On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants

Quentin Summerfield[a] and Mark Haggard[b]

Department of Psychology, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland, United Kingdom

(Received 10 February 1977; revised 19 April 1977)

It has been claimed that a rising first-formant (F1) transition is an important cue to the voiced–voiceless distinction for syllable-initial, prestressed stop consonants in English. Lisker [J. Acoust. Soc. Am. 57, 1547–1551 (L) (1975)] has pointed out that the acoustic manipulations suggesting a role for F1 have involved covariation of the onset frequency of F1 with the duration, and hence the frequency extent, of the F1 transition; he has argued that effects hitherto ascribed to the transition are more properly attributed to its onset. Two experiments are reported in which F1 onset frequency and F1 transition duration/extent were manipulated independently. The results confirm Lisker's suggestion that the major effect of F1 in initial voicing contrasts is determined by its perceived frequency at the onset of voicing and show that a periodically excited F1 transition is not, *per se*, a positive cue to voicing. In further experiments, the relative levels and the frequencies at the onset of voicing of both F1 and F2 were manipulated. The influences on the perception of stop-consonant voicing that resulted were determined specifically by the frequency of F1 and not by its absolute or relative level or by the overall distribution of energy in the spectrum. The results demonstrate a complementary relationship between perceptual cue sensitivity and production constraints: In production, the VOT characterizing a particular stop consonant varies inversely with the degree of vocal-tract constriction, and hence with the frequency of F1, required by the phoneme following the stop; in perception, the lower the frequency of F1 at the onset of voicing, the longer the VOT that is required to cue voicelessness. In this way, the inclusion of F1 onset frequency in the cue repertoire for voicing, and the establishment of the cue trading relationship, reduce the problem of contextual variation that would be met were VOT alone, or some other amalgam of cues, the only basis of the voicing distinction.

PACS numbers: 43.70.Dn

INTRODUCTION

Lisker and Abramson (1964) suggested that the articulatory basis for the voiced–voiceless distinction for stop consonants resides in the relative timing of laryngeal and supralaryngeal articulations. Prestressed, syllable-initial voiced stops in English display temporal coincidence of oral release with the onset of laryngeal vibration. When the onset of vocal cord vibration follows oral release by more than about 40 ms, the stop is voiceless. By translating variation on this articulatory dimension into variation of the parametric input to an acoustic speech synthesizer, Lisker and Abramson (1967) generated VOT continua which spanned the two perceptual categories of voicing for each of the three places of stop production used in English. Phoneme boundaries on these continua occurred close to those values of VOT which optimally segregate voiced from voiceless stops in the productions of speakers of English. Since then, VOT continua have been used extensively as experimental devices. They permit the determination of a phoneme boundary, changes in whose position can be used as sensitive indices of the perceptual consequences of variation of parameters both intrinsic (e.g., Stevens and Klatt, 1974) and extrinsic (e.g., Eimas and Corbit, 1973; Summerfield, 1975a) to the test syllables themselves. However, it has not always been clear which aspects of the stimulus are held to be perceptual cues, given that many of the acoustical parameters so far asserted to possess cue value have tended to covary. Incorporating covariation in a set of stimuli is generally well justified from an articulatory point of view if the objectives of the experiment are linguistic or cognitive. But, if the objectives are psychoacoustical or perceptual, then the use of covarying parameters begs the question of what acoustical parameters are registered and contribute to the perception of the contrast. A precise specification of the perceptually pertinent parameters is important if valid interpretations are to be made of data obtained using various types of continua whose members are said to vary in "VOT."

Using synthetic stimuli, Summerfield and Haggard (1974) artifically varied the temporal separation of the fricated burst from the events which normally follow it: formant transitions and the onset of periodicity. They demonstrated that the temporal interval is indeed a powerful perceptual cue: whether or not it is filled with aspiration. The question remains: Which of the spectral parameters of VOT whose variation is normally correlated with that of the separation interval are also perceptual cues? Stevens and Klatt (1974) suggested that some threshold duration or spectral extent of first-formant (F1) transition may be a psychoacoustically more basic cue to the voiced value of the feature, and that VOT (that is, the temporal separation interval) is grafted onto this through learning in infancy. Summerfield and Haggard (1974) showed that the detectability of transitions in both the first and higher formants, whether or not they were periodically excited, could provide important secondary cues for adults. Lisker (1975) has argued that the simple articulatory basis of VOT (e.g., Lisker and Abramson, 1971) renders it the most general and basic cue, but proposed that if any secondary aspect of the acoustical array related to formant transitions is important, then it is the *onset* frequency of F1 rather than its dynamic spectral properties. Lisker's data show that when the importance of the spectral cues is assessed by trading them against VOT, which in turn affects the values of the secondary transition cues, then VOT does emerge as the most potent perceptual cue. However, his results, based on a nonorthogonally varying stimulus set, implicate the average frequency region of F2 as a functioning secondary cue in addition to F1 onset frequency. The experiments reported here were designed to refine and extend Lisker's conclusion and to reduce the ambiguity by using orthogonally varying stimulus arrays. The matter can be simplified by asking three questions. Does F1 onset cue a voiced percept in inverse relation to its frequency? Is a rising F1 transition a positive cue to voicing independent of its onset frequency? Are spectral influences on the perception of voicing a function only of the frequency of F1 or of the distribution of energy in both F1 and the higher formants? Experiments 1 and 2 were designed to answer the first two of these questions. Experiment 3 was designed to answer the third question.

I. EXPERIMENT 1: CONDITIONS 1 AND 2

In the first condition of experiment 1, the frequency of a fixed-frequency, transitionless F1 was systematically lowered across a set of consonant–vowel (CV) VOT continua. If Lisker's (1975) conclusion is correct, this procedure should increase the probability of a voiced percept at any given VOT. In the second condition, the onset frequency of F1 was held constant *independently of the realized VOT*, while the duration, and consequently the spectral extent of F1 transition following voicing onset, were systematically increased. If a periodically excited F1 transition is, *per se*, a cue to voicing, then this procedure should increase the probability of a voiced percept at any given VOT.

A. Stimuli and procedure

Both conditions of experiment 1 were run interactively with stimuli generated at run-time by a Fonema OVE IIIb serial resonance speech synthesizer controlled by the SPEX program (Draper, 1973) running on a D.E.C. PDP-12 digital computer. Stimuli were exemplars drawn from [g-k^h] VOT continua spanning the VOT range from 0 to +80 ms in 1-ms steps. The closed-loop algorithm controlling stimulus presentation was an implementation of PEST (Taylor and Creelman, 1967) with the following control parameters: deviation limit of the sequential test $(W) = 0.5^2$; starting step size = 16 ms; and terminating step size = 1 ms. These parameters result in an estimate of the p 0.5 point on the psychometric function underlying the physical test continuum; this point corresponds to the phoneme boundary. To achieve a controlled estimate of the position of the boundary, two PEST runs were randomly interleaved with starting points randomly drawn from preselected ranges approximately evenly balanced on either side of the subject's expected phoneme boundary region. The two interleaved runs converged independently from starting points at long and short VOT's. Subjects were unaware of performing in a closed-loop situation. Convergence was continued until the step-size of each run had diminished to less than or equal to 1 ms and the VOT's corresponding to the $p = 0.5$ estimates from each run were within 5 ms of one another. The phoneme boundary position is here defined as the average of these two independent estimates. Previously, open-loop and closed-loop procedures for estimating phoneme boundaries have been compared and shown to produce highly similar results (Summerfield, 1974a).

The stimuli used in each condition were constructed from seven five-formant CV "stimulus types." A stimulus type is that set of synthesis control parameters which generates a stimulus with a VOT of 0 ms. The frequency contours of F2 and F3 did not differ between stimulus types and were constructed with initial formant transitions appropriate for the velar place of articulation. These transitions were linear in frequency/time over their duration of 44 ms. The F2 transition had its onset at 2400 Hz and reached a steady state at 2000 Hz. The F3 transition had its onset at 2600 Hz and reached a steady state at 3000 Hz. F4 and F5 were set to 3500 and 5000 Hz, respectively. The total duration of each stimulus type was 320 ms. The seven stimulus types used in condition 1 were distinguished by the frequencies of their first formants which were set to 200, 225, 250, 275, 300, 350, and 400 Hz. The seven stimulus types used in condition 2 were distinguished by the duration of their F1 transitions; these transitions always onset at 250 Hz and rose linearly at 5 Hz/ms for either 0, 6, 12, 18, 24, 30, or 36 ms after voicing onset. No other synthesis control parameters were varied between stimulus types or conditions. Over the first 80 ms of each stimulus type, the overall amplitude contour was constant and the fundamental frequency (F0) was fixed at 100 Hz so that differences in F0 at voicing onset could not accompany differences in VOT. A stimulus with any VOT in the range 0 to +80 ms could be constructed from any one of the stimulus types by algorithm. The algorithm replaced the periodic excitation prior to the specified VOT with noise excitation 5.5 dB lower and also widened the bandwidth of F1 from 60 to 300 Hz for this portion of the syllable. This procedure reduces the level of aperiodic energy in F1 and thereby simulates the acoustic consequences of coupling to the pharynx and the trachea. The onset of pitch-pulsing was synchronized to the specified VOT by the procedure described by Draper and Haggard (1974). Figure 1 illustrates the differences between the stimulus types used in condition 1 in displays of the formant parameter specifications F1, F2, and F3 of exemplars with VOT's of 0 and +20 ms. Figure 2 displays analogous patterns for the stimuli used in condition 2 and shows that in order to hold the onset frequency of F1 constant as VOT varied, it was necessary to restructure the spectral relation between F1 and the higher formants in a manner that was not representative of any naturally occurring variation.

