for VOT, one might ask whether the two features are equally necessary or perhaps equally sufficient as cues to the contrast, or whether there are F_0 contours for which varying VOT has no effect on labeling behavior, even as there appear to be values of VOT for which varying the F_0 contour has no effect on voicing judgments. The same question can be raised with respect to the Stevens–Klatt hypothesis if, as is reasonably inferred from their argument, they mean to claim for their VTD measure a perceptual importance equal to that determined for VOT.

In the earliest work in this area done at the Haskins Laboratories the pattern feature isolated for primary attention was called "first-formant cutback"; in later studies there the preferred term was "VOT." In all these studies the point was made, more or less insistently, that F_1 attenuation before voicing onset *and* the timing of that onset were to be thought of as acoustic features which *together* are manifestations of a shift in laryngeal state from wide-open and nonvibrating to closed-down and vibrating glottis. The terminological shift from "F_1 cutback" to "VOT" was occasioned by a shift of attention from the perceptual evaluation of synthetic speech patterns to the precise measurement of spectrographic patterns of human vocal-tract speech and to the underlying physiological and articulatory events. In spectrograms of natural speech F_1 cutback is simply very hard to measure; it is neither all that easy to fix the time at which F_1 can be said to have reached full amplitude nor do spectrograms suggest that F_1 amplitude is all that stable. Although it has its difficulties, to be sure, measuring VOT from spectrograms is much more easily accomplished, and by now the published data leave little ground for doubting its usefulness as a basis for distinguishing between stop categories. One might guess that the Stevens–Klatt measure of VTD, which would require fixing the time at which F_1 reaches some criterial intensity level and the time at which it reaches the steady-state frequency of the following vowel, is not one that will be attempted for any large number of spectrograms of natural speech. F_1 cutback and VTD are easily measured in synthetic speech patterns *when those patterns are fabricated with these measures in mind*. If the human listener had only to contend with such patterns it would be so much simpler to describe speech perception. In the case of F_1 cutback and VTD the match between natural speech patterns and synthetic is not easily accomplished, for the reasons just stated; in the case of VOT a very close match indeed has been determined, both for English and several other languages as well.[5] I think the question of determining the match between natural speech and synthetic is an important one, for we know that the match need not be slavishly close for a synthetic stimulus set to be a perceptually satisfactory match to natural speech, satisfactory at least in the sense that the relation between acoustic feature and gross phonemic labeling behavior is the same for the two kinds of stimuli. It may be remembered, for example, that a quite unnatural set of stimuli "accounted for" the do – to contrast by varying F_1 cutback alone, i.e., both VOT and VTD were held constant and in fact equal to zero.[6] This means, one should think, that we may make inferences about the speech-handling capabilities of the sensory–perceptual system from the data of experiments in speech synthesis, but that we must be cautious in asserting just how these capabilities are exercised when natural speech signals are being processed.

Let us come back to the question which serves as the heading to the present discussion. It is thoroughly objectionable if understood as an "either–or" question implying that a satisfactory answer must be one that asserts that one of the feature dimensions, VOT or VTD, plays little or no part in the voicing contrast. But it is reasonable to ask whether VOT or VTD is more important in some sense, and *this* question Stevens and Klatt seem to have answered in favor of VTD, at least as a basis for understanding the behavior of Eimas's infant subjects. I think there are grounds, in particular the data presented here in Figs. 1 and 2, for believing that their VTD measure has somewhat less significance than they would ascribe to it.

Figure 1 represents the labeling responses of 44 phonetically naive University of Connecticut students to stimuli of the types shown schematically in the upper left quadrant. Stimulus type A is composed of a burst and formant-transition configuration appropriate to the velar place of articulation, and the transition is followed by a steady-state formant pattern heard as the vowel /a/. From this basic pattern a set of 13 stimuli was generated by varying VOT together with F_1 onset from a value of 0 to 60 msec in steps of 5 msec, synthesis being accomplished by means of the Haskins Laboratories' parallel resonance synthesizer under computer control. Burst and transition durations were fixed at 20 and 45 msec, respectively. The solid curve in the upper right quadrant of the figure represents percentage /k/ responses as a function of VOT for all 44 subjects tested. In this and the other tests conducted, attention was directed to the initial stop alone, the subjects being informed in advance that the vowel quality was not at issue. The tests were of the usual "forced-choice" kind, with responses limited to /g/ and /k/ because preliminary testing indicated that any other responses were highly unlikely. The point at which responses to pattern A stimuli were divided evenly between /g/ and /k/ falls at just about VOT = 40 msec.

