RELATIVE POWER OF CUES: FO SHIFT VS. VOICE TIMING*

Arthur S. Abramson and Leigh Lisker††

Background

The acoustic features that bear information on the identity of phonetic segments are commonly called cues to speech perception. These cues do not typically have one-to-one relationships with phonetic distinctions. Indeed, research usually shows more than one cue to be pertinent to a distinction, although all such cues may not be equally important. Thus, if two cues, \( x \) and \( y \), are relevant for a distinction, it may turn out that for any value \( x \), a variation of \( y \) will effect a significant shift in listeners' phonetic judgments, but that there will be some values of \( y \) for which varying \( x \) will have negligible effect on phonetic judgments. We say then that \( y \) is the more powerful cue.

A good deal of evidence now exists to show that the timing of the valvular action of the larynx relative to supraglottal articulation is widely used in languages to distinguish homorganic consonants. The detailed properties of the distinctions thus produced depend on glottal shape and concomitant laryngeal impedance or stoppage of airflow, as well as on the phonatory state of the vocal folds. Such acoustic consequences as the presence or absence of audible glottal pulsing during consonant closures or constrictions, the turbulence called aspiration between consonant release and onset or resumption of pulsing, and damping of energy in the region of the first formant, have all been subsumed by us (Lisker & Abramson, 1964, 1971) under a general mechanism of voice timing. In utterance-initial position, the phonetic environment in which consonantal distinctions based on differences in the relative timing of laryngeal and supraglottal action have been most often studied, this phonetic dimension has commonly been referred to as voice onset time or VOT.

Although the acoustic features just mentioned, and perhaps some others, may be said to vary under the control of the single "mechanism" of voice timing, it is of course possible, by means of speech synthesis, to vary them one at a time to learn which of them are perceptually more important. We must not forget, however, that such experimentation involves pitting against one another acoustic features that are not independently controlled by the human speaker.

†Also University of Connecticut.
††Also University of Pennsylvania.

Acknowledgment. This work was supported by Grant HD-01994 from the National Institute of Child Health and Human Development to Haskins Laboratories. An oral version was presented at the Tenth International Congress of Phonetic Sciences, Utrecht, 1-6 August, 1983.

[HASKINS LABORATORIES: Status Report on Speech Research SR-77/78 (1984)]
A relevant feature not so far mentioned is the fundamental frequency (F0) of the voice. If we assume a certain F0 contour as shaped by the intonation or tone of the moment, there is a good correlation between the voicing state of an initial consonant and the F0 height and movement at the beginning of that contour (House & Fairbanks, 1953; but see also O'Shaughnessy, 1979, for complications). After a voiced stop, F0 is likely to be lower and shift upward, while after a voiceless stop it will be higher and shift downward (Lehiste & Peterson, 1961). Although the phenomenon has not been fully explained, it is at least apparent that it is a function of physiological and aerodynamic factors associated with the voicing difference.

The data derived from the acoustic analysis of natural speech can be matched by experiments with synthetic speech that demonstrate that F0 shifts can influence listeners' judgments of consonant voicing (Fujimura, 1971; Haggard, Ambler, & Callow, 1970; Haggard, Summerfield, & Roberts, 1981). Of further interest in this connection is the claim that phonemic tones have developed in certain language families through increased awareness of these voicing-induced F0 shifts and their consequent promotion to distinctive pitch features under independent control in production (Hombert, Ohala, & Ewan, 1979; Maspero, 1911).

Our motivation for the present study was to put F0 into proper perspective as one of a set of potential cues to consonant voicing coordinated by laryngeal timing. After all, our own earlier synthesis (Abramson & Lisker, 1965; Lisker & Abramson, 1970) yielded quite satisfactory voicing distinctions without F0 as a variable. In addition, Haggard et al. (1970) may have exaggerated its importance in the perception of natural speech by their use of a frequency range of 163 Hz, one very much greater than, for example, the range of less than 40 Hz found for English stop productions by Hombert (1975). We set out to test the hypothesis that the separate perceptual effect of F0 is small and dependent upon voice timing, while the dependence of the voice timing effect on F0 is virtually nil. We used native speakers of English as test subjects.

Procedure

Making use of the Haskins Laboratories formant synthesizer, we prepared a pattern appropriate to an initial labial stop followed by a vowel [a]. Variants of this pattern were then synthesized with VOT values of 5, 20, 35, and 50 ms after the simulated stop release. These values were chosen because of earlier work (Figure 1) that determined English voicing judgments for a VOT continuum ranging from 150 ms before release to 150 ms after release. This range of VOT values was sampled at 10 ms intervals, except for the span from 10 ms before release to 50 ms after release, which was sampled at 5 ms intervals. Those stimuli for which voice onset followed release, i.e., to the right of 0 ms on the abscissa, had noise-excited upper formants during the interval between the burst at VOT = 0 and the onset of voice. In the labial data at the top of the figure the perceptual crossover point between /b/ and /p/ falls just after 20 ms of voicing lag. Thus we expected that the extreme values of our more limited range would be heard as unambiguous /b/ and /p/, given an unchanging F0, while the category boundary, lying somewhere between, might be shifted one way or the other as the F0 was varied. In addition to a set of VOT variants having an F0 fixed at 114 Hz, we imposed onset frequencies of 98, 108, 120, and 130 Hz, values commensurate with ranges reported for natural speech (Hombert, 1975; House & Fairbanks, 1953; Lea, 1973; Lehiste & Peter-
Figure 1. English voicing judgments for stops varying in VOT. Below each pair of curves, a histogram (from Lisker & Abramson, 1964) of frequency distributions of VOT in speech. (Reproduced from Lisker & Abramson, 1970.)
son, 1961). That is, the F0 at voicing onset for each variant began at one of those frequencies and shifted upward or downward to a level of 114 Hz where it stayed for the rest of the syllable. These F0 shifts were of three durations, 50, 100, and 150 ms. These fit with our own cursory observations and bracket the value of 100 ms found by Hombert (1975). We recorded the resulting 52 stimuli—two tokens of each—in three randomizations and played the tapes to 11 native speakers of English for labeling as /b/ or /p/. The subjects, three women and eight men, represented a wide variety of regional dialects, ten in the United States and one in Britain.