Six adult subjects performed in the experiment, three in the order condition 1–condition 2, and three in the reverse order. Each was a native speaker of British English and had served previously in experiments involving closed-loop phoneme boundary estimation. Stim-

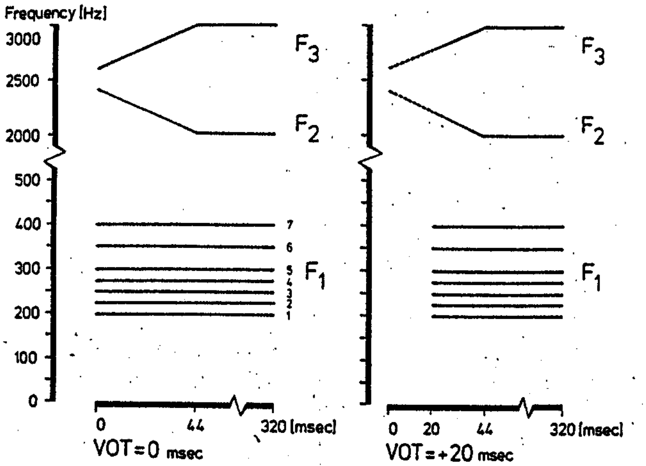

FIG. 1. Schematic spectrograms showing the patterns of the first three formants for the seven stimulus types used in experiment 1, condition 1 in exemplars with VOT's of 0 ms (left) and +20 ms (right). Solid lines indicate periodic and dotted lines aperiodic formant excitation. The stimulus types are distinguished by the frequencies of their transitionless first formants.

uli were presented binaurally through AKG K60 600-Ω headphones to subjects who sat in a sound-damped cubicle. The peak intensity of presentation was constant across subjects at approximately 85 dB SPL for stimuli with 0 ms VOT derived from the two identical stimulus types (types 3 and 1 in conditions 1 and 2, respectively). Subjects were instructed to identify the initial consonant of each stimulus as either [g] or [kʰ] and to indicate their response by pressing one of two buttons labeled "G" and "K." A third button, labeled "?," could be pressed to summon a repetition of the current stimulus. Each subject ran through the whole set of continua twice. In condition 1, three subjects experienced the continua in ascending, followed by descending, order of F1 frequencies, and three in descending, followed by ascending order. The two estimates so obtained were averaged to provide a single estimate for each subject on each continuum. Analogous order balancing was employed in condition 2. The lack of naturalness inherent in the stimulus structure posed no difficulty for listeners, although some subjects reported hearing stimuli with long VOT's and extensive F1 transitions in condition 2, as initiated by the cluster [kʰl] rather than by the single consonant [kʰ].

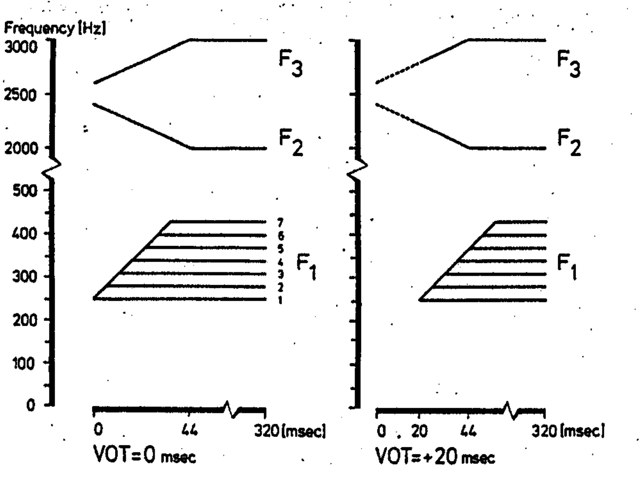

FIG. 2. As Fig. 1 for the seven stimulus types used in experiment 1, condition 2. The stimulus types are distinguished by the duration and extent of their first formant transitions which onset, independently of VOT, at 250 Hz.

TABLE I. Experiment 1: Condition 1. Mean phoneme boundaries in milliseconds of VOT [PBs], and mean differences between boundary estimates [Ds], averaged over estimates by six subjects on seven [g–kʰ] VOT continua differentiated by the frequency of a constant frequency, transitionless first formant (200–400 Hz).

	Number and first-formant frequency (Hz) of stimulus type:–						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	200	225	250	275	300	350	400
[PB's]	33.81	30.99	29.53	28.13	26.64	24.73	22.59
[D's]	1.96	3.67	3.62	3.77	5.22	2.30	2.78

B. Results

Each subject provided two phoneme boundary estimates for each continuum. The average of, and the absolute difference between, the members of each pair provide (a) a measure of the phoneme boundary position, and (b) an indication of its variability, for each subject on each continuum. Mean values, obtained by averaging these measures over subjects, are tabulated in Table I for condition 1 and in Table II for condition 2.[3]

The results of condition 1 support Lisker's (1975) conclusion that the onset frequency of F1 can function as a voicing cue: The data in Table I show that the position of the phoneme boundary averaged across subjects decreases monotonically as the frequency of a transitionless first formant is raised. Only one subject failed to show a systematic decrement. The seven phoneme boundaries from each of the six subjects were examined together in a nonparametric test for monotonic trend (Ferguson, 1966) which gave a value of the normal deviate equal to 6.19, indicating that the trend is significant ($p<0.01$; 2 tailed). The results of condition 2 indicate that variation in F1 transition duration/extent does produce a small effect on the perception of stop voicing. However, it is not in the direction predicted from the arguments of Stevens and Klatt (1974) or Summerfield and Haggard (1974) on the basis of transition detectability. Table II shows that a fall in the value of

TABLE II. Experiment 1: Condition 2. Mean phoneme boundaries in milliseconds of VOT [PBs], and mean differences between boundary estimates [Ds], averaged over estimates by six subjects on seven [g–kʰ] VOT continua differentiated by the durations of their first formant transitions (0–36 ms) which onset at 250 Hz independently of VOT.

	Number and first-formant transition duration (ms) of stimulus type: –						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	0	6	12	18	24	30	36
[PB's]	28.25	28.01	27.59	25.74	26.04	26.81	26.18
[D's]	3.31	3.82	2.74	2.73	1.63	3.46	3.58

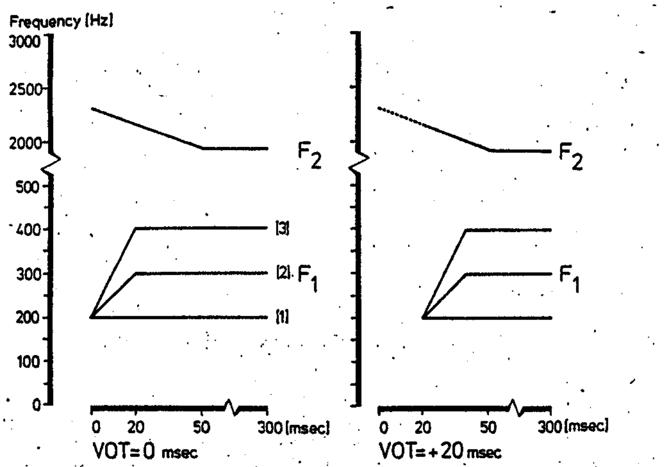

FIG. 3. Schematic spectrograms showing the pattern of the first two formants for stimuli used in experiment 2, condition 1 in exemplars with VOT's of 0 ms (left) and +20 ms (right) in which F1 onsets at 208 Hz. Stimuli were derived from nine VOT continua distinguished by (a) the onset frequencies of their first formants (208, 311, or 412 Hz) and (b) the extent of their first formant transitions (0, 100, or 200 Hz).

VOT at the phoneme boundary occurred as transition duration increased. This trend was shown by four subjects and is also significant ($z=2.58$; $p<0.05$; 2 tailed).

C. Discussion

The results of experiment 1 imply that the critical aspect of F1 for the perception of stop voicing is its perceived frequency at the onset of voicing and suggest that an F1 transition as such does not specifically predispose a voiced percept. However, the relative amplitudes in the outputs of a serial resonance synthesizer are not fixed, but vary according to the formant frequency separations (cf. Fant, 1960). In natural productions, constricting the supralaryngeal vocal tract lowers the frequency of F1 and reduces the amplitudes of the higher formants and the overall intensity of the output. Increasing the frequency of F1 in an OVE synthesizer raises the overall intensity of the output, including the higher formants, so that the distribution of energy in the spectrum increasingly favors higher frequencies. Accordingly, the results of experiment 1 could reflect perceptual sensitivity either to changes in the location of the first spectral peak at the onset of periodicity, or alternatively, to changes in the amplitude of that peak relative to peaks at higher frequencies. To determine which interpretation is more appropriate, a control experiment was run using stimuli generated on a parallel formant synthesizer whose formant amplitudes could be specified individually and for which, therefore, the frequency of F1 and the relative amplitudes of the first three formants could be varied independently.