In the lower left and right quadrants of Fig. 1 are shown the responses of the 19 "best" subjects, those that labeled the largest number of stimuli identically on four exposures, and the six "worst" subjects, who were most nearly random in behavior. Even the "worst" subjects show a crossover value for pattern A stimuli along

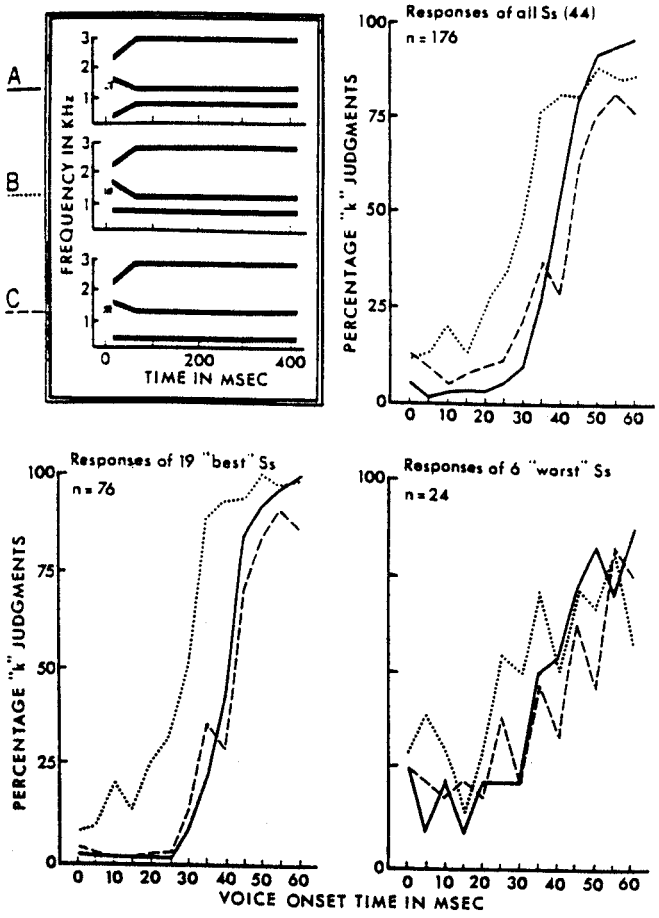

FIG. 1. The k–g contrast: VOT vs first-formant frequency (*n* represents number of judgments for each data point).

the VOT dimension, at about 35 msec.

All the responses to the stimuli of type A can be compared, first of all, with those elicited by patterns of type B, which differs from A only in that the first formant is without any transition, its frequency fixed at the steady-state value of /a/, i.e., at 769 Hz in these particular patterns. The dotted line showing the responses to the B stimuli indicates that the presence of a sharply rising F_1 is not a requirement for a majority of our subjects to report hearing /g/; even the six "worst" subjects gave mostly /g/ responses for VOT less than 25 msec. For the 19 "best" subjects it appears, to be sure, that responses to B stimuli consistently show more /k/ judgments over the entire VOT range than they do in the case of the A patterns; the B curve is displaced to the left by about 10 msec. Moreover, while /k/ judgments reach 100% for large VOT values, /g/ judgments are no better than 90% at VOT = 0. If we ask whether there is any value of VOT for which the pattern difference between A and B is sufficient to shift judgments from mostly /g/ to mostly /k/, the answer is that, for all 44 subjects, there is precisely one value of VOT, namely 35 msec, at which pattern A elicited mainly /g/ (73%) and pattern B mostly /k/ (76%). For all other values of VOT the two patterns were judged, by a greater or lesser majority, to belong to the same stop category. For the "best" subjects the A pattern with VOT = 35 msec yielded 79% /g/ while the B pattern with the same VOT value was scored 88% /k/.

