COGNITIVE PSYCHOLOGY 2, 131–157 (1971)

Discrimination in Speech and Nonspeech Modes[1]

IGNATIUS G. MATTINGLY,[2] ALVIN M. LIBERMAN,[3] ANN K. SYRDAL,[4] AND TERRY HALWES[5]

Haskins Laboratories, New Haven, Connecticut 06510

Discrimination of second-formant transitions was measured under two conditions: when, as the only variation in two-formant patterns, these transitions were responsible for the perceived distinctions among the stop-vowel syllables [bæ], [dæ], and [gæ]; and when, in isolation, they were heard, not as speech, but as bird-like chirps. The discrimination functions obtained with the synthetic syllables showed high peaks at phonetic boundaries and deep troughs within phonetic classes; those of the nonspeech chirps did not. Reversal of the stimulus patterns, producing vowel-stop syllables in the speech context and mirror-image chirps in isolation, affected the speech and nonspeech functions differently. An additional nonspeech condition, presentation of the transitions plus the second-formant steady state, yielded data similar to those obtained with the transitions in isolation. These results support the conclusion that there is a speech processor different from that for other sounds.

For many years, the authors and their colleagues have been interested in the differences in perception between speech and other sounds. That a difference exists is suggested first by the nature of the relation between the perceived phonetic message and the acoustic signal that conveys it: message and signal are linked by a complex code for which there is no parallel in any class of nonspeech sounds; we therefore infer that speech perception is accomplished by a special decoder (Liberman *et al.*, 1967). This complex speech code is not unique, but is, rather, similar in form to the grammatical codes at the higher levels of language: syntax and phonology (Mattingly & Liberman, 1969).

These inferences are supported by experimental results that point more directly to a special mode of perception for speech and suggest that this mode is related to a still broader one that characterizes perception of

[1] This paper incorporates data some of which have been reported earlier by Mattingly *et al.* (1969), Syrdal *et al.* (1970), and Liberman (in press). Support from the National Institute of Child Health and Human Development, the Office of Naval Research, and the Veterans Administration is gratefully acknowledged.

[2] Also University of Connecticut.

[3] Also University of Connecticut and Yale University.

[4] Also University of Minnesota.

[5] Also University of Connecticut.

language in general. Numerous experiments on dichotic listening indicate that the encoded sounds of speech (like the higher levels of language) are normally processed primarily in the left hemisphere of the brain, while nonspeech sounds and the relatively unencoded aspects of speech (such as steady-state vowels) are either processed in the right hemisphere or are not lateralized at all (Kimura, 1964, 1967; Kirstein & Shankweiler, 1969; Shankweiler & Studdert–Kennedy, 1967; Studdert–Kennedy & Shankweiler, 1970).

Other experimental observations imply additional differences in perception between speech and nonspeech. One such observation, which is particularly relevant to the experiments to be reported here, is that the encoded acoustic cues sound very different in and out of speech context. Though the difference has not been precisely measured, its existence is clear enough. When transitions of the second formant, which are sufficient cues for the place distinctions among stop consonants, are presented in isolation, we hear them as we should expect to—that is, as pitch glides or as differently pitched "chirps." But when they are embedded in synthetic syllables, we hear unique linguistic events, [bæ], [dæ], [gæ], which cannot be analyzed in auditory terms. Thus, speech perception cannot be straightforwardly mapped onto the physical dimensions of the speech signal.

There is a more specific sense in which speech perception does not correspond to acoustic reality. If asked to discriminate physically continuous variations in a speech cue, a listener does not hear a continuum of sounds, but, rather, quantal jumps from one sound to another. His discrimination function displays high peaks at phonetic boundaries. These high peaks (and the adjacent troughs) reflect a kind of perception in which the listener hears phonetic units but not intraphonetic variations. In the extreme case, he discriminates no more stimuli than he can absolutely identify. Perception of this sort has been called "categorical" (Liberman *et al.*, 1957; Studdert–Kennedy *et al.*, 1970); it is unusual, if not unique, since, in the perception of nonspeech sounds, many more stimuli can be discriminated than can be identified. Of course, categoricalness is a property of language generally: active–passive and singular–plural, for example, do not admit of degree.

In this paper we shall make use of categorical perception to study the difference between speech and nonspeech. To capture the difference as directly as possible, we will compare listeners' discrimination of the same acoustic variable, once in speech context, where it serves as a cue for a phonetic distinction, and once in nonspeech, where it does not.

Several such comparisons have already been made. In one of these studies (Liberman *et al.*, 1961a), the acoustic variable was the "cutback"

or delay of onset of the first formant, which in initial position is a major cue to the voiced–voiceless distinction. The speech-like stimuli were made on the Haskins Pattern Playback from a series of spectrographic patterns with increasing delay in the onset of the first formant (F1) relative to the onsets of the second and third formants (F2 and F3). Stimuli for which the delay was sufficiently long were heard as [to], other stimuli as [do]. The nonspeech control stimuli were synthesized from inverted versions of these same spectrographic patterns. Thus, the same information was present in both speech-like and control stimuli, but the control stimuli did not sound like speech. The inversion, however, affected the acoustic variable itself; as the authors pointed out,

> ". . . in the control stimuli the formant whose time of onset varied was at a higher frequency than the other two formants, while in the speech stimuli it lay at a lower frequency than the other formants."

Subjects were asked to identify the speech stimuli as [to] or [do] and, in the case of both speech and control stimuli, to discriminate between neighbors along the acoustic series. For a typical S, the speech discrimination function showed a peak at a delay of 20–30 msec, corresponding to the phonetic boundary predicted by the cross-over point of the two identification functions, while the control discrimination function showed no such peak and in fact never rose very far above the chance level.

In the other study (Liberman *et al.*, 1961b) the acoustic variable was the length of the silent interval associated with stop consonants; in intervocalic position this length is a cue to voicing. The speech-like stimuli were synthesized from a series of spectrographic patterns representing a word containing a medial stop, with a silent interval of increasing length: stimuli for which the interval was sufficiently long were heard as *rapid*, other stimuli as *rabid*. Each control stimulus consisted of two bursts of band-limited white noise with the same durations and energy envelopes as the two syllables of a speech stimulus, and separated by a silent interval. The silent intervals matched those of the speech stimuli. As in the [to]–[do] study, Ss were asked to identify the speech stimuli and to discriminate the speech and the control stimuli. The speech discrimination functions showed peaks at the phonetic boundary; the control discrimination functions showed no peaks, and were, in general, lower than the speech functions, but substantially higher than chance.