II. EXPERIMENT 2: CONTROL CONDITIONS 1 AND 2

In the first control condition nine VOT continua were created by combining each of three values of F1 onset frequency with each of three extents of F1 transition. Within each continuum, the onset frequency of F1 was held constant as in condition 2 of experiment 1. If the results of that condition reflect perceptual sensitivity to changes in the onset frequency of F1 then phoneme boundaries should vary here with F1 onset frequency but not with F1 transition extent. In the second control condition, the amplitude of F1 relative to F2 was varied over a 12-dB range across three VOT continua while the spectral specification of the stimuli comprising the continua was unchanged. If the results of experiment 1 reflect perceptual sensitivity to changes in relative formant amplitudes, then phoneme boundaries should shift to shorter VOT's as the intensity of F1 is reduced relative to F2. Alternatively, if the results reflect sensitivity to the frequencies of spectral peaks at the onset of periodicity rather than to their absolute or relative amplitudes then the three boundaries should coincide.

A. Control condition 1: Stimuli and procedure

Nine two-formant [g-k^h] VOT continua were synthesized on the parallel resonance synthesizer at the Haskins Laboratories (Mattingly, 1968). Each continuum consisted of eight 300-ms stimuli which varied in VOT from +15 to +50 ms in 5-ms steps with the onset of pitch-pulsing synchronized to the intended VOT. As VOT increased along each continuum the amplitude of F1 was reduced to zero and F2 was excited by noise. Stimuli with the same VOT in different continua were differentiated only by the frequencies of their first formants. Within any continuum the actual onset frequency of F1 was fixed and did not vary with VOT. Nine continua were created by combining three values of F_1 onset frequency (208, 311, and 412 Hz) with three frequency extents of F1 transition (200, 100, and 0 Hz). The duration of these transitions was 20 ms. (The first-formant frequency parameter changed over five successive 5-ms intervals, reaching a steady state in the fifth interval.) The transition rates were, therefore, 10, 5, and 0 Hz/ms. The transition rate of 5 Hz/ms is the same as that used in condition 2 of experiment 1. The transition duration of 20 ms is longer than the 15 ms which Stevens and Klatt (1974) showed to be the 75% threshold duration for detection of an F1 transition changing at a rate of 8.5 Hz/ms. The acoustic differences among the members of the continua are exemplified in Fig. 3 where the formant parameter specifications of stimuli in which F1 onsets at 208 Hz with VOT's of 0 and +20 ms are displayed.

Two groups of subjects listened to a randomization which included ten occurrences of each of the 72 stimuli. Stimuli were presented binaurally through Grason–Stadler TDH39-300Z headphones at a level of 85 dB SPL (peak deflection). One group of subjects consisted of six members of the research staff of Haskins Laboratories, any of whose residual phonetic naivety was dispelled by a description of the acoustic structure of the stimuli. The other group consisted of nine students attending a Yale University summer school who declared themselves to be phonetically naive. Subjects were instructed to make a forced choice identification of the initial consonant of each stimulus as either [g] or [k^h], but to indicate in addition if the sound which they heard was not a satisfactory exemplar of a CV syllable initiated by either [g] or [k^h].

TABLE III. Experiment 2: Condition 1. Percentages of "G" responses made to the members of nine [g–k^h] VOT continua averaged over ten subjects. Each continuum consisted of eight members ranging in VOT from +15 to +50 ms. The continua were distinguished by the onset frequency of their first formants (208, 311, or 412 Hz) and by their frequency extent of the first formant transitions (0, 100, or 200 Hz).

		First Formant Onset Frequency (Hz)		
		208	311	412
First formant	0	77.8	66.6	33.5
Transition	100	69.2	48.9	34.1
Extent (Hz)	200	60.4	52.9	41.5

A difference of 12.6% between any pair of means is sufficient for *a posteriori* significance at the p≤0.01 level.

B. Control condition 1: Results

Four of the experienced subjects and six of the naive subjects exhibited predictable performance: They reported few instances of stimuli initiated by phonemes other than [g] or [k^h] and reported increasing numbers of [k^h] percepts as VOT increased along each continuum. However, the VOT range +15 to +50 ms was not sufficient to permit the computation of a phoneme boundary for every subject in every condition. Accordingly, the data from condition 1 are summarized in Table III not as phoneme boundary positions but as percentages of [g] responses made by these ten subjects to the eight members of each continuum combined. Figure 4 displays plots of the percentage of [g] responses made to each stimulus in each continuum averaged across these subjects. Each point plots the mean of 100 observations.

Four subjects claimed that more than 25% of the initial consonants were neither [g] nor [k^h]. They heard some stimuli with long VOT's as initiated by palatal affricates (e.g., /ĉi/). Their data were more variable than those of the other subjects. The data of one experienced subject were noise free but will be mentioned no further as he only heard instances of [g].

The numbers of [g] responses afforded each of the 72 stimuli by each of the ten consistent subjects were examined in a three-way univariate analysis of variance with the factors:

(a) subjects (10),

(b) F1 onset frequency (208, 311, or 412 Hz),

(c) F1 transition extent (0, 100, or 200 Hz), and

(d) VOT (15, 20, 25, 30, 35, 40, 45, or 50 ms).

The effects of both the major independent variables and their interaction were significant (F1 onset frequency: $F[2, 18] = 28.64$; $p < 0.01$. F1 transition extent: $F[2, 18] = 11.38$; $p < 0.01$. Interaction: $F[4, 36] = 7.30$; $p < 0.01$). *Post hoc* comparisons made according to the criteria recommended by Scheffe (1959) show that increasing F1 onset frequency both from 208 to 311 Hz, and from 311 to 412 Hz, produced significant decreases in the percentage of [g] responses ($p < 0.05$). Increasing F1 transition extent from 0 Hz to either 100 or 200 Hz, also produced a significant decrease in the percentage of voiced percepts ($p < 0.05$), but no systematic effect resulted from the increase from 100 to 200 Hz of F1 transition extent. The extent to which these results are manifest in individual comparisons may be examined in Table III where a difference of 12.6% between any pair of means is required for *a posteriori* significance at the $p < 0.01$ level.

Overall, the results show that increasing F1 onset frequency reduces the proportion of voiced percepts independently of the characteristics of any following F1 transition. The extent to which the presence of an F1 transition also reduces the proportion of voiced percepts depends on its onset frequency. The effect is largest for onsets at 208 Hz, and diminishes to zero as the onset is raised to 412 Hz.

C. Control condition 2

Two stimuli were added to the continuum used in condition 1 in which F1 had its onset at 311 Hz with 0-Hz F1 transition extent. The extended continuum ranged from +10 to +55 ms of VOT. It was duplicated twice to create a total of three continua in which the level of F1 relative to F2 was +6, 0, and −6 dB.[4] Seven naive subjects listened to a randomization comprising ten instances of each of the 30 stimuli and indicated whether they perceived the initial consonant as [g] or [k^h]. The positions of their phoneme boundaries were estimated by probit analysis (Finney, 1971) and examined in a two-way analysis of variance with the factors:

(a) subjects (7) and

(b) relative formant amplitude (+6, 0, or −6 dB).

Mean boundaries were +36.1, +36.3, and +36.7 ms for the −6, 0, and +6 dB conditions, respectively, and do not differ significantly from one another ($F[2, 12] = 0.093$). Although one subject did show a small increase in boundary position with increasing intensity of F1, two others displayed the reverse pattern. Overall, variation of the relative intensities of F1 and F2 in these continua produced no systematic effect on the decision as to whether the initial stop was voiced or voiceless.

D. Experiment 2: Discussion

The perceptual effects of varying F1 onset frequency in experiment 1 could have been mediated by those co-

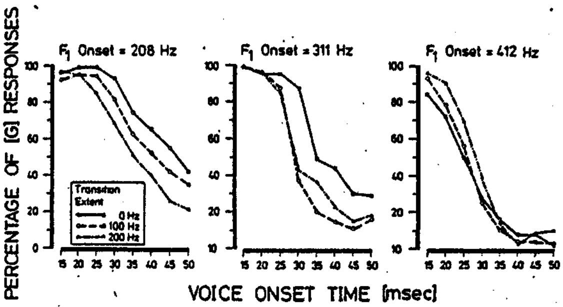

FIG. 4. Results of experiment 2, condition 1 pooled over ten subjects.

TABLE IV. Obtained phoneme boundaries in milliseconds of VOT and boundaries predicted by the equation $Vb=58-100[(\frac{2}{5})\log(\mathrm{F1}^*/200)+(\frac{2}{3})\log(\mathrm{F2}^*/1000)]$ where Vb is the predicted boundary in milliseconds of VOT, and F1* and F2* are the frequencies of the first and second formants at the onset of voicing.

