Perception of sine-wave analogs of voice onset time stimuli

James Hillenbrand
Department of Communicative Disorders, Northwestern University, Evanston, Illinois 60201

(Received 7 February 1983; accepted for publication 6 July 1983)

It has been argued that perception of stop consonant voicing contrasts is based on auditory mechanisms responsible for the resolution of temporal order. As one source of evidence, category boundaries for nonspeech stimuli whose components vary in relative onset time are reasonably close to the labeling boundary for a labial stop voiced–voiceless continuum. However, voicing boundaries change considerably when the onset frequency of the first formant ($F1$) is varied—either directly or as a side effect of a change in $F1$ transition duration. Stimuli consisted of a midfrequency sinusoid that was initiated 0–50 ms prior to the onset of a low-frequency sinusoid. Results showed that the labeling boundary for relative onset time increased for longer durations of a low-frequency tone sweep. This effect is analogous to the $F1$ transition duration effect with synthetic speech. Further, the discrimination of differences in relative onset time was poorer for stimuli with longer frequency sweeps. However, *unlike* synthetic speech, there were no systematic effects when the frequency of a transitionless lower sinusoid was varied. These findings are discussed in relation to the potential contributions of auditory mechanisms and speech-specific processes in the perception of the voicing contrast.

PACS numbers: 43.70.Dn, 43.66.Lj, 43.66.Mk

INTRODUCTION

The voicing distinction in English stops appears, on the surface at least, to be one of the more straightforward contrasts to describe in articulatory, acoustic, and perceptual terms. However, despite the considerable amount of research that has been directed toward the study of this contrast, a number of basic issues remain to be resolved. In production, the stop voicing contrast involves varying the timing of voicing onset relative to the release of the supraglottal articulation—the voice onset time or "VOT" dimension (Lister and Abramson, 1964). The voiced stops /b,d,g/ are produced either with voicing lead (voicing is initiated 50 to 150 ms prior to release) or with a small amount of voicing lag (voicing is initiated 0 to 30 ms following release). The voiceless or "long lag" stops /p,t,k/ are produced with a substantial delay (40 to 120 ms) between articulatory release and voicing onset. In acoustic terms, VOT has been defined as the interval between the burst of noise associated with articulatory release and the appearance of signal periodicity corresponding to the initiation of vocal cord vibration (Lisker and Abramson, 1964).

Abramson and Lisker (1970) synthesized a series of stimuli varying in voice onset time by controlling the interval between a release burst and the initiation of a periodic voice source. The interval between the release burst and voicing onset was filled with aspiration noise that excited only the second and third formants ($F2$ and $F3$). Energy in the region of the first formant ($F1$) did not appear until voicing was initiated, an acoustic feature that has been called "$F1$ cutback" (Liberman *et al.*, 1958; for a more detailed description of these stimuli, see Kuhl and Miller, 1978; Soli, 1983). Abramson and Lisker found that English-speaking subjects changed from hearing voiced stops to hearing voiceless stops when the delay in voicing onset exceeded 20 to 50 ms, depending on place of articulation.

Given the prominent temporal component to the voicing contrast in initial stops, it seems reasonable to ask whether the perception of this distinction might be related to auditory mechanisms responsible for the resolution of temporal order. There is, in fact, some evidence that labeling boundaries for VOT continua are reasonably close to psychophysical boundaries involving judgment of the temporal order of two acoustic events. For example, Hirsh (1959) presented subjects with 500-ms stimuli consisting of two coterminous sinusoids differing in frequency. The relative onset time of the two pure-tone components was varied between -60 ms (low tone leading by 60 ms) and $+60$ ms (low tone lagging by 60 ms). Although not specifically designed to model VOT stimuli, the two-tone patterns between 0 and $+60$ ms are roughly analogous to variations in the $F1$ cutback component of the VOT dimension. Hirsh found that a stimulus onset asynchrony of 15 to 20 ms was required for subjects to judge accurately (75% correct) which of the two tones came first. The minimum onset asynchrony did not change appreciably across a wide range of pure-tone frequency pairs (250–300, 250–1200, 250–4500, 1000–1200, 1000–4500). This 15–20-ms value is only a little smaller than the 25-ms boundary (50% crossover) that Abramson and Lisker (1970) found for a /ba/–/pa/ continuum. However, the value reported by Hirsh is quite different from the 35-ms boundary that Abramson and Lisker reported for a /da/–/ta/ continuum and the 42-ms boundary for a /ga/–/ka/ continuum. A similar study by Pisoni (1977) tested subjects on stimuli consisting of a 500-Hz pure tone representing $F1$ and a 1500-Hz pure tone representing $F2$. Relative onset times were varied between -50 and $+50$ ms. Subjects were asked to label these "tone onset time" or "TOT" stimuli as having one event at onset or two events at onset. Pisoni's results were very similar to those reported by Hirsh: boundaries occurred at about -20 and $+20$ ms (see similar findings by Pastore *et al.*, 1981 and Summerfield, 1982).

 0001-4966/84/010231-10$00.80 © 1984 Acoustical Society of America

The sine-wave analogs discussed to this point have captured only the $F1$ cutback aspect of the VOT dimension. These stimuli did not include an analog of aspiration noise found in long lag stops. Miller *et al.* (1976) generated a stimulus continuum in which the onset of a bandlimited noise, analogous to aspiration, was varied in relation to the onset of a bandlimited 100-Hz pulse train, analogous to $F1$. Miller *et al.* found that a noise lead time of about 15 ms separated "noise lead" and "no noise lead" categories. A value in the 15- to 20-ms range was also reported by Stevens and Klatt (1974) in a task involving the detection of a gap between transient and periodic components of nonspeech stimuli. The stimuli consisted of a 5-ms burst of noise separated from vowel-like formants by silent intervals of varying duration. Stevens and Klatt reported that about 20 ms of silence was required for subjects to detect the presence of a gap between the transient and periodic components of the stimuli.

Results from the nonspeech experiments involving judgments of temporal order by relatively untrained listeners are surprisingly easy to summarize. With a variety of stimulus types, a 15- to 20-ms difference in onset time seems to be required for accurate judgments of temporal order or, in the case of Stevens and Klatt's (1974) task, to detect the presence of a silent gap. Given the approximate correspondence between VOT category boundaries and category boundaries for nonspeech analogs, it is possible to propose a single auditory process that could account for both the speech and nonspeech data. Pisoni (1977), for example, has argued that the locations of category boundaries for VOT, TOT, and noise-buzz stimuli "...reflect a basic limitation on the ability to process temporal-order information" (p. 1360). For all of these stimuli, the psychophysical task can be viewed as judging the onset of one acoustic event relative to the onset of another. Category boundaries in the vicinity of 15 to 20 ms might be related to a difference limen for the judgment of temporal order.

