On the Identification of Sine-Wave Analogues of Certain Speech Sounds*

Peter J. Bailey†, Quentin Summerfield†† and Michael Dorman†††

ABSTRACT

There is no obvious psychoacoustic basis for the perceptual categorization of synthetic stop consonant-vowel syllables according to place of production. Using a task that did not involve overt labeling, we compared the categorization of series of speech sounds formed by varying the onsets of the second and third formant transitions with the categorization of series of analogues of these sounds in which the formants were replaced with frequency- and amplitude-modulated sine-waves. These sine-wave patterns are perceived either as complex tones or as speechlike sounds in which a whistle is initiated by a stop consonant. Different category boundary positions accompanied these two percepts. When the sine-wave series were heard as speechlike, category boundaries were similar to those obtained with the formant stimuli. The demonstration that the perceptual categorization of an acoustic pattern is not determined solely by its spectro-temporal specification is discussed in the context of theoretical accounts of the distinction between speech and nonspeech.

INTRODUCTION

A concern of research that seeks to examine the abilities of human infants and nonhuman animals to perceive speech sounds, is that phonetic categories be dissociated from other auditory categories as the basis for the

*A partial summary of these results was presented at the 93rd Meeting of the Acoustical Society of America, State College, Pennsylvania, June 1977.

†Also Department of Psychology, The University of York, U.K.

††Also The M.R.C. Institute of Hearing Research, The University of Nottingham, U.K.

†††Also Speech and Hearing Clinic, Arizona State University at Tempe.

Acknowledgment: This work was carried out while Peter Bailey and Quentin Summerfield were supported by N.A.T.O. Post-doctoral Research Fellowships from the U.K. Science Research Council. We should like to thank Rod McGuire for his time and programming expertise, Erika Hamm for running the subjects in Experiment II, Michael Studdert-Kennedy for his comments on an earlier draft of this manuscript and, in particular, Alvin Liberman for his hospitality, advice and encouragement.

measured response. Several studies have demonstrated that human neonates categorically perceive voiced and voiceless initial stop consonants (for example, Eimas, Siqueland, Jusczyk and Vigorito, 1971; Lasky, Syrdal-Lasky and Klein, 1975; Streeter, 1976). The interpretation that this ability reflects sensitivity to phonetic categories has been challenged on the grounds that the positions of phonetic and auditory category boundaries were confounded in the stimuli used (Stevens and Klatt, 1974; Kuhl and Miller, 1975; Miller, Wier, Pastore, Kelly and Dooling, 1976; Streeter, 1976; Pisoni, 1977). For example, the members of a continuum formed by varying the relative onset times of two coterminous tones are perceived in three categories with category boundaries corresponding to differences in onset time of about 20 msec (Pisoni, 1977). With tone-onset-times (TOTs) of less than about 20 msec, the onsets of the two tones are perceived as simultaneous; for TOTs of more than 20 msec, the onsets are perceived as successive. These auditory categories of simultaneity and successivity could underly the infant's ability to discriminate synthetic exemplars of voiced and voiceless stop consonants. However, while there are good grounds for contesting the claim that infants discriminate voicing contrasts phonetically, the situation with contrasts of place of production appears to be different. Both Eimas (1974) and Miller and Morse (1976) have demonstrated with similar stimuli, but different methodologies, that infants can discriminate place of production contrasts in initial position categorically, an ability for which a psychoacoustic rationale is less obvious. The stimuli for these experiments were three-formant consonant-vowel syllables generated by a parallel resonance synthesizer and are schematized in Figure 1.

English-speaking adults identify patterns A and B as [gæ] and patterns C, D and E as [dæ] (Pisoni, 1971). Eimas (1974) showed with nonnutritive sucking as a measure of habituation-dishabituation that infants discriminated patterns B and C but not D and E. Miller and Morse (1976) used a heart-rate measure of habituation-dishabituation in a within-subjects design to show that infants discriminate patterns B and C but not patterns A and B or C and D. These authors have noted that the differences in onset frequencies of the second and third formants are the same in the within-category pairs as in the between-category pairs. The results might be explained by arbitrarily according a special status to spectrally diverging patterns (for example, A and B); alternatively, one might contrive the hypothesis that of the five patterns only A and B exceed some threshold for spectral change in the infant's auditory system, although we are not aware of any empirical support for the latter suggestion. Perhaps the most obvious categories for auditory patterns of this kind are those corresponding either to no frequency change or to frequency change in a particular direction; however, the data of Eimas (1974) and of Miller and Morse (1976) suggest that infants do not use such categories. Thus, on the basis of this evidence and in the absence of evidence to the contrary,[1] the tentative conclusion that infants discriminate place of production contrasts phonetically appears to be justified. Nevertheless, we felt that the importance of accurately detailing the ontogeny of human perceptual abilities requires that alternative explanations be

[1] Popper (1972) claimed that discontinuities in the presumed auditory representation of the members of three-formant vowel-consonant place continua underly the location of category boundaries. This claim appears not to be supported by his own data.

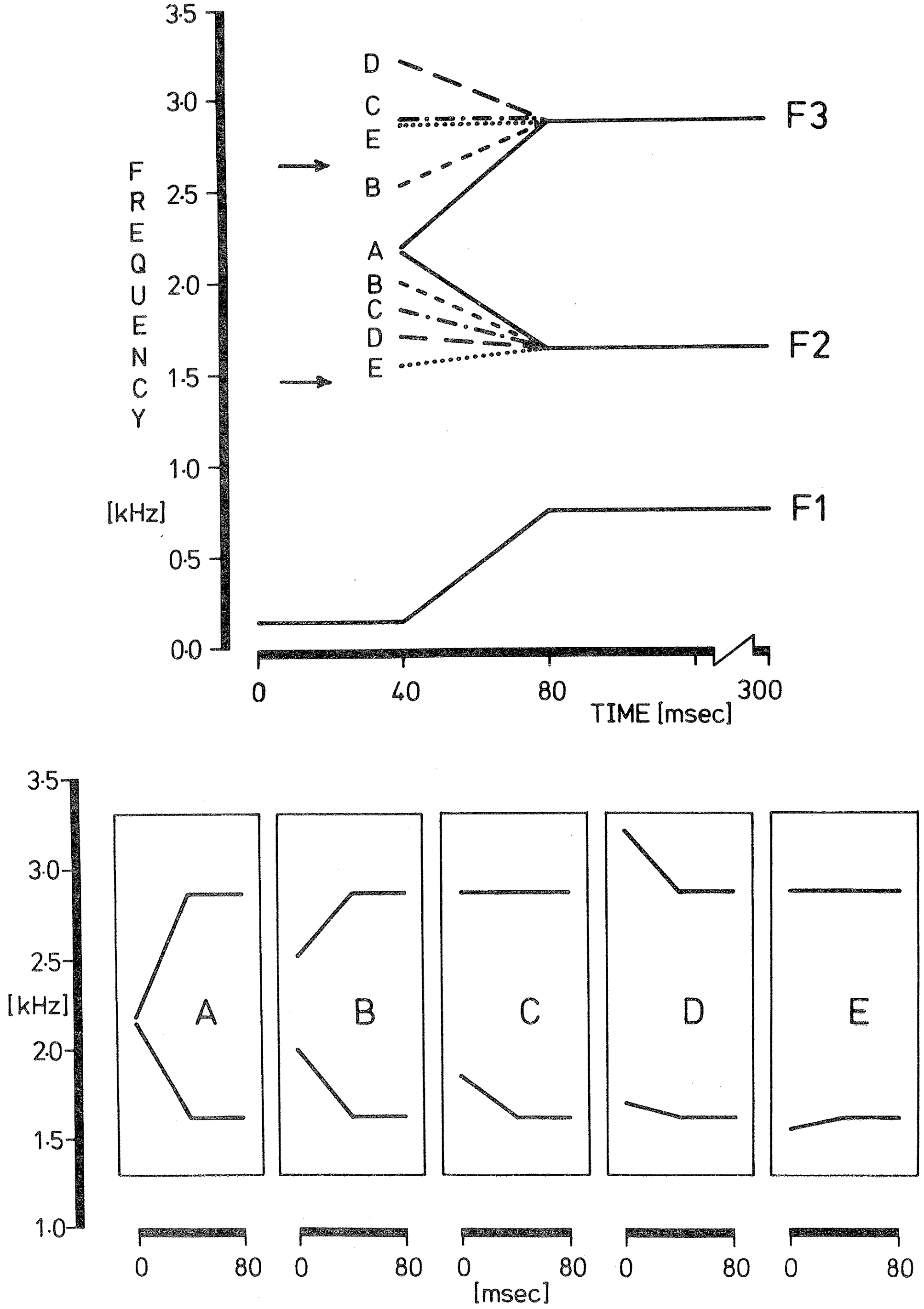

Figure 1: Schematic representations of the five stimulus patterns from Pisoni (1971) used by Eimas (1974) [B,C,D and E] and Miller and Morse (1976) [A,B,C,D]. See the text for details. The lower panel shows the second and third formant transitions in the five patterns individually. For clarity the transitions are represented here as linear segments while in the actual stimuli they were parabolic.

actively eliminated. As a contribution to this end, we wanted to determine whether any psychoacoustic rationale might be found to explain the categorization of three-formant speech patterns into bilabial and alveolar categories.