Continuum	F1*	F2*	Obtained	Predicted	Difference[a]
A	540	1232	39	40	+1
B	769	1232	30	29	−1
C	386	1232	43	41	−2
D	286	1845	35	34	−1
E	412	2000	24	25	+1
Continuum	F1*	F2*	Obtained	Predicted	Difference[b]
1	200	2098	34	37	+3
5	300	2158	27	29	+2
7	400	2194	23	23	0
Continuum	F1*	F2*	Obtained	Predicted	Difference[c]
F1 T-Dur.					
20	645	1200	21	32	+11
25	575	1235	22	34	+12
40	540	1320	30	33	+3
55	478	1375	34	34	0
70	452	1410	40	34	−6
85	427	1445	44	34	−10
100	400	1480	45	34	−11
115	375	1500	46	35	−11

[a] Data from Lisker (1975). The letters A, B, C, D, and E identify five [g–kʰ] continua as in the original paper.

[b] Data from experiment 1: condition 1. Predictions are made for three of the seven continua with F1 frequencies of 200, 300, and 400 Hz.

[c] Data from Lisker *et al.* (1975). Predictions are made for eight [da–tʰa] continua differentiated by F1 transition durations (F1 T-Dur.) of 20, 25, 40, 55, 70, 85, 100, and 115 ms.

variations in relative and overall formant amplitudes which the acoustic theory of speech production predicts, and which an OVE synthesizer produces. Had that been so, no effects should have resulted in experiment 2 from varying the frequency of F1 while holding its absolute and relative amplitude constant, but an appreciable effect should have resulted from varying its amplitude while holding its frequency constant. This was not the case. The opposite pattern was produced and confirms that the critical aspect of F1 for the perceptual categorization of members of VOT continua is its perceived frequency at the onset of voicing, rather than its absolute or relative amplitude.

In control condition 1, the frequency extent of F1 transition was varied while holding its onset frequency fixed. The results of this manipulation confirmed the second finding of experiment 1 that a rising F1 transition following the onset of voicing does not, in itself, increase the probability of a voiced percept. Transitions onsetting at 250 Hz (in experiment 1) and at 208 and 311 Hz (in experiment 2) significantly increased the probability of *voiceless* percepts. The physiological representations of the separation cue and the F1 onset frequency cue could both be influenced by whether voicing onset is accompanied by a rising, as opposed to a steady, F1. If there were less energy in the critical band around the putative onset frequency of an F1 transition than at the onset of a fixed frequency F1, then the separation interval might be perceived as longer and the F1 onset frequency as higher than their respective physical values. The data imply that the perceived onset of F1 in these stimuli is determined by spectrotemporal integration over the duration of the first two or three pitch pulses but that the dependency of F1 onset registration on spectrotemporal integration decreases as physical onset frequencies increase from 200 to 400 Hz.

Experiments 1 and 2 demonstrate that the perceived frequency of F1 at the onset of voicing plays an identifiable role as a spectral parameter influencing the voiced-voiceless decision. They do not determine whether it is correct to impute to the frequency of the F1 peak the entire burden of spectral influence or whether that influence derives from the distribution of energy in the spectrum including both F1 and the higher formants. Lisker (1975) considered this possibility to be unlikely, although the perceived differences between his stimulus types can be economically summarized by expressing the spectral influence as the weighted sum of an effect of F1 and an effect of F2. A dependency of boundary location on the frequencies of both F1 and F2 at the onset of voicing is expressed in the otherwise arbitrary formula

$$\mathrm{Vb}=58-100\{\tfrac{2}{5}[\log(\mathrm{F1}^*/200)]+\tfrac{2}{3}[\log(\mathrm{F2}^*/1000)]\},$$

where Vb is the predicted voicing boundary in milliseconds of VOT and F1* and F2* are the frequencies in Hertz of the first and second formants at the onset of voicing. The values of the constants were derived by trial and error to fit Lisker's (1975) data as shown in Table IV, footnote a. While the fit to Lisker's data is quite good, suggesting a role for F2, and the expression adequately predicts the boundary positions observed here in experiment 1 (see Table IV, footnote b), Table IV, footnote c shows that the equation fails to account for the data of Lisker, Liberman, Erickson, and Dechovitz (1975).

Figure 5 displays a plot of obtained phoneme boundary

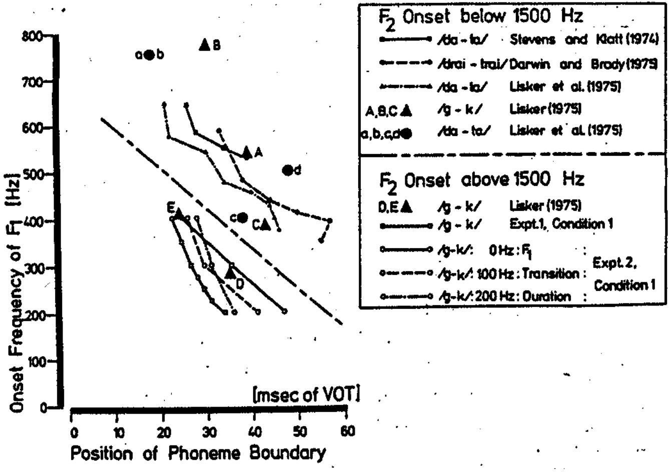

FIG. 5. Plot of the position of the voicing boundary against the onset frequency of F1 for the data sets indicated. The dotted line falling diagonally from left to right segregates the data according to the frequency of F2 at the voicing boundary.

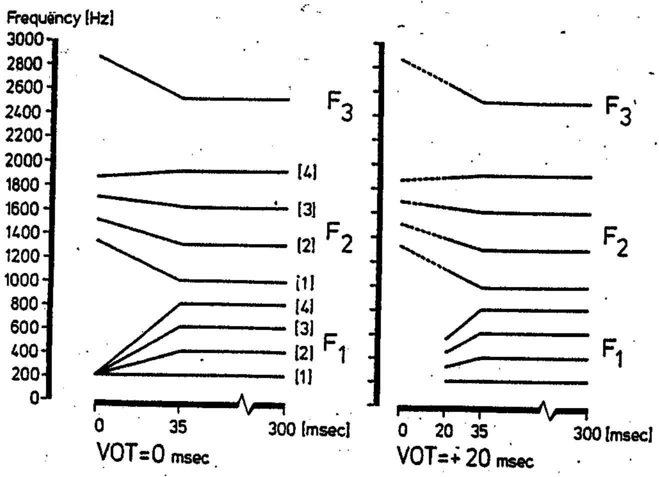

FIG. 6. Schematic spectrograms showing the patterns of the first three formants for the stimuli used in experiment 3 in exemplars with VOT's of 0 ms (left) and +20 ms (right). Sixteen VOT continua were created by combining each of four F1 contours ([1]–[4]) with each of four F2 contours ([1]–[4]). The stimuli included a 10 ms burst centered on 4000 Hz which is not shown.

location as a function of F1 onset frequency for data reported in the present paper and by Stevens and Klatt (1974), Lisker (1975), Lisker *et al.* (1975), and Darwin and Brady (1975). There are two important features of this display. First, the inverse relationship between the onset frequency of F1 and the position of the voicing boundary demonstrated in the present experiments are equally apparent in the other sets of data plotted here. Second, despite the failure of the equation to describe the data of Lisker *et al.*, the remaining data do justify the search for some description of spectral influences that includes the frequency of F2 in addition to that of F1. The dotted line in Fig. 5 falling diagonally from left to right segregates the data according to the frequency of F2 incorporated in the stimuli. Results obtained from stimuli in which F2 was above 1500 Hz fall below this line, those in which F2 was below 1500 Hz fall above the line. The pattern suggests that lowering the frequencies of both F1 and F2 can cause the voicing boundary to shift to longer VOT's. In addition, it appears that the more diffuse the spectrum, the larger is the effect of varying F1 onset frequency.

While this is one explanation for the pattern of data in Fig. 5, it is also possible that the pattern reflects the effects of variations in voicing cues quite different from those considered here [see Klatt (1975) for a review] and the effects of different strategies for synthesis and the use of different groups of listeners. Resolution of these alternatives requires that the same group of listeners categorize the members of a set of VOT continua whose vocalic contexts are characterized by a range of F2 frequencies in combination with a range of F1 frequencies. This was done in experiment 3.

III. EXPERIMENT 3

A. Stimuli and procedure

Sixteen [d–t^h] VOT continua were synthesized on the parallel formant synthesizer at the Haskins Laboratories. The continua included identical synthesis control parameter specifications for F3, F0, and the overall and individual formant amplitudes. They were distinguished only by differences in the frequency contours of their first and second formants. Sixteen continua were formed by combining each of four F1 steady-state frequencies (208, 412, 616, and 818 Hz) with each of four F2 steady-state frequencies (1001, 1306, 1611, and 1917 Hz). This range of formant frequencies includes vowels not found in the English vowel system. Transitions in F1, F2, and F3 were linear in frequency/time over their duration of 35 ms. F1 transitions rose from 208 Hz at stimulus onset to the appropriate steady state. F2 onset frequencies were computed so that the extrapolated trajectories of F2 transitions originated at 1800 Hz 50 ms before syllable onset. The F3 transition had its onset at 2861 Hz and fell to a steady state at 2527 Hz. All stimuli included a fricated burst centered on 4000 Hz and lasting 10 ms from stimulus onset. Each stimulus was 300 ms in duration. Over the first 100 ms the fundamental frequency was constant at 110 Hz. Figure 6 includes schematic displays of the formant parameter specifications of the stimuli.

