RELATIVE AMPLITUDE OF ASPIRATION NOISE AS A VOICING CUE FOR SYLLABLE-INITIAL STOP CONSONANTS*

BRUNO H. REPP
Haskins Laboratories

The present experiments demonstrate that amplitude of aspiration noise (relative to the following periodic portion of the vowel) is a cue for the distinction between voiced and voiceless syllable-initial stop consonants in English, and that it can be traded for voice onset time (VOT). In Experiment I, the category boundary on a synthetic VOT continuum (/da/–/ta/) was found to be a linear function of the amplitude ratio between the aspirated and unaspirated vocalic portions over a 24-dB range. A 1-dB increase in the ratio led to a shortening of the VOT boundary by 0.43 msec., on the average. In Experiment II, the synthetic stimuli were prefixed with a natural 10 msec. release burst, and burst and aspiration amplitudes were varied orthogonally. Both factors affected the voicing boundary in the expected direction but not independently; their interaction was ascribed to backward masking of weak bursts by strong aspiration noise. After accounting for this interaction, the effect of burst amplitude seemed very small compared to that of aspiration amplitude. These results suggest that the amount of aspiration noise is the primary voicing cue in the situation investigated, and that the perception of the noise follows psychoacoustic laws of time-intensity tradeoff. The methodological significance of amplitude parameters in speech synthesis is pointed out.

INTRODUCTION

Many studies have investigated the acoustic cues for various phonetic distinctions, such as voicing or place of articulation. A number of different perceptual cues is known to exist for each such distinction; for example, both voice onset time (VOT) and the onset frequency of the first formant are important voicing cues in syllable-initial stop consonants (Lisker, 1975; Summerfield and Haggard, 1977). Although the acoustic stimulus properties we call cues are generally not independent in articulation, they can be independently manipulated in perceptual experiments, and they are often found to make independent contributions to a given perceptual distinction.

Several recent experiments have investigated trading relations between different cues for the same phonetic contrast: A change in one cue can be compensated for, within limits, by an opposing change in another cue, so that exactly the same phonetic percept results (Bailey and Summerfield, 1978; Massaro and Cohen, 1976, 1977; Repp, Liberman,

* *This research was supported by NICHD Grant HD01994 and BRS Grant RR05596 to the Haskins Laboratories. I would like to thank Patti Price for running Experiment II, Arthur Abramson, Leigh Lisker and Virginia Mann for stimulating discussions, and Alvin Liberman, Dominic Massaro, and especially Arthur Abramson for comments on an earlier draft.*

Eccardt, and Pesetsky, 1978; Summerfield and Haggard, 1977). Such trading relations are a natural consequence of the multiplicity of cues for phonetic contrasts. More importantly, perhaps, these experiments have provided information about the relative importance of different cues in signalling a given phonetic distinction.

In part, the perceptual salience of a given cue may be a direct consequence of its qualitative auditory properties; e.g., for a given phonetic distinction, there may be a natural hierarchy of cues, such that spectral cues are more important than temporal cues, or some spectral cues are more important than others. However, the perceptual weight of a cue also depends on its prominence or clarity relative to the other parts of the signal. Among the quantitative auditory parameters that determine relative prominence are amplitude and bandwidth. These "secondary" parameters have received much less attention in past investigations than the "primary" parameters of duration and frequency. Nevertheless, the former are likely to play an important role in perception. For example, it is almost certain – though rarely pointed out in the literature – that the relative amplitudes of the second and third formants determine the relative weights of their respective transitions as cues for place of articulation (cf. Repp, 1978). By synthesizing speech or by manipulating real-speech tokens, we often disturb the natural cue hierarchy and therefore should pay close attention to factors determining perceptual prominence.

The present experiments examined whether the amplitude relationship between the aspiration noise and the periodic vocalic portion following it constitutes a cue for the perception of the voicing distinction in syllable-initial stop consonants. Voicing contrasts in English are conveyed by a variety of acoustic cues, all of which have been postulated to be consequences of an underlying change in laryngeal timing (Lisker and Abramson, 1964, 1971). Not surprisingly, the temporal aspect of this cue complex has been found to be most prominent perceptually, and many studies have varied the delay between stop release and voicing onset (i.e., VOT)[1] in synthetic syllables to determine the perceptual boundary between the voiced and voiceless categories. Emphasis on the timing aspect has led to a relative neglect of the fact that the interval prior to voicing onset can be filled with aspiration noise, whose very presence is likely to constitute a cue for voicelessness. This must be particularly true in English, where – as every linguist knows – the phonological categories "voiced" and "voiceless" may, for some contexts, represent the phonetic distinction between voiceless unaspirated (occasionally voiced) and voiceless aspirated stops.

Thus, it may be argued that the primary perceptual cue is not the abstract temporal property of "delay in voicing onset" (called "separation" by Summerfield and Haggard, 1974), but the presence and amount of aspiration during that delay, even though these two properties are tightly correlated in natural speech. We might expect, then, that

[1] *It is necessary to distinguish between two meanings of the term VOT: In articulation, it denotes the articulatory gesture of laryngeal adjustment which underlies the manifold acoustic consequences. Following common usage, however, VOT at the acoustic level refers only to the temporal separation between release and voicing onset; it does not include spectral voicing cues, such as F1 onset, that can be independently manipulated in synthetic speech.*

an increase in the amplitude of the aspiration noise relative to the following periodic vocalic portion would increase the salience of this important cue and, thus, increase the probability of classifying an English syllable-initial stop consonant into the voiceless (aspirated) category.