There are a number of ways in which this problem might be approached. Techniques in which specific psychophysical questions are asked using nonhuman animals (for example, Sinnott, Beecher, Moody and Stebbins, 1976) or infants as subjects possess the virtue of directness, but can usefully be supplemented by experimentation with human adults whose communicative abilities allow more detailed questions to be asked. We chose the latter course.

EXPERIMENT I

The object of Experiment I was to determine whether the categorization of acoustic patterns resembling consonant-vowel syllables is a function only of their spectro-temporal specification or whether they are heard as speechlike or as nonspeech.

METHOD

Stimuli

We synthesized two consonant-vowel (CV) continua with the OVE IIIc serial resonance synthesizer at the Haskins Laboratories.[2] Their spectro-temporal specifications are detailed in Figure 2. The continua were composed of three-formant patterns. (The synthesizer produces five formants. The two highest formants were removed by filtering as described below.) The parametric specifications of the first and third formants were the same in each continuum; only the second formant differed between the two. The configuration in the left-hand panel of Figure 2, with the steady state of the second formant (F_2) at 950 Hz, produces a vowel like [o]. The configuration in the right-hand panel, with the steady state of F_2 at 1800 Hz, produces a vowel like [e]. With each vowel a nine-member [b] to [d] continuum was created by covarying the onset frequencies of the second and third formants symmetrically about their steady states in eight 50 Hz steps. The third formant had its steady state at 2500 Hz, and the third formant transition ranged in onset from 2300 Hz to 2700 Hz. In the [bo-do] continuum, the onset of the F_2 transition ranged from 750 Hz to 1150 Hz. In the [be-de] continuum, the onset of the F_2 transition ranged from 1600 Hz to 2000 Hz. The first formant had its onset at 200 Hz and its steady state at 500 Hz. All formant transitions were 35 msec in duration with linear trajectories. The total duration of each syllable was 250 msec.

These parameter values were selected so that both continua would be physically symmetrical but phonetically asymmetrical. This is possible because the relative invariance of the point of constriction in the vocal tract for a particular place of production across vocalic contexts gives rise to a relatively invariant pattern of resonances corresponding to the closed vocal tract for a given place. These "locus frequencies" of approximately 800 Hz

[2] The synthesizer had been modified to ensure that the first pitch pulse occurred at the same point in every syllable.

FIGURE 2

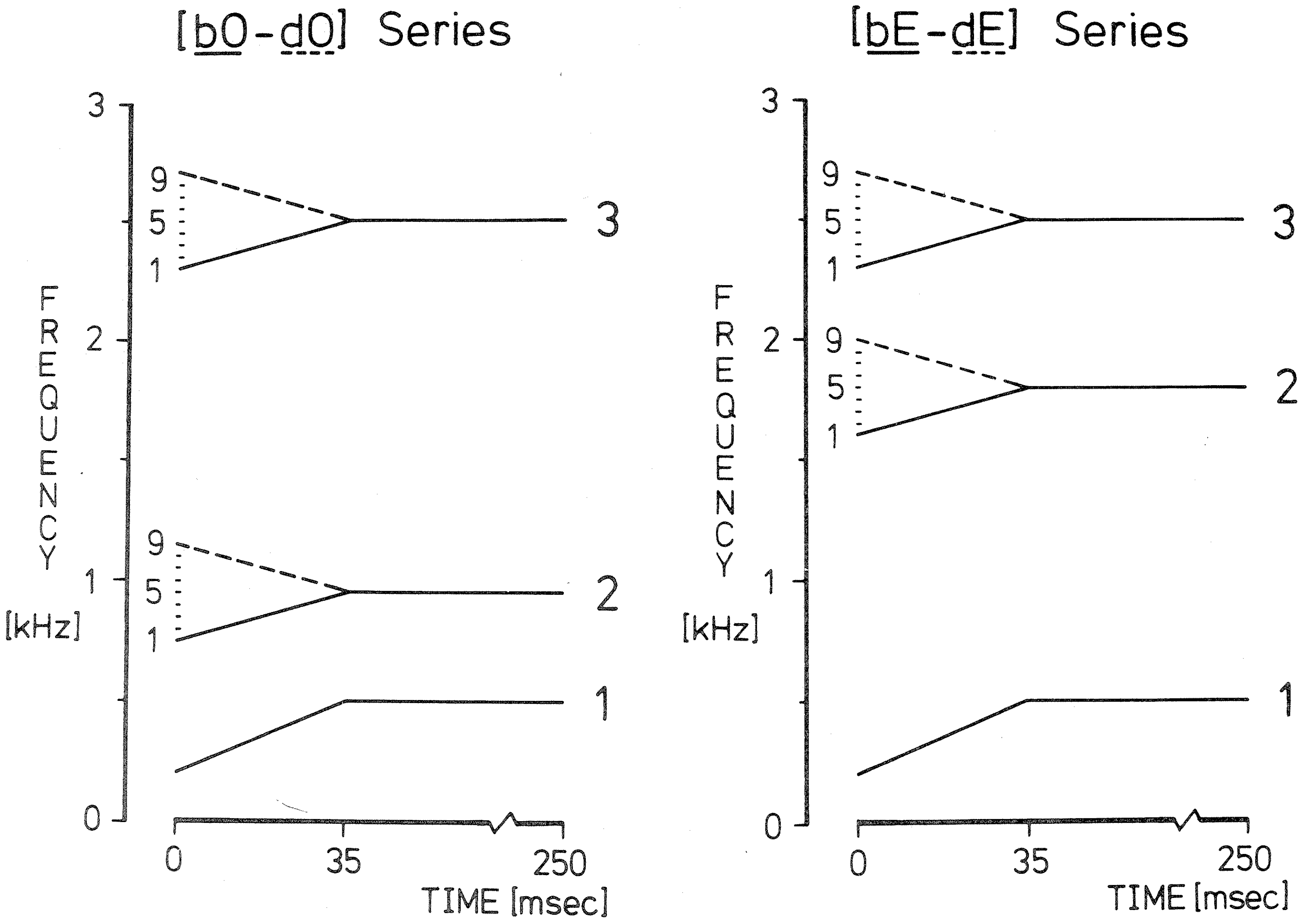

Figure 2: Schematic representations of the stimuli used in Experiment I. See the text for details.

for [b] and 1800 Hz for [d] (Delattre, Liberman and Cooper, 1955) are not realized in the signal. The initial formant transitions represent that portion of the change from the closure frequencies to those corresponding to the steady-state vowel during which there is sufficient airflow to excite the vocal tract. The relative frequencies of locus and steady-state determine the direction of formant transitions characterizing a stop with a particular place of production in the context of the following vowel. We intended the phoneme boundary on the [bo-do] continuum to be associated with falling transitions so that the majority of its members would be heard as [bo]. The inverse should apply to the [be-de] continuum; the majority of its members should be heard as [de]. In this way we intended to dissociate phonetic boundaries from any boundaries that might either accompany flat, as opposed to rising or falling, transitions or simply coincide with the centers of stimulus ranges.

These eighteen stimuli were low-pass filtered with a cut-off at 3.2 kHz and digitized with a sampling rate of 10 kHz. This operation eliminated the fourth and fifth formants. A hardware spectrum analysis was performed on successive 12.8 msec segments of each stimulus over the duration of its formant transitions, and the relative levels of the three formants were measured in each segment. Using these values to specify amplitudes, and using the parameters from the formant synthesis to specify frequencies, we copied the members of both continua by replacing their formants with frequency- and amplitude-modulated sine-waves. These sine-wave patterns were created with a digital software synthesizer and were recorded and redigitized in the same way as the formant stimuli. On first acquaintance these patterns sounded like nonspeech whistles.

Thus, the stimuli for Experiment I formed four continua: two were of vowel-type [o] corresponding to the patterns in the left-hand panel of Figure 2, and two were of vowel-type [e] corresponding to the patterns in the right-hand panel of Figure 2. For each vowel-type, one continuum was constructed from formants and one from sine-waves.

We wished to discover whether psychoacoustic boundaries determined with the sine-wave stimuli would coincide with phonetic boundaries determined with the formant stimuli. This required a task that can provide identification data without involving overt labeling. We used an AXB identification task. On each trial a triad of stimuli was presented. The first and third members of the triad were the end-point stimuli of a continuum, the second member of the triad was a stimulus drawn randomly from that continuum. The task of a listener was to indicate whether this random stimulus was more like the first or more like the last member of the triad. In addition, we wanted to obtain a conventional two-step AXB discrimination function for each continuum. In this test A and B were separated by two steps along the stimulus continuum and X was identical to either A or B.

For the identification test we recorded two sequences of trials for each continuum. The first was a practice sequence of three blocks of eighteen trials involving only the continuum end-points. Feedback was given on each trial in the form of the word 'first' or 'last' included on the tape as appropriate after the response interval. The second sequence consisted of nine blocks of eighteen trials of which the first three were intended to be practice blocks to be discarded before data analysis. In each block all members of the continuum were represented twice in the central position of the

AXB triad, once when A was stimulus #1 and B was stimulus #9, and once with the order reversed. On every trial the interval between the members of the triad was 1 sec, and an interval of 3 secs intervened between successive triads. An extra three seconds separated successive blocks of triads. The same timing pattern was incorporated in the AXB discrimination sequences for which we recorded a sequence of 6 blocks of 14 practice trials with feedback and a sequence of 9 blocks of 14 trials without feedback. Again, the first three of these blocks were intended for practice. For any continuum, each of the seven two-step pairs appeared equally often in each of its four possible orderings.