Each continuum consisted of ten members with VOT's of +5, +10, +15, +20, +25, +30, +35, +40, +45, and +50 ms formed by replacing periodic excitation with noise excitation in F2 and F3 and eliminating energy in F1. The onset of pitch-pulsing was synchronized to the intended VOT in every stimulus.

Ten naive subjects listened to a randomization which included ten instances of each of the 160 stimuli over Grason–Stadler TDH39-300Z headphones at a constant peak intensity of 85 dB SPL. They were instructed to make a forced-choice identification of the initial consonant of each stimulus as either [d] or [t^h] and to indicate their percept by writing "D" or "T." In addition, subjects were instructed to mark with a "?" any response about which they were not confident.

B. Results

Despite being presented with a bizarre array of vowels, most subjects experienced little difficulty in performing the task. While four subjects did indicate that many of their responses to the members of the four continua with F1 set to 200 Hz were guesses, no subject performed inconsistently with stimuli drawn from the other 12 continua.

The data were examined in three ways in different univariate analyses of variance. The first examined the sums of the numbers of "D" responses made to each stimulus by each subject according to the four factors:

(a) subjects (10),

(b) F1 steady state (208, 412, 616, or 818 Hz),

(c) F2 steady state (1001, 1306, 1611, or 1917 Hz), and

(d) VOT (+5, +10, +15, +20, +25, +30, +35, +40, +45, or +50 ms).

Both the main effect of F1 ($F[3, 27]=26.27$; $p<0.01$)

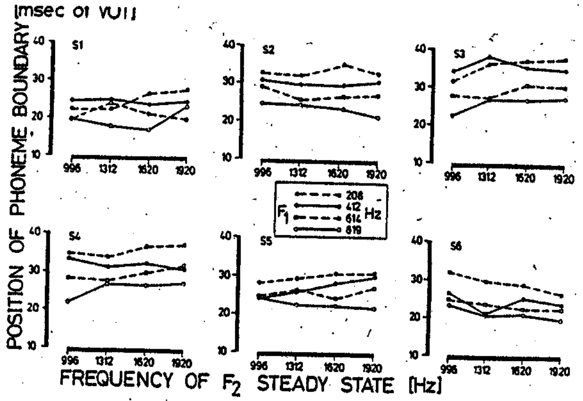

FIG. 7. Results of experiment 3 for six individual subjects who performed consistently on all sixteen continua. For each subject four empirical functions relate the position of the voicing boundary (estimated by probits) to the frequency of the F2 steady state for each of four values of F1 steady state. The functions are essentially horizontal showing no dependency of the position of the voicing boundary on the spectral characteristics of F2.

and its interaction with VOT ($F[27, 243]=13.27$; $p<0.01$) were significant. Neither the main effect of F2 ($F[3, 27]=0.68$; $p>0.2$), nor its interaction with VOT were significant. The data provided by the six subjects who performed consistently on all 16 continua were examined in probit analyses which fitted ogives to the data from each subject for each continuum. Two parameters were estimated for each ogive: the physical stimulus value corresponding to the $p=0.5$ point on the psychometric function and the slope of the probit regression. The first parameter is an estimate of the phoneme boundary. The second varies directly with the standard deviation of the psychometric function underlying the test continuum and hence reflects the slope of the identification function at the boundary. The two parameters were examined in separate analyses with the factors:

(a) subjects (6),

(b) F1 steady state (208, 412, 616, or 818 Hz), and

(c) F2 steady state (1001, 1306, 1611, or 1917 Hz).

Analysis of the 50% intercepts which correspond to the phoneme boundary, showed a significant effect of F1 ($F[3, 15]=35.95$; $p<0.001$), nonsignificant effects of F2 ($F[3, 15]=0.84$; $p>0.2$), and no F1×F2 interaction ($F[9, 45]=0.48$; $p>0.2$). Analysis of the boundary slopes also showed a significant effect of F1 ($F[3, 15]=5.00$; $p<0.025$), nonsignificant effects of F2 ($F[3, 15]=0.05$; $p>0.2$), and no F1×F2 interaction ($F[9, 45]=1.93$; $p>0.1$).

These results may be assessed in relation to the plots in Fig. 7 where boundary position is plotted against the steady-state frequency of F2 for each value of F1 steady-state frequency. Only data provided by the six subjects who performed consistently on all 16 continua are represented. The plots corresponding to each value of F1 onset frequency are horizontal, illustrating the lack of any dependency of boundary position on F2 onset frequency. Means obtained by averaging over these subjects are tabulated in Table V which shows that as the F1 steady state increases in frequency, two things do happen: phoneme boundaries shift to shorter VOT's and the slopes of the probit regressions, and hence of the identification functions at the boundary, become steeper.

TABLE V. Experiment 3. Phoneme boundary positions in milliseconds of VOT averaged over six subjects whose data were internally consistent on all sixteen continua. The continua were distinguished by the frequency of their F2 steady states (1001, 1306, 1611, or 1917) Hz and the frequency of their F1 steady states (208, 412, 616, or 818 Hz). Four values are indicated for each continuum. The first is the position of the average phoneme boundary in milliseconds of VOT [PB]. The second is the average slope of the probit regression line [SL]. Its units are (probit of [+voiced] responses)/ms. The third and fourth values are the frequencies of the first and second formants at the mean phoneme boundary locations (F1* and F2*).

	Continua with F1 = 208 and F1 = 412 Hz: F1 steady state (Hz), F2 steady state (Hz)							
[F1]	208	208	208	208	412	412	412	412
[F2]	1001	1306	1611	1917	1001	1306	1611	1917
Means								
[PB]	30.09	30.47	32.13	31.94	28.83	28.34	28.86	28.92
[SL]	−0.142	−0.158	−0.167	−0.138	−0.187	−0.182	−0.174	−0.167
[F1*]	208	208	208	208	381	381	381	381
[F2*]	1001	1306	1611	1917	1077	1382	1611	1917
	Continua with F1 = 616 and F1 = 818 Hz: F1 steady state (Hz), F2 steady state (Hz)							
[F1]	616	616	616	616	818	818	818	818
[F2]	1001	1306	1611	1917	1001	1306	1611	1917
Means								
[PB]	25.59	25.36	25.73	26.34	22.84	22.24	22.33	22.99
[SL]	−0.188	−0.172	−0.158	−0.169	−0.195	−0.185	−0.195	−0.227
[F1*]	535	535	535	535	611	611	611	611
[F2*]	1077	1382	1611	1917	1077	1382	1611	1917

C. Discussion

It is clear that, overall, the perceived frequency of F2 at the onset of voicing plays an insignificant role in determining how listeners categorize the members of [d–tʰ] VOT continua as voiced or voiceless.[5] It is unlikely that the absence of an F2 effect here, as contrasted with Lisker's (1975) data, results from our use of the alveolar rather than the velar place of production. Comparison of the data from experiment 3 with that plotted in Fig. 5 shows our velar and alveolar data to correspond quite precisely. While Lisker's (1975) data remain anomalous, the present result is congruent with two earlier observations. Summerfield (1974a) varied the durations of syllable-initial F1, F2, and F3 transitions in the members of [ga–kʰa] and [gi–kʰi] VOT continua. This produced a systematic change in the position of the phoneme boundary in /a/ context, where there was an extreme F1 transition whose onset frequency at any given VOT varied with transition duration. However, there was no effect in /i/-context, where, despite a negligible F1 transition, there were appreciable transitions in F2 and F3 whose onset frequencies did vary. Lisker *et al.* (1975) varied the durations of the F2 and F3 transitions independently of that in F1 in the members of a [da–tʰa] continuum. Systematic changes in the position of the voicing boundary resulted from manipulations of F1 but not from those of F2 and F3. The results of experiment 3 augment these earlier findings. They demonstrate that the major spectral influence on the perception of stop-voicing resides in F1 and is not distributed throughout the entire spectrum. Perceptual behavior is explained in terms of the direct acoustic effects of particular vocalic environments on the voicing cues without the invocation of feedback from the phonetic identification of the vowel.

For each steady-state frequency of F2 used in experiment 3, the empirical function relating the position of the phoneme boundary to the onset frequency of F1, if plotted in Fig. 5, would cross the dotted line which purports to segregate results according to the frequency of F2 incorporated in the stimuli. Clearly, a different rationale for the pattern of data in Fig. 5 is required. Its basis may be found in the observation that the different data sets displayed derived from stimuli with different overall durations. The stimuli of Lisker *et al.* (1975) and Darwin and Brady (1975) had durations of 600 ms while those of Lisker (1975) were 450 ms and those used in the present experiments were 300 ms in duration. Summerfield and Haggard (1972) observed that increasing the duration of the steady-state portion of a CV syllable with a fixed VOT increased the probability that the initial consonant would be perceived as voiced. They argued that this finding demonstrated perceptual sensitivity to acoustic covariants of speech rate. We have replicated this finding and found that an increase from 90 to 310 ms in the duration of the vowel in the members of a [biz–pʰiz] continuum shifts the position of the voicing boundary by about seven milliseconds. A simple mechanism which could simulate this effect would scale the duration of the separation interval in a stimulus in relation to the total duration of the syllable, combine the scaled duration with measures of other pertinent cues, and compare the combined cue value with a criterion value to determine the value of the voicing feature. If the effect of manipulating the physical value of another cue, e.g., F1 onset frequency, were assessed by measuring changes in the position of the voicing boundary expressed in terms of the physical value of the separation interval, then the measured effect would increase as the total duration of the stimulus increased. The relation between the present data and that of Lisker *et al.* (1975) and Darwin and Brady (1975) is congruent with this rationale; larger effects of F1 onset frequency variation were produced by these authors' 600 ms stimuli than by our 300 ms stimuli. This explanation remains to be tested and does not account for the patterns of Stevens and Klatt's (1974) and Lisker's (1975) data; those data remain anomalous.

