Evaluation and integration of acoustic features in speech perception

Dominic W. Massaro and Gregg C. Oden

University of Wisconsin, Madison, Wisconsin 53706
(Received 29 September 1978; accepted for publication 27 September 1979)

Identification of synthetic stop consonants as either /bae/, /pae/, /dae/, or /tae/ was examined in two experiments in which the stimuli varied independently on voice onset time (VOT), the consonantal second and third formant (F_2–F_3) transitions and, in experiment 2, the intensity of the aspiration noise during the VOT period. In both experiments, the patterns of the resulting identification probabilities were complex, but systematic, functions of each of the independent variables. Of most interest was the fact that the likelihood of identifying a stimulus to be /bae/ or /pae/, rather than /dae/ or /tae/, was strongly influenced by the VOT as well as by the F_2–F_3 transitions. Analogously, the likelihood of identifying a stimulus to be /bae/ or /dae/, rather than /pae/ or /tae/, depended on the F_2–F_3 transitions as well as on VOT. Three explanations of these results were considered within a fuzzy logical model of speech perception: (1) that there is interaction in the evaluation of acoustic features, (2) that the listener requires more extreme values of acoustic features for some speech sounds than for that of other speech sounds, and (3) that the aspiration noise during the VOT period serves as an independent acoustic feature to distinguish /pae/ and /bae/ from /tae/ and /dae/.

PACS numbers: 43.70.Dn, 43.70.Ve

One of the concepts which has had a profound influence on the study of speech perception during the last quarter of a century is that of distinctive features. Trubetzkoy, Jakobson, and other members of the "Prague school" argued against the idea that the phonemes of a language are minimal units of analysis which cannot be further reduced (Lyons, 1968). According to the Prague school, a phoneme can be characterized by distinctive features that represent its similarities and differences with respect to the other phonemes in the same language. For example, Jakobson, Fant, and Halle (1961) proposed that a small set of orthogonal, binary properties or features were sufficient to distinguish among the larger set of phonemes of a language. Jakobson *et al.* were able to classify 28 English phonemes on the basis of only nine distinctive features. The binary nature of distinctive features means that in the linguistic classification of a phoneme, each feature is either present or absent in an all-or-none fashion. The orthogonality of distinctive features means that the different features are conceptually independent so that, in principle, any combination of features could correspond to a phoneme.

Distinctive feature analysis as perfomed by Jackobson *et al.* (1961) for phonemes or by Chomsky and Halle (1968) for other levels of phonetic and phonological representation is not based entirely on physical measurements of acoustic properties of speech sounds. Rather, it includes the articulatory characteristics which distinguish between pairs of phonemes. For example, the articulatory differences between |p| and |b| are similar to the differences between |t| and |d|. Nevertheless, while originally intended only to capture linguistic generalities, distinctive feature analysis has been widely adopted as a framework for human speech perception. The attraction of this framework is that since these features are sufficient to distinguish among the different phonemes, it is possible that phoneme identification could be reduced to the problem of determining which features are present in any given speech sound. This approach gained credibility with the finding, originally by Miller and Nicely (1955) and since by many others (e.g., Campbell, 1974; Cole and Scott, 1972; Peters, 1963; Singh and Woods, 1971; Wang and Bilger, 1973), that the more distinctive features two sounds share, the more likely they are to be perceptually confused for one another.

Given that distinctive features are not directly manifested in the speech signal, the implicit assumption underlying most perceptual studies is that acoustic characteristics mediate the perception of speech. Following our previous work, the acoustic characteristics functional in speech perception will be referred to as *acoustic* features or acoustic cues (Massaro, 1975; Massaro and Cohen, 1976, 1977; Oden and Massaro, 1978), in contrast to distinctive features, in order to maintain the distinction between the psychoacoustic and linguistic levels of description. Although we do not expect that there is a direct correspondence between distinctive features and acoustic features, the study of the latter is aided by the distinctions made by the former. Thus, for example, one tends to ask what are the acoustic features for the voiced–voiceless distinction rather than what are the acoustic features that distinguish |b| and |p|. As pointed out by Lisker (1975) and others, a plethora of acoustic features are possible for the single linguistic distinction of voicing. One might also expect that a single acoustic characteristic could be relevant to more than one distinctive feature. In the present research, we have found some evidence for the aspiration noise functioning as an acoustic feature both for a place of articulation distinction as well as a voicing distinction.

If this framework is accepted, then two questions are of central concern: (1) Which characteristics of speech sounds actually function as acoustic features; and (2) what are the processes by which this featural information is evaluated and integrated to identify a given

 0001-4966/80/030996-18$00.60

speech sound? The seminal work on the first question was carried out at the Haskins Laboratories using synthetic speech produced by the pattern playback synthesizer (Liberman, Delattre, and Cooper, 1952). These investigators studied the contribution of various acoustic properties to the identification of speech sounds which differ by various different distinctive features and must, therefore, also differ by one or more acoustic features. Liberman, Delattre, and Cooper (1958), for example, demonstrated that for stop consonants in initial position, increases in the time between the onset of the release burst and the onset of vocal cord vibration [voice onset time (VOT)] were sufficient to change the identification from a voiced to a voiceless stop. The distinctive feature of voicing characterizes whether or not the vocal cords are vibrated during a significant portion of a particular sound. Stop consonants with short VOTs were usually identified as voiced stops (|b|, |d|, or |g|) whereas consonants that were identical except for having long VOTs were identified as voiceless stops (|p|, |t|, or |k|). Liberman *et al.* (1958) also found that the presence or absence of aspiration during the VOT period contributed to the identification of stops with respect to voicing, and more recently, it has been found that the onset frequencies of the fundamental (F_0) and the first format (F_1) also cue voicing (Haggard, Ambler, and Callow, 1970; Lisker, 1975; Summerfield and Haggard, 1977).

It should be pointed out that it may be that voice onset time is not actually an acoustic feature itself even though changes in this variable are sufficient to change the identification from a voiced to a voiceless stop. Some other concomitant change or changes in the stimulus may function as the critical feature or features. As an example, Winitz, LaRiviere, and Herriman (1975) disentangled the VOT interval and the aperiodic noise information given by the burst and aspiration during the VOT period. Using real speech, English initial stop-consonant syllables were divided at the onset of vocal-cord vibration, which was identified as periodicity in the waveform. The initial aperiodic portion and the remaining periodic portion were recombined with various time intervals between the segments. As an example, the aperiodic portion from |du| was separated from the periodic portion by an interval that gave the same VOT found in |tu|. In this case the syllable had the burst and aspiration appropriate for |du| and VOT appropriate for |tu|. The results indicated that the burst and aspiration portion accounted for the perceptual judgments. The syllable with the burst and aspiration for |du| was always identified as voiced regardless of the VOT. Similarly, a syllable with burst and aspiration from the voiceless cognate was usually judged as voiceless even though the VOT was shortened to imitate the voiced cognate. This study indicates that the burst and aspiration, rather than simply time, is the functional acoustic cue in the perception of voicing of initial stop consonants.

The acoustic features associated with the distinctive feature of place of articulation in initial stops, which corresponds to the point in the oral cavity (lips, alveolar ridge, or velum) at which the airflow is occluded, include the frequency and direction of the second and third format (F_2 and F_3) transitions and also the burst frequency (Delattre, Liberman, and Cooper, 1955; Harris *et al.*, 1958; Hoffman, 1958). Thus, this research using speech synthesis has been reasonably successful in isolating acoustic cues that are individually sufficient for the distinctions given by particular distinctive features (for reviews, see Cole and Scott, 1974; Darwin, 1976; Klatt, 1975; Massaro, 1975).

In contrast, much less research has been done on the second question and, consequently, comparatively little is known about the processes by which different acoustic features are evaluated and integrated together. Recently, however, a number of researchers have begun to examine the nature of these processes (e.g., Massaro and Cohen, 1976, 1977; Oden, 1978; Oden and Massaro, 1978; Oden and Massaro, 1978; Repp, 1977; Sawusch and Pisoni, 1974). Two particular issues that have been considered with regard to this question are whether acoustic features are perceived in an all-or-none fashion and whether each feature is evaluated independently of the other features. If the linguistic distinctive feature theory were taken literally as a model of human speech perception, then we would expect that acoustic features are perceived to be either present or absent in a given speech sound and that the perception of each acoustic feature would be unaffected by the presence or absence of any other feature. However, considerable evidence has now been accumulated (see Massaro and Cohen, 1976; Oden, 1978; Oden and Massaro, 1978; Paap, 1975, for reviews) that acoustic features are perceived in a continuous manner. For example, listeners in the Samuel (1977) and Carney, Widin, and Viemeister (1977) studies were able to make within category discrimination of sounds differing in voice onset time. If the acoustic featural information were binary, then listeners would not have been able to discriminate two sounds that were identified equivalently.

On the other hand, the issue of acoustic featural independence in human speech perception is still very much unsettled. The early work at the Haskins Laboratory discussed above provided some support for the claim of featural independence in that separate sets of acoustic cues were found to be relevant for the different phonetic features. However, evidence consistent with the idea of featural nonindependence has also been found. For example, Lisker and Abramson (1970) studied the identification of voiced versus voiceless phonemes as a function of VOT for the labial (|b| and |p|), alveolar (|d| and |t|), and velar (|g| and |k|) stop consonants. Voiceless identifications increased systematically with increases in VOT for all three places of articulation but the "boundary," that is, the point at which as many voiced as voiceless identifications were made, varied as a function of place of articulation. The changeover from predominantly voiced to predominantly voiceless sounds occurred at the point where the VOT was about 23 ms for labials, 37 ms for alveolars, and 42 ms for velars.

Two types of featural nonindependence can account for

the results of Lisker and Abramson (1970). First, the exact *degree* of perceived VOT, for example, might also depend on the particular F_2–F_3 transitions, which are cues to the phonetic feature of place. This kind of featural nonindependence means that even a direct measure of the perception of VOT would change systematically with changes in the F_2–F_3 transitions. In this case, one would also expect to find similar changes in nonspeech analogs of these stimuli.