Subjects

Six undergraduates were paid to take part in the experiment. They declared themselves to be phonetically naive, to have normal hearing in both ears and to have learned English as their first language in the U.S.A. They were tested in a quiet room, either singly or in pairs in four two-hour sessions that were held on different days. All stimuli were presented binaurally through headphones at a level of 75 dB.

Procedure

In the first session subjects only listened to the sine-wave stimuli, which, they were informed, differed at their onsets. They described these sounds as nonspeech whistles. They were told that each trial of the experiment would consist of three stimuli and that their task was to decide whether the center stimulus sounded more like the first or more like the last member of the triad. They made their responses by writing down one of the letters 'F' or 'L' on a specially prepared response sheet. In this first session, subjects listened twice to both identification and discrimination practice sequences with feedback. The order of vowel-type was counterbalanced. Initially, subjects found the task difficult, but their performance improved during the session to at least 75 percent correct on the identification task.

In the second session the subjects again only listened to the sine-waves. They performed the discrimination test for each vowel-type and then the identification test for each vowel-type. Each test was preceded by a practice sequence with feedback. Four of the subjects heard the [e] analogues before the [o] analogues, while only two heard the [o] analogues before the [e] analogues. (The experiment had originally been designed for eight subjects, but only six were tested.)

In the third session subjects heard the formant stimuli in the same format, except that they listened to the discrimination tests for a second time at the end of the session. (Having examined the data from session 2, we realized that the number of discriminations in a single administration of the discrimination tests was insufficient to yield interpretable results.)

In the fourth session subjects once again only heard the sine-waves, but at the begining of this session the relationship between the formant and the sine-wave stimuli was explained. After listening to the sine-waves again, all subjects agreed that these stimuli could be heard as initiated by one of the stop consonants [b] or [d]. (In fact, one listener had made this observation without prompting near the end of session 2.) Hearing the stimuli in this new way, subjects found the tasks easier, and during the session reported no tendency for the consonantal percept to disappear.

Results

The data were sorted to yield identification functions and discrimination functions for each subject in each condition. The results obtained in session 3 with the formant stimuli are shown in Figure 3. The graphs in the upper panel correspond to the identification test, those in the lower panel to the discrimination test. The data of all 6 subjects have been pooled to obtain these graphs and each point plots the mean of 72 responses in the identification test and of 144 responses in the discrimination test. In both graphs the solid line corresponds to the functions obtained with the [bo-do] continuum, and the dotted line corresponds to those obtained with the [be-de] continuum. For each continuum, the identification function relates the percentage of times each stimulus was judged to be more like the [b] end-point (that is, stimulus #1) than the [d] end-point (that is, stimulus #9) of its continuum. The stimulus numbers are arrayed along the horizontal axis, the percentage of [b]-like identifications increases along the vertical axis. Our expectations of phonetic asymmetry are at least partially borne out: the two functions do not overlap in the boundary region. The phoneme boundary on the [bo-do] continuum is displaced to the right of the center of the stimulus range; similarly, though to a very much lesser degree, the boundary on the [be-de] continuum is displaced to the left. The discrimination functions do not show any major peaks, but there is a tendency for the [be-de] stimuli to be discriminated better at lower stimulus numbers and for the [bo-do] stimuli to be discriminated better at higher stimulus numbers.

The identification and discrimination functions of Figure 3 can be compared with those for the sine-wave continua from session 2 that are displayed in Figure 4; the parameters of this display are the same as those of Figure 3, except that the data from the one subject who began to hear the stimuli as speechlike in session 2 have been excluded. Thus, each point in the identification function and in the discrimination function plots the mean of 60 responses. Again, two functions of each type have been plotted: the solid line corresponds to the analogues of the [bo-do] continuum, and the dotted line corresponds to the analogues of the [be-de] continuum. The pattern of data in this figure is very different from that displayed in Figure 3. The two sine-wave identification functions largely overlap throughout their ranges, and the 50 percent points on both functions fall close to the centers of the continua. No clear pattern emerges from the discrimination functions which are more variable than those obtained from the formant stimuli; as noted above, only half as many discrimination data per point were collected with the sine-wave stimuli.

Normal-ogive psychometric functions were fitted to the identification data of each subject in each condition using probit analysis (Finney, 1971). An estimate of the position of the phoneme boundary was obtained by computing the 50 percent point on each fitted function. For the formant stimuli, the mean of the boundaries on the [bo-do] continuum was 5.57, and on the [be-de] continuum it was 4.54. The difference between these means, although in the predicted direction, is not significant when assessed in a one-way analysis of variance ($F_{1,5}=3.13; p < 0.1$). Nevertheless, it is larger than the difference

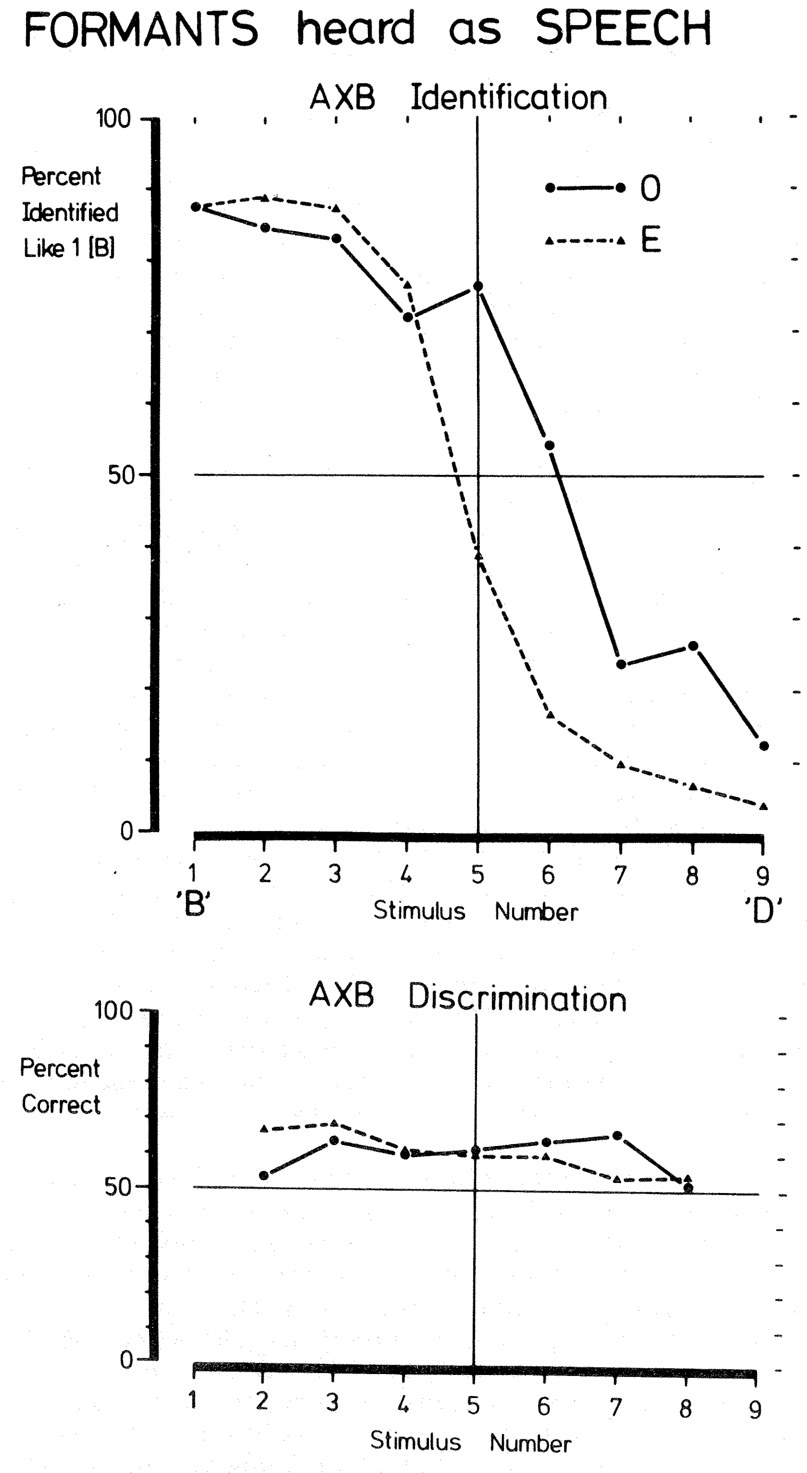

Figure 3: Identification and discrimination functions from Session 3 in Experiment I.

between the means obtained with the sine-wave stimuli that were (for the five subjects whose data are illustrated in Figure 4) 4.47 for the [bo-do] analogues and 4.42 for the [be-de] analogues. The difference between these means is not significant ($F_{1,4}$=0.01;p > 0.2).

The failure of the two formant continua to produce significant differences in mean boundaries is surprising. We can trace this outcome to two factors: the members of the [bo-do] continuum were not identified very systematically, and there was little asymmetry in the way the [be-de] continuum was identified. These factors reflect our failure to estimate stimulus parameters appropriately for naive listeners. However, it is possible to carry out a less stringent test of the asymmetry hypothesis by determining whether the members of the [bo-do] continuum were perceived as more [b]-like than the members of the [be-de] continuum. Using the number of [b]-like responses made to each stimulus by each subject as the dependent measure, an analysis of variance with the factors Subjects[6] x Vowel-types[2] x Stimuli[9] was performed. The effect of vowel-type was significant ($F_{1,5}$=12.31;p < 0.025), showing that overall, a greater number of [b]-like responses were made to the members of the [bo-do] continuum.