Two earlier studies have given some attention to the perceptual effects of aspiration in syllable-initial stops. Winitz, LaRiviere, and Herriman (1975) actually varied aspiration intensity in one of their experiments, but their results were not clear-cut, due in part to ceiling effects. Summerfield and Haggard (1974) found that presence v. absence of aspiration noise following the release burst had little (or even a paradoxical) effect on voicing perception. However, they did not report the amplitude of their noise source, which may have been relatively weak.

Experiment I was conducted to investigate the perceptual trading relation between relative amplitude and duration of aspiration (i.e., VOT) as joint voicing cues. The design of the experiment also permitted a second question to be asked, viz., whether changes in overall (absolute) stimulus amplitude within a 12-dB range affect voicing perception.

EXPERIMENT I

Method

Subjects. Eight subjects participated. They included the author, a research assistant, and six paid volunteers (Yale undergraduates) who had participated in previous experiments using synthetic syllables and had proven to be reliable listeners. The author is a native speaker of German who has lived in the United States for 10 years;[2] the other seven subjects are all native speakers of American English.

Stimuli. The stimuli were generated with the OVEIIIc serial resonance synthesizer at Haskins Laboratories. All stimuli were stop-consonant-vowel syllables perceived as either /da/ or /ta/. Their total duration was 300 msec. Fundamental frequency was constant at 125 Hz over the first 84 msec. and then fell linearly to 90 Hz at offset. The initial formant transitions were stepwise-linear and 48 msec. in duration; F1 rose from 285 to 771 Hz, F2 fell from 1543 to 1233 Hz, and F3 fell from 3019 to 2520 Hz. The duration of the synthesis time frames was 4 msec.

A ten-member VOT continuum was created by replacing periodic excitation with noise and simultaneously increasing the bandwidth of F1 to its maximum (thereby essentially eliminating F1). The amplitude of the noise source (which is independent of the amplitude of the periodic source in the OVEIIIc synthesizer) was constant at a nominal value 5 dB higher than that of the voiced source. However, the two amplitude parameters of the synthesizer are specified on unrelated scales, and the effective amplitude of the aspirated portion was about 20 dB below that of the following periodic

[2] *Linguist colleagues of the author are of the opinion that for syllable-initial voicing distinctions in stop consonants there should be little or no difference between German and English, at least in the dialects of the subjects. Anyway, there was no significant difference between the author's results and the subjects' results.*

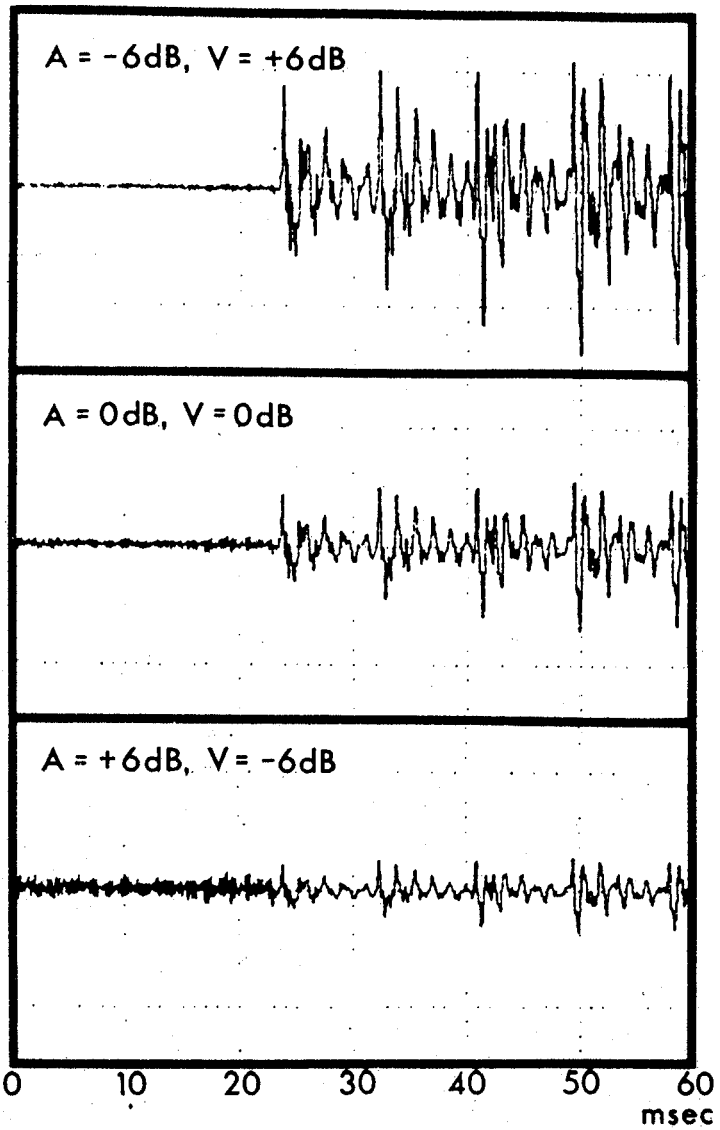

Fig. 1. Oscillograms of the first 60 msec. of a representative stimulus (VOT = 24 msec.) in three conditions. The baseline condition is shown in the center panel; the two conditions with the most extreme aspiration-to-periodic-portion amplitude (A/V) ratios are shown in the top and bottom panels, respectively.

portion of the vowel, as determined by later measurements of the synthesizer output. The periodic source was turned on 8 msec. (one pitch period) before voicing onset but kept at a minimal amplitude. This procedure insured that the second pitch pulse, which marked the true onset of voicing, had full amplitude. The ten stimuli thus generated had VOTs ranging from 8 to 44 msec. in 4-msec. steps. They had no special release bursts at onset.