One potential problem with this hypothesis is that VOT boundaries can be changed systematically by introducing variations in stimulus dimensions other than relative onset time. For example, VOT boundaries occur at longer relative onset times when: (1) the duration of the first-formant transition is increased (Stevens and Klatt, 1974; Lisker, 1975; Lister *et al.*, 1977; Summerfield and Haggard, 1977), (2) the frequency of the first formant is lowered (Lisker, 1975; Summerfield and Haggard, 1977), (3) the relative intensity of aspiration noise is lowered (Repp, 1979), and (4) the fundamental frequency at voicing onset is increased (Haggard *et al.*, 1970, 1981; Massaro and Cohen, 1976). If VOT boundaries are, in fact, related to a threshold for temporal order, it should be possible to demonstrate the effects listed above using nonspeech analogs, e.g., either noise-buzz or TOT stimuli. This report will focus on the first two of these effects: the tendency for VOT boundaries to take on larger values as the duration of the first-formant transition is increased and, related to this, the tendency for VOT boundaries to take on larger values as the onset frequency of the first formant is lowered.

The $F1$ transition duration effect was first reported by Stevens and Klatt (1974), who showed that the crossover from voiced to voiceless occurred at longer VOTs as the duration of the $F1$ transition was increased (see similar findings by Lisker *et al.*, 1977; Summerfield and Haggard, 1977). Stevens and Klatt concluded that the voiced–voiceless distinction in initial stops is cued not simply by VOT but also by, "...the presence or absence of a significant and rapid change in the spectrum at the onset of voicing" (p. 654). There are, however, other ways to characterize the difference between two stimuli having the same VOT but different $F1$ transition durations. Figure 1 compares schematic representations of the first 100 ms of two stimuli with identical 20-ms VOTs but different $F1$ transition durations. These stimuli differ in three ways: the stimulus at left has a shorter $F1$ transition, a higher rate of frequency change in $F1$, and a higher $F1$ starting frequency. Experiments by Lisker (1975) suggest that the shift in the VOT boundary is controlled by the frequency of the first formant at voicing onset rather than the extent of frequency change in the first formant. Lisker's findings were replicated in a more extensive study by Summerfield and Haggard (1977). The stimuli for this study were constructed in such a way that $F1$ transition duration, rate of frequency change in $F1$, and the onset frequency of $F1$ were independently controlled. The results of these manipulations showed that, "...the major effect of $F1$ in initial voicing contrasts is determined by its perceived frequency at the onset of voicing... (A) periodically excited $F1$ transition is not, *per se*, a positive cue to voicing" (p. 435).

There are two very different ways to view the influence of $F1$ onset frequency on VOT boundaries. The view taken by Summerfield and Haggard (1977) suggests that relative onset time (the "separation cue" in their terminology) and $F1$ onset frequency are *independent* cues, each of which contributes to the subject's voiced–voiceless decision. These two cues, and perhaps others, are weighted in some fashion and combine in a "trading relation" similar to others that have been discussed in speech perception (Repp, 1982). The influence of $F1$ onset frequency might be acquired as a result of experience in listening to speech, since low $F1$ onset frequencies are generally associated with short-lag VOT values, and relatively high $F1$ onset frequencies are typically associated with long-lag VOT values. An alternate possibility is that relative timing is the only relevant cue to the voiced–voiceless contrast, but that the perception of relative onset time is influenced by other parameters, such as $F1$ onset frequency. In other words, it is possible that, for as yet undetermined reasons, judgments of simultaneity versus successivity are poorer for stimuli with low-frequency first formants. Ac-

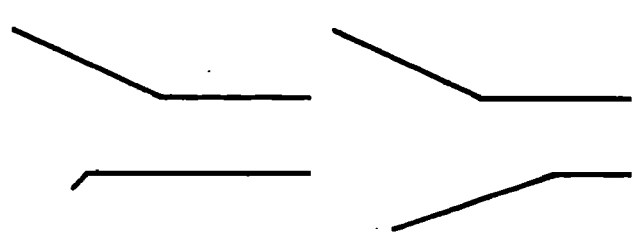

FIG. 1. Schematic spectrograms of the first 100 ms of two stimuli with identical 20-ms voice-onset times but different first-formant transition durations. Note that the stimuli differ not only in $F1$ transition duration but also in $F1$ onset frequency and in the rate of frequency change in $F1$.

cording to this hypothesis, the sole cue to voicing is the "separation cue" and the influence of $F1$ onset frequency is best described as a parameter dependency rather than a trading relation between two separate cues.

The "parameter dependency" argument essentially suggests that the $F1$ onset frequency effect has a relatively straightforward auditory basis. If this is true, it should be possible to produce an analogous effect with nonspeech sounds. On the other hand, if the effect is the result of a speech-specific trading relation between VOT and $F1$ onset frequency, there is no reason to expect an analogous effect with nonspeech stimuli. A recent study by Summerfield (1982) tested the effects of frequency variations on category boundaries for a synthetic VOT continuum and for two nonspeech analogs of VOT. One of the nonspeech analogs was a TOT continuum similar to the one used by Pisoni (1977) and the other was a noise-buzz or "noise onset time" (NOT) continuum similar to the one used by Miller *et al.* (1976). The parameter that was varied was the frequency of the first formant in the case of the VOT stimuli, the frequency of the lower sinusoid in the case of the TOT stimuli or the frequency of a bandlimited pulse train in the case of the NOT stimuli. The first formants, and $F1$ analogs, were synthesized without transitions and were set to either 200, 300, 400, or 500 Hz. Summerfield found that VOT boundaries shifted to longer relative onset times as $F1$ was lowered, but there were no systematic frequency effects for the TOT or NOT stimuli. Summerfield suggested that, "...the role of $F1$ in the perception of voicing does not have a purely auditory basis... (other) factors, such as production constraints or arbitrary processes of cultural development, appear to be required to account for the positions of voicing boundaries..." (p. 51).

In the present study four experiments were run that examined temporal and spectral effects on the perception of nonspeech stimuli whose components vary in relative onset time. Two experiments tested for an analog of the $F1$ transition duration effect reported by Stevens and Klatt (1974), Summerfield and Haggard (1974), and Lisker *et al.* (1977). The stimuli were similar to the TOT series used by Pisoni (1977), except that frequency sweeps were introduced to represent formant transitions. Experiment 1 examined subjects' labeling of stimuli varying in relative onset time as a function of the duration of the frequency sweep in the sinusoid representing $F1$. Experiment 2 tested subjects on the same kinds of stimuli but used an ABX discrimination procedure. Experiments 3 and 4 were designed, as was Summerfield's (1982) study, to examine an analog of the $F1$ onset frequency effect reported by Lisker (1975) and Summerfield and Haggard (1977). The stimuli were similar to those used in experiments 1 and 2 except that there was no frequency movement in the lower tone. The parameter was the frequency of the lower tone. Experiment 3 tested subjects' labeling of the stimuli and experiment 4 tested discrimination of stimuli varying in relative onset time.