The results of Lisker and Abramson can also be interpreted as reflecting modification of the perception of VOT by experience since the obtained boundaries agreed rather well with VOT measurements of natural speech (Klatt, 1975; Lisker and Abramson, 1964). If listeners allow the place information to influence the perception of voicing, it is natural to expect that, whatever the underlying mechanism, the influence will mirror natural speech. Since voiceless stops tend to have longer VOTs as place of articulation is changed from labial to alveolar, the listener may require a longer VOT in order to perceive an alveolar sound as voiceless than that required for a labial sound. Haggard (1970) made this assumption and proposed that the perception of voicing is dependent on the prior analysis of place information.

However, the results of Lisker and Abramson's experiments are somewhat equivocal because of a confounding of acoustic characteristics in their stimuli: In addition to changing the onset of the F_2–F_3 transitions in order to cue different places of articulation, the transition trajectories and durations were also changed as was the duration and trajectory of F_1. [Kuhl and Miller (1978) provide a more detailed description of these stimuli.] Although these changes were made to obtain more naturally sounding tokens, they may also have provided additional acoustic features and thereby have been directly responsible for the changes in the voicing boundary as a function of place. For example, the F_1 transition rose more slowly in the alveolar than in the labial stimuli so that the F_1 onset frequency for a given VOT was lower for alveolar than for labial stops. This property would tend to make the alveolars be perceived as more voiced since a low F_1 frequency at the onset of vocal-cord vibration cues a voiced sound (Lisker, 1975). This would explain the observed shorter voicing boundary for labial than for alveolar stops. However, when Miller (1977) replicated the Lisker and Abramson study with stimuli that held F_1 constant with changes in place, significant though much smaller boundary changes were found: 24.55, 27.80, and 29.30 ms VOT for the labial, alveolar, and velar stops, respectively.

Miller (1977) also examined the identification of labial versus alveolar phonemes as a function of F_2–F_3 transitions for nasal, voiced, and voiceless stop consonants. The results reveal that the boundary between labial and alveolar stops was located at higher values of F_2–F_3 onset frequencies as the stops were changed from nasal to voiced to voiceless. This means that a speech sound with a given F_2–F_3 frequency was more often identified as labial when the stop was voiceless than when it was voiced. Since Miller attempted to prevent concomitant changes along other acoustic dimensions, it is natural to interpret the boundary shifts observed for one dimension as a function of a second dimension as supporting the nonindependence of feature evaluation in speech perception.

However, an alternative explanation for the boundary change results is that the features are evaluated independently but that the observed interaction results from the way in which the featural information is integrated according to the specifications of the relevant alternatives in memory. According to the logic of information integration theory (Anderson, 1974), in order to determine the precise nature of the integration processes, it is necessary to independently vary the acoustic features of VOT and F_2–F_3 transitions and to examine the pattern of the interaction over the range of VOT and F_2–F_3 transitions. Therefore, to examine this question, Oden and Massaro (1978) crossed five levels of VOT with seven levels of F_2–F_3 transitions and had subjects identify the sounds as |b|, |d|, |p|, and |t|.

Figure 1 presents the percentage of times the stimuli were identified as voiced phonemes as a function of VOT; the parameter is the level of the F_2–F_3 transitions. Figure 2 presents the percentage of times the stimuli were identifed as labial phonemes as a function of the F_2–F_3 transitions; the parameter is VOT. As can be seen in these figures, there was very little effect of the F_2–F_3 transitions on the identification of voicing but a large effect of VOT on the identification of place. Figure 2 shows that the sound became more labial with increases in VOT. This result replicates

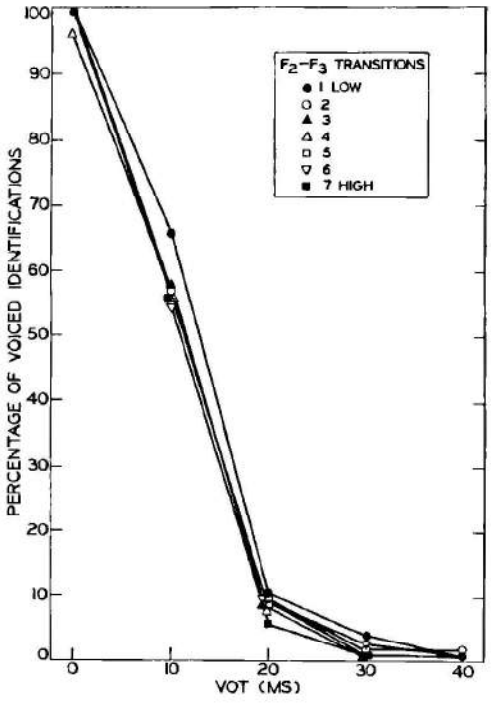

FIG. 1. Percentage of voiced identifications as a function of VOT; the level of F_2–F_3 transitions is the curve parameter (data from Oden and Massaro, 1978).

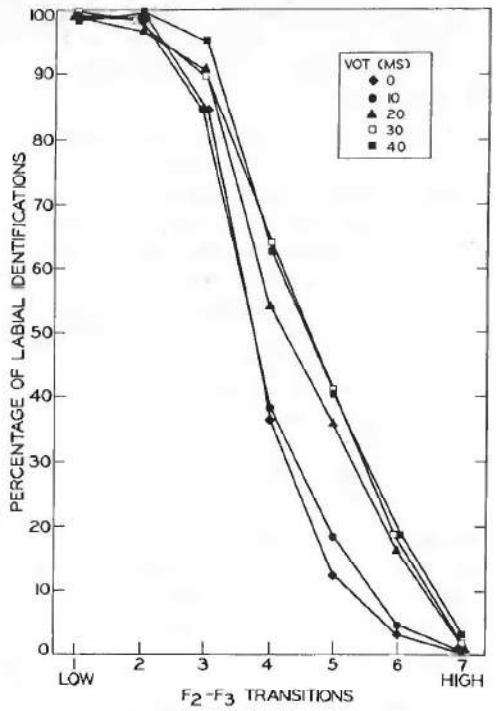

FIG. 2. Percentage of labial identifications as a function of the level of the F_2–F_3 transitions; VOT is the curve parameter (data from Oden and Massaro, 1978).

Miller's (1977) finding of a higher place boundary for voiceless than for voiced sounds.

There are at least two ways in which VOT could produce the place boundary shift directly but still not require the nonindependence of feature evaluation. First, it could be that for some phonemes, more extreme values of the features are required than for others. One way for this to come about could be through the listener's experience with natural speech. For example, Fant (1973, Chap. 11) points out that, with most vowels, the locus of F_2 at the instant of release is higher for |t| than it is for |d|. Thus, since a high F_2-locus is a cue to an alveolar stop, it would be reasonable for listeners to expect |t| to have high F_2 onset frequencies relative to |d|. If listeners use this information, then for a given level of F_2 onset frequency, they should make fewer |t| identifications than they would if they did not have this expectation of high F_2 values for |t|. Producing fewer |t| responses would mean that the place boundary would be shifted toward the alveolar end when the speech sounds are voiceless.

The second possible explanation for the boundary shifts is that changes in a single stimulus dimension may influence more than one acoustic feature. Since any manipulation of an arbitrarily defined stimulus dimension will also produce changes along other dimensions, it may be that some of the covarying changes influence other acoustic features. For example, increasing VOT not only increases the time between burst onset and vocal-cord vibration onset but also increases the total amount of aspiration noise. The aspiration

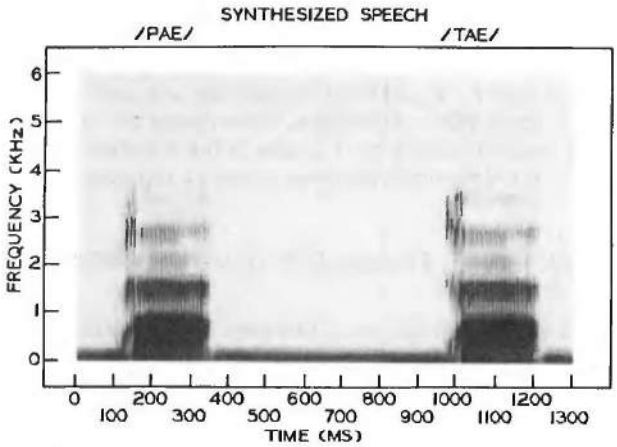

FIG. 3. Spectrograms of two synthesized speech sounds from the present experiment 1.

may function as an independent acoustic feature. It is possible that the aspiration during the VOT period of most synthetic speech without bursts (see Fig. 3) is more representative of the burst and aspiration of a voiceless labial (|p|) than a voiceless alveolar (|t|) (see Fig. 4). Figure 4 shows that the burst and aspiration noise of |tae| is higher in frequency than the burst and aspiration noise of |pae|. Also, the total amount of low-frequency noise appears to be greater in |pae| than in |tae|. If so, longer VOTs would produce more low-frequency aspiration which would produce a sound more like |pae| than |tae| and as a consequence, cause the place boundary to shift toward the alveolar end with long VOTs. Thus, by this explanation, the interaction is "built into" the stimuli and is produced even though the evaluation of the acoustic feature of F_2–F_3 transitions does not depend in any way on the values of the acoustic feature of aspiration.

Each of these explanations may be expressed quantitatively within the framework of a model of speech feature integration developed by Oden and Massaro (1978). In this model, each phoneme[1] is defined by a prototype which is represented by a proposition in long

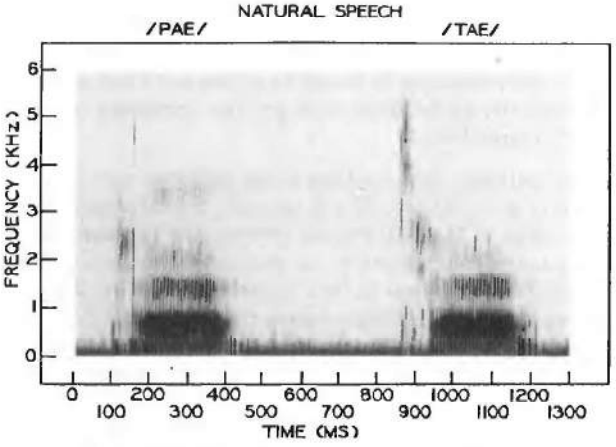

FIG. 4. Spectrograms of two naturally occurring speech sounds.

term memory. Each prototype proposition is a logical expression whose terms correspond to acoustic features. For example, among other things a prototypical |t| has high F_2–F_3 starting frequencies and does not have a short VOT. Therefore, if we ignore for now the other characteristics of |t|, then in the simplest version of the model this phonome would be represented as[2]

T: (HIGH F_2–F_3 TRANSITIONS) AND [NOT (SHORT VOT)]. (1)

The words *high* and *short* are used in the prototypes for ease of exposition. The listener would require a more precise entry describing the ideal values for the F_2–F_3 transitions and the VOT. The values would also have to be relative values in order to allow for the normalization that occurs for different speakers and for different rates of speaking. Having specific ideal values also allows for a mismatch when, for example, the F_2–F_3 transitions are too high.