All stimuli were digitized at 10 kHz using the Haskins Laboratories pulse code modulation system. From the digitized waveforms, eight additional stimulus series were constructed by independently amplifying or attenuating the aspirated and periodic portions of the original stimuli. Changes in amplitude in either stimulus portion were achieved by means of a computer instruction after placing a cursor at the onset of the first true pitch pulse. These manipulations resulted in a total of nine stimulus series, each a 10-member VOT continuum. Relative amplitudes of the aspirated portion of either –6, 0, or +6 dB (relative to the original stimuli) were orthogonally combined with relative amplitudes of the periodic portion of either –6, 0, or +6 dB (relative to the original stimuli). Thus, the stimulus ensemble included both a 12-dB range in absolute stimulus amplitude (from the –6/–6 to the +6/+6 stimulus series; the slash symbolizes the partition into aspirated and periodic stimulus portions), and a 24-dB range in the relative amplitudes of aspirated and periodic portions (from the –6/+6 series at one extreme to the +6/–6 series at the other). This range is illustrated in Fig. 1 which shows oscillograms of the first 60 msec. of a representative stimulus (VOT = 24 msec.) in the –6/+6, 0/0, and +6/–6 conditions. Given a true amplitude ratio between periodic and nonperiodic portions of about 20 dB in the 0/0 condition, the total range extended from 8 dB (+6/–6 condition) to 32 dB (–6/+6 condition).

Each of the nine stimulus series yielded a basic test unit of 28 stimuli, since the 10 stimuli in each series were recorded with the following frequencies: 1, 2, 3, 4, 4, 4, 4, 3, 2, 1. Thus, four times as many responses were collected for the center stimuli (which were likely to bracket the voicing boundary) than for the endpoint stimuli (which were likely to be reliably classified as voiced and voiceless, respectively). The complete stimulus sequence contained $9 \times 28 = 252$ stimuli in random order, with interstimulus intervals of 2 sec. Two such sequences were recorded. Some practice stimuli (endpoint stimuli at different overall intensities) preceded the first series.

Procedure. Each subject participated in two sessions, the second session being an exact replication of the first. The task was to identify in writing each syllable as beginning with either a D or a T. All in all, each subject listened to 4 blocks of 252 stimuli, providing a total of 16 responses to each stimulus with one of the four critical VOTs (20-32 msec.) in the voicing boundary region.

The tape was played back on an Ampex AG-500 tape recorder, and the subjects listened binaurally over Telephonics TDH-39 earphones. The amplitude was calibrated by means of a series of rapidly repeated /da/ syllables; the peak amplitude for this calibration series on a Hewlett-Packard voltmeter was about 80 dB SPL, which approximates the absolute amplitude of the periodic portion. The lowest intensity of the aspiration noise in the course of the experiment was about 54 dB SPL and thus well above the (unmasked) detection threshold.

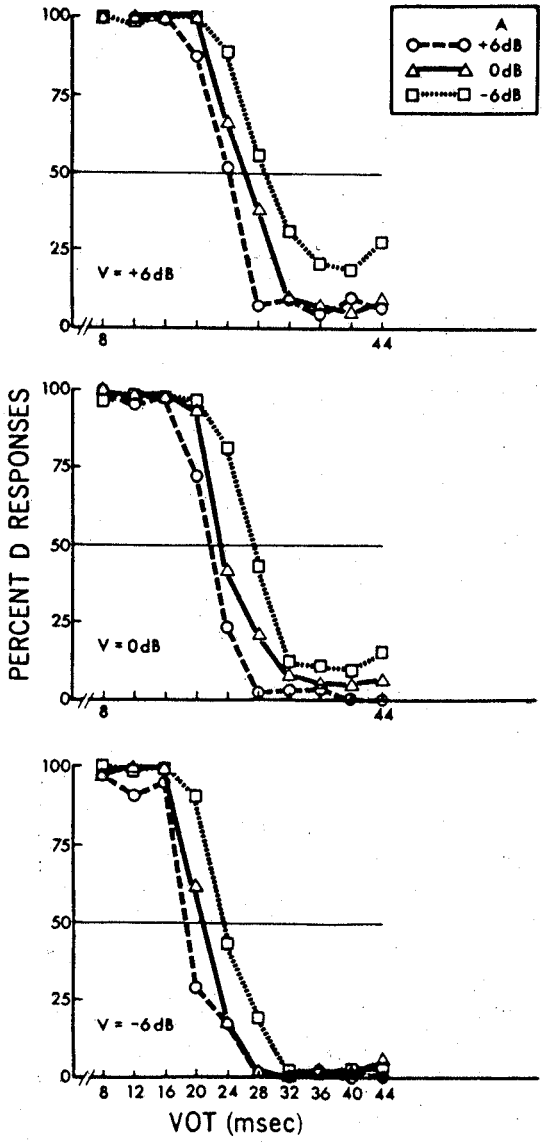

Fig. 2. Percentage of D responses as a function of VOT at three different relative amplitudes of aspiration noise (A) and three different relative amplitudes of the periodic portion (V).

Results

The results are graphically displayed in Fig. 2. The three panels show the effect of varying the amplitude of the aspiration noise (A) at each of the three levels of the amplitude of the periodic portion (V). Comparing the response functions within each panel, it can be seen that the number of D responses decreased (and that of T responses increased) as A increased, in accord with the predictions. Comparing the functions across panels, it is evident that the number of D responses decreased as V decreased. Since a decrease in V increased A relative to V (the A/V ratio), this result was equally in line with the predictions.

Two separate two-way analyses of variance were conducted, one on the total frequencies of voiced (D) responses in each stimulus series, and one on the voicing boundaries (50% intercepts of the labeling functions). Both analyses gave similar results, and since some of the individual boundary estimates were unreliable, only the results of the first analysis will be reported here. This analysis showed highly significant effects of both A, $F(2, 14) = 78.3$, $p < 0.001$, and V, $F(2, 14) = 37.6$, $p < 0.001$, but no significant interaction between these two factors, $F(4, 28) = 1.0$.