Both of these studies indicated that perception of the relative timing of two acoustic events was different depending on whether the difference in timing cued a distinction between two speech sounds. In the case of speech there were peaks in the discrimination functions at the phonetic boundaries; in the case of nonspeech there were not. Moreover, the

results indicated that the peaks in the speech represented, by comparison with nonspeech, a sharpening of discrimination at phonetic boundaries, not a reduction of discrimination within the phonetic category. Although these results are suggestive, their interpretation is complicated by the fact that the acoustic variable was the same for speech and nonspeech only in a derived sense: the time intervals between two sounds were identical, but the sounds themselves were different.[6]

The purpose of the two experiments reported here was to provide a more appropriate nonspeech context for the comparison with speech. To that end, we examined the perception of the second-formant transition. Unlike the timing cues of the earlier studies, the second-formant transition is itself an actual acoustic event. The problem of devising an appropriate nonspeech control context thus becomes much more straightforward. In fact, it is possible to use the simplest context of all: isolation. As we have noted, second-formant transitions distinguish [b], [d], and [g] in speech context, but in isolation sound like chirps.[7]

EXPERIMENT I

The purpose of the first experiment was to compare the discrimination of F2 transitions in stop-vowel syllables and in isolation.

Method

Stimuli. The Haskins Laboratories computer-controlled synthesizer (Mattingly, 1968) was used to produce the stimuli of the experiment. A stimulus to be synthesized is specified by time functions for each of the several parameters of the synthesizer (e.g., F0, the fundamental frequency; F1, the first-formant frequency, and so on). Each of these functions is represented by a series of digital values stored in computer

[6] An interesting and somewhat relevant experiment, in which the speech and nonspeech context were determined not by the stimuli but by the *Es*' instructions to the Ss, has been carried out by Cross and Lane (1964). Presented with synthetic speech stimuli of marginal realism, one group of Ss was told that they were being tested in speech-sound discrimination, while another group was told that the test had to do with discrimination of tones. The discrimination functions obtained with the first group showed peaks at the phonetic boundaries; the discrimination functions for the other group did not.

[7] This method of comparing speech and nonspeech was suggested by Kirstein's (1966) pilot study, in which she used isolated second formants with both an initial transition and a following steady state—what we have called "bleats" in this paper. In an experiment applying detection theory to categorical perception, Popper (1967) included a same–different discrimination test using bleats with final transitions in the [b]–[d] range. The d' function obtained can be compared with that for a test with the same Ss using speech stimuli. The results of both Kirstein and Popper are consistent with the results reported here.

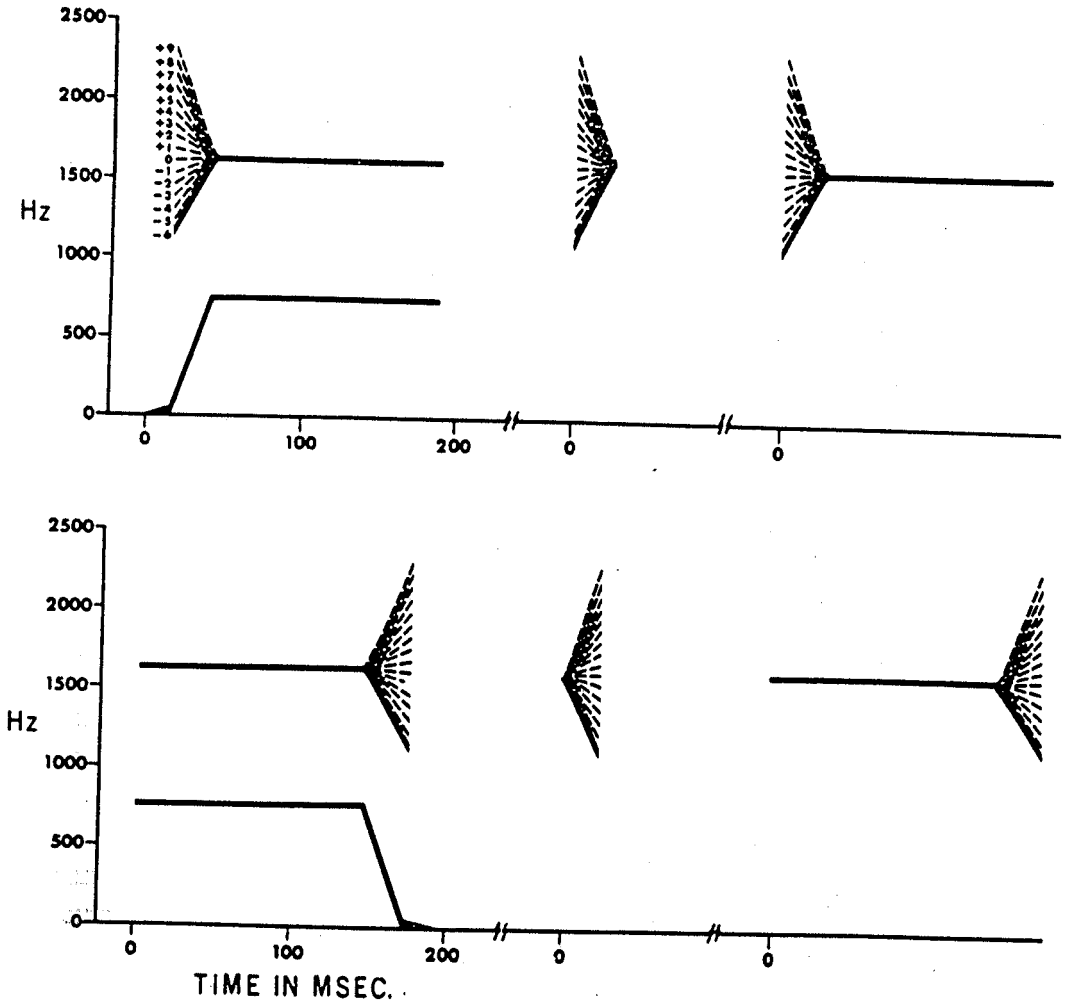

FIG. 1. Top: stimuli for forward condition, with initial transitions. Syllables (left), chirps (center), bleats (right). Bottom: similar stimuli for backward condition, with final transitions.

memory. To produce the stimulus, a set of values, one for each parameter, is transmitted every 5 msec by the computer to the synthesizer.

The two sets of stimuli used in Experiment I are shown in Fig. 1, top left and center. (The other stimuli shown in Fig. 1 were used only in Experiment II.) The set at top left are speech stimuli and consist of 16 syllables, each beginning with a voiced stop and ending with the vowel [æ]. In all the syllables, the fundamental frequency is constant at 90 Hz, and only the first and second formants of the synthesizer are used. A 15-msec period of closure voicing, represented by a low-amplitude F1 at 150 Hz, is followed by a 40-msec transitional period during which the two formants move toward the steady-state frequencies to [æ]: F1 = 740 Hz, F2 = 1620 Hz. The steady-state period of the stimulus is 190 msec long. Throughout the stimulus, the two formants are of equal amplitude. The F1 transition always starts at 150 Hz. The experimental

variable is the starting point of the F2 transition. This is varied in 15 approximately equal steps from 1150 to 2310 Hz. In Fig. 1, top left, the level transition, for which the starting point is 1620 Hz, is labeled 0; transitions with higher (or lower) starting points are labeled positively (or negatively) with reference to the level transition. Depending on the starting point (and therefore the slope) of the F2 transition, these stimuli are heard as [bæ], [dæ], or [gæ].