In the elaboration of the model, this logical expression would be more complex in ways which would reflect the particular hypothesis. For example, under the hypothesis that the phoneme |t| requires more extreme values of the F_2–F_3 transitions the phoneme prototype might be represented as

T: [QUITE (HIGH F_2–F_3 TRANSITIONS)]
AND [NOT (SHORT VOT)], (2)

where the modifier "quite" expresses the extremity of this feature in the prototype. On the other hand, under the hypothesis that both F_2–F_3 transitions and aspiration during the VOT period contribute to place perception, the prototype proposition might be

T: [NOT (LOW-FREQUENCY ASPIRATION)]
AND (HIGH F_2–F_3 TRANSITIONS) AND
[NOT (SHORT VOT)], (3)

where T is now characterized by not having low-frequency aspiration as well as the previously specified features. The acoustic feature of aspiration would refer primarily to the amount of low-frequency aspiration whereas the acoustic feature of the F_2–F_3 transitions refers primarily to the location of the transitions. Therefore, it is possible to manipulate these two features independently by changing either the duration or the intensity of the aspiration and the frequency of the F_2–F_3 transitions.

The different descriptions in the phoneme prototypes given in Eqs. (2) and (3) are the only manifestation in this model of the differences between the two explanations under consideration. In each case, phoneme identification is assumed to be a matter of choosing the phoneme prototype which comes closest to matching the speech sound. This involves, first of all, feature evaluation, which determines the degree to which each respective feature is present in the stimulus. Feature evaluation is often represented as the application of a predicate to the stimulus to determine whether it is true that the required property is present. In this case, since it may be more or less true that the property is present in the stimulus, feature evaluation is represented by the application of fuzzy predicates (Goguen, 1969; Zadeh, 1975).

After the features have been evaluated, the featural information must be combined for each phoneme according to the specification of its prototype in order to determine how well that phoneme matches the stimulus. This involves evaluating the logical expression which defines the phoneme prototype to arrive at the degree to which it is true that the stimulus is an instance of the phoneme. Since the featural information consists of fuzzy truth values, these are combined by the use of fuzzy logical operators, which are assumed to be multiplication for conjunction, powering for intensification (for modifiers such as "quite"), and subtraction from one (perfect truth) for negation (see Oden, 1977; Oden and Hogan, 1977).

Thus, for example, in the simple version of the model [Eq. (1)] the degree of match of the phoneme |t| to a particular stimulus S would be given by the expression referred to as the matching function for |t|:

$$T(S) = \mathrm{HFT} * (1 - \mathrm{SV}), \tag{4}$$

where HFT represents the degree to which it is true that the stimulus has high F_2–F_3 transitions and SV represents the corresponding value for short VOT. The matching function under the hypothesis that |t| requires more extreme alveolarity would also correspond directly to its associated prototype description [Eq. (2)]:

$$T(S) = \mathrm{HFT}^q * (1 - \mathrm{SV}), \tag{5}$$

where q represents the degree of intensification for the modifier "quite." Similarly, under the aspiration feature hypothesis, the matching function corresponding to the prototype expressed by Eq. (3) would be

$$T(S) = [(1 - \mathrm{LFA}) * \mathrm{HFT}] * (1 - \mathrm{SV}), \tag{6}$$

where LFA represents the degree to which it is true that the stimulus has low-frequency aspiration.

Finally, following the rationale of Luce's (1959) choice model it is assumed that the probability of identifying a stimulus to be a particular phoneme is equal to the *relative* degree to which that phoneme matches the stimulus compared to the degree of match of the other phonemes under consideration. For example, if a person must identify a speech sound as either |t|, |p|, |d|, or |b|, then the probability that |t| will be chosen is given by

$$p(t|S) = \frac{T(S)}{T(S) + P(S) + D(S) + B(S)}, \tag{7}$$

where $P(S)$, $D(S)$, and $B(S)$ are the values representing the goodness of match of the respective phonemes to the stimulus and are given by expressions analogous to Eqs. (4), (5), or (6) depending on the particular hypothesis.

The purpose of the present experiments is to examine the proposed alternative explanations for boundary shifts in more detail than was possible in previous re-

search. The first experiment is a direct extension of Oden and Massaro's (1978) study to include a finer manipulation of VOT. In the previous study, VOT was varied in 10-ms steps which appears to have been too large to produce a sufficient number of stimuli for which the voicing of the sound was ambiguous. As Fig. 1 shows, only one level of the voicing factor resulted in stimuli which were sometimes identified as voiced phonemes and sometimes as voiceless. Consequently, the experiment may not have been sensitive enough to detect small boundary shifts as a function of F_2–F_3 transitions. In addition, the excellent description of the data given by the fuzzy logical model may have been due in part to the relatively large number of unambiguous stimuli in the experiment. Levels of independent variables which lead to consistent zero or one response probabilities are uninformative with respect to the integration of featural information and, thus, do not provide a rigorous test of the model. To remedy this shortcoming, the number of levels of VOT was increased in the present study from five to seven by varying VOT in 5- rather than 10-ms steps. The F_2–F_3 transition frequencies were also changed slightly to produce a larger number of ambiguous stimuli.

I. EXPERIMENT 1

A. Method

1. Stimuli

The 49 test stimuli were generated factorially by combining seven levels of F_2–F_3 onset frequencies with seven levels of VOT. The initial F_2 and F_3 frequencies for the seven levels are given in Table I. The VOT could take on the values 10, 15, 20, 25, 30, 35, or 40 ms.

The stimuli were produced on line during the experiment by a speech synthesizer (formant series resonator FONEMA OVE-IIId) under the control of a PDP-8/L computer (Cohen and Massaro, 1976). Each stimulus was specified as a series of lists of parameter vectors. Each parameter vector specified the target value of a parameter, the transition time, and whether the transition was to be linear or negatively accelerated. Each list specified the amount of time until the next list would take control. Time values were specified and parameters calculated in 5-ms increments.

The speech sounds were consonant–vowel syllables (CVs) using stop consonants and the vowel |ae| as in bat. The CV can be represented as a consonant transition followed by a final vowel. Figure 5 gives a schematic diagram of one of the CVs used in the experiment. The values of F_1, F_2, F_3, F_4, and F_5 for the final vowel were 734, 1600, 2851, 3500, and 4000 Hz, respectively, and the amplitude of the buzz source simulating vocal cord vibration (AV) was set at 16 dB.[3] The onset of the voicing energy for the consonant was instantaneous and the offset of the final vowel was a linear fall off which took 30 ms. For the CV transition, F_1, F_2, and F_3 moved from the initial frequencies to those of the final vowel configuration in 40 ms, following a negatively accelerated path. The initial value of F_1 at the onset of the sound was always 200 Hz. The F_4 and F_5 frequencies remained constant at 3500 and 4000 Hz, respectively. A period of aspiration (AH) was created during the VOT period by sending the noise source through the vowel formants at 14 dB. The fundamental frequency (F_0) was set at 133 Hz, and fell linearly to 126 Hz during the last 100 ms of the final vowel. The final vowel was always 170 ms from the end of the CV transition to the beginning of the final fall of AV. Figure 6 shows a spectrogram of the stimulus shown schematically in Fig.

TABLE I. The F_2 and F_3 formant frequencies for the seven stimulus levels of place of articulation in the present experiment.

	Place	F_2	F_3
Labial	1	1345	2397
	2	1425	2614
	3	1510	2770
	4	1600	2934
	5	1695	3020
	6	1745	3109
Aveolar	7	1796	3200

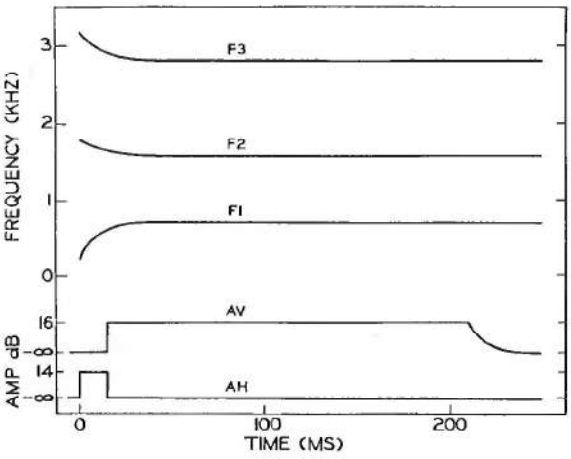

FIG. 5. Schematic diagram of the parameters of a synthesized speech sound.

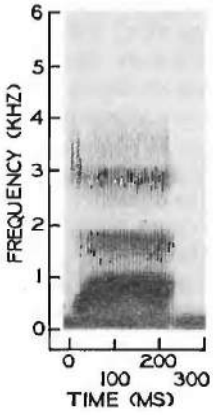

FIG. 6. Spectrogram of the synthesized speech sound represented schematically in Fig. 5.

5. All stimuli used in these experiments had this basic form.

The output of the speech synthesizer was bandpass filtered 20–5000 Hz (KHRON-HITE model 3500 R), amplified (McIntosh model MC-50), and presented over KOSS PRO-4AA headphones at 65 dB SPL (A) during the steady-state vowel and 60 dB SPL (A) during the aspiration period. The subjects sat in individual sound attenuated rooms.

2. Procedure

Each trial began with the presentation of a CV, selected randomly without replacement within blocks of 49 trials. The subjects were told to identify the CV as one of the four sounds, |bae|, |dae|, |pae|, or |tae|, and to make the best guess possible if they were not sure. The subject responded by pressing one of four buttons, labeled "B", "D", "P", and "T". The subjects were given 2 s to make their response. The end of the response interval was indicated by displaying "***" on a light-emitting diode display for 250 ms. The next trial began 1 s later.