The response functions differed not only in their 50% crossovers, but also in their tails, especially at longer VOTs. For this reason, a curve-fitting procedure was not considered appropriate, and average voicing boundaries were estimated by simple linear interpolation. Fig. 3 summarizes the results in terms of boundary estimates derived from Fig. 2. The solid lines connect equal levels of V. Their negative slope indicates the effect of A, whereas their vertical separation indicates the effect of V on the voicing boundary. Both effects appear to be approximately linear, and the parallelism of the lines confirms the absence of a statistical interaction between the A and V effects. The dashed lines connect points of equal A/V ratio, differing only in overall amplitude. These lines are approximately horizontal, suggesting that overall amplitude – within the range investigated – did not affect the voicing boundary.

The most concise summary of the data is contained in Fig. 4. There we see the voicing boundary as a function of A/V ratio. The data points fall almost exactly on a straight line with a slope of –0.43. Thus, within the 24-dB range investigated here, the voicing boundary was a linear function of A/V ratio, with a 1-dB increase in the ratio leading to a shortening of the boundary by 0.43 msec., on the average.

Discussion

The results of this experiment have both methodological and theoretical implications. On the methodological side, they demonstrate the importance of considering amplitude relationships in speech synthesis. In creating synthetic syllables, and VOT continua in particular, investigators have generally restricted their attention to the temporal and spectral parameters of the stimuli. Amplitude settings are often chosen in a more or less arbitrary fashion, and they are rarely reported in the literature. Uncontrolled variations in amplitude relationships in synthetic stimuli may account for some differences in perceptual boundaries from one study to the other (cf. Repp, 1978), as well as for lack of agreement between perception and production results (cf. Lisker and

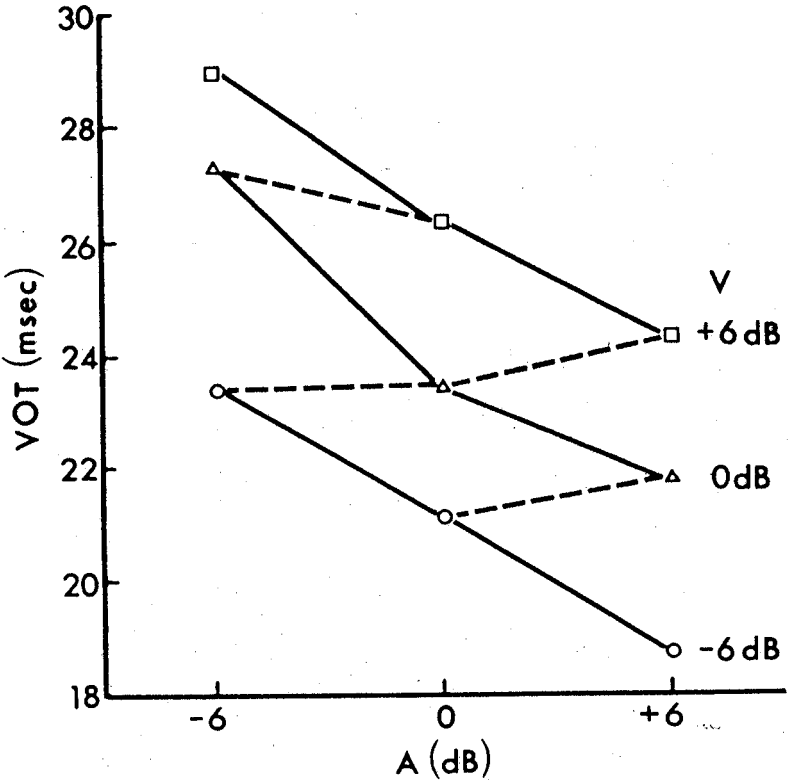

Fig. 3. Effects of A and V on the voicing boundary (in msec. of VOT).

Abramson, 1970). The present results suggest that amplitude relationships deserve as much attention in speech synthesis as temporal and spectral parameters.

Amplitude has also been neglected in the analysis of natural speech. A number of investigators have measured VOTs in natural speech without paying attention to amplitude relationships.[3] Current knowledge of articulatory mechanisms suggests that speakers should have little control over A/V ratio.[4] Any increases in A due to an increase in air flow probably would also increase V and thus leave the A/V ratio unchanged. Therefore, A/V ratio may be expected to vary only randomly in the production of voiceless stops

[3] *Dissertation research currently being conducted by Barbara Moslin at Brown University may be the only exception. At this time, I have not seen a sufficiently detailed account of that work to do it justice in the present context.*