The second set of stimuli (Fig. 1, top center) are the nonspeech controls. They consist simply of transitions identical to those of the first set, but with the closure voicing, the steady state of F2, and all of F1 absent. In the first set—that is, in the syllables—the transitions were the only cues to point of articulation. In the second set the transitions have been removed from their speech contexts and do not sound at all like speech. To most listeners they sound like chirps, and it is not hard, at least in the case of the more extreme members of the set, to tell whether a chirp is rising to a higher or falling to a lower frequency.

Procedure. With the aid of the synthesis system, the digital parametric representations of all the stimuli were stored on a disc file, and the tests required for the various experiments were then automatically compiled and recorded. The tests included an identification test for the syllables and discrimination tests for the syllables and for the chirps.

The purpose of the identification test was to determine where, and how reliably, the *S* placed the phonetic boundaries. It consisted of 160 syllables in 10 groups of 16. Each of the different syllables occurred once in each group, and each group was differently randomized. The *S*'s task was to identify each of the 160 syllables as beginning with [b], [d], or [g].

To find out how well the Ss could discriminate the stimuli, we used an oddity method: each item in the test consisted of a triad in which one member of a pair of stimuli to be discriminated occurred once, and the other, twice; the *S*'s task was to select the odd stimulus. For each pair there are six ways in which a triad can be ordered. Pairs of stimuli two steps apart along the continuum of Fig. 1 were to be discriminated; for each set there are 14 such pairs. Each test consisted of the 84 possible triads in 6 groups of 14. Each stimulus pair was used to form one triad in each group. The assignment of the six triad orderings to the six groups was separately randomized for each pair; the order of pairs within a group was separately randomized for each group. There were four differently randomized forms of the discrimination test. The tests for syllables and chirps were made in the same way.

The tests were presented to the Ss over headphones. The gain on the tape recorder was set so that Ss could listen to the syllable stimuli comfortably; this same gain setting was used for the chirps.

For each S, there were five experimental sessions on five separate days. On each day the S was given different forms of the discrimination test for the syllables and different forms of the discrimination test for chirps, in random order. Altogether, he received all four forms of the syllable discrimination test twice and all four forms of the chirp discrimination test twice. Thus, for each stimulus comparison, each S gave 48 judgments. The identification test was given once on each of the first, second, fourth, and fifth days. Each stimulus was presented for judgment 10 times on each identification test; there was then a total of 40 judgments per stimulus.

Subjects. There were seven Ss, all undergraduate students at the University of Minnesota and all paid volunteers. None was told the purpose of the experiment.

Results

In Fig. 2 are the results for two of the seven Ss, chosen on a basis to be described later. The upper portion of the block for each S plots his

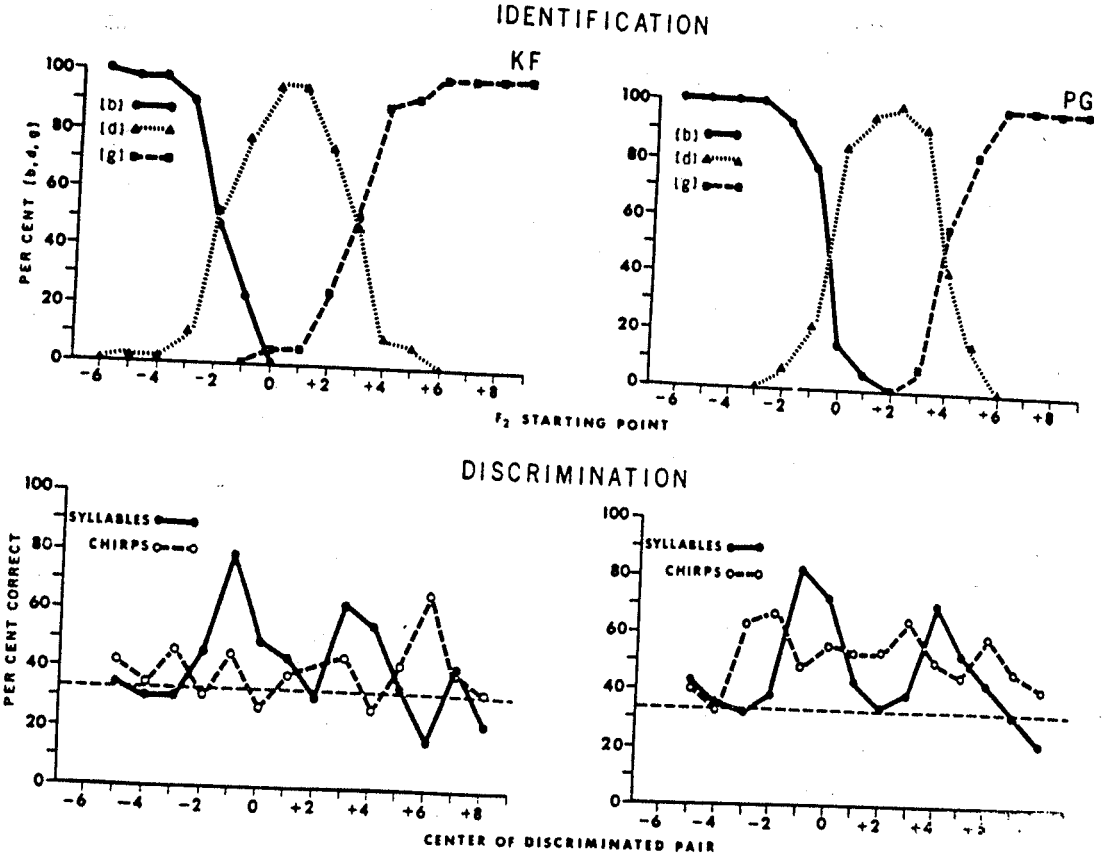

FIG. 2. Identification functions for syllables and discrimination functions for syllables and chirps for two Ss.

identification functions for [b], [d], and [g]: the abscissa represents the stimuli ordered according to the series of F2 starting points; the ordinate represents the percentage of response for each of the three stops.

Both of the Ss shown sorted the stimuli cleanly into the three phonetic categories. The areas of uncertainty are small by comparison with those where the Ss apply the phonetic labels with consistency. Of the seven Ss, six yielded identification functions approximately as reliable as those shown; moreover, agreement among the Ss in the location of the phonetic boundaries is almost perfect. One S did very poorly on the identification; he labeled stimuli inconsistently, and there was substantial overlap in the identification functions for the three stops. We have rejected all the data from this S because we suspect that he did not hear the synthetic patterns very well as speech; if he did not, then a comparison of the way he perceived speech and nonspeech stimuli, which is the purpose of this experiment, becomes meaningless.