A similar analysis was carried out on the sine-wave data provided by the five subjects who heard the sine-waves as nonspeech sounds, in which the effect of vowel-type was not significant ($F_{1,4}$=0.69;p > 0.1). This is consistent with the pattern of data in Figure 4 and confirms the tendency for sine-wave continua to be divided symmetrically when heard as nonspeech sounds. This result is not materially influenced by the fact that our stimulus continua are only physically symmetrical when frequency is represented on a linear scale. Asymmetries introduced during peripheral auditory transmission, by the critical band mechanism (Scharf, 1970) for instance, would be likely to enhance the discriminability of rising, as oppposed to falling transitions, a distinction that is not apparent in our data. What is not revealed is the reason why subjects divide the sine-wave continua in the way they do. There are at least three possibilities: listeners may detect the presence or absence of spectral change in the higher resonances; they may distinguish rising from falling transitions; or they may simply divide the continua into two approximately equal ranges. However, none of these strategies can account for the asymmetrical categorization of the formant continua. This could either be correlated with the different spectral properties of formants and sine-waves, or with the way these stimuli are heard. The condition run in the fourth session of the experiment in which the sine-waves were heard as speechlike goes some way to dissociating the two possibilities.

The identification and discrimination data obtained in the fourth session are displayed in Figure 5. The pattern of these data is very different from that in Figure 4. Here, where the sine-waves were heard as speechlike, the identification and discrimination functions corresponding to the [o] and [e] analogues no longer overlap. Moreover, as with the formant stimuli, the majority of the [bo-do] analogues were heard as [b]-like, while the majority of the [be-de] analogues were heard as [d]-like. In this condition, the discrimination functions were reasonably smooth and consistent and were similar to those obtained with the formant stimuli. As before, boundaries were estimated by probit analysis. The mean boundary obtained with the [bo-do] analogues corresponded to a stimulus number of 5.65. That obtained with the [be-de] analogues was 3.82. The difference between these means is

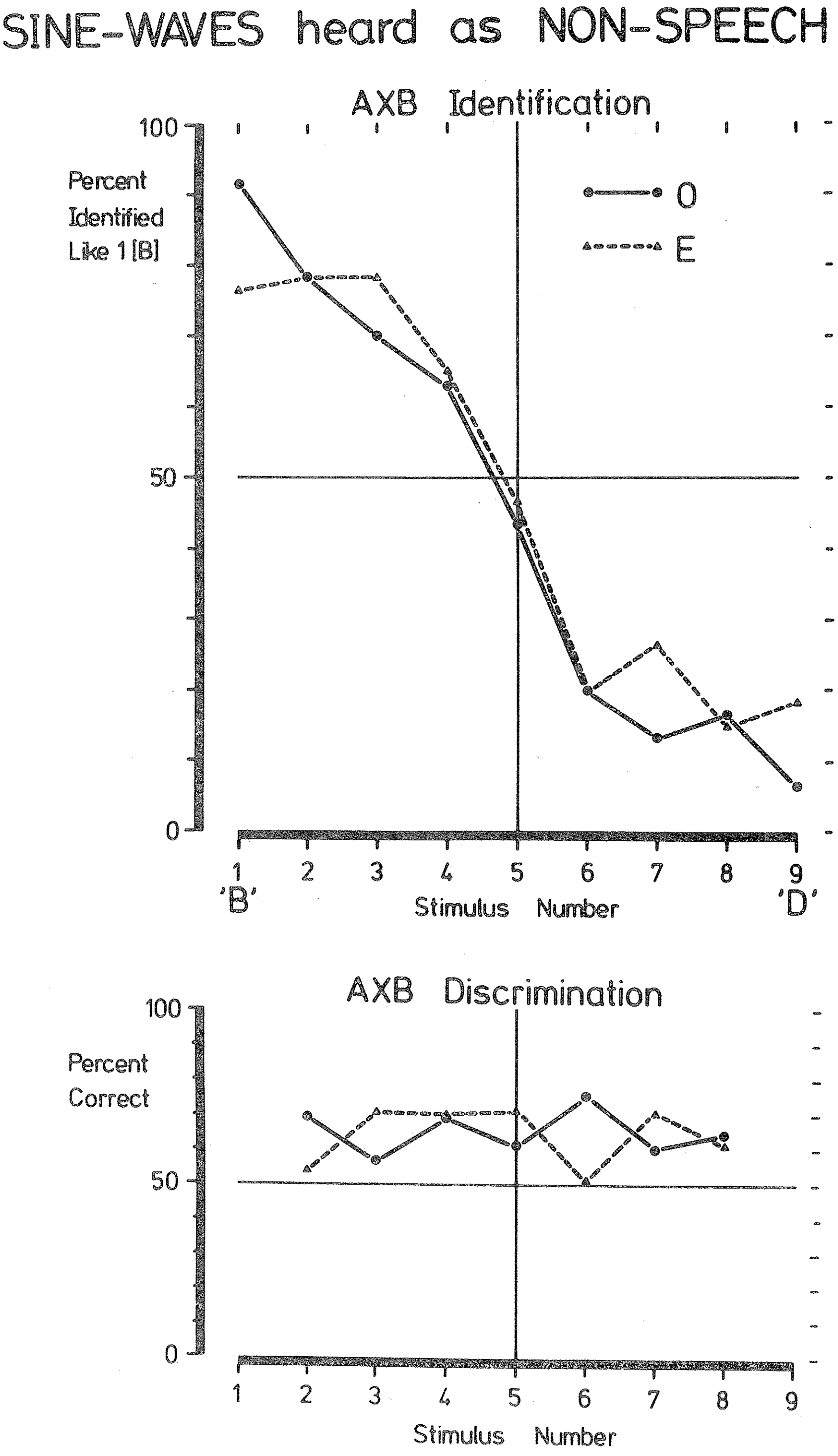

Figure 4: Identification and discrimination functions from Session 2 in Experiment I.

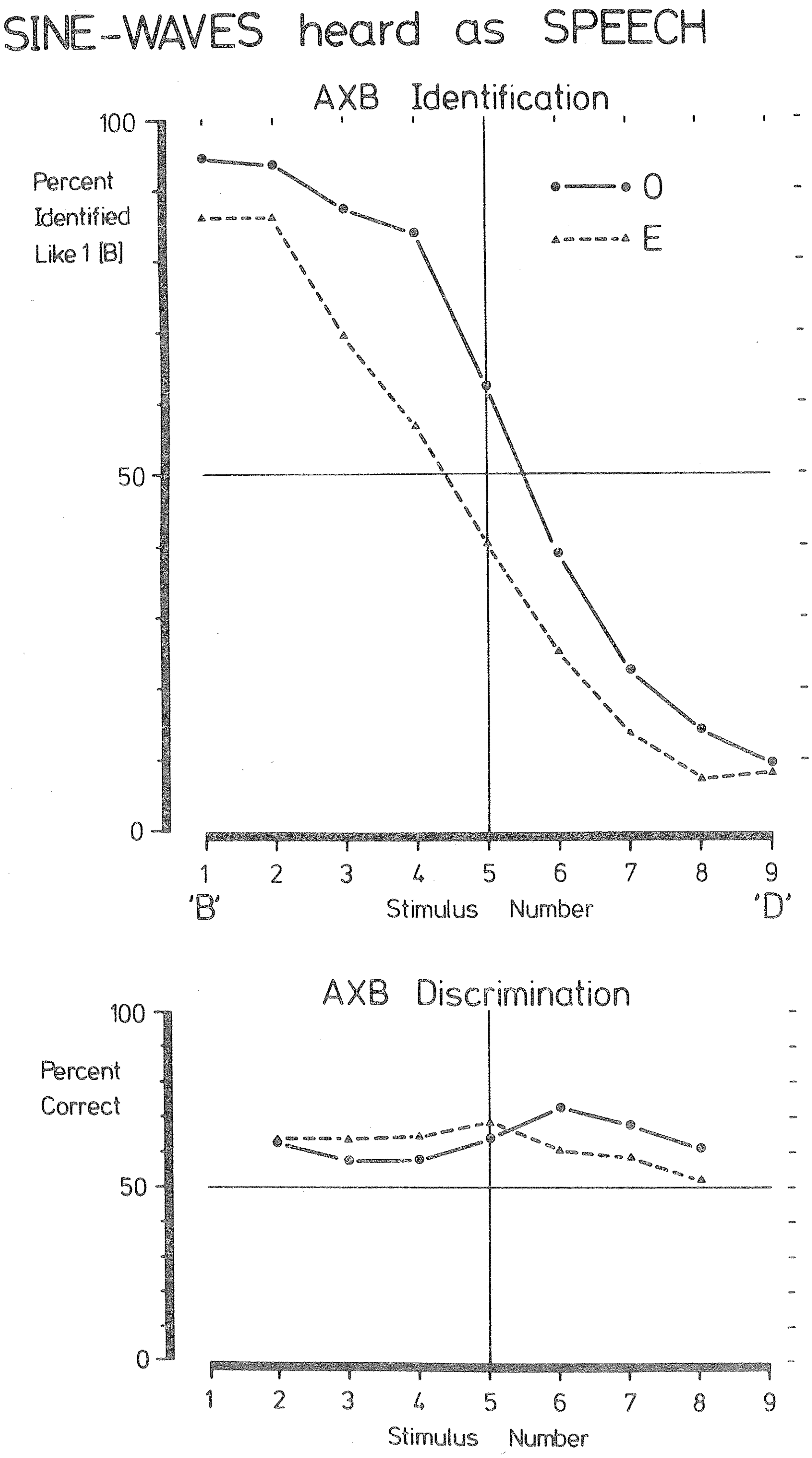

Figure 5: Identification and discrimination functions from Session 4 in Experiment I.

significant ($F_{1,5}$=18.35;$p < 0.001$). The pattern of data obtained in session 4 when the sine-waves were heard as speechlike is clearly more akin to the pattern corresponding to the formant stimuli than to that obtained in session 2 when the sine-waves were heard as nonspeech whistles. This correspondence suggests that the results are due not so much to the spectral structure of the stimuli, but rather to the way in which they are heard. Formants and sine-waves produce similar patterns of data when both are heard as speech.