Pattern C resembles B in having a straight F_1; it differs in that the frequency of the formant is very near (386 Hz) the onset frequency of the bent F_1 of pattern A (361 Hz), so that its auditory quality is rather [ɯ] than [a]. The effect of this lowering of the F_1 frequency on the /g/–/k/ boundary is most clearly visible in the responses of the "best" subjects: for small VOT values as many /g/ responses were elicited by pattern C as by A, *despite the absence of any F_1 frequency shift in C*. In fact, it would seem as though pattern C responses differ from A mainly in that at higher VOT values the former elicited somewhat fewer /k/ judgments. In other words, it might be inferred that the lower steady-state F_1 frequency is a more strongly pro-/g/ cue than the absence of an F_1 frequency shift is pro-/k/. It must of course be remembered that pattern C, with post-transition F_1 and F_2 frequencies of 386 and 1232 Hz, respectively, is heard as a stop followed by a vowel other than /a/, but we should presume that an adequate theory of stop voicing perception must be able to account for more than a single vowel context. The data for patterns A, B, and C suggest that it is not so much F_1 frequency shift as simply F_1 onset frequency that favors /g/. A low F_1 frequency tells the listener that the mouth is not very open, whether or not it is very soon to be more open.

Figure 2 presents labeling data for two more patterns which generally resemble types B and C, but whose post-

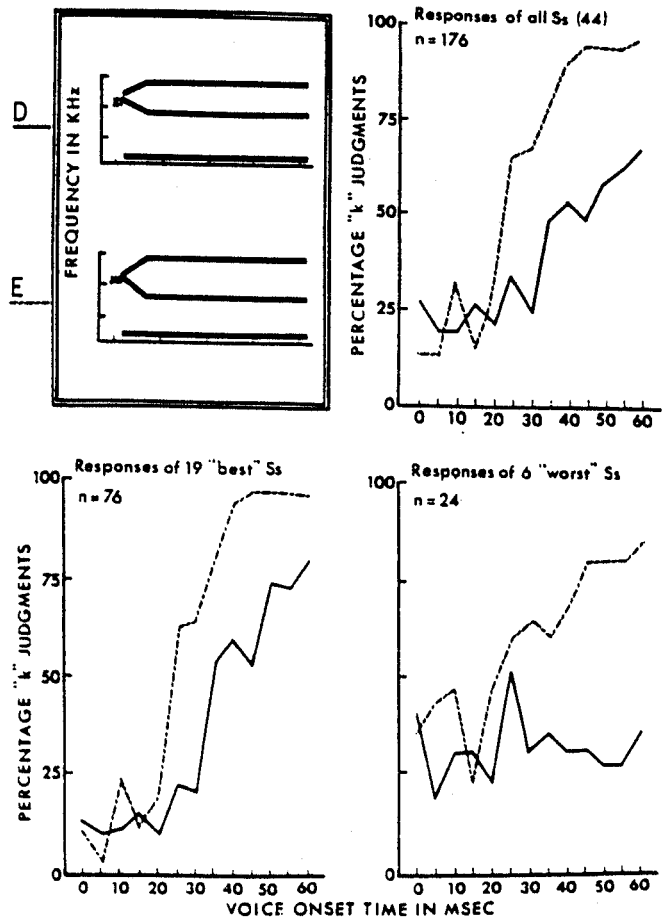

FIG. 2. The k–g contrast: VOT vs first-formant frequency (*n* represents number of judgments for each data point).