The lower portion of each block in Fig. 2 plots the S's discrimination functions for syllables (solid line) and for chirps (dashed line). Each point along the abscissa corresponds to the stimulus pair whose members are the stimuli one step higher and one stop lower in the series than the stimulus represented by the corresponding point in the abscissa of the identification test plot. The ordinate is the percentage of correct discriminations for each pair. The horizontal broken line at 33% represents the level of discrimination expected by chance.

For the syllables, the discrimination function shows peaks near the phonetic boundaries indicated by the identification functions for each S. Since the boundaries are constant from S to S, the locations of the peaks are likewise constant. The peaks for [b]–[d] boundaries are generally somewhat higher than those for [d]–[g] boundaries. Away from phonetic boundaries, the discrimination functions are at or near chance.

The chirp discrimination functions are quite different. There are no peaks in discrimination at points corresponding to the phonetic boundaries. Both Ss have a peak at +6, but we believe that this is to be attributed to an artifact resulting from a previously unremarked shortcoming of the synthesizer: its pitch generator was free-running, so that the occurrence of the first pitch pulse of a chirp (or indeed, of any other stimulus) could lag by as much as half a pitch period (6.5 msec) behind the nominal starting point. The synthesizer parameter values change stepwise; for the more extreme stimuli, for which F2 moves rapidly, there would, therefore, be substantial variation in the actual initial frequency, as well as in the duration, of the transition in the different tokens of the "same" stimulus. Such variation was, of course, randomized across these several tokens. However, inspection of the tokens for

Stimuli +5 and +7 (discrimination of which produced the peak at +6), reveals that the variations were unbalanced in such a way that careful listeners could discriminate accurately on the basis of differences in duration or exaggerated differences in F2 starting point. That this is, in fact, the cause of the peak is indicated by the results of later experiments in which we synchronized the pitch pulses and the peak at +6 disappeared.

The two Ss whose results are shown in Fig. 2 were chosen to illustrate the extremes in the general level at which the chirps were discriminated. One of them (KF) discriminates the chirps at a level only slightly above chance, except at +6; the other (PG) does considerably better. In general, the variation among Ss in level of discrimination, and also in the shape of the function, was greater for the chirps than for speech.

In Fig. 3 is a plot of the pooled discrimination data of the six (out of

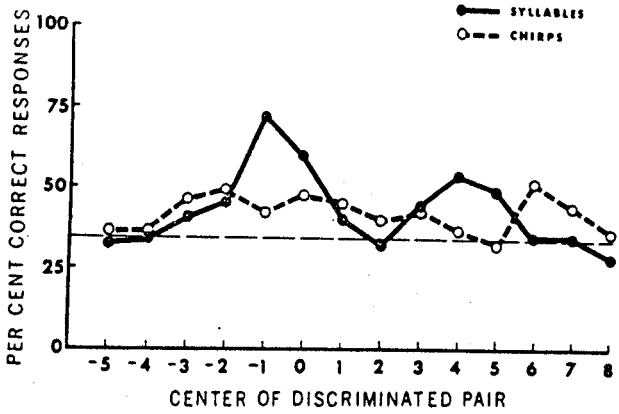

FIG. 3. Pooled discrimination functions for syllables and chirps for six Ss.

seven) Ss who identified the syllables well. The chirp discrimination function and the syllable discrimination function are clearly different. The chirp function is low (except for the peak at +6) but above chance. The syllable function shows peaks near phonetic boundaries and is at or near chance away from phonetic boundaries. The Ss' perception of the second-formant transition apparently depends on whether they are listening in the speech mode.

EXPERIMENT II

The second experiment was prompted by the observation, made in one of the studies with synthetic speech, that the F2 transition is a less powerful cue to place of articulation in final position than in initial position (Liberman *et al.*, 1954). Though this difference reflected directly only the relative difficulty of identifying the transitions, it is reasonable

to suppose that discrimination might also be different in final position than in initial. Preliminary experiments have since suggested that this is so. As with so many findings in speech perception, the question arises whether this difference is to be accounted for psychoacoustically or whether it is, rather, a consequence of the special processing that the speech signal undergoes. If the explanation is psychoacoustic, then we should expect that the F2 transitions in nonspeech context—that is, the chirps—would also be differently discriminated in final position. The second experiment was designed to provide relevant data. In it we have compared the discriminability of the F2 transitions in initial and final positions when they are, in one condition, cues for speech and, in another, not.

The second experiment was intended also to determine whether possible reservations about the chirp control are justified. It might be argued that this control is faulty: when the F2 is in initial position in the syllable, the vowel steady state may provide a reference that is, of course, absent in the chirp. When F2 is in final position in the syllable, as in this experiment, the steady state may provide a reference and, conceivably, a fatigue effect. Therefore, we introduced in Experiment II an additional set of nonspeech control stimuli (Fig. 1, top right). These stimuli have not only the various second-formant transitions, as do the chirps, but also the second-formant steady state. Naive Ss do not commonly hear these as speech. We have called them "bleats."

Six sets of stimuli were required for the experiment: F2 transitions in initial and final positions in two-formant syllables; F2 transitions in isolation in "initial" and "final" positions (chirps); and F2 transitions attached to steady-state second formants in initial and final positions (bleats). The syllables and chirps with initial F2 transitions were produced as in Experiment I; the bleats with initial transitions were produced by synthesizing two-formant syllables with F1 turned off. The production of the stimuli was better controlled than in Experiment I. The synthesizer was made to produce its first pulse at the start of every stimulus, instead of randomly, so that each token of a stimulus had exactly the same duration and frequency excursion, thus eliminating the basis for the pile-up of correct discriminations at +6 in the first experiment. It was not necessary to produce separate sets of stimuli with final F2 transitions, since these stimuli (Fig. 1, bottom) were equivalent to the available stimuli in reverse temporal order. Thus, tests requiring stimuli with initial transitions were run by playing the test tapes forward; tests requiring stimuli with final transitions were run by playing these same tapes backward.

Procedure. The formats of the identification test (for the syllables) and

the discrimination test (for the syllables, chirps, and bleats) were the same as in Experiment I. The S's task included the oddity judgment (selecting the one stimulus of each triad that he thought different from the other two) used in Experiment I and, in addition, a confidence rating. For the purposes of the confidence rating, the S was asked to estimate the correctness of each discrimination judgment on a three-point scale. These estimates were then treated according to a method developed by Strange and Halwes (in press) and successfully applied by them to increase the sensitivity of discrimination measures of the voiced–voiceless distinction. By their method, the confidence-rating score for each discriminated pair is determined by multiplying the number of correct responses for which the S used a particular confidence rating by a weight assigned to this rating; summing these products over all ratings; and dividing by the number of trials per pair to give a number between 0 and 1. The weight is equal to $(3p - 1)/2$, where p is the ratio, for all pairs in a given testing condition, of the number of correct responses for which a particular confidence rating was used, to the total number of responses for which this rating was used. Thus, the weight for a rating is 0 when the level of discrimination over all pairs is at chance ($p = 1/3$); and 1 when discrimination is perfect ($p = 1$). The advantage of the confidence rating is that it permits a reliable approximation of a S's discrimination function with fewer responses per stimulus pair than if only the correctness or incorrectness of his responses is considered.