Results

The overall results are shown in Figure 2. The three panels are for the durations of F0 shift. The abscissa of each panel shows the four VOT values, while the ordinate gives the percentage identified as /p/ for each VOT. The coded line standing for the variants with a flat F0 of 114 Hz is, of course, a plot of the same data in all three panels. The 50% perceptual crossover point for the flat F0 falls at about 25 ms of VOT. This is consistent with the results for the more finely graded series of stimuli in Figure 1. Indeed, for all conditions in Figure 2, it is VOT that is the main causative factor, regardless of F0, with perceptual crossovers in the region of the VOT of 20 ms. With hindsight we can say that additional stimuli with VOTs of 15 and 25 ms would have given more precision. At the same time, we do note effects of the fundamental frequency shifts: In each panel there is much spread of data points for 35 ms, and none for 50 ms.

In Figure 3 we focus on the results for the stimuli with a VOT of 20 ms, the one that shows the major effect of F0 shifts. For each of the four F0 onsets we see the percentage of /p/ responses. The coded lines stand for the three durations of F0 shift. A rather general upward trend in /p/ responses is evident as F0 onset rises. A two-way analysis of variance yielded a significant main effect for F0 onset, $F(3, 30) = 36.45, p < 0.001$, and a strong interaction between shift-duration and F0 onset for each duration, $F(6, 60) = 6.00, p < 0.01$.

Figure 4 focuses on the F0 onset of 130 Hz, the one that had the highest number of /p/ identifications. The /p/ responses for this F0 onset at all four VOT values are shown. Coded lines stand for the three shift durations; the flat F0 plot, marked "no shift," is repeated from Figure 2. It is once again obvious that the major effect is at the VOT of 20 ms, with the deviation from "no shift" increasing with greater shift duration.

The spread of points at the VOT of 5 ms in Figure 4, although much smaller than that at 20 ms, made us look for significant effects in individual cells of the confusion matrix underlying all our plots. That is, wherever we found apparent effects of fundamental frequency at VOT values other than 20, the locus of the main effect, we did a one-tailed t-test for significant deviations from 100%. All such suspicious clusters of responses were at VOT values of 5 ms and 30 ms; for the former, we expected 100% /b/ identifications and for the latter, 100% /p/ identifications. We found three such significant deviations; all of them at the VOT of 5 ms: (1) 120 Hz onset and 50 ms duration, $t(10) = 2.70, p < 0.01$, (2) 130 Hz onset and 100 ms duration, $t(10) = -2.51, p < 0.025$, (3) 130 Hz onset and 150 ms duration, $t(10) = 2.799, p < 0.01$. No such significant deviations were found at the VOT values of 35 ms and 50 ms.
Figure 2. Effects of F0 shifts on identification of VOT variants as English labial stops.
Figure 3. Effects of F0 shifts on VOT of 20 ms.

Figure 4. Effects of VOT and shift durations on onset of 130 Hz.
Conclusion

We conclude that there is a modest effect of fundamental frequency shifts on judgments of consonant voicing even within more natural ranges of F0 perturbation than those in Haggard et al. (1970). This is much like the results obtained in the investigation of Thai in an attempt at determining the plausibility of arguments on the rise of distinctive tones (Abramson, 1975; Abramson & Erickson, 1978).

Although they too used a more natural F0 range, Haggard et al. (1981) used an experimental design and stimuli that were somewhat different from ours; their aims were also rather different. To the extent that their data and ours are comparable, they support each other.

If, for the sake of considering the question of relative power of acoustic cues in the perception of a phonetic distinction, we separate fundamental-frequency shifts from the other cues linked to the dimension of voice timing, voice onset time is clearly the dominant cue. Only VOT values that are ambiguous with a flat F0 are likely to be pushed into one labeling category or the other by F0 shifts in a forced-choice test. Finally, there are values of VOT that are firmly categorical; they cannot be affected by F0. There are, however, no values of fundamental frequency that cannot be affected by voice onset time.

References


Footnote

The normal ranges of F0 variation linked to consonant voicing, not only in citation forms but especially in running speech (Lea, 1973; O'Shaughnessy, 1979), have still not been well described. We hope to report soon on our current study of this matter with different sentence intonations as a variable.