We should be explicit about what we mean when we say that the sine-waves could be heard as 'speechlike'. We do not mean that after repeated exposure to these sounds and to their formant analogues listeners were able to identify their onset frequencies and infer a correspondence to [b] or [d]. Rather, it is our own experience and that described by our listeners, that the sine-wave patterns can 'name themselves' (cf. Studdert-Kennedy, 1976, p. 244). One hears an initial [b] or [d] followed by a whistle. We say that the percept is 'speech-like' then because, while the stop is compellingly phonetic, what follows it is not. In order to determine to what extent this result was idiosyncratic of these six listeners or was a function of exposure to formant stimuli in Session 3, we ran a second experiment.

EXPERIMENT II

For this experiment we synthesized a single [ba-da] continuum. Again we copied these formant stimuli with frequency- and amplitude-modulated sine-waves. We intended the new continuum to be both more natural and more asymmetrical than the continua used in Experiment I. As a result, more listeners heard the sine-wave analogues as speechlike without prompting or prior exposure to the formant stimuli. Thus, we were able to divide our subjects into two groups on the basis of their descriptions of their initial perception of the sine-wave stimuli.

METHOD

Stimuli

A single 11-member [ba-da] continuum was created with the OVE IIIc synthesizer. The total duration of each stimulus was 250 msec, and the duration of the syllable-initial formant transitions was 40 msec. The first formant had its onset at 350 Hz and rose linearly to a steady state at 750 Hz. The second and third formants had their steady states at 1000 Hz and 2500 Hz, respectively. The onset of the F_2 transition ranged from 650 Hz to 1350 Hz in ten 70 Hz steps; the onset of the F_3 transition ranged from 2250 Hz to 2750 Hz in ten 50 Hz steps. Thus, as in Experiment I, the frequency transitions in this continuum ranged symmetrically about the steady states. The stimuli were low-pass filtered at 3.2 kHz and digitized at a sampling rate of 10 kHz. The waveforms were analyzed with a hardware spectrum analyzer, and the relative levels of the formants measured over the durations of the formant transitions. As before, these measurements were used to control the levels of three frequency-modulated sine-waves produced by a digital synthesizer. In this way a sine-wave analogue of each member of the [ba-da] continuum was created.

We recorded two AXB identification tests with 110 trials in each, one with formant stimuli and one with sine-wave stimuli. The format of these tests was the same as that used in Experiment I, except that only 0.5 sec intervened between successive members of each triad. No discrimination tests were administered in this experiment.

Subjects and Procedure

Thirty subjects were tested. They were drawn from the members of a course in speech and hearing at Arizona State University. The AXB task was explained to the subjects and they were told that they would hear stimuli constructed from sine-waves, but they were told nothing about the relation between the sine-waves and possible speech sounds. The test with the sine-waves was administered first, followed by the test with formant stimuli. At the end of each test, subjects were instructed to write down a description of the stimuli.

Results

Somewhat fortuitously, 15 listeners said that they heard the sine-waves as nonspeech whistles or tones, while 15 listeners heard them as speechlike. The latter group described the sine-waves as being initiated by either a stop consonant or semi-vowel with bilabial or alveolar place of production.[3]

Figure 6 displays the AXB identification data for the group who said that they heard the sine-waves as nonspeech sounds. The dotted line plots the function obtained with the sine-wave stimuli; the solid line plots the function obtained with the formant stimuli. The formant continuum was divided asymmetrically into two sharply segregated categories. The sine-wave continuum, on the other hand, was divided less asymmetrically and into two less sharply defined categories. The contrast between the way the continua were categorized is exemplified by the responses to stimulus #6, the stimulus with flat transitions in the second and third resonances. When represented by formants, this stimulus was identified as like the [b]-endpoint of its continuum on 98 percent of its presentations; when represented by sine-waves, it was identified in this way on only 56 percent of its presentations. However, it is clear that the sine-wave continuum was not divided into two equal ranges, in contrast to the results of Experiment I. Figure 7 displays the identification functions for the group who said they heard the sine-waves as speechlike. These subjects also divided the formant continuum asymmetrically. Figure 7 suggests that sine-waves heard as speechlike are categorized in a similar way to formant stimuli. This similarity is exemplified by a comparison of responses to stimulus #6 in each continuum: when represented by formants, this stimulus was identified as like the [b] end-point of its continuum on 88 percent of its presentations; when represented by sine-waves, it was identified this way on 85 percent of its presentations.

We examined the statistical reliability of these observations in several ways. First, two measures were extracted from psychometric functions fitted to the identification data of each subject in each condition. One was the

[3] Of the fifteen subjects who heard the sine-wave stimuli as speech sounds, five perceived [b] and [d], six [w] and [d], two [w] and [l], one [w] and [z] and one [w] and [y].

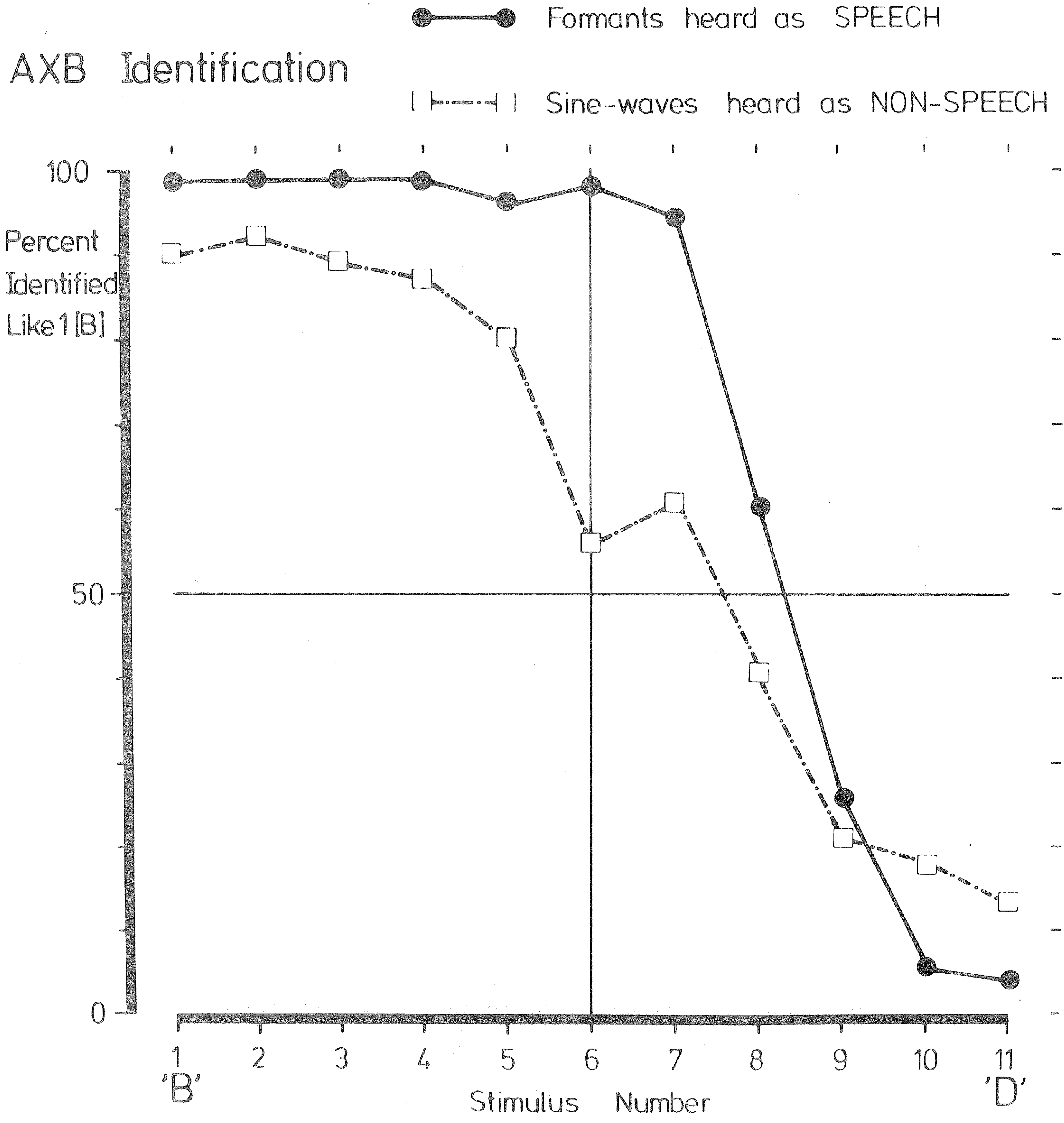

Figure 6: Identification functions for the group who heard sine-waves as speech-like in Experiment II.