I. EXPERIMENT 1

A. Stimuli

The stimuli consisted of a midfrequency sinusoid representing $F2$ and a low-frequency sinuosid representing $F1$. To simplify stimulus descriptions, the lower-case designations "$f1$" and "$f2$" will be used when referring to sinusoidal analogs of the first and second formants. The onset time of $f1$ relative to $f2$ was varied between 0 and + 50 ms in 10-ms steps. The mid- and low-frequency components were terminated simultaneously at 600 ms. Figure 2 shows the 0- and 50-ms endpoints of one of the three stimulus continua that was synthesized. The primary difference between these stimuli and those used by Hirsh (1959) and Pisoni (1977) was the introduction of frequency sweeps to represent formant transitions. The frequency movements here were modeled after formant transitions for an alveolar stop followed by an /a/ vowel. The midfrequency sinusoid started at 1700 Hz then fell linearly over the next 50 ms to a steady-state value of 1240 Hz. The low-frequency component started at 200 Hz and rose linearly to a steady-state value of 720 Hz. Since the purpose of this experiment was to model the $F1$ transition duration effect, three TOT continua were synthesized that differed in the duration of the $f1$ frequency sweep. The $f1$ transition was either 25, 50, or 75 ms in duration. All of the stimuli were synthesized using a computer program that allows control of the frequency, amplitude, and phase of sinusoidal waveforms (Hillenbrand, 1981). The signals were calculated at a 10-kHz sample rate with 14 bits of amplitude resolution. The intensity of $f2$ was − 3 dB relative to $f1$, approximating formant levels for an /a/ vowel. The delay in the onset of $f1$ was achieved simply by setting the amplitude of the low-frequency component at zero for 0, 10, 20, 30, 40, or 50 ms. This method is analogous to the one used to vary "$F1$ cutback" in synthetic speech. Use of this method means that stimuli that differ in TOT also differ in $f1$ onset frequency. The only exceptions will be comparisons in which neither of the stimuli show frequency movement in the lower tone (e.g., in the 25-ms transition series, the 30-, 40-, and 50-ms stimuli will differ only in relative onset time). The two sinusoidal components were generated independently and were gated on and off with 15-ms linear rise and fall times. The

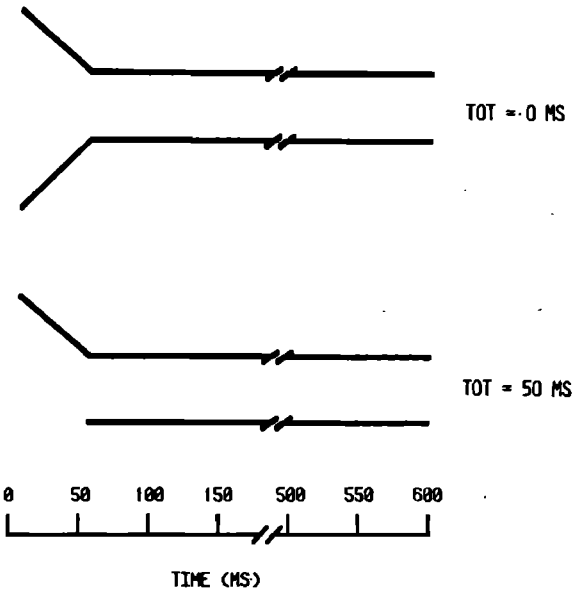

FIG. 2. Schematic representations of the 0- and 50-ms endpoints of a stimulus continuum varying on tone-onset time (TOT).

mid- and low-frequency components were then summed to form a single stimulus. Each stimulus was processed by an intensity modification program which ensured that all stimuli were equal in overall rms intensity (Prall, 1981).

B. Subjects and procedures

Subjects were 30 Northwestern University students with no reported history of hearing or speech problems. Ten subjects were run in each of the three transition duration conditions. Subjects were told that they would hear "simple musical sounds" consisting of a high-pitch tone and a low-pitch tone. They were asked to judge whether the high- and low-pitch tones had simultaneous onsets or whether the high-pitch tone started first. Subjects were given training with feedback on the 0- and 50-ms endpoints in blocks of 120 trials. Training continued until performance was 90% or better on the endpoints. Twenty-four subjects met the training criterion in one block of 120 trials and the remaining six subjects required one additional block. The identification task consisted of 40 presentations of each of the six TOT stimuli in pseudorandom order. No feedback was provided. At the end of the listening session subjects were asked to provide brief descriptions of the stimuli. Presentation of stimuli and the collection and analysis of responses were under the control of a laboratory computer equipped with a high-speed disk drive and a 14-bit D/A converter. At the output of the D/A converter, the signals were low-pass filtered at 3 kHz, amplified, attenuated, and delivered diotically over matched TDH-49 headphones. Signals were adjusted to peak at 75 dBA using a 6-cc coupler and sound-level meter.

C. Results and discussion

Figure 3 shows the results of the identification tests. The graph shows percent identification as "simultaneous" as a function of tone onset time. The parameter is the duration of the $f1$ transition. Each function represents the pooled results from ten subjects. It can be seen that the category boundary for tone onset time increases for longer durations of the low-frequency tone sweep. The category boundaries, which were calculated by linear interpolation of the 50% crossover, are 18, 24, and 29 ms for the 25-, 50-, and 75-ms conditions, respectively. A one-way analysis of variance showed that the shift in the category boundary was significant [$F(2,27) = 13.7$, $p < 0.01$]. The general pattern here is very similar to the one reported for synthetic speech by Stevens and Klatt (1974), Lisker *et al.* (1977), and Summerfield and Haggard (1977). It should be noted, however, that the category boundaries for these nonspeech analogs are by no means identical to those for synthetic speech sounds with similar transition durations. For example, Lisker *et al.* (1977) reported category boundaries of 21, 32, and 40 ms for a synthetic /da/–/ta/ series with transition durations of 25, 55, and 70 ms, respectively. There are a number of factors that might account for the differences in boundary locations, such as differences in the range of relative onset time values and a wide variety of stimulus differences. However, this fits in with a general pattern in which category boundaries for nonspeech analogs of VOT occur consistently at shorter relative onset times than the synthetic speech continua on which they were modeled (e.g., Miller *et al.*, 1976; Pisoni, 1977; Summerfield, 1982).