IV. GENERAL DISCUSSION

A. Trading relationships in production and perception

These results identify the perceived frequency of the first formant at the onset of voicing as the critical spectral parameter influencing the perceptual categorization of members of VOT continua. They have shown that a larger value of the separation interval, the purely temporal component of VOT, is required for the perception of a voiceless stop when F1 has a low onset frequency (indicating greater vocal tract constriction) and *vice versa*. This trading relationship corresponds elegantly with one in production.

In production, oral release gestures of differing extents made by the same articulators nevertheless tend to require the same length of time (e.g., Kent and Moll, 1969; Perkell, 1969). It is observed that VOT varies inversely with both the rate at which the oral release gesture is made and with the degree of vocal tract constriction required by the phoneme following the stop. Thus, longer VOT's characterize velar stops, compared to alveolars, compared to bilabials (Lisker and Abramson, 1964); VOT's tend to be longer before the vowel /i/ than before /a/ (Klatt, 1975; Summerfield, 1975a); VOT's are longer in stop + /r/ + vowel and stop + /l/ + vowel environments than in stop + vowel environments (Lisker, 1961; Klatt, 1975).[6] It is not entirely clear why this relationship occurs in production. A relatively constricted vocal tract both increases the acoustic load on the glottal source (Flanagan and Landgraf, 1968), and may also retard the attainment of the transglottal pressure drop necessary for vocal cord vibration (Van den Berg, 1968). Klatt (1975) claims, however, that passive aerodynamics can only contribute to variations in VOT observed in productions of voiced stops since in voiceless productions the supraglottal pressure established during the occlusive phase is entirely dissipated during the fricative portion of the stop-release and is at atmospheric level at the time when vocal cord adduction is initiated. He suggests that, to offset the inherently low frequency of F1 when stops are produced before a close vowel or a lateral, the timing of glottal adduction relative to oral release could be actively delayed.[7] It is fairly parsimonious to postulate such learned compensation in production. Perceptual sensitivity to the summed

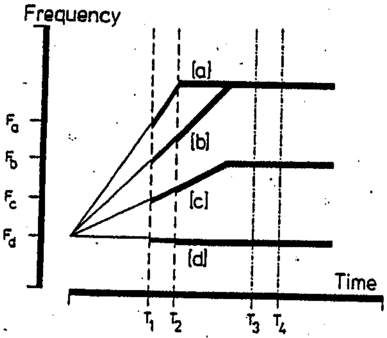

FIG. 8. Schematic descriptions of four syllable-initial first formant contours ([a], [b], [c], [d]) which could be incorporated in the members of different VOT continua. Were voicing to onset at time T_1 first formant onset frequencies of F_a, F_b, F_c, and F_d Hz would result.

cue values of separation interval and F1 onset frequency is already required, whatever the habits of production may be. By pooling measures of these two cues at a low level, or registering the relationship between them directly (cf. Lappin, Bell, Harm, and Kottas, 1975), the noninvariance problem for perception is reduced. This perceptual summation should apply equally in the speaker's perception of his own productions. As a *quid pro quo*, production could be expected to develop vowel contingent modifications to delay adduction in order to permit a general criterion value of the summed measure to characterize phoneme boundaries in most circumstances. Possibly, small passive aerodynamic effects of the adjacent vowel upon voicing onset occur in unstressed syllables, while larger delays result from controlled adduction delay in stressed syllables.

The identification of the role of F1 onset frequency permits the rationalization of a group of previously reported results. In Fig. 8, four F1 transition contours which might be incorporated in the members of synthetic VOT continua are schematized. Transitions [a] and [b] differ in duration, while transitions [b] and [c] differ in spectral extent. Contour [d] evinces no transition. Were voicing to onset at time T_1 ms, F1 onset frequencies of F_a, F_b, F_c, and F_dHz would result. The diagram exemplifies, as Lisker *et al.* (1975) have emphasized, that variation in either the temporal duration or the frequency extent of an F1 transition results in covariation of F1 onset frequency at any given VOT. Thus, effects previously attributed to F1 *transitions* following experimental manipulation of either transition duration (Stevens and Klatt, 1974; Summerfield, 1974a), or frequency extent (Summerfield and Haggard, 1974), where the F1 steady state was fixed, are more appropriately ascribed to variation in F1 onset frequency. Similarly, phoneme boundaries on VOT continua involving the vowel /i/ (with a low frequency F1 in the vowel and hence little or no F1 transition) fall at longer VOT's than do those on continua with the vowel /a/ (with a high frequency F1 in the vowel and a potentially extensive F1 transition) (Cooper, 1974; Summerfield, 1974a; 1975b); that finding is rationally explained by the necessarily lower F1 onset frequency in /i/ context. (Compare contours [b] and [d] in Fig. 8.) These results would be paradoxical if the transition were considered to be a cue to voicedness; the paradox led Summerfield and Haggard (1974) to consider a possibility which they otherwise acknowledged to be unparsimonious, namely, that the perceptual weightings of measures of the temporal and spectral aspects of VOT might be conditioned by vocalic context. With F1 onset frequency identified as the critical spectral parameter, there is no need for such feedback, and the voicing decision may be reached without reference to the category of phoneme following the stop. (See also Darwin and Brady, 1975.) Further methodological implications are reviewed below in Sec. IVC.

The results obtained here may reflect the effects of another, less influential, spectral parameter in addition to F1 onset frequency. The schematic displays in Figs. 1, 2, 5, and 8 show that the constraints which were applied to the acoustic structure of the stimuli necessarily resulted in covariation of the frequency of the F1 *steady state* with, in different conditions, either the onset frequency of F1, or the extent of the F1 transition. Increases in both these latter variables raised the probability that a stop-consonant characterized by a particular VOT would be perceived as voiceless. Thus, the results exhibit a correlation between the frequency of the F1 steady state and the probability of a voiceless percept. Experiment 1 showed that there is not a strong causative relationship between the two. However, the results do not eliminate the possibility that there may be some influence. Stevens and House (1963) noted that the contour of F1 in the vocalic portions of natural CV syllables is lower in frequency following voiced, as opposed to voiceless, stops reflecting the increase in vocal tract length that results from the lower position of the larynx in voiced productions, (e.g., Ewan and Krones, 1974). This aspect of articulatory behavior increases the spectral difference in F1 at voicing onset between voiced and voiceless productions. It remains to be determined whether an additional perceptual effect derives from the coarticulated variation in the F1 steady state.

B. First formant transitions and first formant onsets

The failure of an F1 transition to cue voicing in adults raises doubts about Stevens and Klatt's (1974) suggestion as to its perceptual primacy for the perception of voicing contrasts in infants. Such wariness is reinforced by two recent findings. First, demonstrations of the categorical perception of the members of continua formed by varying the relative onset times of noise and buzz segments(Miller, Pastore, Wier, Kelley, and Dooling, 1976) and pairs of sine waves (Pisoni, 1977) have confirmed Hirsh's claim (Hirsh, 1959; Hirsh and Sherrick, 1961) that a natural psychoacoustic boundary between the perception of successive and simultaneous coterminous acoustic events occurs at a temporal offset of about 17 ms. Although, as the results of the present experiments show, the perception of voicing contrasts involves the registration of the spectral concomitants of the interval between release and voicing onset, psychoacoustic considerations may well dictate why a temporal interval is the basis of the voicing distinction in general (whether positive or negative values of VOT are involved) and why in particular many of the world's languages place a category boundary between VOT's of 0 and +40 ms. The second difficulty for the supposed primacy of transitions comes from a developmental study by Simon (1974). He showed that children older than eight years do not categorize any members of a "Goat–Coat" VOT continuum as initiated by [g] unless they contain a low F1 onset frequency. Children younger than five years, on the other hand, indicate that they have perceived [g] in the absence of the spectral cue and appear to be primarily sensitive to variation in the temporal cue. These results support Lisker's assertion of the primacy of the temporal aspect of VOT and suggest that it is the ability to detect the spectral cue that is learned.

At present, it is not clear whether infants' behavior in discriminating members of VOT continua (cf. Eimas *et al.*, 1971; Streeter, 1976) represents a psychoacoustic ability to distinguish successive from simultaneous acoustic events, or a phonetic ability to distinguish voiced from voiceless stops (Pisoni, 1977). The alternatives could be dissociated by experimenting with VOT continua (e.g., [gri–k^{h}ri]) on which the phoneme boundary, by virtue of a low F1 onset frequency, occurred at a considerably longer VOT than the simultaneity–successivity threshold. Would infants display enhanced discrimination near the psychoacoustic boundary, the phonetic boundary, or both?[8]

C. Implications for studies using stimuli drawn from VOT continua

The demonstration that the temporal and spectral components of VOT may be traded for one another and that, by implication, each possesses perceptual potency in cueing the voicing distinction, has methodological import for studies whose stimuli are drawn from VOT continua.