The subjects participated in two 20-min sessions on each of two consecutive days. Between sessions the subjects took 5–10 min breaks. Five blocks of 49 trials were presented in each session. Unknown to the subjects, the first block of each session was not recorded. Thus, a total of 16 observations were collected for each subject for each of the 49 stimuli. Before the first session of the first day, the subjects responded to an additional practice block of 49 unscored trials.

3. Subjects

The subjects were 11 introductory psychology students who chose to participate for extra course credit. Subjects were run in groups of three or four.

B. Results and discussion

Figure 7 presents the percentage of times the stimuli were identified to be voiced phonemes[4] (|bae| or |dae|) plotted as a function of VOT; F_2–F_3 transitions is the curve parameter. An analysis of variance was carried out on the proportion of voiced identifications. The results show that VOT is a sufficient cue for voicing; the data go from completely voiced at a 10-ms VOT to completely voiceless at a 40-ms VOT, $F(6,60) = 154, P < 0.001$. Phoneme identifications were most ambiguous with respect to voicing at the 20 and 25 ms VOTs. Although VOT was a sufficient cue to voicing, Fig. 7 shows that F_2–F_3 transitions also had a consistent influence on voicing: The percentage of voiced phoneme identifications increased as the onset frequencies of the F_2–F_3 transitions decreased, $F(6,60) = 6.96, p < 0.001$. The significant interaction of VOT and the F_2–F_3 transitions, $F(36,360) = 3.54, p < 0.001$, reflects the fact that the influence of the F_2–F_3 transitions was largest at the intermediate values of VOT. The significant boundary shift contrasts with our previous finding of no voicing shift (Fig. 1) or the findings that the

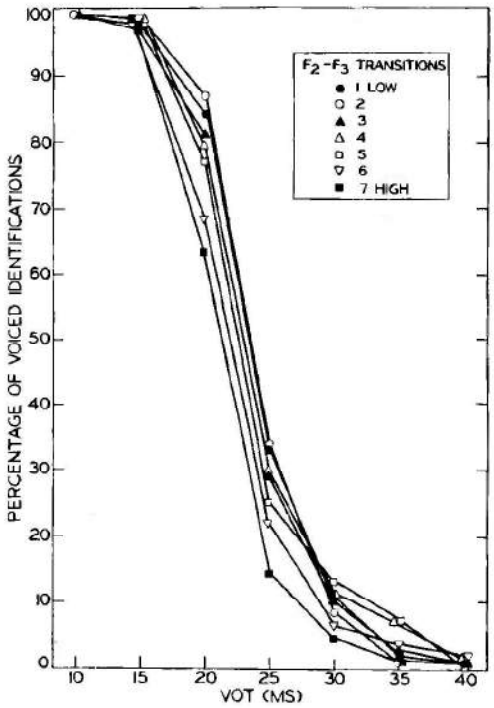

FIG. 7. Percentage of voiced identifications (/bae/ or /dae/) as a function of VOT; the level of F_2–F_3 transitions is the curve parameter.

voicing boundary is at shorter values of VOT for labial stops than it is for alveolar stops (Lisker and Abramson, 1970; Miller, 1977).

Figure 8 presents the percentage of times the stimuli were identified to be labial phonemes (|bae| or |pae|) as a function of the F_2–F_3 transitions; VOT is the curve parameter. An analysis of variance was carried out on the proportion of labial identifications. Replicating previous findings, phoneme identification with respect to place changed from labial to alveolar with increases in the onset frequencies of the F_2–F_3 transitions, $F(6,60) = 391, p < 0.001$. Sounds with low F_2–F_3 onset frequencies were consistently heard as labial phonemes whereas those with high frequencies were heard as alveolar phonemes. Intermediate frequencies gave more ambiguous identifications. Although the F_2–F_3 transitions were sufficient cues to place, VOT also had a relatively large influence on place, especially at the intermediate levels of the F_2–F_3 transitions, $F(6,60) = 25.2$ and $F(36,360) = 7.4$, both $p < 0.001$. Decreasing the VOT decreased the likelihood of a labial identification. The place boundary was at level 3 of the F_2–F_3 transitions for a VOT of 15 ms whereas it was at level 4 for a VOT of 35 ms. The decrease in the percentage of labial identifications with decreases in VOT was highly consistent except for an inversion with VOT of 10 ms. This general result agrees with those of Miller (1977) and Repp (1977). Any possible explanation for the reversal at a VOT of 10 ms is secondary to the more important issue of the dependence of labial identifications on VOT.

One important question is whether the shift in the

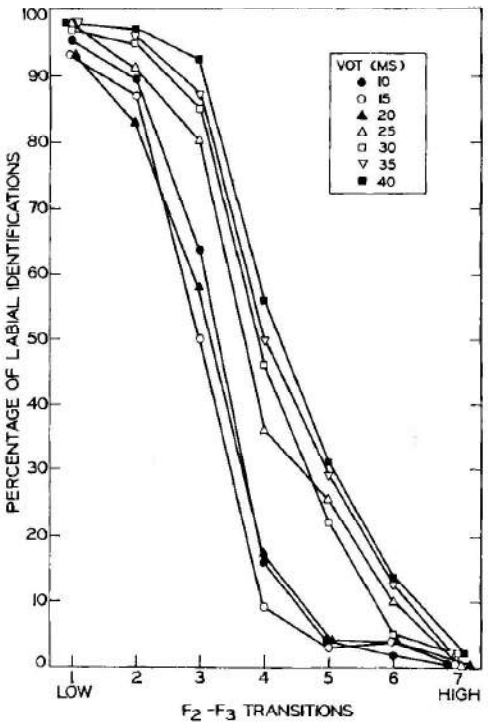

FIG. 8. Percentage of labial identifications (/pae/ or /bae/) as a function of F_2–F_3 transitions; VOT is the curve parameter.

place boundary is a relatively continuous function of VOT rather than simply an abrupt change at the voicing boundary. Figure 8 shows a relatively gradual shift from left to right with increases in VOT although the largest shift is between the 20 and 25 ms VOTs. These two VOTs lie on opposite sides of the voicing boundary shown in Fig. 7. To allow a closer examination of this question, the percentage of labial identifications as a function of VOT for each of the 11 subjects is given in Table II. Although most of the subjects show a relatively large shift between two adjacent VOTs, they also show significant shifts between most of the other adjacent VOT values. These data extend Miller's result by demonstrating that the shift in the boundary is a relatively continuous function of VOT rather than simply a change at the voicing boundary. This suggests that the boundary shift may be caused by the perception of VOT directly rather than by whether or not the sound is identified as voiced. That is, in contrast to Haggard's (1970) idea that the perception of one dimension is modified by the *phonetic* value of another dimension, it appears that the identification of phonemes with respect to one dimension is directly influenced by the *perceptual* value of another dimension.

TABLE II. The percentage of labial identifications as a function of VOT for each of the 11 subjects in experiment 1.

	VOT (ms)						
Subject	10	15	20	25	30	35	40
1	46	38	36	48	48	53	55
2	49	40	43	56	56	60	55
3	27	33	36	39	46	49	45
4	32	38	34	54	54	58	52
5	51	42	37	40	46	53	71
6	40	41	46	69	67	63	65
7	32	29	35	54	46	53	56
8	38	32	38	46	43	49	52
9	43	40	46	51	49	53	55
10	27	19	18	40	49	53	59
11	43	36	41	42	47	46	52
Average	39	35	37	49	50	54	56

In order to examine the nature of the interaction between the VOT factor and the F_2–F_3 transition factor in more detail, it is necessary to look at the pattern of the data separately for each of the response alternatives. The four panels of Fig. 9 present the percentage of |bae|, |pae|, |dae|, and |tae| identifications, respectively, as a function of the F_2–F_3 transition and of VOT. The levels along the abscissa are not equally spaced but rather have been adjusted to be proportional to the differences between the respective marginal means across the levels of the F_2–F_3 transitions. The differences were computed separately for each of the four response alternatives and then averaged over response types so that all four of the panels have the same spacing along the abscissa.

In general, the pattern of the results shown in Fig. 9 replicates that obtained by Oden and Massaro (1978): For each response alternative, the pattern is that of a gradually diverging fan of curved lines. The fact that the lines form a diverging fan indicates that the information about the features associated with the two independent variables are combined in a multiplicative rather than additive (see Massaro and Cohen, 1976) or other (see Oden, 1977) fashion for each of the candidate phonemes. Furthermore, the fact that the lines are generally curved rather than straight indicates that something in addition to a simple multiplication of independently evaluated featural values is involved since the simple fuzzy logical model predicts straight lines when the abscissa is spaced in the manner of Fig. 9 (see Oden and Massaro, 1978). That is, in these graphs, it is the curvature of the lines which reveals the presence of the effects which have previously been discussed in terms of boundary shifts.

For example, we saw in Fig. 8 that a VOT of 40 ms led to many more identifications of labial phonemes than did a VOT of 15 ms. In Fig. 9, this effect corresponds to the fact that with a 40-ms VOT, there are fewer |tae| responses than would be obtained with the simple multiplication of independent acoustic featural values, whereas with a 15-ms VOT, there are more |dae| responses than would be obtained with this simple model. In other words, compared to the fan of straight lines pattern that the simple fuzzy logical model would produce, Fig. 9 reveals that the curves for |tae| are convex downward and those for |dae| are convex upward so that overall there were more labial responses with longer VOTs than with shorter VOTs