[4] *Arthur S. Abramson and Leigh Lisker, personal communication.*

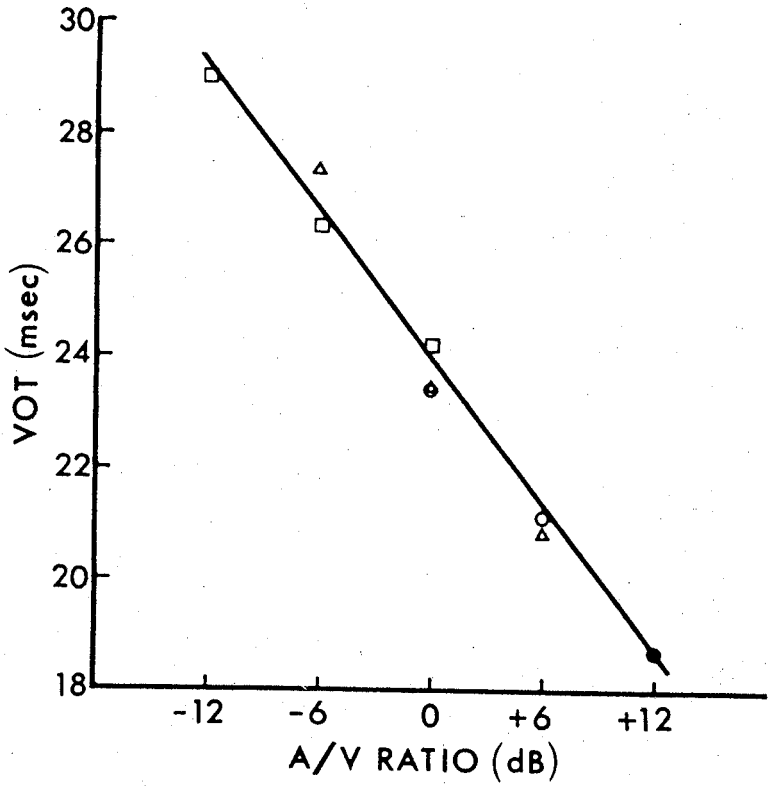

Fig. 4. The trading relation between VOT and A/V ratio as joint voicing cues.

and not to be correlated with VOT. Still, the relevant acoustic measurements remain to be done.

Turning now to theoretical issues, we note that the linear function in Fig. 4 represents a perceptual trading relation between A/V ratio and VOT, comparable to similar trading relations found between multiple cues for other phonetic distinctions (Massaro and Cohen, 1976, 1977; Repp *et al.*, 1978; Summerfield and Haggard, 1977). One of the most important theoretical issues in speech perception is to explain why acoustically quite diverse cues frequently contribute to the same phonetic percept and can be traded off against each other. The perceptual integration of these cues is often difficult to explain on purely auditory grounds. Repp *et al.* (1978) concluded on the basis of their findings that the only plausible explanation for the observed trading relations lies in the fact that the diverse cues are all consequences of the same articulatory act. This articulatory hypothesis also provides a valid explanation for why A/V ratio should be a voicing cue in English. If the articulation of voiced stops is contrasted with that of voiceless

stops, it is true that voiced stops generally have little aspiration noise, so that their A/V is naturally low. Voiceless stops, on the other hand, have a sizeable segment of aspiration noise in some contexts. This trivial observation implies a binary correlation between the voicing feature and A/V ratio in natural speech that may be sufficient to explain the perceptual trading relation. As pointed out above, A/V ratio is probably not correlated with VOT if only voiceless stops are considered.

While the articulatory hypothesis just outlined does provide a rationale for the present perceptual findings, it is not the only plausible explanation. Unlike certain other trading relations that seem difficult to explain in auditory terms (Bailey and Summerfield, 1978; Repp *et al.*, 1978), the effect of A/V ratio on the voicing boundary may very well have a psychoacoustic basis: The perceived duration of the aspirated portion may increase with A/V ratio, the increase in A/V ratio may elevate the noise above a masked threshold, or it may reduce the noise duration required for an accurate temporal-order judgment. (For relevant discussions, see Divenyi and Danner, 1977; Homick, Elfner, and Bothe, 1969; Kuhl and Miller, 1978; Miller *et al.*, 1976; Pisoni, 1977.) Pastore[5] has noted that the aspiration durations commonly used in VOT studies lie well within the range of auditory time-intensity tradeoff, so that an increase in intensity should be perceptually equivalent to an increase in duration. If, moreover, the perception of the noise is limited by backward masking due to the following periodic portion, any time-intensity tradeoff would presumably be enhanced. The occurrence of such a tradeoff in the perception of VOT rules out one form of psychoacoustic explanation according to which the abstract property of temporal separation (between stimulus onset and voicing onset, as measured in the signal) is the major voicing cue. Rather, the results suggest that the critical cue is aspiration *energy;* the voicing boundary then represents the point at which this energy exceeds the listener's response criterion. This criterion may (but need not) coincide with a masked detection threshold for the noise portion.

This argument in favor of a psychoacoustic mechanism does not diminish the plausibility of articulatory explanations for tradeoffs between acoustically quite different portions of the speech signal. For example, an articulatory rationale seems to be required to account for the perceptual integration of F1 onset frequency and VOT as joint cues to voicing (Summerfield and Haggard, 1977). However, the present tradeoff between amplitude and duration of aspiration noise occurs between two aspects of the same acoustic segment, and such trading relations are more likely to take place at the level of auditory processing.

EXPERIMENT II

The results of Experiment I must be qualified by the fact that the synthetic stimuli did not have any release bursts. Burstless synthetic stimuli have been used in numerous experiments, and the methodological conclusions reached above apply to these studies

5 *Pastore, R.E. Psychoacoustic factors in speech perception. To appear in a book edited by P.D. Eimas and J.L. Miller. This chapter provides an excellent review of the relevant issues.*

in particular. However, it is possible that, when the release is more clearly marked by a plosive burst, aspiration amplitude decreases in perceptual salience. This is especially suggested by the study of Summerfield and Haggard (1974) who found no perceptual effect of presence v. absence of aspiration following a release burst (in /–a/ context). Perhaps, the variations in aspiration amplitude in burstless stimuli serve only to mark the moment of stimulus onset more or less clearly. In order to investigate this hypothesis, a second experiment was conducted in which the stimuli had an initial release burst. The amplitudes of that noise burst and of the following aspiration noise were varied orthogonally to reveal their respective effects on voicing perception.