All Ss were given (1) the syllable identification test, once in the forward and once in the backward condition, (2) the syllable discrimination test, three forms forward and three forms backward, and (3) one of the two nonspeech discrimination tests, three forms forward and three forms backward. The chirps served as nonspeech controls for half the Ss, the bleats for the other half. For each S, there were three separate test sessions on three separate days. Each chirp S took a different form of each of the four discrimination tests (forward and backward, syllables and chirps) each day in a different random order. In the case of the bleat Ss, however, since the bleats were more like syllables than the chirps, we thought it wiser to protect the Ss' naivete by presenting all the bleat tests before all the syllable tests. During the first day and a half, therefore, each bleat S took three forms of the forward and backward bleat tests; during the remaining day and a half, he took three forms of each of the two speech tests. Thus, for each discrimination test, there were 18 judgments per stimulus pair for each S. The syllable identification test in the forward condition was given to all Ss at the end of the second day, and the identification test in the backward condition at the end of the third day.

Subjects. There were 11 Ss, all undergraduate students at the University of Minnesota and all paid volunteers. None was told the purpose of the experiment. Three Ss were eliminated because of their inability to identify the syllables accurately. Data were provided, then, by eight Ss, four in each of the two experimental subgroups (chirps and bleats).

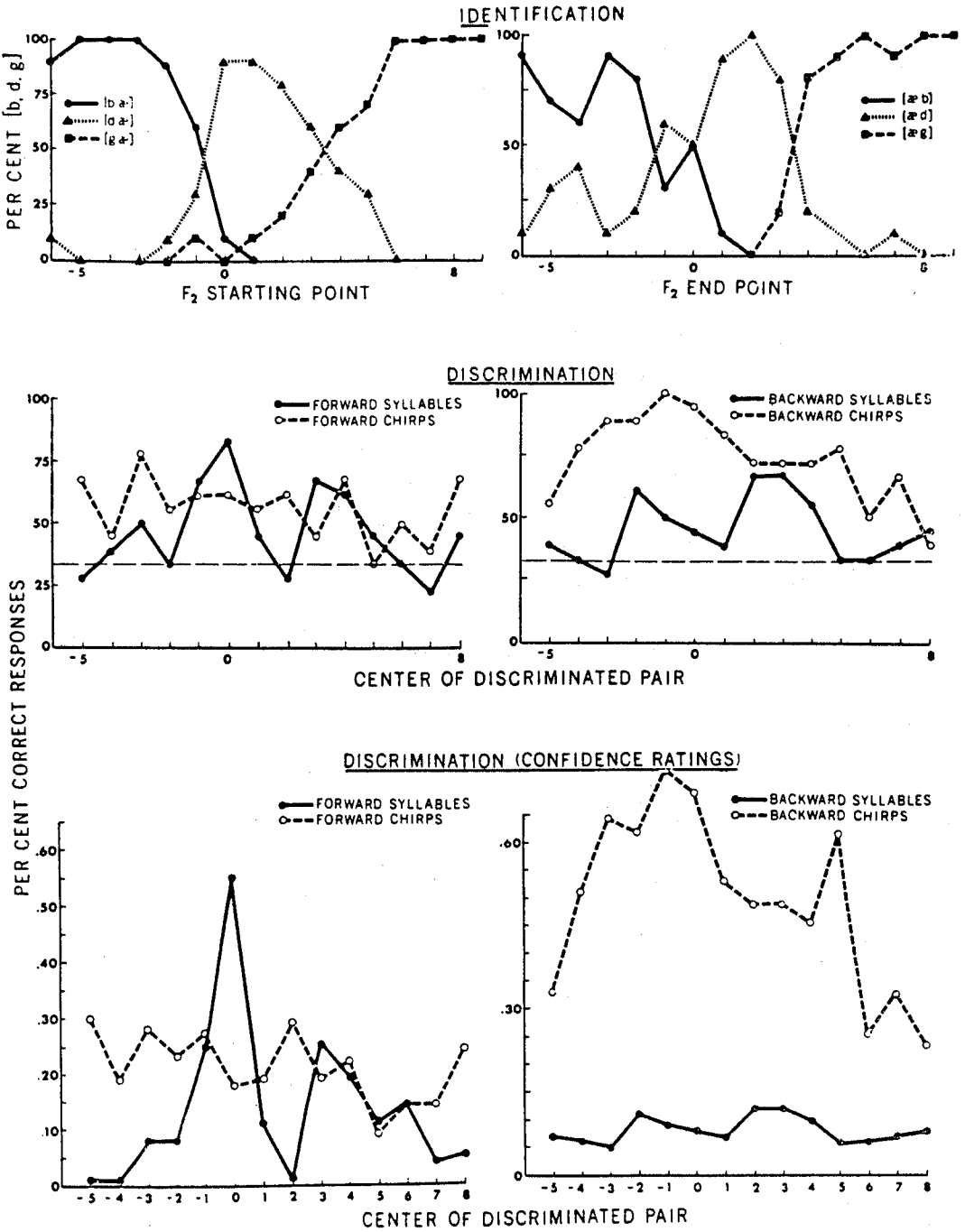

FIG. 4. Identification function for syllables and discrimination functions for syllables and chirps for one S.

Results

In Fig. 4 are the results for one typical S in the syllable-chirp half of Experiment II. In the left-hand column are the results for the forward condition and in the right-hand column the results for the backward condition. The topmost graphs show his identification functions; the middle graphs, his discrimination functions without regard to his confidence ratings; and the lowest graphs, his discrimination functions, taking into account the confidence ratings.

Figure 5 shows syllable and chirp discrimination functions based on pooled data for all four Ss. The upper portion of the figure shows the forward condition; the lower portion, the backward condition.

For the forward condition, the results are consistent with the first experiment. Discrimination functions for syllables peak at the phonetic boundaries implied by the identification functions, but tend toward random elsewhere. Discrimination functions for chirps appear to have no relation to discrimination functions for syllables. The characteristic peaks and troughs of syllable discrimination are even more pronounced in the confidence-rating analyses; on the other hand, the adventitious peaks of the chirp functions tend to be leveled. Still, chirp discrimination levels for all four Ss are clearly above random. One exceptional

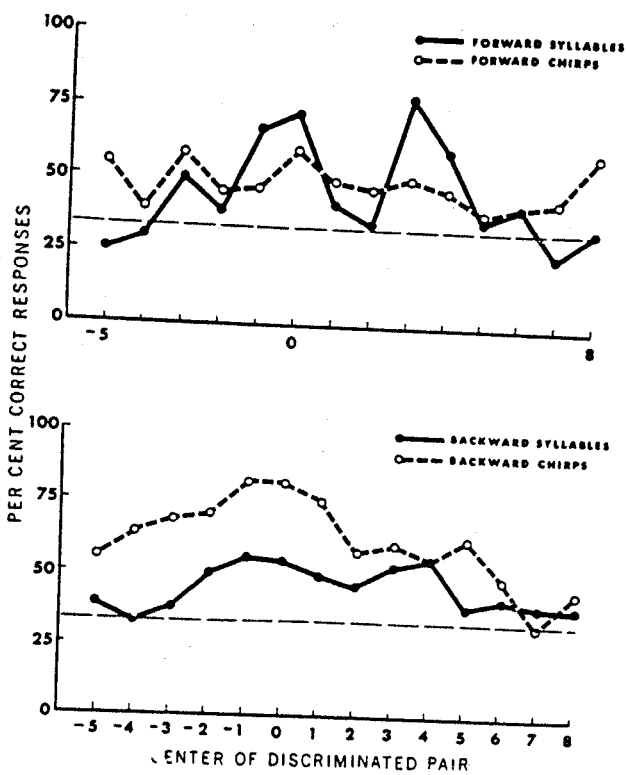

FIG. 5. Pooled discrimination functions for syllables and chirps for four Ss.