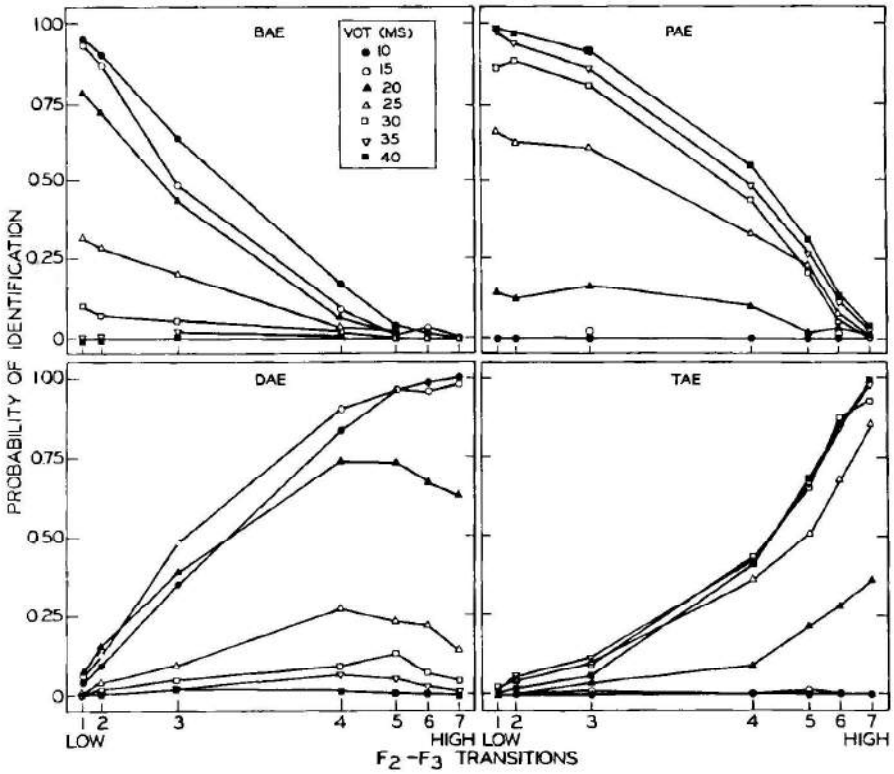

FIG. 9. Percentage of /bae/, /pae/, /dae/, and /tae/ identifications as a function of VOT and F_2–F_3 transitions. Note that the spacing along the abscissa is roughly proportional to the spacing of the marginal means across the levels of the F_2–F_3 transitions.

for any particular level of the F_2–F_3 transitions factor. Furthermore, Fig. 9 makes it apparent that this effect is a general, systematic effect which holds over the entire stimulus range.

This particular pattern of curvature for each of the different phonemes was also obtained in the previous study (Oden and Massaro, 1978). However, the curvature in the present results is considerably more pronounced, especially for the |dae| and |tae| responses. In fact, as can be seen in the lower left panel of Fig. 9, with VOTs of 20–35 ms, the curves for the |dae| identification percentages are nonmonotonic: Some of the intermediate F_2–F_3 onset frequencies produce more |dae| responses than do the highest frequencies. This is a very reliable effect, showing up in the data for all 11 subjects at either the 20 or 25 ms VOT levels and for more than half of the subjects at both. In fact, the only instances which did not produce these effects were for VOT values at the individual subject level for which the |dae| responses were either at the floor or ceiling of the probability scale.

The nonmonotonic curves for |dae| may at first glance seem unreasonable since high rather than intermediate F_2–F_3 onset frequencies are characteristic of alveolar phonemes as is reflected in the fact that the stimuli which were identified to be |dae| 100% of the time have the highest F_2–F_3 onset frequencies. However, on further inspection of Fig. 9, it becomes clear that the percentage of |dae| identifications must increase in this way because of the plummetting decrease in |tae| identifications as the F_2–F_3 onset frequencies are decreased. Thus, it may well not be that the |dae| prototype matches the stimulus better absolutely for the intermediate than for the higher onset frequencies but rather that the |tae| prototype matches *much* less well for the intermediate than for the higher frequencies so that the *relative* goodness of match of |tae| decreases markedly and the *relative* goodness of match of |dae| actually increases at the intermediate frequencies. In other words, this particular effect appears to be the result of the fact that decreasing the F_2–F_3 onset frequencies causes a greater decrease in the goodness of match to the stimulus for |tae| compared to the corresponding decrease for |dae|. This can be described naturally as resulting from |tae| requiring more extreme values of the F_2–F_3 onset frequencies. Decreases in the degree to which it is true that the stimulus has high F_2–F_3 onset frequencies produce even greater decreases in the degree to which it is true that it has *quite* high F_2–F_3 onset frequencies.

This underlying phenomenon can account for some of the influence of VOT on the percentage of labial phoneme identifications. If |bae| and |pae| are represented in their prototype as having equivalent F_2–F_3 onset frequencies, then the place boundary for voiced stimuli will differ from that for voiceless stimuli since |tae| is represented as having higher F_2–F_3 onset frequencies relative to |dae|. It is also interesting to note that this same underlying process could account for much of the influence of F_2–F_3 transitions on the percentage of voiced phoneme identifications. As can be seen in Fig. 7, this effect was obtained only for VOTs from 20 to 35 ms and is primarily a matter of the highest F_2–F_3 onset frequencies leading to fewer voiced phoneme identifications. This corresponds directly to the decrease in |dae| responses in this region of the F_2–F_3 transitions dimension (see Fig. 9).

On the other hand, it should be emphasized that this single phenomenon cannot account for all of the interactive effects in the data. For example, if there were no other processes at work, the percentage of |bae| responses for completely labial stimuli at a given level of VOT would always be the same as the percentage of |dae| responses for completely alveolar stimuli at the same level of VOT since both response probabilities would depend only on the VOT feature value. This relationship is clearly not found in the present data as can be seen in Fig. 9 for VOTs of 20 and 25 ms. The probability of a |bae| response was significantly greater than a |dae| response at these two VOTs. Therefore, some other process must be involved as well. Among the possibilities are that there are modifiers in other phoneme prototypes or that aspiration also functions as an acoustic feature. These hypotheses will be evaluated using the quantitative pattern of the data.

1. Quantitative assessment of the models

To provide a more precise evaluation of the merits of the alternative hypotheses, several versions of the fuzzy logical model were fit to the mean data using subroutine STEPIT (Chandler, 1969) with a least-squares criterion of fit. As a baseline against which to compare the other models, the simple fuzzy logical model, in which only F_2–F_3 transitions and VOT features are used as in Eqs. (1) and (4), was fit to the data and resulted in a root mean squared deviation (RMSD) of the predictions of the model from the data of 0.080. This degree of goodness of fit is fairly good considering that 147 independent data points are being fit using 14 parameters, one for each of the seven levels of each of the two independent variables.[5] Table III gives the values of the 14 parameters. However, as has already been observed, the data deviate systematically from the fan of straight lines that this model predicts for the plots in Fig. 9. In general, systematic deviations are as diagnostic of the inadequacy of a model as is a poor fit which could be simply due to noisy data.

To evaluate the relative ability of the various alternative hypotheses to account for the overall pattern of the data, three general classes of model were devised and fit to the data. These were (a) a model incorporating featural nonindependence, (b) models including modifiers in the pattern prototype descriptions, and (c) a model in which aspiration serves as an independent acoustic feature relevant to the place of articulation distinction.

TABLE III. Parameter values for the simple fuzzy logical model fit to the data of experiment 1.

VOT (ms)	VOT cue	F_2–F_3 transition		F_2–F_3 cue
10	0.999	(low)	1	0.031
15	0.991		2	0.075
20	0.764		3	0.253
25	0.270		4	0.661
30	0.088		5	0.824
35	0.036		6	0.928
40	0.012	(high)	7	0.995

The first of these types of models, the featural nonindependence model, was quantified within the framework of the fuzzy logical model by assuming that the perception of the F_2–F_3 transitions was directly influenced by the VOT value. This means that the HFT value in Eq. (4) can change not only with the level of the F_2–F_3 transitions but also with the level of VOT. Given that the nature of nonindependence is not known, a free parameter for HFT must be estimated for each of the 49 stimulus conditions. In addition, seven parameter values must be estimated for the SV values which can change as a function of VOT. The fit to the data was improved, relative to the simple fuzzy logical model, resulting in a RMSD of 0.032. Table IV gives the parameter values. Assessing this improvement in the description of the data will be delayed until the other descriptions of boundary shifts are tested.

The second type of model considered was the "prototype modifiers" model. The qualitative evidence presented in the discussion of the results is consistent with the idea that a modifier on the F_2–F_3 transitions feature for |tae| is required. Therefore, a single exponent parameter was incorporated into the simple model to represent such a modifier as illustrated in Eq. (5). This resulted in an RMSD of 0.045 which is quite good being little more than half of that for the simple model, even though only one more parameter was used. Table V gives the parameter values for this description of the data. Thus, this greatly better fit achieved by only one additional parameter verifies the importance of including a place modifier in the |tae| prototype.

However, it was also observed in the discussion of the results that there were other systematic effects to be accounted for in the data. The hypothesis of additional prototype modifiers was considered by developing a model using 22 parameters: The 15 used in the previous model plus an additional seven for the other possible modifiers. Table V gives the parameter values. The RMSD for this model with all eight prototype modifiers (the "complex" fuzzy logical model of Oden and Massaro, 1978) was 0.023. Thus, this model provides a better fit to the data than the nonindependence model and with 34 fewer parameters.

The third type of model is based on the idea that aspiration can function as an independent acoustic fea-

TABLE IV. Parameter values for the featural nonindependence model fit to the data of experiment 1.

F_2–F_3 cue VOT (ms)	F_2–F_3 transition 1	2	3	4	5	6	7
10	0.045	0.095	0.358	0.839	0.962	0.984	0.999
15	0.058	0.124	0.503	0.911	0.974	0.969	0.999
20	0.047	0.149	0.446	0.897	0.964	0.942	0.940
25	0.039	0.092	0.177	0.564	0.709	0.897	0.999
30	0.033	0.039	0.118	0.503	0.748	0.955	0.996
35	0.001	0.034	0.119	0.475	0.700	0.869	0.992
40	0.012	0.026	0.066	0.440	0.690	0.863	0.981
VOT cue	0.998	0.989	0.774	0.261	0.086	0.029	0.005

TABLE V. Parameter values for three of the versions of the fuzzy logical model fit to the data of experiment 1.