Method

Subjects. The author and seven new subjects (including one graduate research assistant, one colleague, and five paid student volunteers, all native speakers of American English) participated. The paid subjects were less experienced in listening to synthetic speech than those in Experiment I; however, this was made up for by the improved quality of the stimuli.

Stimuli. A new 10-member VOT continuum was synthesized, similar to the basic stimulus series in Experiment I (0/0 condition) but ranging in VOT from 0 to 36 msec. in 4-msec. steps. These stimuli were digitized and then prefixed with a 10-msec. alveolar release burst, so that all VOTs were increased by 10 msec. The burst was taken from the digitized waveform of a natural-speech /da/ pronounced by the author. Using the same methods as in Experiment I, the amplitude of the 10-msec. burst portion (B) and the amplitude of the following, synthetic aspirated portion (A) were varied orthogonally in three steps (–6, 0, +6 dB relative to the baseline stimuli), leading to a total of nine stimulus series. The amplitude of the periodic portion was not varied in this experiment. The baseline (0/0, the slash now denoting the division between burst and aspiration) and the extreme conditions (+6/–6 and –6/+6) are illustrated by the oscillograms in Fig. 5.

Procedure. Design, stimulus tapes, and procedure were all analogous to Experiment I, with the sole exception that all data were collected in a single session.

Results

The effects of the two amplitude factors, B and A, are shown in Fig. 6 which is exactly analogous to Fig. 2, with B taking the place of V. It can be seen that, in general, there was a decrease in D responses as A increased (within panels) and as B increased (across panels). However, the effects seemed less systematic than in Experiment I.

A 3 × 3 analysis of variance of the total frequencies of voiced (D) responses in the nine stimulus series yielded highly significant effects of B, $F(2, 14) = 14.0, p < 0.001$, of A, $F(2, 14) = 32.4, p < 0.001$, and a significant B × A interaction, $F(4,28) = 19.3, p < 0.001$. The results were extremely consistent across subjects, as indicated by the high significance levels. The interaction is more clearly represented in Fig. 7 which plots voicing boundaries derived by linear interpolation from Fig. 6. Surprisingly, B had its largest effect when A was at its highest level; when A was low, B had only a negligible effect.

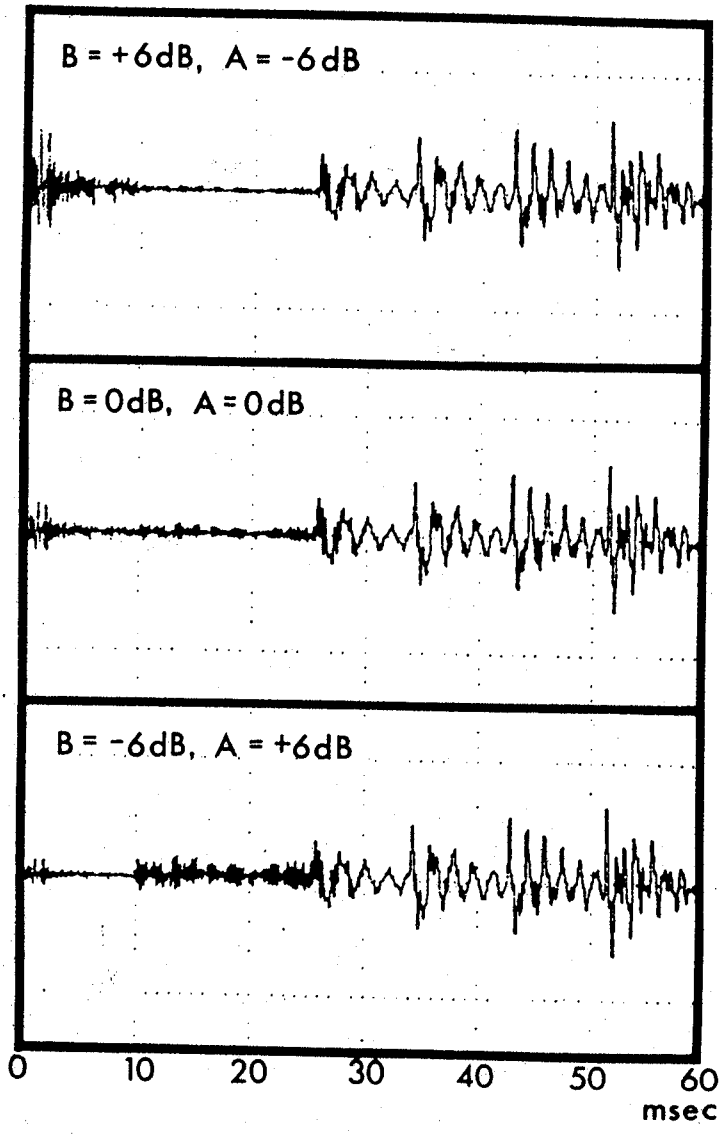

Fig. 5. Oscillograms of the first 60 msec. of a representative stimulus (VOT = 26 msec.) in three conditions. The baseline condition is shown in the center panel; the two conditions with the most extreme burst-to-aspiration amplitude (B/A) ratios are shown in the top and bottom panels, respectively.