S has a much higher overall level of chirp discrimination than that of the S shown in Fig. 4 (or, indeed, of any of the other Ss). That S also has chirp peaks at the same points as his speech peaks, 0 and +3; in his confidence-rating analysis the peak at 0 becomes more pronounced by comparison with the one at +3.

Syllables, as expected, are much less consistently identified in the backward than in the forward condition. There is also a certain tendency, shown by all Ss, for the cross-over point for [d]–[g] to move to the right, increasing the range over which Ss tended to hear [d]. Poorer identification functions predict lower peaks in the discrimination functions, and indeed, for the S shown in Fig. 4 and for all other Ss, syllable discrimination peaks are lower in the backward condition. The confidence-rating analysis accentuates this difference between the two conditions. But while the peaks are lower, the troughs are not so deep. The difference in both peaks and troughs is obvious in the pooled data of Fig. 5.

Unlike the syllables, chirps are clearly much better discriminated in the backward than in the forward condition. This is true of all Ss though the absolute level of performance varies among Ss just as in the forward condition. The discrimination functions for the backward chirps for two Ss are as good as their backward syllable discrimination functions, and for the two other Ss, including the one for whom data are given in Fig. 4, the chirp functions are substantially better than the syllable functions at every point along the abscissa. All four backward chirp functions have their highest peak in the −1, 0, +1 range, but Ss tend to have idiosyncratic peaks elsewhere. The confidence-rating analyses emphasize the difference between forward and backward chirps and between backward chirps and backward syllables, and accentuate the peaks near 0. The improved discrimination of chirps in the backward condition, and the tendency to peak in the −1, 0, +1 range, are apparent from comparison of the forward and backward chirp functions in Fig. 5. In short, perception of chirps differs greatly from perception of syllables in the backward as well as in the forward condition; and the increase in discrimination induced by reversing the chirps does not appear to parallel the similarly induced change in perception of syllables.

The results for the bleat Ss are quite similar to those for the chirp Ss. In Fig. 6 are the data obtained from a typical S, arranged as in Fig. 4. Pooled data for all four Ss, showing discrimination functions for syllables and bleats, are shown in Fig. 7 (cf. Fig. 5). Discrimination of syllables is high at phonetic boundaries, near random elsewhere; identification is more consistent and discrimination of the boundaries better in the forward condition than in the backward condition. In fact, for these Ss

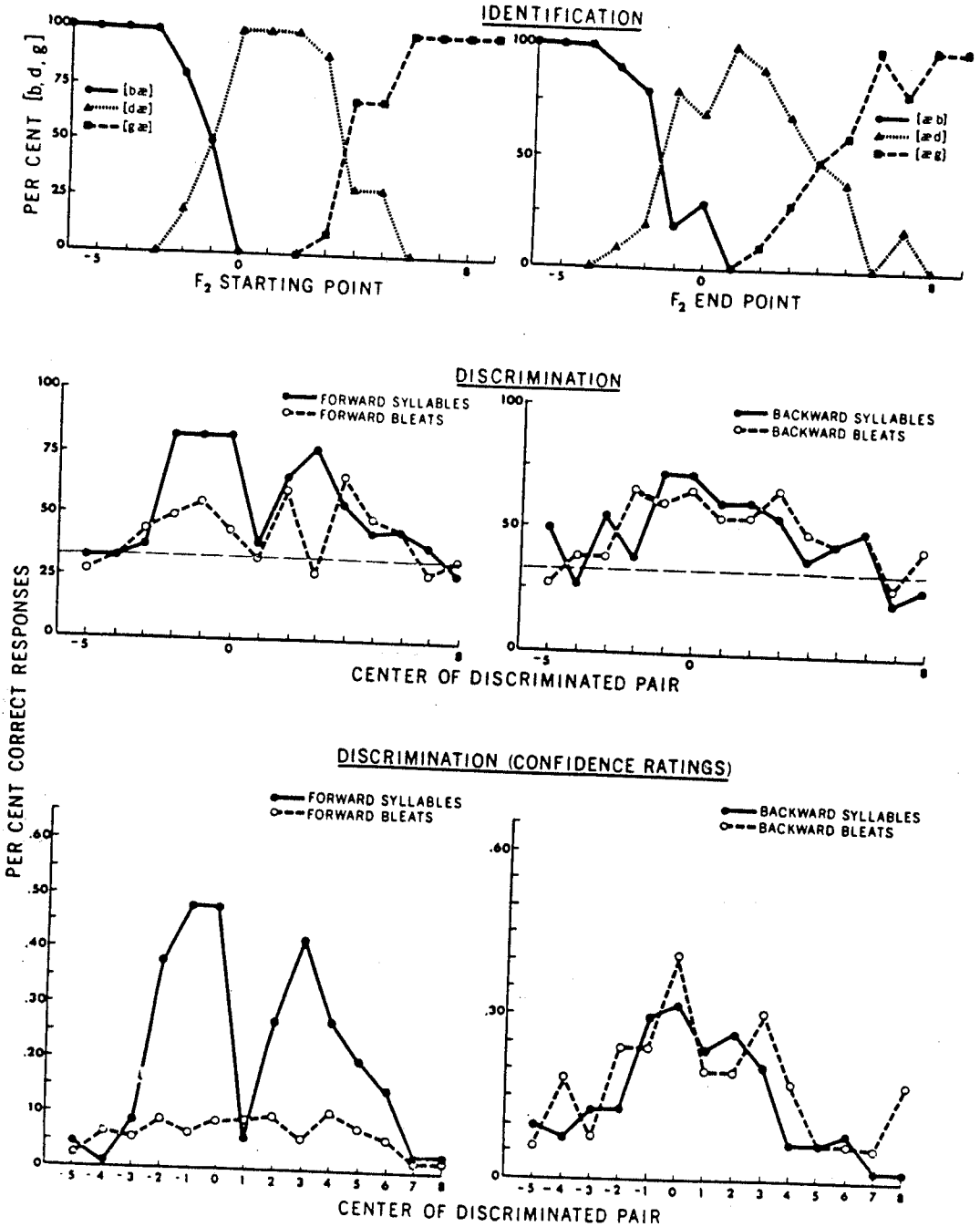

FIG. 6. Identification function for syllables and discrimination functions for syllables and bleats for one S.

the backward syllable discrimination function has lost its bimodal shape and its characteristic troughs, and looks not unlike the backward chirp function.