transition F_1 and F_2 have different frequencies, so that vowel qualities other than [a] and [ɯ] are heard. Pattern D, with a straight F_1 at 286 Hz, elicited a lower percentage of /k/ judgments than any of the other patterns tested; the six "worst" subjects in fact gave mostly /g/ responses for all but a single value of VOT (oddly enough at 35 msec). This behavior is understandable if we suppose that the low F_1 onset frequency is a strong voicing cue. Even so, the failure of the "worst" subjects to report /k/ for high VOT values is troublesome to a theory that bases the voicing contrast exclusively on VOT; it is, however, at least as damaging to the Stevens–Klatt hypothesis that absence of a shifting F_1 is a negative cue for voicing. That hypothesis would *not* lead us to predict the observed behavior in response to the pattern D stimuli; rather should we have expected failure to report /g/ for any of the test stimuli (of types B and C as well as D) over the entire VOT range of variation. But it is perhaps dangerous to put too much emphasis on the behavior of subjects singled out because of their generally unstable labeling performance. For the subject group as a whole it is hard to explain why /k/ judgments failed to reach even 70% for the D pattern with maximum VOT value; it seems implausible to suppose that the differences between C and D responses are based on the F_1 frequency difference alone. Of course pattern D also differs from C in having an F_2 whose steady-state frequency is considerably higher (i.e., 1845 Hz), and we might entertain the notion, harebrained on the face of it, that the raised F_2 is responsible for the massive shift to /g/ on the part of the "worst" subjects and the general bias toward /g/ displayed by the group as a whole, particularly noticeable for high VOT values. Pattern E disposes of such a hypothesis, however, since it has an F_2 whose steady-state frequency of 1772 Hz is almost as high as F_2 of pattern D. With a steady-state F_1 at the midrange value of 412 Hz, pattern E elicited solidly /k/ responses for high VOT values, while at the low end of the VOT range there was a preponderance of /g/ responses, though not more than for D and rather fewer than for A, B, or C. If we compare VOT crossover values for the four patterns with steadystate F_1 we find overall mean values which range from a minimum at about VOT = 23 msec for pattern *E* to a maximum at about 43 msec for C. This 20-msec difference is possibly large enough to deserve explanation, and if the literature is any guide, we should like to implicate F_1. With respect to this feature, however, patterns C and E differ by only 26 Hz, which may be considered a difference that is only doubtfully detectible. The most obvious difference between the two patterns is in their second formant configurations, but to consider this the basis for the response difference means to suppose that raising F_2 favors /k/, a supposition that makes no more sense, one should think, than the contrary hypothesis generated by the comparison of the "worst" responses to patterns C and D. The only thing left to say at this point is that I have nothing plausible to suggest to explain the 20-msec crossover difference in the responses to the C and E stimulus sets, and that work is continuing with the object of ensuring that the data just reported are indeed replicable, and if that is so, to determine more precisely the effects on stop voicing of F_1 and F_2 both separately and jointly.

To sum up, our data suggest that the presence of a voiced F_1 transition is not a requirement for stops to be heard as /b, d, g/. None of the data presented tells us, to be sure, whether *absence* of voiced F_1 transition is a requirement for English initial /p, t, k/. Of course a pattern with VOT equal, let us say, to −50 msec, and with F_1 beginning at that point with a low frequency and rising transition, would hardly be found in natural speech. More to the point, however, is the fact, based on data now in preparation, that such patterns are not heard as /b, d, g/ + vowel, but as /p, t, k/ − some other phonetic segment (perhaps /l/ more than anything else) + vowel. A sharply rising F_1, moreover, is most likely to occur in sequences of stop and a vowel with a high F_1; with the vowels /i/ and /u/, for example, such a feature is much less evident? Unless infants learn their stop voicing distinctions primarily from exposure to stops before /a/ or /æ/ (and possibly they do!), it seems doubtful that VTD, certainly a highly context-sensitive dimension, triggers a built-in device, while VOT, rather less context sensitive,[8] does not. Moreover the notion that VTD triggers a basic mechanism, for all its appeal, suggests that English owes its important position in the present-day world to the fact that very many languages seem perversely *not* to exploit it: Spanish-speaking children, for instance, must presumably learn to ignore information provided by this F_1 transition detector as one aspect of their process of language acquisition. If we say that the English-speaking learner calls the initial prestress stop /p, t, k/ if he detects aspiration, and that his detection of this feature rests significantly on the absence of F_1 transition after voice onset, then we may ask why Hindi-speaking listeners seem to require longer VOT intervals than do American listeners before they will report hearing voiceless aspirated stops.[9] It is not necessary to look far afield for languages which do not exploit the VTD dimension; English itself contrasts voiceless inaspirates and voiced stops medially, and VOT does a fair job of separating them. Where VOT fails, VTD does not help.

*This paper was presented orally before the annual meeting of the American Association of Phonetic Sciences held in St. Louis, MO on 5 November 1974.

[1]K. N. Stevens and D. H. Klatt, "Role of Formant Transitions in the Voiced–Voiceless Distinction for Stops," J. Acoust. Soc. Am. 55, 653–659 (1974).