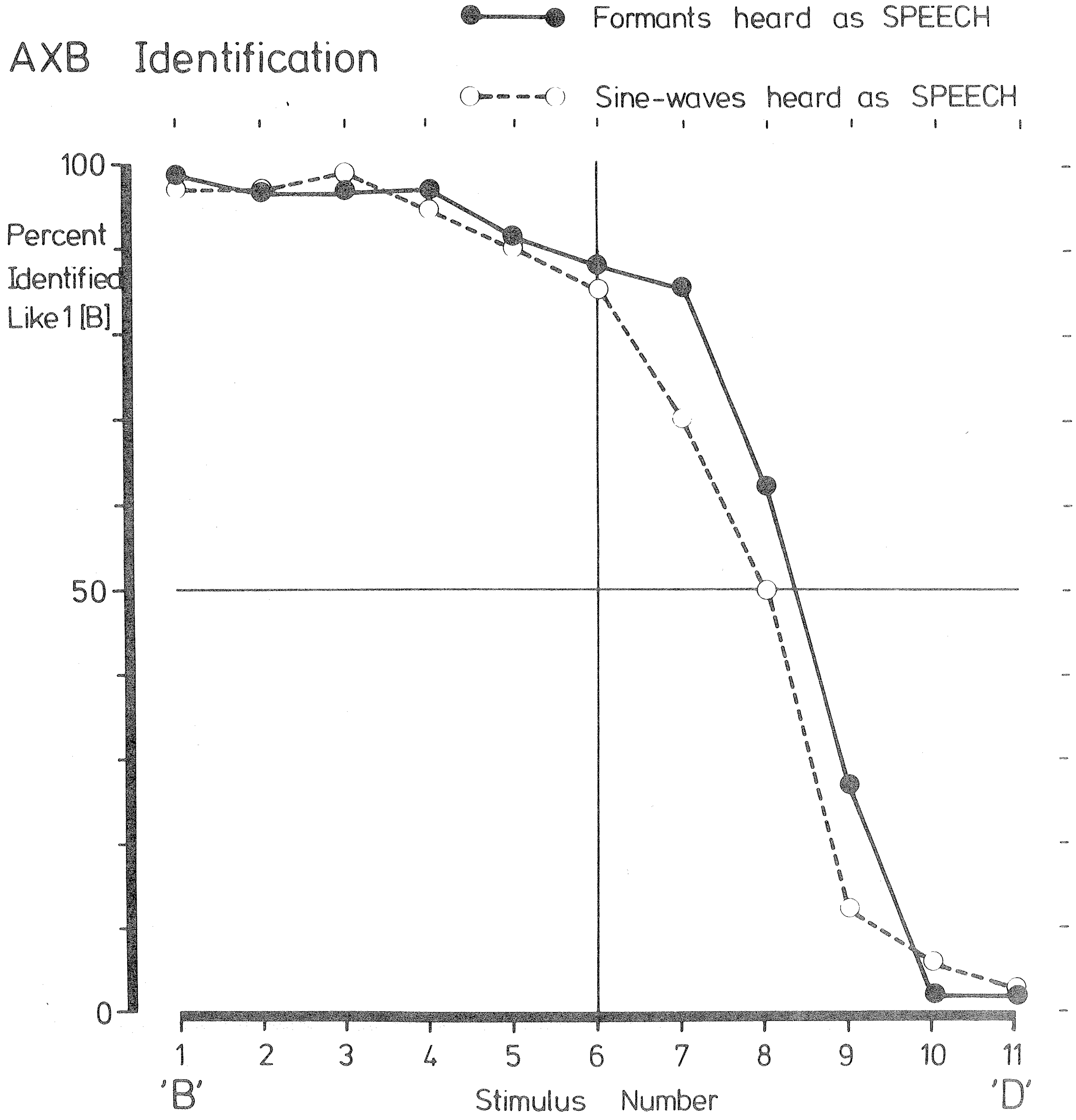

Figure 7: Identification functions for the group who heard sine-waves as non-speech in Experiment II.

stimulus number corresponding to the 50 percent point on the fitted function: this provides an estimate of the position of the phoneme boundary. The other was the slope of the probit regression line, a parameter that is directly related to the slope of the identification function in the boundary region. For the group who heard the sine-waves as nonspeech, category boundaries corresponded to stimulus numbers of 7.01 for the sine-wave stimuli and 8.25 for the formant stimuli. For the group who heard the sine-waves as speechlike, the equivalent values were 7.57 and 7.93. (The central stimulus in each continuum was number 6.) The differences between these means were assessed in an analysis of variance that showed that the overall difference between the boundaries on the two continua is significant ($F_{1,28}$=12.74;$p < 0.01$); however, the size of this difference does not differ significantly between the two groups of subjects ($F_{1,28}$=3.79;$0.1 > p > 0.05$).

A similar analysis was undertaken with probit regression line slopes as the dependent measure, which showed that identification functions produced by sine-wave stimuli were flatter than those produced by formant stimuli ($F_{1,28}$=13.27;$p < 0.01$). However, in this case, the size of the effect does differ between the two groups ($F_{1,28}$=6.44;$p < 0.025$). To ensure that this interaction does not result from differences between the slopes of the formant functions, we carried out analyses comparing the two groups for the formants and sine-waves independently. Only the sine-wave slopes differ between the groups ($F_{1,28}$=4.24;$p < 0.05$), confirming that flatter slopes were produced when sine-waves were heard as nonspeech.

Despite the failure of the first of these dependent measures to distinguish the condition in which the sine-waves were heard as nonspeech from the other conditions, this distinction is implied by the patterns of data in Figures 6 and 7. In an attempt to specify how the two groups of subjects differ, we carried out two further analyses, using as the dependent measure the number of [b]-like responses made by each subject to each stimulus. For the group who heard the sine-waves as speechlike, the mean percentages of [b]-like responses made to the formant and sine-wave continua were 69.34 percent and 64.00 percent (difference = 5.34 percent). For the group who heard the sine-waves as nonspeech sounds, the means were 70.91 percent and 58.97 percent (difference = 11.94 percent). In an analysis with the factors Groups (2)xContinua(2), the interaction between these two factors was significant ($F_{1,28}$=4.59;$p < 0.05$), showing that the difference between the formant and the sine-wave means is significantly larger for the group who heard the sine-waves as nonspeech. However, we cannot conclude from this that significantly fewer [b]-like responses were made to the sine-wave stimuli, because the groups also differed in the number of [b]-like responses that they made to the formant stimuli. Therefore, we ran analyses to compare the means for the sine-waves and formants independently. Although the two formant means do not differ significantly ($F_{1,28}$=1.32;$p > 0.2$), neither do the two sine-wave means ($F_{1,28}$=2.55;$p > 0.1$).

Discussion

The results of Experiment II are consistent with one aspect of the results of Experiment I. In both experiments, the categorization of sine-wave stimuli depended upon how the sine-waves were heard and, when the sine-waves were heard as speechlike, more closely approximated the categorization of formant stimuli. Taken together, these results indicate that the way in which a sound is categorized is not simply a function of the spectral structure of that sound, but also relates to how that sound is heard. However, there is one respect in which the results of the two experiments do differ. In Experiment I, the sine-wave continua were divided symmetrically when heard as nonspeech. In the equivalent condition of Experiment II, the continuum was not divided symmetrically.

It is possible that these different results follow from differences in the proximities of the first and second resonances in the stimuli used in the experiments. The first resonance was closer to the second resonance in the [ba-da] continua used in Experiment II than it was in either the [bo-do] or the [be-de] continua used in Experiment I. If the upward spread of masking from the first resonance has the effect that equal differences in the onset frequency of the second resonance become more discriminable as the onset frequency of the second resonance rises, then the result found when the sine-waves were heard as nonspeech in Experiment II might be expected. One reason for questioning this explanation, however, is that in the sine-wave continuum, as in the formant continuum used in Experiment II, the intensity of the second resonance at its onset frequency increased as its onset frequency was lowered. This might have been expected to counteract the effects of masking.

GENERAL DISCUSSION

Let us first consider the results obtained when the sine-wave continua were not heard as speechlike. The equivocal outcomes of our two experiments provide evidence both for and against an auditory discontinuity underlying the location of category boundaries in the perception of place of production. This unsatisfactory situation must be resolved by further experimentation. However, the demonstration of symmetrical categorization of sine-wave continua when heard as nonspeech in Experiment I, is consistent with the results of experiments on the discrimination of formant transitions in nonspeech contexts. Mattingly, Liberman, Syrdal and Halwes (1971) found no discrimination peaks corresponding to the location of phoneme boundaries when second formant transitions extracted from the members of a [bæ-dæ-gæ] continuum were presented in isolation. We have already noted that there is no obvious auditory strategy for dividing continua of frequency transitions that can account consistently for the way subjects categorize formant and sine-wave continua when these are heard as speech. We now turn to these data.

To a greater or lesser extent, category boundaries on formant and sine-wave continua, when heard as speechlike, were placed asymmetrically. In Experiment I the [be-de] sine-wave continuum and both formant continua were categorized asymmetrically by five listeners out of six, while the [bo-do] sine-wave continuum was categorized asymmetrically by four listeners out of six. The results of Experiment II, on the other hand, are unequivocal in showing asymmetrical categorization of the [ba-da] continuum. All thirty subjects categorized the formant continuum asymmetrically, while fourteen out of the fifteen subjects who heard the sine-wave continuum as speechlike categorized it asymmetrically.