To facilitate comparison of the results with chirps and bleats, we

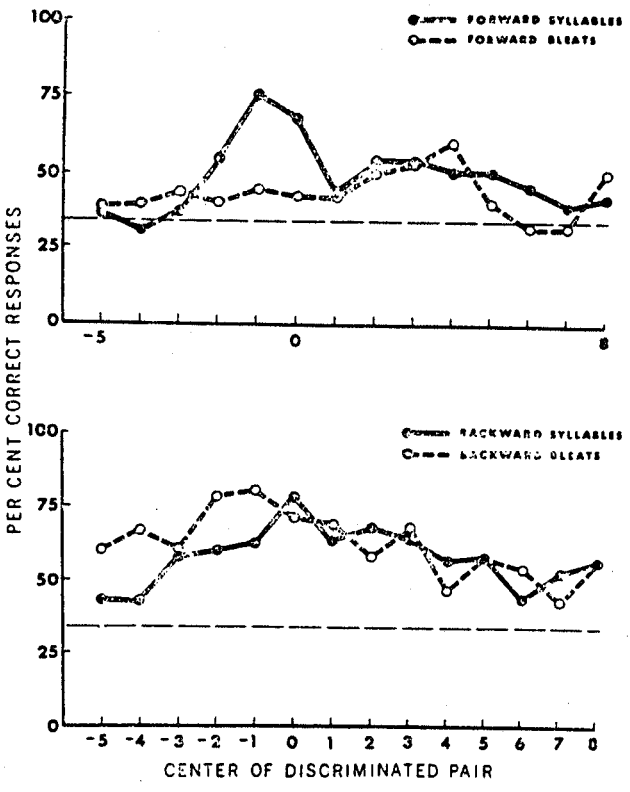

FIG. 7. Pooled discrimination functions for syllables and bleats for four Ss.

have presented together in Fig. 8 the pooled data for these two non-speech controls. The discrimination functions for the bleats parallel those for the chirps: the functions in the forward condition are above random, low and irregular, while the functions in the backward condition are considerably higher and show peaks in the −1, 0, +1 range. As with the backward chirps, individual Ss (including the S shown in Fig. 6), show idiosyncratic peaks in their backward bleat functions, but there is no sign in either forward or backward chirp or bleat functions of an artifact such as gave trouble in Experiment I. However, in the forward condition, discrimination of the chirps is somewhat better than discrimination of the bleats.

At this point, we must consider whether there is any difference between the discrimination functions for the chirps and those for the bleats which would lend plausibility to the argument that the comparison between chirps and speech is in one respect or another unfair. Had we found that bleats were discriminated better than chirps in either forward or backward conditions, we might have supposed that the absence of a steady-state second formant at a constant frequency in the chirp stimuli made them more difficult to perceive than the syllable stimuli. No such result was obtained; in fact, forward bleats are not discriminated

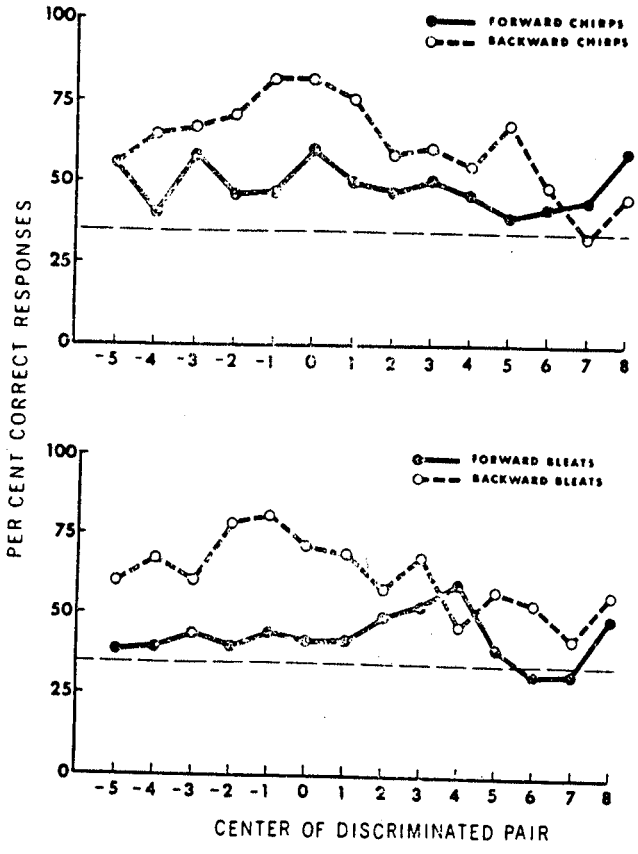

FIG. 8. Comparison of pooled chirp and pooled bleat discrimination functions.

quite so well as forward chirps. (This is probably attributable to the fact that bleat Ss took all the nonspeech discrimination tests first.) Had we found that backward chirps were discriminated better than backward bleats, with no comparable improvement in the forward condition, we might have supposed that the absence of a fatiguing steady state in the chirps made them easier to perceive than the syllables. Though our bleat control was imperfect, since it is still possible to argue that fatigue might be induced by the presence of the steady states of both first and second formants, the least that can be said is that the outcome of the bleat experiment does not encourage such an argument. Chirps are discriminated at the same level as bleats in the backward condition. Since the shapes of the corresponding chirp and bleat functions are similar, the effect of the second-formant steady state can probably be ignored; and it will be convenient for purposes of our discussion to pool the results for the two groups of Ss in Experiment II, as in Fig. 9 and Fig. 10.

Let us sum up the results of Experiment II, referring to Figs. 9 and 10. In forward condition, the speech discrimination function shows peaks

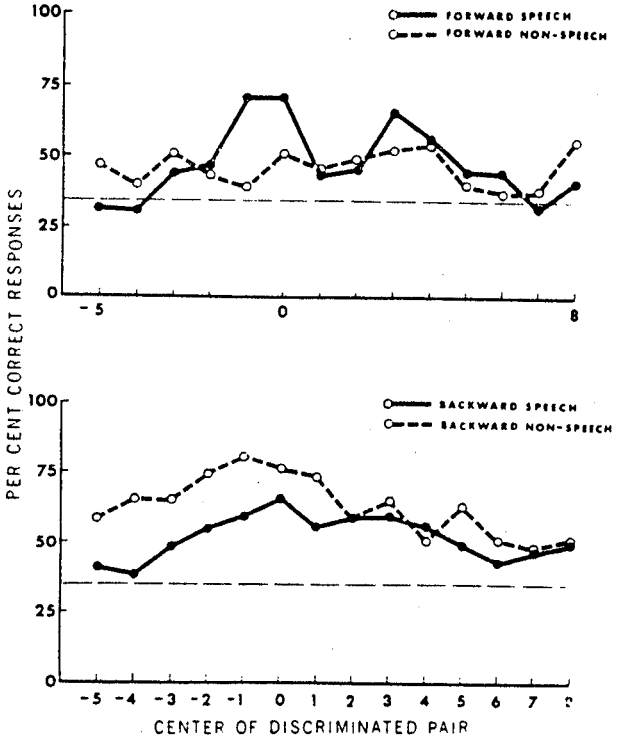

FIG. 9. Comparison of speech and nonspeech discrimination functions.