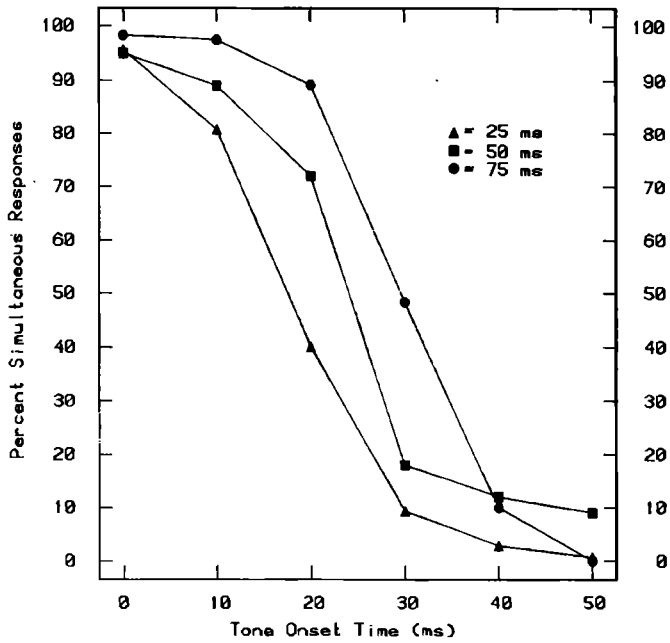

FIG. 3. Identification results for stimulus continua with $f1$ transition durations of 25 ms (triangles), 50 ms (squares), and 75 ms (circles). Each function represents the pooled results from ten subjects. From left to right, the category boundaries are 18, 24, and 29 ms.

With one exception, descriptions of the stimuli provided by subjects at the end of the session gave no indication that they heard the signals as speechlike. Subjects described the stimuli as sounding like "telephone tones," "electronic game sounds," "synthesizer music," "beeps and boops," "electric organ pitches," and other nonspeech sounds. One subject, however, described the stimuli as "sometimes like a note and sometimes like part of a syllable like 'd.'"

II. EXPERIMENT 2

It would be possible to explain the findings of experiment 1, and the $F1$ transition duration effect produced with synthetic speech, by postulating that, for as yet undetermined reasons, the ability of the auditory system to resolve temporal order is poorer for stimuli with longer $F1$ transitions. To gain more specific information about this hypothesis, a separate group of subjects was tested on the discrimination of small differences in relative onset time for TOT stimuli varying in $F1$ transition duration. An ABX discrimination procedure was used to determine the minimum stimulus onset asynchrony necessary for subjects to detect the difference between simultaneous and nonsimultaneous onsets.

A. Stimuli

Three new TOT continua were synthesized that were virtually identical to the 25-, 50-, and 75-ms transition duration series used in experiment 1. The only difference was that between tone onset times of 0 and 20 ms the step size was 2 ms instead of 10 ms. All other details of stimulus generation were identical to those described for experiment 1.

B. Subjects and procedures

Subjects were three adults with no reported history of hearing or speech problems. Two of the subjects were naive to the purpose of the experiment; the author served as the third subject. Subjects were tested with an ABX discrimination procedure. In all cases the 0-ms stimulus was compared with one of the other stimuli in which the low tone was delayed relative to the high tone. The experiment was run in seven 50-min sessions. A single training session was followed by two sessions at each of the three transition durations. The training sessions compared the 0- and 50-ms endpoints. Four kinds of trials could occur: 0-50-0, 0-50-50, 50-0-50, 50-0-0. The interstimulus interval was 0.5 s and the intertrial interval was 1.5 s. Subjects were told that the first two stimuli would always be different and were asked to judge whether the third stimulus was identical to the first stimulus or the second stimulus. Feedback was provided on all trials. Subjects were given 120 training trials at each transition duration. Each subject was given a different order of presentation and each performed at 90% or better on the three transition conditions. Discrimination testing began on the second session. The ABX trials were identical in format to those presented during training except that the tone onset time of the asynchronous stimulus was varied from a maximum of 50 ms to a minimum of 2 ms. In all cases, stimuli with asynchronous onsets were compared with the 0-ms stimulus. Trials were presented in blocks of 40 with all trials in a block comparing the same pair of stimuli. Feedback was provided on every trial. The first three blocks of each session were 50, 20, and 14 ms. Thereafter, the tone onset time of the asynchronous stimulus of each successive block was decreased by 2 ms. All trials in a given session were from the same transition duration condition. The subjects were given two sessions of testing (720 trials) at each $f1$ transition duration, and each subject received a different ordering of the conditions.

C. Results and discussion

Results from each subject, and combined data from the group, are shown in Fig. 4. Each graph shows percent correct discrimination as a function of the tone onset time of the asynchronous stimulus; the parameter is the duration of the low-frequency tone sweep. Results are shown for onset asynchronies between 2 and 14 ms. Each data point in the functions for individual subjects is based on 80 responses. It can be seen in the group data, and in the data from each individual subject, that discrimination performance was poorer for stimuli with longer frequency sweeps. For all three subjects, performance was best for the stimuli with 25-ms sweeps and worst for the stimuli with 75-ms sweeps.