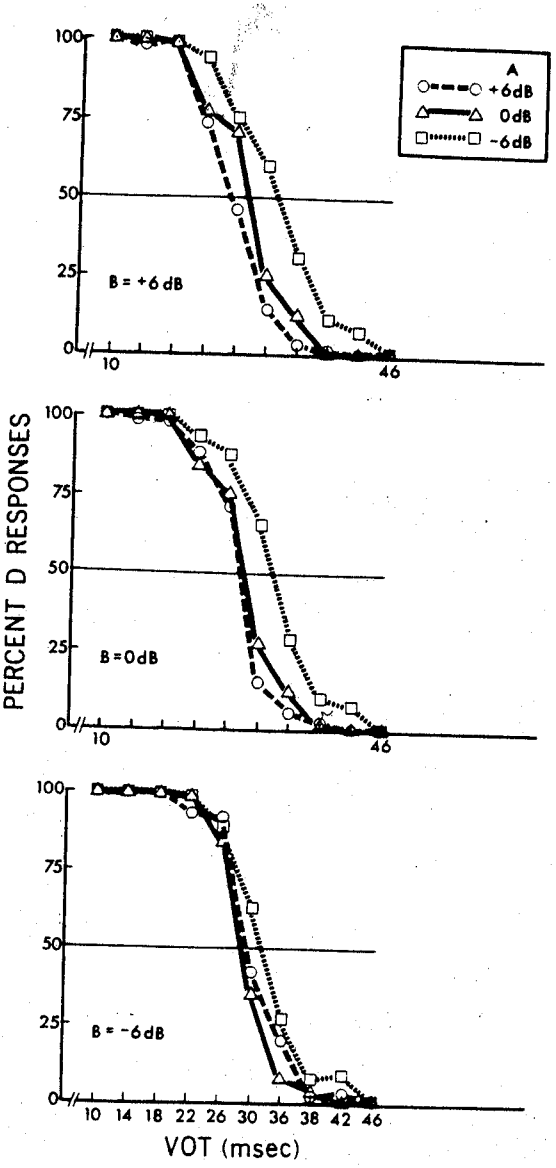

Fig. 6. Percentage of D responses as a function of VOT at three different relative amplitudes of aspiration noise (A) and three different relative burst amplitudes (B).

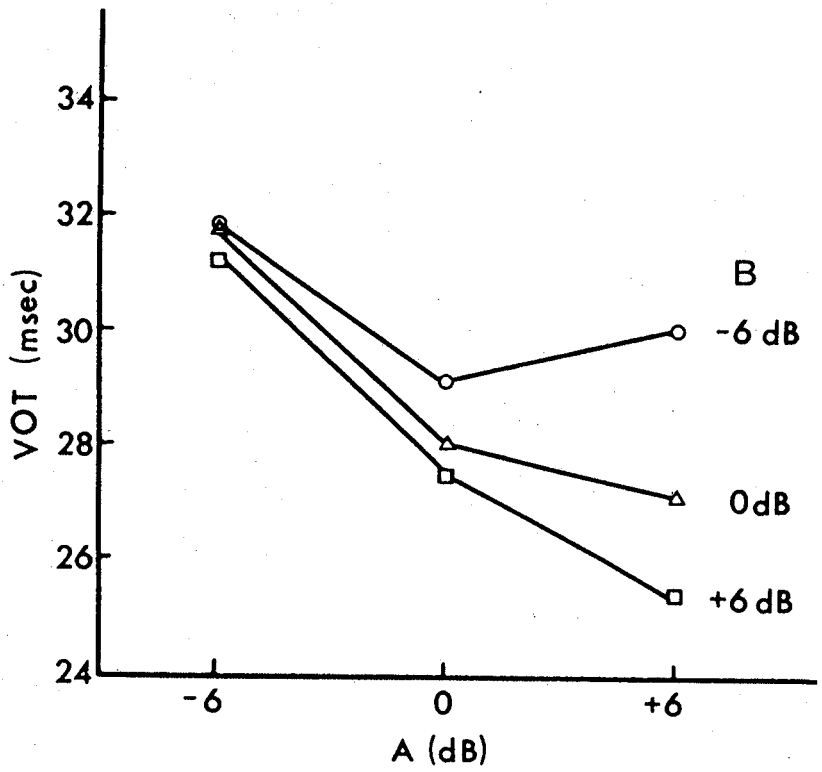

Fig. 7. Effects of A and B on the voicing boundary (in msec. of VOT).

Similarly, A had its strongest effect when B was high; when B was low, A had no systematic effect.

It should be noted that, despite the interaction, the noise amplitude effect of Experiment I was almost exactly replicated: Considering only amplitude changes in the *total* noise portion preceding voicing onset (–6/–6, 0/0, and +6/+6 conditions), we find that the corresponding voicing boundaries in Experiment II fell approximately on a straight line with a slope of –0.50, as compared with a slope of –0.43 in Experiment I. It is also evident from a comparison of Figs. 3 and 7 that the voicing boundaries in Experiment II were about 5 msec. longer than in Experiment I, suggesting that the 10-msec. burst was perceptually equivalent to only about 5 msec. of aspiration noise, perhaps due to its nonuniform amplitude contour (cf. Fig. 5).

Discussion

Experiment II fulfilled its purpose by demonstrating that aspiration amplitude has an effect on voicing judgments even when a burst is present, particularly (and paradoxically) when the burst is strong. Thus, the hypothesis that aspiration amplitude in burstless stimuli has its perceptual effect solely by masking the onset of the stimulus can be safely rejected. Rather, the total noise portion seems to be perceptually significant.

Experiment II also demonstrated an effect of burst amplitude (B). This effect is in agreement with the observation that voiced stops tend to have somewhat weaker bursts (relative to the following vocalic portion) than voiceless stops in English (Zue, 1976). Thus, there may be an articulatory basis for the perceptual effect of B. However, we must seriously consider psychoacoustic explanations for the present results, particularly to explain the curious finding that B had an effect only when A was high (and vice versa). This finding seems paradoxical when considering what would happen when B is gradually reduced to minus infinity. Since the effect of A was already small and nonsystematic at the lowest burst intensity used here (–6 dB), no effect of A would be predicted in burstless stimuli, in contradiction to Experiment I.