Parameter	Single prototype modifier	Eight prototype modifiers	Aspiration as an independent feature
VOT cue			
VOT level: 10	1.000	0.939	0.999
15	0.998	0.920	0.992
20	0.723	0.646	0.769
25	0.211	0.442	0.268
30	0.070	0.334	0.087
35	0.062	0.264	0.033
40	0.015	0.189	0.007
F_2–F_3 cue			
F_2–F_3 transition			
level: 1 (low)	0.056	0.063	0.015
2	0.129	0.118	0.038
3	0.450	0.300	0.178
4	0.708	0.567	0.655
5	0.878	0.693	0.833
6	0.948	0.829	0.943
7 (high)	1.000	1.000	0.998
Aspiration cue			
VOT level: 10			0.273
15			0.183
20			0.207
25			0.606
30			0.631
35			0.677
40			0.705
Feature modifiers:			
VOT of /t/		1.82	
VOT of /d/		3.01	
VOT of /p/		3.00	
VOT of /b/		2.95	
F_2–F_3 of /t/	7.82	2.73	
F_2–F_3 of /d/		1.06	
F_2–F_3 of /p/		1.27	
F_2–F_3 of /b/		3.03	

ture. It was argued that the aspiration of our synthetic speech without bursts is more representative of the burst and aspiration of labials than of alveolars. Accordingly, within the framework of the fuzzy logical model, the prototypes for D and T were defined as not having low-frequency aspiration as in Eqs. (3) and (6) and the prototypes for B and P were defined as having low-frequency aspiration. Seven parameter values were estimated for the aspiration feature (which was allowed to change with changes in VOT) in addition to the 14 parameter values for the VOT and F_2–F_3 features. The RMSD for this model was 0.034, an improvment equal to the nonindependence model and with 35 fewer parameters. The parameter values for this model are given in Table V.

Comparing the success of all of the models considered, both the prototype modifiers model and the independent aspiration feature model were able to provide a very good fit with a fairly small number of parameters. In fact, the model with eight prototype modifiers provided the best fit to the data of all of the models and did so with only eight more parameters than the simple fuzzy logical model that was used as a baseline for comparison.

In comparing the two most successful models, it should be emphasized that the fact that the inclusion of the eight prototype modifiers results in a better fit to the data than does the inclusion of aspiration as a feature can by no means be taken as conclusive evidence that aspiration does not also serve as an acoustic feature. Not surprisingly, when place values for aspiration are included in the model along with all eight prototype modifiers, the fit to the data improves still further: RMSD = 0.020 using 29 parameters. Of course, one must be cautious when evaluating an improvement of this sort since some of the improvement may just be due to absorbing more of the noise with the extra parameters. However, the fit is better than that of the nonindependence model with 56 parameters. Furthermore, the aspiration parameter values which resulted from fitting this model covered a sizeable range (from 0.21 to 0.63) and as predicted, were monotonic functions of VOT with longer aspiration leading to more labiality. Thus, it seems unlikely that these parameters could be only fitting error. Rather, they appear to be capturing some systematic trend

other than that accounted for by the prototype modifiers. Nevertheless, a more definitive evaluation of the possible role of aspiration in perception requires a direct manipulation of the amount of aspiration in the stimuli. Such a manipulation was performed in experiment 2.

II. EXPERIMENT 2

The major intent of the first experiment was to extend and quantify previous observations of shifts in the place and voicing boundaries for stop consonants. It is clear from the data that although the probability of a labial identification is primarily dependent on the F_2–F_3 transitions, it also varies systematically with the VOT. Similarly, the identification of phonemes with respect to voicing varies systematically with the F_2–F_3 transitions even though it is primarily dependent on VOT. Furthermore, the fuzzy logical model of speech perception provides a good account for these data including the boundary shifts while still maintaining independent evaluation of the acoustic features.

However, while there was direct evidence in support of the hypothesis that at least part of the boundary shift effects is due to some phoneme prototypes requiring more extreme values for specific acoustic features, it is still not clear whether or not the use of aspiration as an acoustic feature relevant to the place distinction is also involved. If aspiration functions in this way, then varying the degree of aspiration in some other way than VOT should systematically change the value of the feature. Furthermore, if increasing the amount of aspiration by increasing the VOT influences the feature value, then presumably increasing the amount of aspiration by increasing its intensity should also influence the feature value. Therefore, to obtain more specific information on the role of aspiration in stop consonant identification, a second experiment was performed in which the intensity of the aspiration during the VOT period was varied factorially with VOT and the F_2–F_3 transitions.

A. Method

1. Stimuli

The stop CV stimuli were produced in the same manner and had the same structure as those in experiment 1. Eighty-four unique speech sounds were produced by the independent variation of seven levels of F_2–F_3 onset frequencies, three levels of VOT, and four levels of aspiration intensity. The seven F_2–F_3 transition values were identical to those in experiment 1, the VOT values were 15, 25, and 35 ms, and the four aspiration intensity values were 28, 34, 39, and 51 dB SPL (A). Figure 10 presents spectrograms of the stimuli for each of the four aspiration intensities with the fourth level of the F_2–F_3 transitions and a VOT of 35 ms.

2. Procedure

The procedure was identical to experiment 1 except that the stimuli were sampled randomly without replacement from the population of 84 stimuli. The sessions consisted of four blocks of 84 trials.

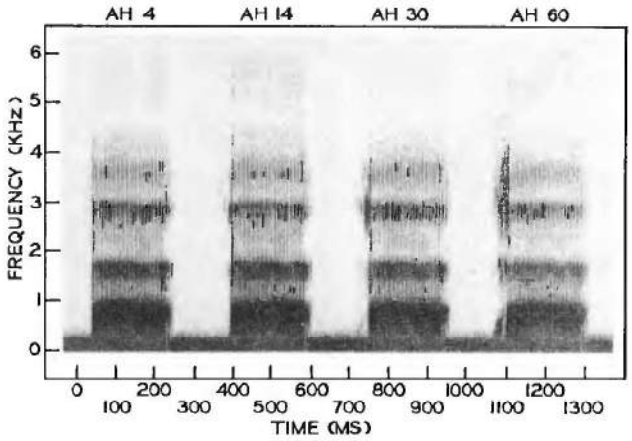

FIG. 10. Spectrograms of the stimuli for each of the four aspiration intensities with the fourth level of place of articulation and a VOT of 25 ms. Note that the AH values refer to the synthesizer's parameters of aspiration intensity.

3. Subjects

Eight subjects sampled from the same population as those in experiment 1 were tested for two sessions on each of two days.

B. Results and discussion

The primary result of interest in the present study is how phoneme identification with respect to place depends on VOT and aspiration intensity. Figure 11 presents the percentage of labial phoneme identifications as a function of F_2–F_3 transitions; VOT is the curve

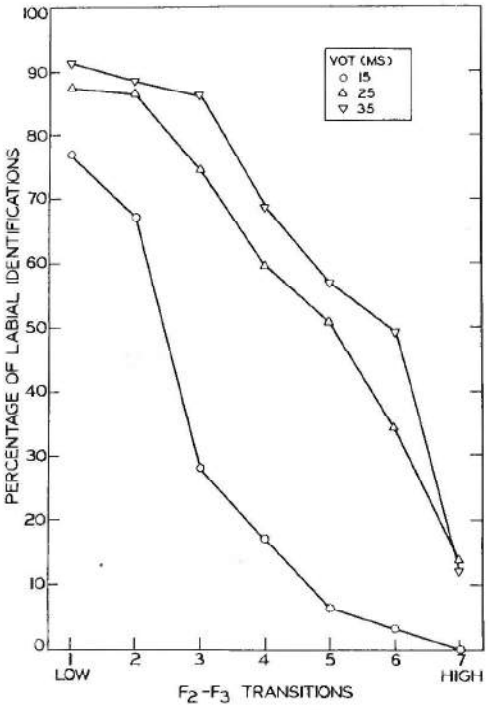

FIG. 11. Percentage of labial identifications as a function of the level of the F_2–F_3 transitions; VOT is the curve parameter.

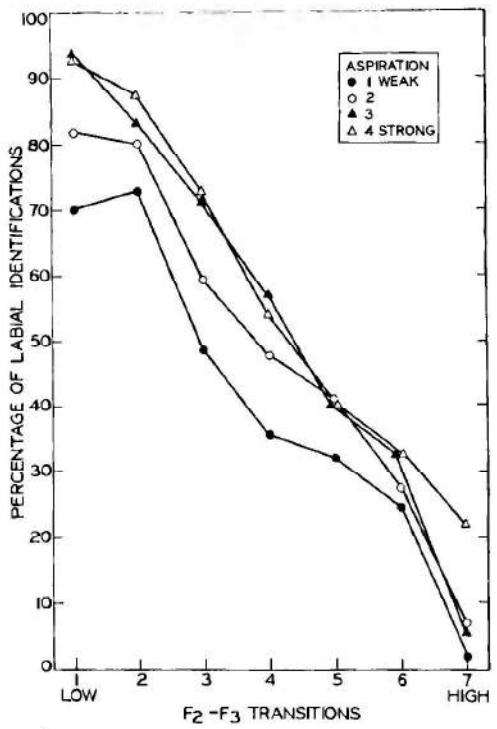

FIG. 12. Percentage of labial identifications as a function of the level of the F_2–F_3 transitions; aspiration level is the curve parameter.

parameter. In replication of experiment 1, there are very large effects of both variables on place identifications with the sounds being more likely to be identified as labial for low F_2–F_3 transitions and long VOTs. An analysis of variance of the percentage of place identifications gave significant effects of the F_2–F_3 transitions, $F(6,42)=145$; and of VOT, $F(2,14)=45.5$, both $p<0.001$. Although the F_2–F_3 transitions were very important for the place judgment, a VOT of 35 ms resulted in a place boundary about $3\frac{1}{3}$ levels further toward the alveolar end than that with a VOT of 15 ms.

Figure 12 presents the percentage of labial phoneme identifications as a function of F_2–F_3 transitions; aspiration level is the curve parameter. The place boundary is dependent on aspiration intensity, $F(3,21)=13.5$, $p<0.001$ although the boundary shifts are not as large as those as a function of VOT (see Fig. 11). However, it might be that a wider range of intensity levels would produce shifts that were more comparable to those found for VOT. In any event, this result provides direct evidence that aspiration intensity affects the place boundary. Furthermore, the direction of the shift is what was predicted by the aspiration feature hypothesis: Since aspiration in synthetic speech appears to be more similar to the burst and aspiration of labial than alveolar stop consonants, increasing aspiration intensity makes the stimuli sound more labial.

1. Quantitative tests of the models

The simple fuzzy logical model, in which the feature of F_2–F_3 transitions functions as a cue to the place distinction and VOT-aspiration feature functions as a cue to the voicing distinction, was fit to the data using a unique parameter for each of the 12 combinations of VOT and aspiration intensity along with seven parameters for each of the levels of F_2–F_3 transitions. Table VI presents the parameter values. The resulting RMSD of 0.137 is markedly poorer in absolute terms than that for the simple model in experiment 1 but is still fairly good considering that 252 independent data points were fit using 19 parameters.