[2]P. D. Eimas, E. R. Siqueland, P. Jusczyk, and J. Vigorito, "Speech Perception in Infants," Science 171, 303–306 (1971).

[3]F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some Experiments on the Perception of Synthetic Speech Sounds," J. Acoust. Soc. Am. 24, 597–608 (1952).

[4]M. Haggard, S. Ambler, and M. Callow, "Pitch as a Voicing Cue," J. Acoust. Soc. Am. 47, 613–617 (1970) report findings for which they provide no very clear interpretation. The fundamental frequency contour is said to serve as a stop category cue in synthetic speech patterns that are described as "ambiguous between /bi/ and /pi/" (p. 613), and while not all subjects responded unequivocally to the stimulus set, the authors express the belief that F_1 cutback is possibly no more robust a cue. They make no reference to VOT, and neither F_1 cutback nor VOT values are specified for their

test stimuli. A clearer picture of the relation between F_0 contour and VOT is presented in O. Fujimura, "Remarks on Stop Consonants—Synthesis Experiments and Acoustic Cues," in *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fisher-Jorgensen*, L. L. Hammerich, R. Jakobson, and E. Zwirner, Eds. (Akademisk Forlag, Copenhagen, 1971). From his data Fujimura concludes that F_0 plays a subsidiary role in the voiced–voiceless distinction among English initial stops.

[5]Data for English, French, Spanish, Thai, and Korean speakers can be found in one or more of the following.

A. S. Abramson and L. Lisker, "Voice Onset Time in Stop Consonants: Acoustic Analysis and Synthesis," 5th Int. Congr. Acoust., Liège, 1–4 (1965).

A. S. Abramson and L. Lisker, "Voice Timing in Korean Stops," *Proc. 7th Int. Congr. Phon. Sci., Montreal*, 1971, (Mouton, The Hague, 1972), pp. 439–446.

A. Caramazza, G. H. Yeni-Komshian, E. B. Zurif, and E. Carbone, "The Acquisition of a New Phonological Contrast: The Case of Stop Consonants in French–English Bilinguals," J. Acoust. Soc. Am. **54**, 421–428 (1973).

L. Lisker and A. S. Abramson, "The Voicing Dimension: Some Experiments in Comparative Phonetics," *Proc. 6th Int. Congr. Phon. Sci., Prague*, 1967, 563–567 (Academia, Czechoslovak Acad. Sci., Prague, 1970).

L. Williams, "Speech Perception and Production as a Function of Exposure to a Second Language," unpublished Ph. D. dissertation, Harvard U. (1974).

[6]A. M. Liberman, P. C. Delattre, and F. S. Cooper, "Some Cues for the Distinction between Voiced and Voiceless Stops in Initial Position," Lang. Speech **1**, 153–167 (1958).

[7]E. Fischer-Jorgensen long ago reported for Danish /b, d, g/ that "a movement of the first formant is evident only when it is high, i.e. in open vowels" ["Acoustic Analysis of Stop Consonants," Miscellanea Phonetica II, 42–59 (1954), p. 45].

[8]This is controversial, however. Lisker and Abramson, "Some Effects of Context on Voice Onset Time in English Stops," Lang. Speech **10**, 1–28 (1967), claim than VOT is unaffected by the particular vowel following the stop (p. 15), but a contrary finding is reported by D. Klatt, "Voice-Onset Time, Frication and Aspiration in Word-Initial Consonant Clusters," Res. Lab. Electron. Q. Prog. Rep.—MIT **109**, 124–136 (1973). In agreement with Klatt is W. E. Cooper, "Contingent Feature Analysis in Speech Perception," Percep. Psychophys. **16**, 201–104 (1974), who reports a somewhat higher VOT crossover value for /bi/–/pi/ than for /ba/–/pa/, in experiments with synthesized syllables. The Cooper data are difficult to interpret in the absence of detailed information about the transitional configurations involved.

[9]Statements to this effect by Hindi-speaking subjects are not in themselves strong evidence, but they are consistent with VOT measurements on Hindi reported in Lisker and Abramson, "A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements," Word **20**, 384–422 (1964).