Where F1 transition duration is held unnaturally constant across continua which represent articulations in which it would normally vary, the positions of phoneme boundaries should not vary. Darwin and Brady (1975) synthesized [de–t^{h}e] and [dri–t^{h}ri] continua with identical parametric specifications of F1. The perceptual identification functions for the two continua differed slightly, but in the reverse direction from that to be expected if the boundary locations were determined by phonetic class: Boundaries on the [dri-t^{h}ri] continuum occurred at shorter VOT's than those on the [de–t^{h}e] continuum. Lisker *et al.* (1975) synthesized [ba–p^{h}a], [da–t^{h}a], and [ga–k^{h}a] continua with identical transition specifications for F1. Boundaries on these three continua coincided, in contrast to those obtained in Lisker and Abramson's original (1967) study where the duration of the F1 transition covaried naturally with place of production.[9]

If VOT continua involve cutback of the duration/frequency extent of an F1 transition, then variation in VOT over the duration of this transition (e.g., between times t_1 and t_2 in Fig. 8) will alter the physical values of *both* cues. Equivalent variation beyond the end of the transition (e.g., between times t_3 and t_4), or on continua not involving an F1 transition (e.g., between either t_1 and t_2 or between t_3 and t_4 on contour [d]), will only vary the value of the separation cue. If, as the results of the present experiments suggest, the decision as to the value of the voicing feature may be represented as being based on a combination of analogue measures of these two cues and others (Summerfield, 1974b; 1975c; cf. Hoffman, 1958; Haggard, 1974), then the perceptual effect of a particular change in VOT will depend upon the magnitude of the change in the combined value of the cues that it produces. A VOT shift that changes the physical values, and hence the perceptual measures, of both cues should produce a larger perceptual effect than should one which only varies the value of the separation cue. It is likely, in addition, that the perceptual scaling of the temporal separation component of VOT for values greater than the simultaneity–successivity threshold approximates Weber's law (Abel, 1972; Miller *et al.*, 1976). As a result of both these factors, the perceptual effect of a change in VOT of fixed size should diminish as the absolute VOT on which that change is centered increases. The perceptual consequences of the two factors have not been dissociated, although effects have been observed which reflect their joint operation. Pisoni and Lazarus (1974) carried out 4IAX discrimination tests of members of a [ba–p^{h}a] continuum involving syllable-initial formant transitions of 50-ms duration. They noted that discrimination of stimuli differing in VOT by 20 ms was more accurate in the voiced range of VOT's from 0 to 40 ms, where the physical values of both cues were changing, than in the voiceless range above 40 ms. Similarly, Summerfield (1975c) measured phoneme boundary widths, defined as the difference between the VOT's corresponding to 25% and 75% voiced responses for each of eight subjects on a [ga–k^{h}a] continuum which was synthesized with an extensive rising F1 transition of 60-ms duration and on a [gi–k^{h}i] continuum which was synthesized with no F1 transition. Boundary width, in this definition, relates inversely to discrimination in the boundary region and should reflect the rate of change of the combined value of the two cues at the boundary. Mean phoneme boundaries occurred at +29.0 ms in /a/ context and at +41.6 ms in /i/ context. Mean boundary widths were 6.6 ms in /a/ context and 10.5 ms in /i/ context. Each of the eight subjects displayed larger boundary widths on the [gi–k^{h}i] continuum than on the [ga–k^{h}a] continuum. Similarly, estimates of the slope of the psychometric functions underlying the continua in experiment 3 decreased significantly as mean phoneme boundary location increased. In all these studies therefore, discrimination of VOT differences was best (a) at shorter as opposed to longer VOT's, and (b) when the change in VOT to be discriminated varied both the separation interval and the onset frequency of the first formant.

An implication of these observations is that the size of the change in the position of the phoneme boundary on a VOT continuum induced by a given difference in some contextual variable will be greatest when the induced change occurs at a large mean VOT and only varies the duration of the separation cue. It will be smallest when the change occurs at short VOT's and varies both the onset frequency of F1 *and* the duration of the separation cue. Summerfield (1975b) measured the size of shifts in the phoneme boundary on VOT continua caused by variation in the syllabic rate of phrases which introduced test syllables drawn from the continua. On continua synthesized with the vowel /i/ (where F1 was low in frequency and there was only a small F1 transition) phoneme boundaries fell at longer VOT's and larger phoneme boundary shifts were measured than on continua with the vowel /a/ (where there was an extensive F1 transition). These observations confirm the deductions outlined above concerning discriminability and lend force to recent warnings by Abramson (1976) that the VOT dimension, though a simple temporal continuum when viewed in articulatory terms, involves variation in a complex set of acoustic parameters whose relative availability is a function of both absolute VOT and phonetic context. The interpretation of data obtained with stimuli drawn from such continua is only valid if it takes this complexity into account.

V. SUMMARY AND CONCLUSIONS

The experiments reported here permit two conclusions:

(1) The perceived onset frequency of F1 is the critical spectral parameter included in the repertoire of cues to the voicing decision for syllable-initial prestressed stop consonants in English. The spectral influence derives only from F1, not from the spectrum comprising F1 and the higher formants.

(2) A periodically excited, rising first formant transition is not, *per se*, a positive cue to voicing when its onset frequency is controlled.[10]

In perception, the temporal separation component of VOT and the F1 onset frequency component may be traded one for the other: the lower the frequency of F1 at the onset of voicing, the longer the separation interval required to produce a voiceless percept. This trading relationship parallels one in production where VOT varies inversely with the degree of vocal tract constriction, and hence with the frequency of F1, required by the phoneme following the stop consonant.

The greater role of F1 onset frequency than of F1 transition here does not imply that transition characteristics are never important in speech perception. A rising first formant at the onset of a pattern of formant frequencies signals an obstruent articulation and is more likely to predispose a consonantal percept than is a fixed-frequency transitionless first formant onsetting at the same frequency. Such a rapid spectral change need not be confined to the spectrum above 1 KHz as Stevens (1975) suggests. It is something to which an F1 transition, relieved of the burden of characterizing (+voiced), contributes.

VI. ACKNOWLEDGMENTS

Experiment 1 was conducted in the Department of Psychology at the Queen's University of Belfast, Northern Ireland with the support of grant AT/2058/021/HQ from the Joint Speech Research Unit, U.K. and grant B/RG/1466 from the Science Research Council, U.K. It was reported as "First formant onset frequency as a cue to the voicing distinction in pre-stressed, syllable-initial stop-constants," in *Speech Perception* No. 5, pp. 25–33. (Progress Report, Department of Psychology, The Queen's University of Belfast.) This paper was written and the later experiments were carried out at the Haskins Laboratories, New Haven, CT with the support of grant HD-01994 from NICHD while Quentin Summerfield was supported by a NATO postdoctoral research fellowship. We should like to express our appreciation to Alvin Liberman for his generous hospitality and encouragement, and to Bruno Repp, Peter Bailey, Gary Kuhn, and David Pisoni for their criticisms of earlier drafts of this manuscript.

[a] Present address: Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510, U.S.A.

[b] Present address: The Institute of Hearing Research, The Medical School, The University of Nottingham, Nottingham NG7 2UH England, United Kingdom.

[1] With reference to the acoustics of production, the term "VOT" will refer to the time interval between the onset of the occlusion release transient and the onset of quasi periodicity. With reference to continua of synthetic stimuli, the term "VOT" will refer to the interval between the onset of the stimulus (which may or may not include a burst) and the onset of periodic excitation. During this interval the presence of noise excitation in F2, F3, and the higher formants and the absence of energy in F1 is implied. The term "separation interval" will refer only to the temporal aspect of VOT.

[2] With $W = 0.5$ the PEST algorithm is simplified. The Wald sequential decision test is obviated and a change in stimulus value occurs after every response.

[3] Data for individual subjects are reproduced in full in "First formant onset frequency as a cue to the voicing distinction in pre-stressed, syllable-initial stop-consonants," Speech Perception No. 5, pp. 25–33, 1976. (Progress Report, Department of Psychology, The Queen's University of Belfast.)

[4] This range is less than that which occurred across the stimuli used in experiment 1, condition 1 which was about 25 dB. It is representative of the range which occurs in natural speech, however, Peterson and Barney (1952) report a range of 16 dB across the English vowels produced by adult male talkers.

[5] However, see Draper and Haggard (1974), Sawusch and Pisoni (1974), and Repp (1976) for discussions of effects on the perception of place and voicing deriving from the microstructure of F2 and F3 transitions as opposed to the macroscopic effect sought here.

[6] Lisker (1961) reports VOT's measured in syllable-initial voiceless stops preceding a vowel and preceding /r/ + vowel

[p^h]: +61 ms, [p^{h}r]: +89 ms;
[t^h]: +64 ms, [t^{h}r]: +110 ms; and
[k^h]: +77 ms, [k^{h}r]: +107 ms.

Klatt (1975) reports similar data for voiceless plosives and the following data for voiced plosives:

[b]: +7 ms, [br]: +12 ms;
[d]: +14 ms, [dr]: +29 ms; and
[g]: -+23 ms, [gr]: +32 ms.