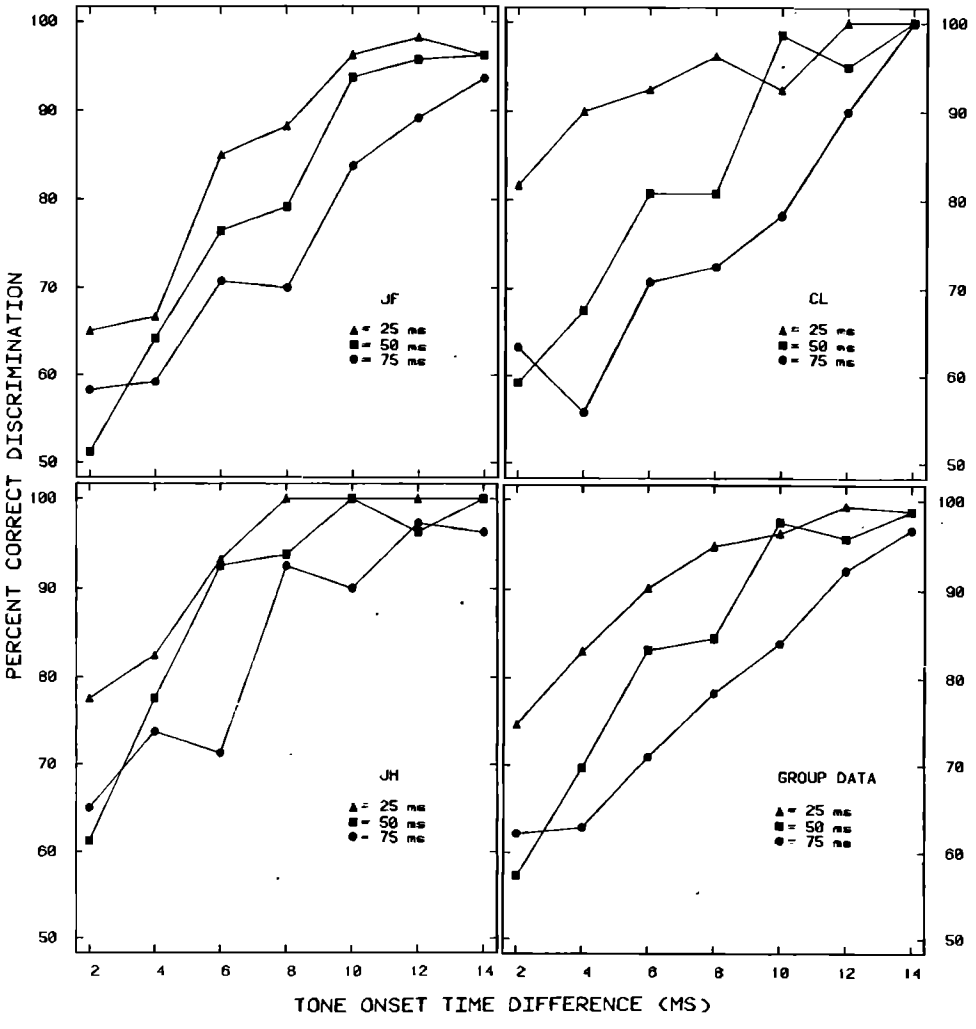

FIG. 4. Percent correct ABX discrimination as a function of the difference in tone-onset time between the A and B stimuli. The $f1$ transition duration is 25 ms (triangles), 50 ms (squares), or 75 ms. The graph in the lower right represents pooled results from the three individual subjects. (Subject JH is the author.)

On the basis of these results, it appears as though temporal order resolution is better for stimuli with abrupt transitions. However, it is possible that what appears to be a temporal effect is actually due to the increased availability of spectral information for stimuli with abrupt transitions. This possibility will be explored in the General Discussion (Sec. V).

III. EXPERIMENT 3

The discrimination findings from experiment 2 are consistent with the idea that the threshold for temporal order is influenced by $F1$ transition duration. Recall, however, that Lisker (1975) and Summerfield and Haggard (1977) have shown that the basis of the $F1$ transition duration effect is spectral rather than temporal. In other words, stimuli with long $F1$ transitions cross over at longer VOTs because the starting frequency of the first formant is relatively low. The purpose of experiment 3 was to determine if an analogous effect could be produced with TOT stimuli that did not have an $f1$ transition. Stimuli were synthesized using the same procedures described for experiment 1. As in the previous experiments, the midfrequency sinusoid started at 1700 Hz and fell linearly over the first 50 ms to a steady-state frequency of 1240 Hz. There was no frequency movement in the lower tone. Three TOT continua were generated; the frequency of the lower tone was set at either 250, 450, or 750 Hz. Tone onset times were varied between 0 and 50 ms in 10-ms steps. Apparatus, calibration methods, and procedures for the labeling task were identical to those described for experiment 1. Ten normal hearing adults were tested at each frequency condition.

A. Results and discussion

Results of the labeling tasks are shown in Fig. 5. Each function represents pooled data from ten subjects. Based on the synthetic speech findings, the expectation was that the 750-Hz series would cross over first and the 250-Hz series would cross over last. This clearly was not the case. The boundary values were 26 ms at 250 Hz, 18 ms at 450 Hz, and 22 ms at 750 Hz. Analysis of variance revealed a significant effect for the frequency of the lower tone [$F(2,27) = 4.6$, $p < 0.05$]. Newman–Keuls *post-hoc* analyses showed that only the 450- and 750-Hz conditions were significantly different. As in experiment 1, descriptions of the stimuli provided by subjects gave no indication that they heard the sounds as speechlike. These labeling results are in general agreement with those reported by Summerfield (1982) and Hirsh (1959) in the sense that frequency effects on TOT stimuli seem to be different from frequency effects on VOT stimuli. The frequency values used in the present study are sufficiently different from those used by Summerfield and Hirsh that a straightforward comparison of boundary values is probably not wise. However, the general conclusion is the same: unlike synthetic speech, there is no evidence that labeling boundaries for TOT stimuli take on larger values as the frequency of the lower component is decreased.

As a further check on these findings, a separate group of five subjects was run on the same sort of labeling task. The only difference was that, unlike the identification task described above, each subject was tested on all three frequency conditions. Subjects were tested on three separate days and each subject received a different ordering of the conditions. As before, the identification task consisted of 120 trials of training on the 0- and 50-ms endpoints followed by 240 trials of identification testing (40 trials at each tone onset time). The general pattern of results was similar to that described above. Category boundaries were 26 ms at 250 Hz, 21 ms at 450 Hz, and 25 ms at 750 Hz. Again, there was no evidence for an inverse relation between the TOT category boundary and the frequency of the lower tone.

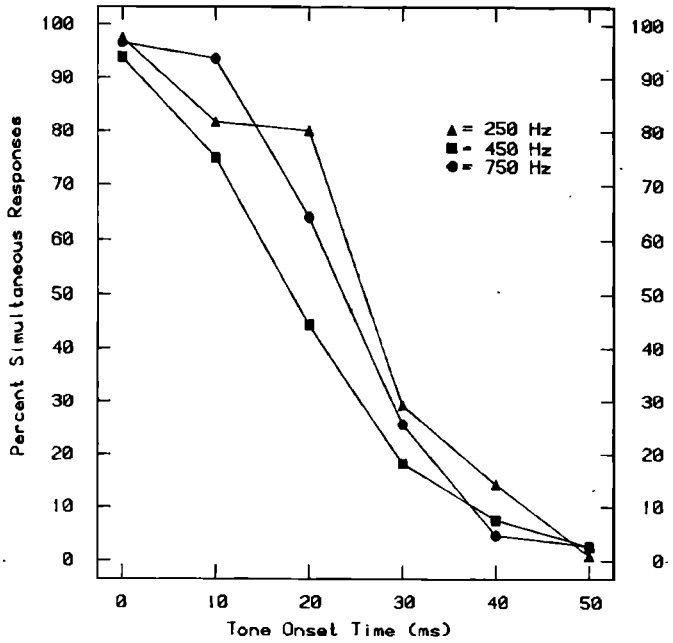

FIG. 5. Identification results for stimulus continua with $f1$ set to 250 Hz (triangles), 450 Hz (squares), or 750 Hz (circles). Each function represents the pooled results from ten subjects. The category boundaries are 26, 18, and 22 ms for the 250-, 450-, and 750-Hz conditions, respectively.