A nonindependence model was not fit to the data since a prohibitive number of parameters would have to be estimated. If both VOT and aspiration intensity are

TABLE VI. Parameter values for the simple fuzzy logical model and the complex model with eight prototype modifiers fit to the data of experiment 2.

VOT-aspiration cue for voicing distinction: VOT (ms)	Aspiration intensity: 1	2	3	4
15	0.965	0.976	0.998	0.969
25	0.504	0.421	0.353	0.170
35	0.185	0.144	0.046	0.022

F_2–F_3 cue for place distinction	F_2–F_3 transitions: 1	2	3	4	5	6	7
	0.159	0.203	0.396	0.535	0.644	0.726	0.919

Feature modifiers	
VOT-aspiration of /t/	0.91
VOT-aspiration of /d/	1.51
VOT-aspiration of /p/	1.70
VOT-aspiration of /b/	0.61
F_2–F_3 of /t/	2.36
F_2–F_3 of /d/	0.78
F_2–F_3 of /p/	0.61
F_2–F_3 of /b/	2.76

TABLE VII. Parameter values for the complex model assuming that aspiration functions an independent feature for the place distinction.

VOT-aspiration cue for voicing distinction VOT (ms)	Aspiration intensity 1	2	3	4
15	0.978	0.980	0.995	0.971
25	0.507	0.419	0.348	0.161
35	0.200	0.149	0.037	0.014

VOT-aspiration cue for place distinction VOT (ms)	Aspiration intensity 1	2	3	4
15	0.117	0.166	0.224	0.202
25	0.433	0.634	0.692	0.784
35	0.593	0.710	0.789	0.783

F_2–F_3 cue for place distinction	F_2–F_3 transitions 1	2	3	4	5	6	7
	0.071	0.102	0.348	0.533	0.664	0.766	0.939

assumed to influence the cue value of the F_2–F_3 transitions, 84 parameters would be necessary to describe the 12 VOT-aspiration conditions times the seven F_2–F_3 transitions conditions. And, of course, 12 additional parameters would be necessary to describe the cue values of the VOT-aspiration conditions, giving a total of 96 in all.

The model with prototype modifiers was fit to the data by taking the parameter values given by the fit of the simple model and including eight additional parameters to represent prototype modifiers (see Table VI). The fit was greatly improved, RMSD = 0.058.

The model representing aspiration as a cue to the place distinction was fit to the data by expanding the simple model to include 12 additional parameters to represent the place cue value of each VOT and aspiration intensity combination. Table VII presents the parameter values. In this case, the fit was also improved to an analogous degree, RMSD = 0.053.

Including *both* the prototype modifiers and the aspiration as a cue to the place distinction resulted in very little additional improvement, RMSD = 0.050. Thus, the two complex models of prototype modifiers and aspiration place cues appear to account for the same quantitative aspects of the data pattern and do so essentially equally well, even though they are based on widely different psychological principles.

III. GENERAL DISCUSSION

The present article has examined the processes by which features are evaluated and integrated during speech perception. The particular focus of the experiments has been on the processes which underly the interactive effects that result from the manipulation of VOT and F_2–F_3 transitions of stop consonants. Three explanations for these effects have been considered: (1) nonindependence of feature evaluation, (2) modifiers in the phoneme protoypes, and (3) aspiration cues to the identification of place. The following sections will discuss additional considerations with respect to each of these explanations.

A. Featural nonindependence

One type of featural nonindependence examined in the Introduction was Haggard's (1970) suggestion that the evaluation of VOT is influenced by the prior phonetic decision about whether the phoneme is labial or alveolar. Strong evidence against this notion is the finding of boundary shifts of labial identifications as a function of VOT. If a phonetic decision about place of articulation influences the perception and evaluation of VOT, then the place decision must be made before VOT is evaluated. It follows that under this hypothesis, VOT cannot influence the place judgment and yet it does.

However, there are other possible types of featural nonindependence. One such possibility is that there are complex low-level auditory interactions so that, for example, the perceptual realization of VOT is modified by the F_2–F_3 transition frequencies. Analogously, the perceptual realization of the F_2–F_3 transitions could be modified by VOT. According to this view, VOT and F_2–F_3 transitions maintain their role as the acoustic features of voicing and place, respectively, but the value of each feature is dependent on the stimulus value of the other. As an example of such an interaction in nonspeech stimuli, changing the perceived hue of a color from green to blue by changing the wavelength also changes the perceived brightness since we are less sensitive to wavelengths in the blue than in the green part of the visible spectrum. An experiment in color perception which would be analogous to the present studies would independently vary the wavelength and intensity of the color and the results would presumably show that wavelength not only influences the perception of hue but also the perception of brightness.

The strongest evidence to date in favor of a feature-interaction description of speech perception is that obtained from research with chinchillas by Kuhl and

Miller (1978). These animals are of particular interest because they have auditory capabilities that are very similar to those of humans. Kuhl and Miller trained chinchillas to discriminate and respond differentially to synthetic voiced and voiceless alveolar stops. After training they were tested in a generalization paradigm on the alveolar VOT continuum of Lisker and Abramson (1970). The resulting labeling functions were nearly identical to those obtained with human listeners. The animals were also tested on the labial and velar VOT continua and the chinchillas produced labeling functions very similar to those produced by humans: Most importantly, the chinchilla VOT boundary increased as place of articulation changed from labial to alveolar to velar. In the Introduction of the present article, it was argued that the boundary shifts observed by Lisker and Abramson (1970) might have been due to the contribution of other covarying independent cues to voicing, such as the onset frequency of F_1. However, given that chinchillas, being alinguistic, must have been responding to some perceptual realization of VOT rather than to the abstract feature of voicing, then the onset frequency of F_1 should not have affected the boundary unless it directly affected the perceptual realization of VOT. It remains a possibility, however, that humans and chinchillas produce the same pattern of data as a result of different processes.

The difficulty with the auditory interactions hypothesis is that, without further specification of the nature of the underlying psychoacoustic processes which could produce such interactions, it is capable of predicting any conceivable pattern of results and cannot, therefore, be disproven. It is even possible that all of the results of the present experiments are produced by auditory interactions and that the mathematical structure of the fuzzy logical model also closely reflects low-level acoustic processes. However, there is no single clear interpretation of the components of the model in terms of psychoacoustic processes whereas as a model of speech feature integration it has a straightforward and natural interpretation in terms of the psychological theory from which it was directly derived. Consequently, until some independent motivation for the auditory interactions hypothesis is provided, it may be less preferable than the independent feature evaluation hypotheses and, as the present experiments have shown, it is not necessary to appeal to featural nonindependence in order to account in detail for the types of boundary shifts which have been obtained so far.

B. Phoneme prototype modifiers

The second explanation for the boundary shift effects is that they result from the fact that more extreme values on some acoustic features may be required for certain phonemes compared to that required for other phonemes. This has a natural representation in the fuzzy logical model of Oden and Massaro (1978) in terms of modifiers of the acoustic features in the prototypes of the specific phonemes. For convenience, we express these modifiers by the use of words such as "quite" or "very" but this should not be construed to mean that surface words occur in the prototypes; rather, the modifiers in the actual prototypes would be abstract underlying semantic intensifiers which may or may not correspond directly to the meaning of any particular surface word.

A somewhat peripheral attraction of the prototype modifiers hypothesis is that it emphasizes the central importance for the fuzzy logical model of the cognitive nature of the phoneme prototypes. In contrast to feature evaluation which is assumed to be performed by low-level automatic sensory-perceptual processes (below the level of primary recognition in the information processing model of Massaro, 1975), the integration processes follow the logical recipe specified by the prototypes which are assumed to be stored in long term memory. Consequently, it is natural to think of the prototypes as being relatively changeable rather than being immutable components of the sensory hardware. While presumably not directly under conscious control, changes in the modifiers of specific acoustic features in specific phonemes would be a simple way for the human speech perception system to adjust the identification of phonemes in order to more accurately reflect the structure of the stimuli being identified. It is easy to imagine, for example, that the learning of new dialects or the compensation for specific types of environmental noise might be accomplished in this way.

More long term adjustments to the structure of one's language may, in part, account for the particular configuration of phoneme prototype modifiers that is necessary to produce the obtained boundary effects. It could be the case that the independent acoustic cues of VOT and F_2–F_3 transitions are somewhat correlated in natural speech. Experience with the stop consonants in natural speech would allow the listener to modify the prototype for each stop to incorporate whatever correlation exists. Evidence for a correlation of VOT and F_2–F_3 transitions in natural speech was given by Lisker and Abramson (1964) who found that the average VOTs of voiceless stops differed for the three places of articulations; VOTs tend to increase as place of articulation moves from labial to alveolar to velar. Incorporating this information into the prototypes in long term memory would allow the listener to require a labial stop to have a shorter VOT than the analogous alveolar stop in order to hear it as voiced.

However, Lisker and Abramson's (1964) measurements would imply that listeners should hear a sound with an intermediate VOT as voiceless if it is labial but as voiced if it is alveolar and this empirical result was not obtained here although it was in the Lisker and Abramson (1970) and to some extent in the Miller (1977) study. In the present case, though, the effects of the voicing modifiers are probably overwhelmed by the larger effect due to the extreme modifier on place for the |tae| prototype.

A more local change in the modifiers may have occurred if the subjects attempted to adjust their identifications to come closer to choosing each phoneme an equal number of times. That subjects might do this would by no means make the obtained data pattern ar-

tifactual since presumably both this adjustment and also the integration processes used after making the adjustment would reflect everyday, naturally occurring processes involved in normal speech perception. However, since the marginal response probabilities in our experiments are highly unequal, differing in some cases by as much as two to one, if such an adjustment does occur it is by no means complete.

In any case, the nonmonotonicity of the curves for the intermediate VOT levels with |dae| responses in experiment 1 provides qualitative support for the existence of modifiers of acoustic features in the phoneme prototypes. In addition, the good fit of the fuzzy logical model with prototype modifiers to the data of experiment 1, as well as the success of this model in accounting for the data of Oden and Massaro (1978), provides some degree of quantitative support for this hypothesis.