It is noteworthy that a putatively voiced, syllable-initial [gr] can be characterized by a VOT almost twice as large as the simultaneity threshold (Hirsh, 1959) which has been invoked as a psychoacoustic basis for the voicing distinction in English (e.g. Miller *et al.*, 1976; Pisoni, 1977).

[7] We and our colleague Peter Bailey have recently measured periods of devoicing and VOT's in productions of [p^h], [t^h], and [k^h] before [i], [a], [ri], and [ra] in bisyllables such as [bəp^{h}ri]. Total periods of devoicing (i.e. the time from the disappearence of periodicity in the waveform at approximately the moment of stop closure to its re-emergence at voicing onset) tend to be more invariant than either the period of de-

voicing preceding oral release or the VOT itself. Possibly observed covariations of VOT with the degree of vocal tract constriction required by the following phoneme reflect an active process in which it is the moment of oral release that is varied within a fixed time frame of adduction–abduction.

[8] The value of this test would be nullified if the psychoacoustic simultaneity threshold varied as a function of the frequency of the lower spectral component of the stimulus. This possibility is currently under investigation.

[9] A small place-voicing correlation, equivalent to a shift in the VOT boundary of about ±2 ms, remains even when all acoustic differences between stimuli are neutralized (see Draper and Haggard, 1974; Sawusch and Pisoni, 1974; Repp, 1976; Miller, in press).

[10] Not all aspects of the present results are entirely novel. Liberman, Delattre, and Cooper (1958) noted that cutting-back F1 changed the values of two correlated variables: the onset time of F1 relative to F2 and F3; and the onset frequency of F1. They demonstrated that relative onset time has perceptual significance independent of onset frequency. Whether F1 onset frequency had independent perceptual significance was not reported at the time. The intervening years have enabled us to bring more sophisticated synthesis, psychophysical methods, and both psychological and articulatory interpretations to the classical problem of specifying the cues.

Abel, S. M. (1972). "Discrimination of temporal gaps," J. Acoust. Soc. Am. 52, 519–524.

Abramson, A. S. (1976). "Laryngeal timing and consonant distinctions," Haskins Lab. Stat. Report Speech Res. SR-47, 105–112.

Berg, J. W. van den (1968). "Mechanisms of the larynx and the laryngeal vibrations," in *Manual of Phonetics*, edited by B. Malmberg (North–Holland, London), pp. 278–308.

Cooper, W. E. (1974). "Contingent feature analysis in speech perception," Percept. Psychophys. 16, 201–204.

Darwin, C. J., and Brady, S. A. (1975). "Voicing and juncture in stop-/r/ clusters," J. Acoust. Soc. Am. 57, S24(A).

Draper, G. (1973). "SPEX: a system to run speech perception experiments," Proceedings of the 9th DECUS Europe Seminar (Digital Equipment Users Society, Maynard, MA), pp. 89–93.

Draper, G, and Haggard, M. P. (1974). "Facts and artifacts in feature interdependence," in *Preprints of the Stockholm Speech Communications Seminar*, edited by G. Fant (Almqvist and Wiksell, Uppsala), Vol. 3, pp. 67–75.

Eimas, P. D., and Corbit, J. D. (1973). "Selective adaptation of linguistic feature detectors," Cogn. Psychol. 4, 99–109.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). "Speech perception in infants," Science 171, 303–306.

Ewan, W. G., and Krones, R. (1974). "Measuring larynx movement using the thyroumbrometer," J. Phonetics 2, 327–336.

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

Ferguson, G. A. (1966). *Statistical Analysis in Psychology and Education* (McGraw–Hill, London).

Finney, D. J. (1971). *Probit Analysis* (Cambridge University, New York).

Flanagan, J. L., and Landgraf, L. (1968). "Self-oscillating source for vocal-tract synthesizers," IEEE Trans. Audio Electroacoust. AU-16, 57–84.

Haggard, M. P. (1974). "The perception of speech," in *The Physics and Psychology of Hearing*, edited by S. Gerber (Saunders, Philadelphia).

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

Hirsh, I. J., and Sherrick, C. E. (1961). "Perceived order in different sense modalities," J. Exp. Psychol. 62, 423–432.

Hoffman, H. S. (1958). "Study of some cues in the perception of the voiced stop consonants," J. Acoust. Soc. Am. 30, 1035–1041.

Kent, R. D., and Moll, K. L. (1969). "Vocal tract characteristics of the stop cognates," J. Acoust. Soc. Am. 46, 1549–1555.

Klatt, D. H. (1975). "Voice onset time, frication and aspiration in word-initial consonant clusters," J. Speech Hear. Res. 18, 686–706.

Lappin, J. S., Bell, H. H., Harm, O. J., and Kottas, B. (1975). "On the relation between time and space in the visual discrimination of velocity," J. Exp. Psychol. 1, 383–394.

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

Lisker, L. (1961). "Voicing lag in clusters of stop plus /r/," Haskins Laboratories Final Report on Speech Research and Instrumentation (unpublished).

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

Lisker, L, and Abramson, A. S. (1964). "A cross-language study of voicing in initial stops: acoustical measurements," Word 20, 324–422.

Lisker, L., and Abramson, A. S. (1967). "The voicing dimension: some experiments in comparative phonetics," in *Proceedings of the Sixth International Congress of Phonetic Sciences, Prague*, 1967 (Academia, Prague, 1970), pp. 563–567.

Lisker, L., and Abramson, A. S. (1971). "Distinctive features and laryngeal control," Language 47, 767–785.

Lisker, L., Liberman, A. M., Erickson, D. M., and Dechovitz, D. (1975). "On pushing the voice onset time boundary about," Haskins Labs. Stat. Rep. Speech Res. SR-42/43, 257–264.

Mattingly, I. G. (1968). "Synthesis by Rule of General American English," Doctoral dissertation (Yale University). (Supplement to Haskins Lab. Stat. Rep. Speech Res.)

Miller, J. D., Pastore, R. E., Wier, C. C., Kelly, W. J., and Dooling, R. J. (1976). "Discrimination and labeling of noise–buzz sequences with varying noise-lead times," J. Acoust. Soc. Am. 60, 410–417.

Miller, J. L. (1977). "The perception of voicing and place of articulation in initial stop consonants: evidence for the nonindependence of feature processing," J. Speech Hear. Res. (in press).

Perkell, J. S. (1969). *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (MIT, Cambridge).

Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

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.

Pisoni, D. B., and Lazarus, J. M. (1974). "Categorical and noncategorical modes of speech perception along the voicing continuum," J. Acoust. Soc. Am. 55, 328–333.

Repp, B. H. (1976). "The voicing boundary as a function of F2 and F3 transitions and fundamental frequency," J. Acoust. Soc. Am. 60, S91(A).

Sawusch, J. R., and Pisoni, D. B. (1974). "On the identification of place and voicing features in stop consonants," J. Phonetics 2, 181–194.

Scheffe, H. (1959). *The Analysis of Variance* (Wiley, New York).

Simon, C. (1974). "Some aspects of the development of speech production and perception in young children," in *Preprints of the 1974 Stockholm Speech Communication Seminar*, edited by G. Fant (Almqvist and Wiksell, Uppsala), Vol. 4, 7–14.

Stevens, K. N. (1975). "The potential role of property detectors in the perception of consonants," in *Auditory Analysis and the Perception of Speech*, edited by G. Fant and M.A.A. Tatham (Academic, London), pp. 303–330.

Stevens, K. N., and House, A. S. (1963). "Perturbation of vowel articulations by consonantal context: An acoustical study," J. Speech Hear. Res. 6, 111–128.

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.

Streeter, L. A. (1976). "Language perception of 2-month old infants shows effects of both innate mechanisms and experience," Nature 259, 39–41.

Summerfield, A. Q. (1974a). "Processing of cues and contexts in the perception of voicing contrasts," in *Preprints of the 1974 Stockholm Speech Communication Seminar*, edited by G. Fant (Almqvist and Wiksell, Uppsala), Vol. 3, pp. 77–86.

Summerfield, A. Q. (1974b). "Towards a detailed model for the perception of voicing contrasts," Speech Perception No. 3, 1–26. (Progress Report, The Department of Psychology, The Queen's University of Belfast).

Summerfield, A. Q. (1975a). "How a detailed account of segmental perception depends on prosody and vice-versa," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer–Verlag, New York), pp. 51–68.

Summerfield, A. Q. (1975b). "Cues, contexts and complications in the perception of voicing contrasts," Speech Perception No. 4, 99–130. (Progress Report, The Department of Psychology, The Queen's University of Belfast.)

Summerfield, A. Q. (1975c). "Information-processing analyses of perceptual adjustments to source and context variables in speech, "Doctoral dissertation, The Queen's University of Belfast, (unpublished).

Summerfield, A. Q., and Haggard, M. P. (1972). "Speech rate effects in the perception of stop voicing," Speech Syn. Percept. No. 6, 1–12 (Progress Report, Psychological Laboratory, University of Cambridge.)

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, 279–294.

Taylor, M. M., and Creelman, C. D. (1967). "PEST: efficient estimates on probability functions," J. Acoust. Soc. Am. 41, 782–787.