The finding that phoneme boundaries are not tied to particular directions of formant movement is not restricted to our data (for example, Liberman, Delattre, Cooper and Gerstman, 1954; Delattre, et al., 1955; Pisoni, 1971; Blumstein, Stevens and Nigro, 1977). For instance, the two arrows in Figure 1 indicate the onset frequencies of the second and third formants at the mean [b-d] phoneme boundary found by Pisoni (1971). Here, for the vowel [æ], both transitions are rising at the boundary. This pattern can be contrasted with that found by Blumstein et al. (1977) for the vowel [a], where the boundary was characterized by a slightly rising transition in F_3 and a slightly falling transition in F_2. What explanation can account for the placement of category boundaries on continua of synthetic syllables that vary in place of production?

One explanation has emphasized the correspondence of general properties of the acoustic signal to discrete phonetic categories (Stevens, 1975). For instance, Blumstein et al. (1977) observe that: "Acoustic energy at the stimulus onset is spread or 'diffuse' for the [d] and [b] stimuli and is concentrated in a narrow frequency region or is 'compact' for the [g] stimuli. These properties may be described in terms of the onset characteristics and the following spectral changes: a diffuse-rising pattern characterizing [b], a diffuse-falling pattern for [d] and a compact-spreading pattern for [g]" (p. 1036). We find it reasonable to suppose that the abilities to produce and perceive speech have coevolved in such a way that maximally different acoustic patterns support the information for discrete categories of articulatory events. However, the data reported in the present paper and those reviewed above, are not compatible with the idea that there is a one-to-one isomorphism between the registration of these acoustic properties and the perception of phonemic identity. Whether or not this isomorphism exists in the acoustic description of natural productions, the need to account for the perception of synthetic speech remains. For example, we need to characterize the commonality which underlies a bilabial percept, whether this results from frequency-modulated sine-waves, synthetic formant transitions, or the rich acoustical structure of natural speech. Indeed, we feel that an account of phonetic perception should be based on a rationalization of sensitivity to that commonality, rather than on an enumeration of sensitivities to specific acoustical elements in the speech stream.

An integral component of this rationalization is the dissociation of phonetic from nonphonetic perception. The original focus of our interest in these experiments was the realization that naive listeners, who initially describe the sine-wave patterns as whistles, come to perceive them phonetically, and that the change seems to be irreversible.[4] What underlies this changing percept of an unchanging stimulus? This question could be answered in many ways. One answer is provided by a class of models in which the perceptual system detects the same array of sensory information but processes that information in different ways. Two examples of this type of solution are considered below. We shall contrast them with an alternative view in which no explicit distinction is drawn between different modes of processing; the changing percept results from a change in the organization of attention to information in the signal.

[4] In an experiment by Cutting (1974) using sine-wave stimuli somewhat like those used in the present experiment, listeners apparently did not report that the stimuli took on a phonetic character. The differences in stimuli and procedure between his experiment and ours render the difference in outcome difficult to interpret.

Most accounts of speech perception, confronted with the apparent uniqueness both of the speech signal and of its perceptual consequences, have dichotomized sounds into two classes: speech and nonspeech. Having construed the dichotomy, they have had to explain how the perceptual process achieves this classification. As noted above, we can identify two types of solutions to the problem: in one, the decision about speech-likeness is assumed to be explicit and directive; in the other, the decision is implicit and passive. According to the former, speech-likeness is supposedly marked by specific acoustical attributes that, if detected in an initial stage of auditory analysis, direct the signal to a special phonetic processor. For example, Stevens and House (1972), following House, Stevens, Sandel and Arnold (1962), remark that "The listener need not be set for speech prior to his hearing the signal; his prepared state is triggered by the presence of a signal that has appropriate acoustic properties" (p. 13). We note that this account could be buttressed to explain the change from hearing sine-waves as nonspeech to hearing them as speechlike if supplemented with a variable criterion for the acceptability of the evidence provided by the first stage. An initially conservative setting of the criterion could be relaxed to achieve intelligibility within the context specified by an experiment. However, although the data can be explained in this way, we can question the general approach on two grounds. First, attempts to identify acoustical trigger features empirically have not been successful (Haggard, 1971; Allen and Haggard, 1977). Second, and fundamentally, it is hard to see how the phonetic processor could have evolved without the omniscience of the initial classification stage. Given that, why should the phonetic processor have evolved at all? Similar objections can be raised to the suggestion made by Cutting (1974) that a high-level decision as to the status of the signal might censor the output of a phonetic processor when the signal is insufficiently speechlike.

The other type of solution to the problem of distinguishing speech from nonspeech sounds starts by suggesting that phonetic and generalized auditory analyses are accorded in parallel to all acoustic inputs. Phenomenal perception corresponds to whichever process achieves a satisfactory analysis. For instance, Liberman, Mattingly and Turvey (1972) have suggested that "...the incoming signal goes indiscriminately to speech and nonspeech processors. If the speech processors succeed in extracting phonetic features, then the signal is speech; if they fail, then the signal is processed only as nonspeech" (pp. 323-324). Clearly this approach depends upon a characterization of the acoustical representation of phonetic features. Our demonstration of the perceptual duality of sine-wave stimuli can be accommodated in this model by a provision for context-sensitive adjustment in the specification of adequate stimuli for detectors of phonetic features. But, given such provision, we wonder whether the speech processor would ever abandon the search for phonetic features to admit a nonspeech solution.

We suggest that both of the preceding accounts of the distinction between speech and nonspeech are either inadequate or incomplete because they fail to capture an important inherent characteristic of the speech signal. In particular, they do not achieve explanatory adequacy because they assume that the information in the signal that must underpin phonetic perception is a specification only of discrete acoustic elements in the three-dimensional metric of frequency, amplitude and time, and not of the origin of the elements expressed in the four-dimensional metric of a three-dimensional vocal tract undergoing continuous reconfiguration over time. Those accounts which commence with only a three-dimensional specification of the signal suppose that speech perception is mediated by knowledge of the way vocal tracts behave. For example, Stevens and House (1972) suggest that "After processing by peripheral auditory structures, some attributes of an incoming auditory pattern are then, as it were, looked up in the dictionary of auditory-articulatory correspondences" (p. 54); a similar well-developed view presented in detail elsewhere (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967) is illustrated by Mattingly and Liberman (1969): "In effect, the key that the listener has available to him is an articulatory model that relates the phonetic message to the signal" (p. 102). Seen in this way, perception consists of interpreting elements by imposing structure upon them, but note that this structure derives from constraints embodied in an internal articulatory model. A perceptual model of this kind would seem to involve at least two stages: in the first, a sequence of acoustic elements must be segregated and detected; in the second, these elements must be interpreted, presumably to reconstruct the information encoded in the sequential properties of the signal. Knowledge of vocal tract behavior may assist the first stage but it governs the second stage. While we have no doubt that speech perception is inextricably tied to the origin of the signal in a vocal tract, we wonder whether a process of fractionation followed by reintegration would best capture the information endowed to the signal by the continuous articulatory flow of a dynamic vocal tract.

In the following, we attempt to sketch in very general terms how an alternative account of the distinction between speech and nonspeech sounds might be developed. It will become clear that this account fits naturally into the wider context of a view of the perception of speech that might loosely be described as ecological. We shall discuss the general view while acknowledging that our data are only indirectly supportive of it.

In the natural world, sounds result from the participation of three-dimensional structures in events that occur over time. We suppose that the evolution of sensitivity to sound pressure variation progressed by a developing facility in identifying events that produce sounds, not just sounds *per se* (Gibson, 1966). When they specify events, acoustic signals describe not only their source but also what that source is doing. We do not yet have a sophisticated account of how this information is represented in the speech signal, but acoustic variation corresponding more or less directly to vocal tract cavity size variation can be identified, and perceptual sensitivity to it demonstrated (Kuhn, 1975). We wish to examine the possibility that the perception of acoustic patterns, whether speech or nonspeech, is properly described by registration of the *coherence* between information specifying the source of a sound and that specifying the transformation wrought upon the source. What we understand by coherence may be illustrated in a visual analogy. When a man runs, he structures light in such a way that both his identity as a man and his act of running are specified. When we perceive him running, we detect the coherence of these specifications; we do not first perceive the actor in order that we may interpret the elements of his act. Similarly we do not perceive speech by imposing articulatory structure upon an otherwise unstructured array of elements; rather, we perceive that structure because it is specified in the organization of the elements. This suggests an answer to the question of what is a speech sound: a pattern of sound may be perceived as speech if it specifies coherently its source as a human vocal tract partaking in a physiologically permissible act of articulation. The registration of coherence is analogous to perceiving the solutions to a set of simultaneous equations. The equations provide structure and coherence for the solutions, but no one solution necessarily mediates the attainment of any other.

We require a more precise specification of what these notions entail, but we find them an appealing account of the way our listeners heard the sine-wave stimuli. When sine-waves were heard as nonspeech sounds, we suppose that listeners attended to the elements in the acoustic array but not to their organization. In hearing them as speechlike, on the other hand, they attended both to the elements _and to their organization_ (cf. Polanyi, 1969), that together specify, albeit in a highly reduced form, a 'vocal tract' undergoing a bilabial or an alveolar articulation. Those familiar with R. C. James' photograph reproduced in Lindsay and Norman (1972, p. 8) will recognize that the foregoing analogously describes both the initial perception of the picture as a random array of dark and light areas and the subsequent perception of a Dalmatian dog walking in dappled sunlight. Both hearing sine-waves as speechlike and seeing the Dalmatian are compelling perceptions. Perhaps the search for coherence in stimulus information is a general goal of perceptual systems, guided and rewarded by the attainment of clarity (Gibson, 1966). We have already noted that when our listeners switched to hearing sine-waves as speechlike, their identification functions became more consistent and more categorical.