The puzzle is resolved as follows[6]: Note that if B is minimal, the original burst duration (10 msec.) must be subtracted from the VOT. If B is gradually lowered, it is likely that the burst suffers increasing backward masking by the following aspiration noise, which reduces the effective burst duration (and thus the effective VOT) until no burst is perceived any longer. The results in Fig. 7 have been plotted on a scale of *physical* VOT (as measured in the acoustic signal), but we do not know the *psychological* VOT perceived by the listener after the occurrence of auditory interactions between the signal components. It seems reasonable to assume that the psychological VOT was reduced in those stimuli in which a weak burst was followed by a strong aspiration noise, so that the data points for the –6/+6, –6/0, and 0/+6 conditions should be shifted downward on the VOT scale (cf. Fig. 7). Such a shift would tend to eliminate the B × A interaction and generate a pattern of results similar to that of Experiment I, viz., three parallel lines.

One consequence of this interpretation is that the effect of B appears to be negligible once a correction for the hypothetical masking effect is applied. Varying B helps the burst evade backward masking by a strong aspiration noise, but it seems to have little direct cue value for the voiced-voiceless distinction in English.[7] What remains is a strong effect of A, in accordance with Experiment I. It is possible to give this effect an interpretation similar to that of B in Experiment II: The aspiration noise may suffer backward masking from the powerful periodic portion that follows, and changes in amplitude in either stimulus portion may change the effective temporal relation between them, perhaps

[6] *Virginia Mann helped guide me toward this interpretation.*

[7] *It is entirely possible that burst intensity constitutes an important cue when aspiration is absent, as in the voiced-voiceless distinction for unaspirated stops in certain languages that feature such a distinction and, for that matter, in English for intervocalic stops beginning unstressed syllables.*

by affecting neural transmission times to the centers of speech perception. One might predict that A would lose its cue value once it exceeds the values at which backward masking occurs, but since such values were presumably not reached in the present experiments, the consistent effects of A in all conditions do not contradict a backward masking interpretation. Thus, the hypothesis that the significant voicing cue is the abstract temporal relation between the onsets of nonperiodic and periodic portions may be resurrected if it is understood that this timing relation is a perceptual one, determined at a stage following psychoacoustic interactions between the signal components.

To conclude, the present results should not be taken as a demonstration that VOT perception, or voicing perception, or perhaps even all of speech perception can and should be explained by psychoacoustic principles alone. They do suggest, however, that the alternative extreme view, that speech perception is to be understood entirely in terms of apprehending the behavior of an articulatory system, is similarly untenable. Instead of fostering another artificial dichotomy, it seems more reasonable to assume that speech perception involves mechanisms at a number of different levels. Psychoacoustic factors often seem prominent because of the psychoacoustic problems that we create for listeners by manipulating speech signals in certain arbitrary ways. Almost certainly, some of the auditory processes revealed in experiments such as the present ones play a role in natural speech perception. However, it is equally true that the guiding principle of speech perception at higher levels is likely to be found in the articulatory origin of the auditory signal.

REFERENCES

BAILEY, P.J. and SUMMERFIELD, Q. (1978). Some observations on the perception of [s]+stop clusters. *Haskins Labs. Status Report on Speech Research*, SR-53 (Vol. 2), 25-60.

DIVENYI, P.L. and DANNER, W.F. (1977). Discrimination of time intervals marked by brief acoustic pulses of various intensities and spectra. *Perception and Psychophysics*, **21**, 125-42.

HOMICK, J.L., ELFNER, L.F. and BOTHE, G.G. (1969). Auditory temporal masking and the perception of order. *J. acoust. Soc. Amer.*, **45**, 712-18.

KUHL, P.K. and MILLER, J.D. (1978). Speech perception by the chinchilla: identification functions for synthetic VOT stimuli. *J. acoust. Soc. Amer.*, **63**, 905-17.

LISKER, L. (1975). Is it VOT or a first-formant transition 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. (1970). The voicing dimension: some experiments in comparative phonetics. *Proc. 7th Int. Congr. Phonetic Sci., Prague, 1967* (Prague), 563-67.

LISKER, L. and ABRAMSON, A.S. (1971). Distinctive features and laryngeal control. *Language*, **47**, 767-85.

MASSARO, D.W. and COHEN, M.M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/–/si/ distinction. *J. acoust. Soc. Amer.*, **60**, 704-17.

MASSARO, D.W. and COHEN, M.M. (1977). Voice onset time and fundamental frequency as cues to the /zi/–/si/ distinction. *Perception and Psychophysics*, **22**, 373-82.

MILLER, J.D., WIER, C.C., PASTORE, R.E., KELLY, W.J. and DOOLING, R.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times. *J. acoust. Soc. Amer.*, **60**, 410-17.

PISONI, D.B. (1977). Identification and discrimination of the relative onset time of two component tones: implications for voicing perception in stops. *J. acoust. Soc. Amer.*, **61**, 1352-61.

REPP, B.H. (1978). Stimulus dominance in fused dichotic syllables. *Haskins Labs. Status Report on Speech Research*, SR-55/56, 133-48.

REPP, B.H., LIBERMAN, A.M., ECCARDT, T. and PESETSKY, D. (1978). Perceptual integration of temporal cues for stop, fricative, and affricate manner. *J. exp. Psychol. (HPP)*, **4**, 621-37.

SUMMERFIELD, A.Q. and HAGGARD, M.P. (1974). Perceptual processing of multiple cues and contexts: effects of following vowel upon stop consonant voicing. *J. Phonetics*, **2**, 279-95.

SUMMERFIELD, 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.

ZUE, V.W. (1976). Acoustic characteristics of stop consonants: a controlled study. *Lincoln Lab. Tech. Rep.* 523. M.I.T., Lexington, Massachusetts.