C. Aspiration feature

The third explanation for the boundary shift effects is that they may result from the existence of an aspiration feature for the place dimension which covaries with the manipulation of VOT cues for the voicing dimension. In particular, the present article examined the possibility that the aspiration which occurs during the VOT period may cue the perception of place so that changing the amount of aspiration by changing the VOT significantly influences the identification of phonemes with respect to place. Qualitative evidence was provided for this hypothesis by experiment 2 in that manipulation of the intensity of the aspiration was found to cause a shift in the place boundary. Thus, this shows that aspiration must serve as a direct cue to place rather than merely having an indirect effect on place through its interaction with voicing produced by the prototype modifiers. This effect can also account for the place boundary shifts obtained by Miller (1977) and Repp (1977).

In addition to the role of aspiration, other acoustic features could play some part in producing boundary shifts. Analogous to the VOT manipulation providing an independent feature to place, it might be assumed that the F_2–F_3 transition frequencies provide an additional feature to voicing. Stevens and Klatt (1974) showed that significant F_1 and F_2 transitions after the onset of voicing provide additional evidence for the identification of voiced phonemes. In a following study, Lisker (1975) showed that the effect of the F_1 transition was actually due to the onset frequency of F_1; a low onset frequency provides an important feature to a voiced sound. If this result could be generalized to F_2 then a low F_2 might provide an acoustic feature to voicing. Some measure of support for this hypothesis is provided by other results of Lisker (1975) which indicate that a higher starting frequency for F_2 increases the voiceless quality of velar stop consonants. However, it is possible that differences in F_1 may also account for some of these differences. It is also possible that the frequency information from F_1 and F_2 is partially smeared and, therefore, the frequencies of both are important. A low F_2 providing an independent feature to a voiced sound would be in qualitative agreement with the finding that the voicing boundary occurs at shorter VOTs for low F_2–F_3 transitions than for high F_2–F_3 transitions. Much experimental work will be required to assess the role of all of the possible acoustic features of both place and voicing.

D. Conclusions

The present experiments have provided some evidence that both phoneme prototype modifiers and also an aspiration feature to the place distinction are involved in producing the observed dependency between the variables of VOT and F_2–F_3 transition in phoneme identification. In addition, the success of the fuzzy logical model in providing a quantitative account of the present data demonstrates that the observed interactive effects do not necessitate the abandonment of the assumption of independent acoustic feature evaluations.

ACKNOWLEDGMENTS

This research was supported by National Institute of Mental Health Grant MH 19399 and by National Science Foundation Grant BSN77-15820. Michael Cohen provided assistance in performing this research and Michael Cohen and Lola Lopes provided helpful comments and suggestions. The authors would also like to acknowledge the useful comments of an anonymous reviewer. Requests for reprints should be sent to Dominic Massaro or Gregg Oden, Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706.

[1]Given that only acoustic characteristics of the consonants are varied in the present experiments, we refer to the prototypes as *phoneme* prototypes for expository purposes only. Other evidence indicates that the prototypes must be, at least, at the syllable level (Derr and Massaro, 1978; Massaro, 1972, 1975, 1979). The assumption of syllable prototypes contrasts with the commonly accepted notion of phonetic or phonemic prototypes in which phonetic or phonemic decisions mediate speech perception. Although we may occasionally refer to "the perception of voicing" as a shorthand for "the identification of the stimulus as a voiced speech sound," in the present model, the perception of voicing qua voicing does not mediate syllable recognition. Similarly, stating "aspiration functions as an acoustic feature to place" only means that the aspiration feature distinguishes between labial and alveolar phonemes, not that the listener processes place in the course of syllable identification. Accordingly, the conscious realization that two sounds differ in voicing is presumed to occur only with postperceptual introspection (see Paap, 1975, pp. 175–177 for an illuminating discussion of this issue).

[2]The representation of the voiceless VOT level as NOT (SHORT VOT) is not meant to be a claim about which feature is marked and which unmarked since we are not sure. If the voiceless end of this dimension is psychologically unmarked then it would be represented as LONG VOT and its opposite as NOT(LONG VOT).

[3]The amplitudes given here are the OVE-IIId amplitude parameters used, and not the levels at the subjects' ears. They

are given for the benefit of those readers who are familiar with the OVE.

[4]The use of the terms voiced and labial phonemes is for ease of expositon and should not imply that these linguistic concepts are psychologically real.

[5]Given the range of F_2–F_3 transitions used in the present experiments, it was assumed that all stimuli would be either labial or alveolar. Therefore, only the degree of labiality, L_i, is needed to be estimated directly for each F_2–F_3 transition level and the degree of alveolarity, A_i, was computed as $1-L_i$.

Anderson, N. H. (1974). "Information integration theory: A brief survey," in *Contemporary Developments in Mathematical Psychology*, edited by D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Freeman, San Francisco), Vol. 2.

Campbell, H. W. (1974). "Phoneme recognition by ear and by eye: A distinctive feature analysis," dissertation, University of Nijmegen, Netherlands.

Carney, A. E., Widin, G. P., and Viemeister, N. F. (1977). "Noncategorical perception of stop consonants differing in VOT," J. Acoust. Soc. Am. 62, 961–970.

Chandler, J. P. (1969). "Subroutine STEPIT—Finds local minima of a smooth function of several parameters," Behav. Sci. 14, 81–82.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row, New York).

Cohen, M. M., and Massaro, D. W. (1976). "Real-time speech synthesis," Behav. Res. Meth. Instrum. 8, 189–196.

Cole, R. A., and Scott, B. (1972). "Distinctive feature control of decision time: Same-different judgments of simultaneously heard phonemes," Percept. Psychophys. 12, 91–94.

Cole, R. A., and Scott, B. (1974). "The phantom in the phoneme: Invariant cues for stop consonants," Percept. Psychophys. 15, 101–107.

Darwin, C. J. (1976). "The perception of speech," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), Vol. VII.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. 27, 769–773.

Derr, M. A., and Massaro, S. W. (1978). "The contribution of vowel duration, F_0 contour, and frication duration as cues to the /juz/–/jus/ distinction," WHIPP Report No. 8, Wisconsin Human Information Processing Program, Madison, September 1978 (unpublished).

Fant, G. (1973). *Speech Sounds and Features* (MIT, Cambridge, Mass.).

Goguen, J. A. (1969). "The logic of inexact concepts," Syntheses 19, 325–373.

Haggard, M. P. (1970). "The use of voicing information," Speech Syn. Percept. 2, 1–15.

Haggard, M., Ambler, S., and Callow, M. (1970). "Pitch as a voicing cue," J. Acoust. Soc. Am. 47, 613–617.

Harris, K. S., Hoffman, H. S., Liberman, A. M., Delattre, P. C., and Cooper, F. S. (1958). "Effect of third-formant transitions on the perception of the voiced stop consonants," J. Acoust. Soc. Am. 30, 122–126.

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

Jakobson, R., Fant, C. G. M., and Halle, M. (1961). *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates* (MIT, Cambridge, Mass.).

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

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., and Cooper, F. S. (1952). "The role of selected stimulus variables in the perception of the unvoiced stop consonants," Am. J. Psychol. 65, 497–516.

Liberman, A. M., Delattre, P., and Cooper, F. S. (1958). "Distinction between voiced and voiceless stops," 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, 384–423.

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

Luce, R. D. (1959). *Individual Choice Behavior* (Wiley, New York).

Lyons, J. (1968). *Introduction to Theoretical Linguistics* (Cambridge University, Cambridge, England).

Massaro, D. W. (1972). "Preperceptual images, processing time, and perceptual units in auditory perception," Psychol. Rev. 79, 124–145.

Massaro, D. W. (Ed.) (1975). *Understanding Language: An Information Processing Analysis of Speech Perception, Reading and Psycholinguistics* (Academic, New York).

Massaro, D. W. (1979). "Issues in Speech Perception," *Proceedings of the Ninth International Congress of Phonetic Sciences*, Institute of Phonetics, University of Copenhagen, Copenhagen, Denmark, 1979 (unpublished).

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

Massaro, D. W., and Cohen, M. M. (1977). "The contribution of voice-onset time and fundamental frequency as cues to the /zi/–/si/ distinction," Percept. Psychophys. 22, 373–382.

Miller, G. A., and Nicely, P. E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 26, 338–352.

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

Oden, G. C. (1977). "Integration of fuzzy logical information," J. Exper. Psychol. Human Percept. Perform. 3, 565–575.

Oden, G. C. (1978). "Integration of place and voicing information in the identification of synthetic stop consonants," J. Phonetics 6, 83–93.

Oden, G. C., and Hogan, M. E. (1977). "A fuzzy propositional model of negation on semantic continua," paper presented at the meeting of the Midwestern Psychological Association, Chicago, May 1977 (unpublished).

Oden, G. C., and Massaro, D. W. (1978). "Integration of featural information in speech perception," Psychol. Rev. 85, 172–191.

Paap, K. (1975). "Theories of speech perception," in *Understanding language: An Information Processing Analysis of Speech Perception, Reading, and Psycholinguistics*, edited by D. W. Massaro (Academic, New York).

Peters, R. W. (1963). "Dimensions of perception for consonants," J. Acoust. Soc. Am. 35, 1985–1989.

Repp, B. H. (1977). "Dichotic competition of speech sounds: The role of acoustic stimulus structure," J. Exp. Psychol. Human Percept. Perform. 3, 37–50.

Repp. B. H. (1977). "Interdependence of voicing and place decisions," Haskins Laboratories, New Haven CT, September 1977 (unpublished).

Samuel, A. G. (1977). "The effect of discrimination training on speech perception: Noncategorical perception," Percept. Psychophys. 22, 321–330.

Sawusch, J. R., and Pisoni, D. B. (1974). "On the identification of place and voicing features in synthetic stop conson-

ants," J. Phonetics **2**, 181–194.
Singh, S., and Woods, D. R. **(1971)**. "Perceptual structure of 12 American English vowels," J. Acoust. Soc. Am. **49**, 1861–1866.
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, W., and Haggard, M. **(1977)**. "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," J. Acoust. Soc. Am. **62**, 435–448.
Wang, M. D., and Bilger, R. C. **(1973)**. "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.
Winitz, H., LaRivere, C., and Herriman, E. **(1975)**. "Variations in VOT for English initial stops," J. Phonetics 3, 41–52.
Zadeh, L. A. **(1975)**. "The concept of a linguistic variable and its application to approximate reasoning—II," Inf. Sci. 8, 301–357.