IV. EXPERIMENT 4

The labeling task in experiment 3 did not show the expected relationship between $f1$ frequency and TOT category boundaries. Identification tasks, however, can be subject to considerable amounts of response bias. Since the ABX discrimination procedure is much less sensitive to response bias, three subjects were tested on their ability to discriminate TOT stimuli with $f1$ frequencies of 250, 450, and 750 Hz. Two of the subjects, CL and JH (the author), had participated in the previous ABX task. A new set of stimuli was generated that was identical to that described for experiment 3 except that between 0 and 20 ms the step size was 2 ms instead of 10 ms. All other details concerning procedures and apparatus were identical to those described for the ABX technique used in experiment 2.

A. Results and discussion

Discrimination results from each subject, and pooled results for the group, are shown in Fig. 6. According to the synthetic speech data, discrimination results ought to be poorest for the 250-Hz condition and best for the 750-Hz

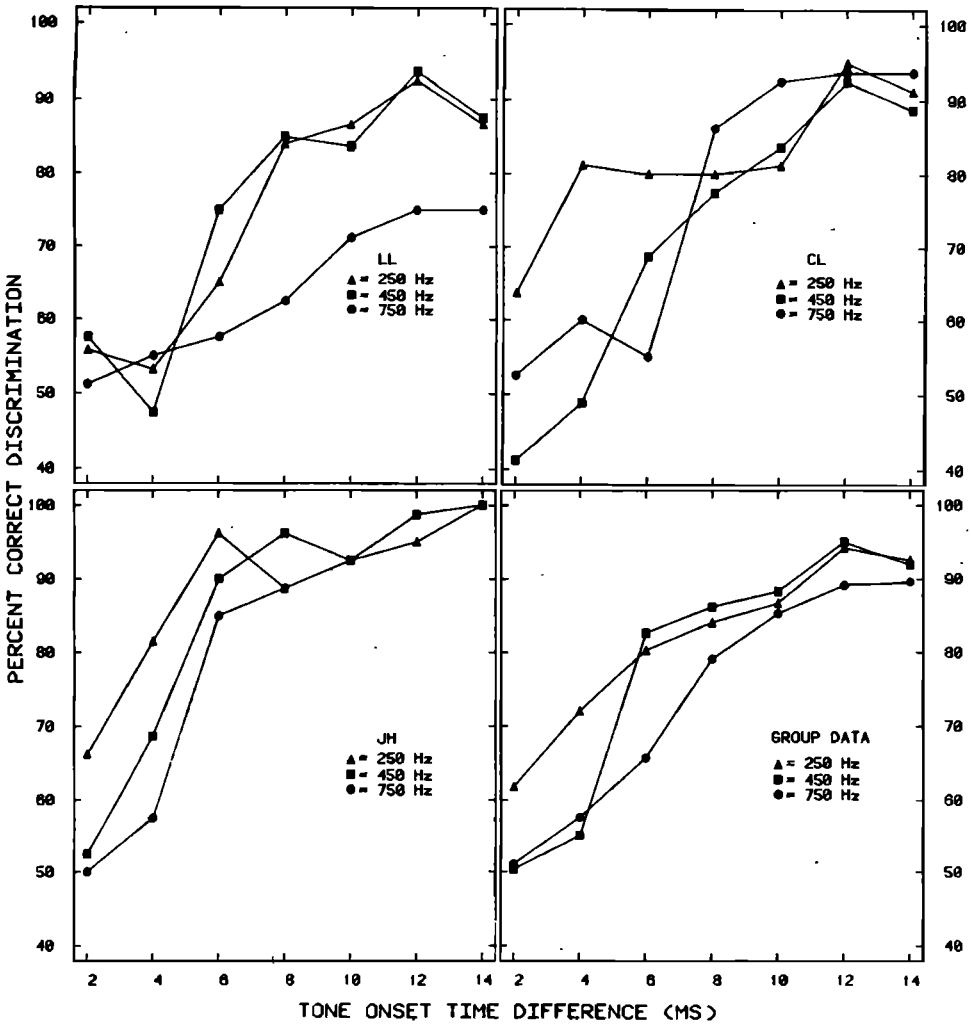

FIG. 6. Percent correct ABX discrimination as a function of the difference in tone onset time between the A and B stimuli. The frequency of a transitionless $f1$ is either 250 Hz (triangles), 450 Hz (squares), or 750 Hz (circles). The graph in the lower right represents the pooled results from the three individual subjects. (Subject JH is the author.)

condition. There is no evidence for this pattern in the discrimination results. In fact, the pattern shown in the group data is exactly opposite: performance was best at 250 Hz and worst at 750 Hz. However, since each subject showed a different pattern of results, it is unclear how meaningful the group data are. The main conclusion to be drawn from the discrimination results is the failure to find any support for the idea that the discrimination of simultaneity–successivity contrasts becomes poorer as the frequency of the low component is decreased.

V. GENERAL DISCUSSION

The results of these experiments are puzzling from several points of view. Experiments 1 and 2 appear to support an auditory account of the $F1$ transition duration effect on VOT boundaries while experiments 3 and 4 seem to argue against an auditory explanation of the $F1$ onset frequency effect. Experiment 1 showed that, as for VOT stimuli, category boundaries for TOT stimuli tend to occur at longer relative onset times as the duration of the frequency sweep in the lower component is increased. The discrimination results reported in experiment 2 appear to support the idea that the simultaneity–successivity threshold is poorer for stimuli with longer transitions in the low tone. These results would tend to support an auditory account of the $F1$ transition duration effect for VOT stimuli. In addition, these findings suggest that the influence of $F1$ transition duration is best described as a parameter dependency rather than a trade-off between two separate cues. In other words, experiments 1 and 2 make it appear as though relative onset time is the primary cue to VOT distinctions but that, for undetermined reasons, the resolution of temporal order is poorer for stimuli with longer frequency sweeps in the low-frequency component.