We have suggested that adult listeners perceive speech directly by obtaining information about articulation from an acoustic waveform. In the introduction to this paper, we reviewed experiments that show that infants and adults have similar sensitivities to place of production contrasts. Do infants, like adults, perceive articulatory events? The arguments above lead us to suppose that they do, and that the tendency to search for the coherence in sounds that specifies the events that produce them is an innate predisposition. For a human being, there would be considerable utility in a genetic endowment to detect the particular coherence found in the productions of human vocal tracts. Clearly, further data from three types of currently ongoing experimentation are required to evaluate these assertions. The first is the endeavor to specify how articulatory events structure sound in perceptually accessible ways. This may be achieved through examination of the correspondence between perceptual sensitivity and particular acoustic events in the speech signal (for example, Kuhn, 1975 and the present experiment), but a more fruitful way to specify the metric of the information in speech sounds may be to specify first the metric of articulatory dynamics (Fowler, 1977). The second type of experimentation seeks to delimit the sensitivity of the neonate to acoustic patterns having articulatory relevance (for example, Eimas, 1974; Miller and Morse, 1976) and to plot its ontogenetic refinement (for example, Simon, 1977). The third is experimentation with nonhuman animals reared in controlled environments that assesses perceptual sensitivity in circumstances where, we should predict, there is no innate predisposition to detect information about human articulation (for example, Kuhl and Miller, 1975; Sinnott et al., 1976). The evaluation of these data will be facilitated if attention is given to ensuring that the stimuli used in such experiments dissociate psychoacoustic and phonetic categories as the basis for the measured response.

SUMMARY

In the two experiments reported here, we attempted to determine whether a 'psychoacoustic' basis exists for the classification of continua of synthetic speech sounds that vary in place of articulation. What we mean by a psychoacoustic basis would be the existence of some attribute of the auditory system that predisposes the categorization of acoustic patterns into groups bearing a one-to-one correspondence with their phonetic labels. We studied the categorization both of continua of three formant CV syllables and of sounds modeled on these in which formants were replaced by frequency- and amplitude-modulated sine-waves. Our first experiment produced no support for the psychoacoustic explanation. The sine-wave continua were divided symmetrically into two halves with boundaries corresponding to flat initial transitions. The formant continua, on the other hand, were divided asymmetrically with boundaries corresponding either to rising or falling transitions. However, the results of the second experiment were equivocal: both formants and sine-waves were divided asymmetrically, although the formants were categorized more consistently and tended to be categorized more asymmetrically. Nevertheless, taken together, the results of the two experiments suggested that different information was detected when the sine-wave stimuli were heard as nonspeech and when the formant stimuli were heard as speech. This feeling was endorsed by the finding that the sine-wave stimuli could be perceived as speechlike: that is, they could be heard as initiated by a clear bilabial or alveolar consonant. When heard as speechlike, sine-waves were categorized more like formant stimuli, both more consistently and more asymmetrically than when they were heard as whistles. Thus the pattern of results appeared to relate not so much to the spectral structure of the stimuli, as to the way in which the stimuli were heard.

The perceptual duality of these sine-wave patterns provoked us to scrutinize existing accounts of the difference between speech and nonspeech. These accounts imply that speech is perceived when the acoustic elements in a sound stream can be interpreted by reference to an internalized representation of the vocal tract. We examined an alternative that supposes that the acoustic signal completely specifies the articulatory event that produces it. This account suggests that a sound is perceived as speech when it specifies coherently its source as a human vocal tract participating in a physiologically permissible act of articulation. We find this description of phonetic perception to be underspecified but appealing. From this orientation, the task both of the perceiver and of the experimenter is to determine how the acoustic signal specifies articulatory events.

REFERENCES

Allen, J. and M. P. Haggard. (1977) Perception of voicing and place features in whispered speech: a dichotic choice analysis. Percept. Psychophys. 21, 315-322.

Blumstein, S. E., K. N. Stevens and G. N. Nigro. (1977) Property detectors for bursts and transitions in speech perception. J. Acoust. Soc. Am. 61, 1301-1313.

Cutting, J. E. (1974) Two left-hemisphere mechanisms in speech perception. Percept. Psychophys. 16, 601-612.

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

Eimas, P. D. (1974) Auditory and linguistic processing of cues for place of articulation by infants. *Percept. Psychophys. 16*, 513-521.

Eimas, P. D., E. R. Siqueland, P. W. Jusczyk and J. M. Vigorito. (1971) Speech perception in infants. *Science 171*, 303-306.

Finney, D. J. (1971) *Probit Analysis*. (Cambridge, U.K.: Cambridge University Press).

Fowler, C. A. (1977) Timing control in speech production. Ph.D. thesis, University of Connecticut.

Gibson, J. J. (1966) *The Senses Considered as Perceptual Systems*. (Boston: Houghton Mifflin).

Haggard, M. P. (1971) Encoding and the REA for speech signals. *Quart. J. Exp. Psychol. 23*, 34-45.

House, A. S., K. N. Stevens, T. T. Sandel and J. B. Arnold. (1962) On the learning of speech-like vocabularies. *J. Verbal Learn. Verbal Behav. 1*, 133-143.

Kuhl, P. A. and J. D. Miller. (1975) Speech perception by the chinchilla: voiced-voiceless distinction in alveolar plosive consonants. *Science 190*, 69-72.

Kuhn, G. M. (1975) On the front cavity resonance and its possible role in speech perception. *J. Acoust. Soc. Am. 58*, 428-433.

Lasky, R. E., A. Syrdal-Lasky and R. E. Klein. (1975) VOT discrimination by four to six and a half month old infants from Spanish environments. *J. Exp. Child Psychol. 20*, 213-225.

Liberman, A. M., P. C. Delattre, F. S. Cooper and L. J. Gerstman. (1954) The role of consonant-vowel transitions in the perception of stop and nasal consonants. *Psychol. Monogr. 68*, no. 8 (whole no. 379).

Liberman, A. M., F. S. Cooper, D. P. Shankweiler and M. Studdert-Kennedy. (1967) Perception of the speech code. *Psychol. Rev. 74*, 431-461.

Liberman, A. M., I. G. Mattingly and M. T. Turvey. (1972) Language codes and memory codes. In *Coding Processes in Human Memory*, ed. by A. W. Melton and E. Martin. (New York: Winston), pp. 307-334.

Lindsay, P. H. and D. A. Norman. (1972) *Human Information Processing: An Introduction to Psychology*. (New York: Academic Press).

Mattingly, I. G. and A. M. Liberman. (1969) The speech code and the physiology of language. In *Information Processing in the Nervous System*, ed. by K. N. Leibovic. (New York: Springer), pp. 97-118.

Mattingly, I. G., A. M. Liberman, A. Syrdal and T. Halwes. (1971) Discrimination in speech and nonspeech modes. *Cog. Psychol. 2*, 131-157.

Miller, C. L. and P. A. Morse. (1976) The "Heart" of categorical speech discrimination in young infants. *J. Speech Hearing Res. 19*, 578-589.

Miller, J. D., C. C. Wier, R. Pastore, W. J. Kelly and R. J. Dooling. (1976) Discrimination and labelling of noise-buzz sequences with varying noise-lead times: an example of categorical perception. *J. Acoust. Soc. Am. 60*, 410-417.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. (Ph. D. thesis, University of Michigan.) [*Supplement to Haskins Laboratories Status Report on Speech Research*].

Pisoni, D. B. (1977) Identification and discrimination of the relative onset times of two component tones: Implications for voicing perception in stops. *J. Acoust. Soc. Am. 61*, 1352-1361.

Polanyi, M. (1969) Knowing and being. In *Knowing and Being*, ed. by M. Greene. (Chicago: University of Chicago Press), pp. 123-137.

Popper, R. D. (1972) Pair discrimination for a continuum of synthetic voiced stops with and without first and third formants. *J. Psycholing. Res. 1*, 205-219.

Scharf, B. (1970) Critical bands. In Foundations of Modern Auditory Theory, ed. by J. V. Tobias, vol. 1. (New York: Academic Press), pp. 157-202.

Simon, C. (1977) Cross-language study of speech pattern learning. J. Acoust. Soc. Am. 61, S64(A).

Sinnott, J. M., M. D. Beecher, D. B. Moody and W. C. Stebbins. (1976) Speech sound discrimination by monkeys and humans. J. Acoust. Soc. Am. 60, 687-695.

Stevens, K. N. (1975) The potential role of property detectors in the perception of consonants. In Auditory Analysis and the Perception of Speech, ed. by G. Fant and M. A. A. Tatham. (New York: Academic Press), pp. 303-330.

Stevens, K. N. and A. House. (1972) Speech perception. In Foundations of Modern Auditory Theory, ed. by J. V. Tobias, vol. 2. (New York: Academic Press), pp. 1-62.

Stevens, K. N. and D. H. Klatt. (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.

Studdert-Kennedy, M. (1976) Speech perception. In Contemporary Issues in Experimental Phonetics, ed. by N. J. Lass. (New York: Academic Press), pp. 243-294.