A different picture emerges from experiments 3 and 4 in which the frequency of a transitionless sinusoid representing $F1$ was varied. Recall that these experiments were needed because, for VOT stimuli, Lisker (1975) and Summerfield and Haggard (1977) demonstrated that the $F1$ transition effect essentially operates by varying the frequency of $F1$ at voicing onset. If the $F1$ onset frequency effect has a straightforward auditory explanation, there should have been an inverse relation between TOT boundaries and the frequency of the lower sinusoid. There was no evidence for this relationship either in the identification data or the ABX discrimination results, confirming previous findings by Hirsh (1959) and Summerfield (1982).[1] These experiments suggest that the influence of $F1$ onset frequency on VOT boundaries is best described as a trading relation between two separate cues. The failure to find an analog of the $F1$ onset frequency effect with nonspeech sounds would tend to support a phonetic account of this effect. As was discussed previously, it is

possible that a low $F1$ onset frequency biases subjects toward hearing a voiced stop because of a learned association between low $F1$ onset frequencies and short lag VOT values. A learned effect of this type would obviously be speech-specific; there would be no reason to predict an analogous effect for nonspeech stimuli.

Simon and Fourcin (1978) have presented cross-language developmental evidence that seems to support the idea that the role of $F1$ in voicing contrasts is acquired. English- and French-speaking children were presented with synthetic stimuli varying in VOT. Some of the stimuli were synthesized with $F1$ transitions and some were synthesized with flat first formants. The results led Simon and Fourcin to conclude that, "English children learn to make use of the $F1$ transition feature around five years, whereas French children never use it as a voicing cue" (p. 925). On the assumption that the role of $F1$ in voicing contrasts is acquired, these findings would be quite sensible. Since French makes a distinction between prevoiced stops and voiceless unaspirated stops, a low $F1$ onset frequency would be associated with both voicing categories. If sensitivity to the contour of $F1$ is acquired, there should be evidence of development in the English but not the French children. There are, however, several reasons for caution regarding the conclusions drawn by Simon and Fourcin. First, close inspection of their data for the English children shows that a transitionless $F1$ reduced the probability of voiced responses at all ages except two and perhaps three years. At these ages, the children were quite inconsistent in their responses to all of the stimuli. Second, the two-year-old English children were tested on a labial contrast ("ball"–"Paul") while the older children were tested on a velar contrast ("goat"–"coat"). Last, and perhaps most serious, the French children were tested on yet a third contrast: an alveolar place of articulation with a voicing opposition in two stops ("dodo"–"Toto"). The VOT range for this contrast was -30 to $+30$ instead of the -10 to $+60$ (/b/–/p/) and 0 to $+70$ (/g/–/k/) VOT ranges used with the English children. It is obviously very difficult to separate perceptual-learning effects from those that might be due to differences in the stimuli or to differences in the attentional abilities of the children at different ages.

These interpretive difficulties aside, it might be instructive to assume for the moment that the $F1$ onset frequency effect is, in fact, a speech-specific acquired phenomenon. On this assumption, a problem that remains is to reconcile the failure to find a nonspeech analog of the $F1$ onset frequency effect with the data from experiments 1 and 2 showing rather strong evidence for an analog of the $F1$ transition duration effect. One possible explanation of the discrimination data is based on the fact that differences in $f1$ onset frequency accompany differences in TOT only for the stimuli with transitions in the lower sinusoid.[2] For stimuli with $f1$ transitions, tokens with longer TOTs will also have higher $f1$ onset frequencies. Further, for any difference in TOT, the difference in $f1$ onset frequency will be larger for stimuli with short transitions. For example, comparing 0- vs 10-ms TOT stimuli would show a 208-Hz difference in $f1$ onset frequency in the 25-ms series but only a 69-Hz difference in the 75-ms series. Assuming that subjects use both spectral and temporal information in the discrimination task, pairs of stimuli with a given difference in TOT ought to be more discriminable at shorter $f1$ transitions. Findings from a recent study of voice onset time discrimination by Soli (1983) are consistent with this interpretation. Soli found that the discrimination of voice onset time differences was most accurate in the region of the Abramson and Lisker (1970) VOT continuum in which relative onset time and $F1$ onset frequency covary.[3]

Although this explanation seems to apply rather well to the discrimination data, it is still not clear how the labeling data should be interpreted. If the onset frequency of the lower component does not independently affect temporal order resolution, then why do both TOT and VOT boundaries increase with longer transition durations? And why do VOT boundaries take on larger values as the frequency of a transitionless $F1$ is lowered, while an analogous effect is not seen for TOT stimuli? It might be possible to explain these findings by first assuming that the effect of $F1$ onset frequency on VOT boundaries is the result of a learned association between short-lag VOTs and low $F1$ onset frequencies. Low $F1$ onset frequencies might bias listeners toward hearing voiced stops, causing them to require longer VOTs before shifting to voiceless responses. The analogous effect on TOT boundaries might have occurred because subjects heard the TOT stimuli used in experiment 1 as speechlike.

Since an analog of the $F1$ onset frequency effect was not found in experiment 3, this explanation would require that the stimuli with frequency sweeps in the lower tone sound more speechlike than the stimuli with transitionless $F1$ analogs. Descriptions of the stimuli provided by the subjects gave no indication that they heard either set of stimuli as speechlike. However, this does not rule out the possibility that the stimuli with frequency sweeps in the low tone produced a "phonetic processing" effect that was below the level of conscious awareness. A paired-comparison listening task provided a preliminary test of this possibility. Fifteen subjects were presented with pairs of stimuli consisting of one token from experiment 1 (with $f1$ transitions) and one token from experiment 3 (without $f1$ transitions). The stimuli in each pair were matched for tone onset time but were combined in all other ways. Subjects were asked to judge which of the stimuli sounded more like a speech sound. Subjects chose the stimulus with frequency movement in the lower sinusoid on 68% of the trials. It should be noted, however, that nearly all of the subjects remarked that they felt that their judgments were arbitrary and that none of the stimuli sounded particularly speechlike. Further, although this preference for stimuli with $f1$ transitions is statistically significant, it is not as large as one might expect given the very substantial differences in the way subjects behaved on the labeling tasks with these two sets of stimuli. Additional experiments are underway to study this problem in greater detail.

One final point should be made regarding the general strategy and limitations of experiments comparing the perception of a speech contrast with the perception of nonspeech analogs of that contrast. When experiments of this type find that speech and nonspeech stimuli behave in different ways, there is a temptation to implicate a strong role for

phonetic knowledge or other speech-specific processes. Conversely, there is a temptation to conclude that the phenomenon under study has no straightforward auditory basis. However, a frequently overlooked alternative is that some critical attribute of the speech contrast was not faithfully modeled in the nonspeech analogs. The problem, of course, is that improving the physical match between the speech stimulus and its analog often results in a stimulus that is easily heard as speech. This obviously creates the possibility of engaging perceptual and cognitive mechanisms that are ordinarily reserved for speech recognition. Given this inherent limitation of experiments with nonspeech analogs, studies using nonhuman listeners become especially interesting. For example, Kuhl and Miller (1978) used a shock-avoidance procedure to train chinchillas to respond differently to the 0- and 80-ms endpoints of a synthetic /da/–/ta/ continuum. On subsequent generalization trials, the animals were exposed to stimuli with intervening VOTs and to stimuli at all three places of articulation. Category boundaries were very similar to those produced by adult English-speaking subjects. Of particular interest to the present discussion, category boundaries for the chinchillas shifted to longer VOTs as place of articulation changed from labial to alveolar to velar. This place-dependent shift in VOT boundaries for the Abramson and Lisker (1970) stimuli is quite consistent in human listeners and seems to be related, in part at least, to differences in the contour of the first formant (for further discussion of place effects on VOT boundaries, see Kuhl and Miller, 1978; Summerfield, 1982; Miller, 1977; Massaro and Oden, 1980; Soli, 1983). If this is the case, the Kuhl and Miller results with chinchillas would seem to provide evidence in support of an auditory explanation for the role of $F1$ in the perception of voicing contrasts. It should be noted, however, that the labial, alveolar, and velar stimuli generated by Abramson and Lisker differ in other dimensions besides the contour of the first formant and, further, that the place-related changes in VOT boundaries are by no means perfectly understood for these stimuli. Data from nonhuman listeners for stimuli varying in $F1$ onset frequency, and for nonspeech analogs of the type used in the present study, might help to clarify the relative contributions of auditory and phonetic processes in the perception of voicing contrasts.

ACKNOWLEDGMENTS

This work was supported by NIH Biomedical Sciences Support Grant 5 S05 RR07028. Gratitude is expressed to Chris Prall for his technical assistance and to Bruce Smith, Richard Pastore, and Chris Darwin for comments on earlier drafts.

[1]Divenyi and Sachs (1978) reported frequency effects in a gap discrimination study that appear to conflict with the present findings and with those of Hirsh (1959) and Summerfield (1982). Divenyi and Sachs found that gap discrimination became poorer as the frequency separation between two masker tones increased from 0 to 2 octaves. The effect was seen primarily when the gap of the reference stimulus was relatively short (10–20 ms). It should be noted, however, that the Divenyi and Sachs study examined the discrimination of unfilled rather than filled intervals. Further, Divenyi and Sachs asked listeners to discriminate gaps of different durations; the present study and those of Hirsh and Summerfield involved simultaneity–successivity judgments.

[2]I am grateful to Quentin Summerfield for suggesting this interpretation to me.

[3]This hypothesis makes one very clear prediction that, unfortunately, is very difficult to evaluate in the discrimination data from the present study. If the discrimination advantage for stimuli with shorter transitions is due to the increased availability of spectral information, then overall performance should have been better for the stimuli with transitions (experiment 2) as compared to the transitionless stimuli (experiment 4). Comparison of the group data from experiments 2 and 4 (Figs. 4 and 6) finds no support for this prediction. However, two of the subjects in experiment 4 had participated previously in experiment 2. It is possible that practice effects obscured differences that might have been due to the presence versus absence of transitions.

Abramson, A. S., and Lisker, L. (1970). "Voice-timing perception in Spanish word-initial stops," J. Phonet. 1, 1–8.

Divenyi, P. L., and Sachs, R. M. (1978). "Discrimination of time intervals bounded by tone bursts," Percept. Psychophys. 24, 429–436.

Haggard, M. P., Ambler, S., and Callow, M. (1970). "Pitch as a voicing cue," J. Acoust. Soc. Am. 31, 613–617.

Haggard, M. P., Summerfield, A. Q., and Roberts, M. (1981). "Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading $F0$ cues in the voiced–voiceless distinction," J. Phonet. 9, 49–62.

Hillenbrand, J. (1981). "SINESYN: A FORTRAN program for the synthesis of sinusoids varying in frequency, amplitude, and phase," unpublished Tech. Rep., Northwestern University, Evanston, IL.

Hirsh, I. J. (1959). "Auditory perception of temporal order," J. Acoust. Soc. Am. 31, 759–767.

Kuhl, P. K., and Miller, J. D. (1978). "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli," J. Acoust. Soc. Am. 63, 905–917.

Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). "Some cues for the distinction between voiced and unvoiced stops in initial position," Lang. Speech 1, 153–167.

Lisker, L. (1975). "Is it VOT or a first-formant transition detector?," J. Acoust. Soc. Am. 57, 1547–1551.

Lisker, L., and Abramson, A. S. (1964). "A cross-language study of voicing in initial stops: Acoustical measurements," Word 20, 384–422.

Lisker, L., Liberman, A. M., Erickson, D. M., and Dechovitz, D. (1977). "On pushing the voice-onset-time (VOT) boundary about," Lang. Speech 20, 209–216.

Massaro, D. W., and Cohen, N. M. (1976). "The contribution of fundamental frequency and voice onset time to the /zi–si/ distinction," J. Acoust. Soc. Am. 60, 704–717.

Massaro, D. W., and Oden, G. C. (1980). "Evaluation and integration of acoustic features in speech," J. Acoust. Soc. Am. 67, 996–1013.

Miller, J. L. (1977). "Nonindependence of feature processing in initial consonants," J. Speech Hear. Res. 20, 519–528.

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: An example of categorical perception," J. Acoust. Soc. Am. 60, 410–417.

Pastore, R. E., Harris, L. B., and Kaplan, J. K. (1981). "Temporal order identification: Some parameter dependencies," J. Acoust. Soc. Am. 71, 430–436.

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. Am. 61, 1352–1361.

Prall, C. W. (1981). "MODAUD: A computer program for measurement and modification of digitized audio signals," unpublished Tech. Rep., Northwestern University, Evanston, IL.

Repp, B. H. (1979). "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," Lang. Speech 22, 173–189.

Repp, B. H. (1982). "Phonetic trading relations and context effects: New evidence for a phonetic mode of perception," Psychol. Bull. 92, 81–110.

Simon, C., and Fourcin, A. J. (1978). "Cross-language study of speech pattern learning," J. Acoust. Soc. Am. 63, 925–935.

Soli, S. (1983). "The role of spectral cues in discrimination of voice onset time differences," J. Acoust. Soc. Am. 73, 2150–2165.

Stevens, K. N., and Klatt, D. H. (1974). "Role of formant transitions in the voiced–voiceless distinction for stops," J. Acoust. Soc. Am. 55, 653–659.

Summerfield, Q., and Haggard, M. P. (1977). "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," J. Acoust. Soc. Am. 62, 435–448.

Summerfield, Q. (1982). "Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops," J. Acoust. Soc. Am. 72, 51–61.