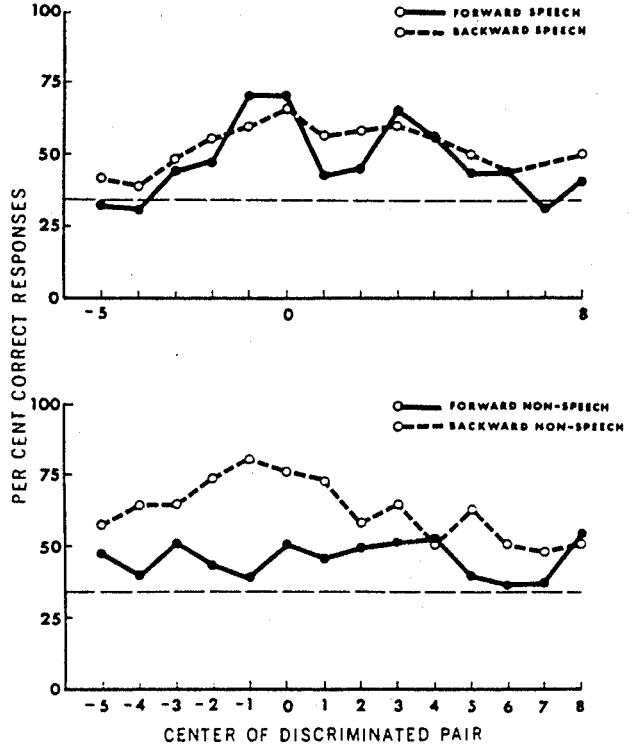

FIG. 10. Comparison of forward and backward discrimination functions.

at phonetic boundaries and troughs within phonetic categories. The nonspeech function shows no such peaks or troughs; it is irregular and low, though above random. In backward condition, the level of discrimination for speech is about the same as in forward condition, but the function has all but lost the peaks and troughs. The nonspeech function peaks near zero; it is higher than the speech function and much higher than the nonspeech function in forward condition. Thus, speech and nonspeech differ in each condition (Fig. 9) and the change of conditions affects speech in one way and nonspeech in another (Fig. 10).

GENERAL DISCUSSION

There are three different classes of phenomena to be accounted for: the responses of Ss to the chirps and bleats which served as nonspeech control stimuli; the responses to the speech-like stimuli; and the differences in the response to the corresponding speech-like and nonspeech stimuli.

We must first attempt to interpret the results for the nonspeech stimuli. For convenience we will speak of chirps, but it will be seen that the argument applies just as well to the bleats. Since, surprisingly, we have been able to find only one psychoacoustic study of dynamically varied resonances (Brady *et al.*, 1961), against which we could check our conclusions, this interpretation must be considered as highly tentative.

For each of several stimuli similar to our chirps, with various durations and initial and final frequencies, Brady *et al.* asked their Ss to adjust the frequency of a steady-state resonance until it sounded most like the test stimulus. The Ss showed a very pronounced tendency to select a steady-state frequency approximately equal to the final frequency of the chirp. It seems plausible to infer that for some reason Ss find it easier to estimate the final frequency of a chirp than its frequency at some earlier moment. If so, we should expect to find, as we do in the present experiment, that a discrimination task in which the stimuli differed most in their final frequencies and not at all in their initial frequencies (the backward condition) would be easier than a task for which the reverse was true (the forward condition).

But we cannot go on to assume that in our experiment Ss discriminate simply by comparing the three estimated frequencies of each oddity triad. Suppose that Ss were given a chirp discrimination test in which both the initial and final frequencies of the chirps were varied. Before a S could compare the three chirps in a triad, it would be necessary for him (1) to estimate the frequency at some fixed time during each of the three chirps; and (2) to determine the slope of each chirp. How-

ever, in the special case where either the initial or the final frequencies of all chirps are constant throughout a test, either (1) *or* (2) would give the *S* sufficient information to discriminate the stimuli from one another.

Which of these two methods are the Ss using? In the case of the backward chirps it seems clear that Ss are using method (2). They discriminate best those pairs of stimuli straddling values −1, 0, +1, i.e., pairs having negative and zero, negative and positive, and zero and positive slopes, respectively. It is not surprising that these three special cases of slope comparison should prove easy. On the other hand, these pairs of stimuli have no particular significance in terms of method (1), comparison of frequencies.

With respect to the forward chirps, no similar conclusion can be drawn. Performance was in general too poor to reveal any significant pattern, although one of the four Ss has a peak at +1 and another at 0 and the highest peak of the pooled data is at 0. But if we make the assumption that Ss are comparing slopes in the case of forward as well as backward chirps, a further inference, about the way Ss determine slopes, is possible.

Conceivably, a *S* might estimate the slope directly. Alternatively, he might estimate the frequency at two different moments during the signal t and $t + \Delta t$ (or possibly just the difference between these two frequencies), and compute $(f_t - f_{t+\Delta t}/\Delta t$. Computing the slope in this way does not, of course, involve the kind of frequency estimation required for method (1): it is not necessary to hold t constant for estimates for all three members of a triad. Moreover, in the case of the backward chirps, he could then let $t = 0$ and take advantage of the fact that for this value of t, f_t is a constant; in the case of the forward chirps, similarly, he could let $t + \Delta t = 40$ msec.

Now if the *S* estimates the slope directly, he should do as well with forward as with backward chirps. If he computes the slope, this will not necessarily be the case, since the computational process is not the same. For the backward chirps, the *S* can choose Δt freely (his optimal choice is 40 msec), and knows its value at t. For the forward chirps, on the other hand, either the *S* must compute $\Delta t = 40$ msec $- t$, or he must, before t, choose Δt and compute $t = 40$ msec $- \Delta t$, or he must wait until $t + \Delta t$ to measure Δt. If these constraints make it more difficult for the *S* to evaluate t or Δt, his slope computations would in turn be affected. Thus, there is a second reason why we should expect the backward chirps to be better discriminated. Not only is the final frequency of a chirp apparently easier to estimate than its initial frequency, but, also, the time estimation required to compute the slope is easier when

the initial frequency is known to be constant and the final frequency varied than in the reverse case.

Brady *et al.* pointed out the conflict between their result and the much greater cue value of the second-formant transition in initial than in final position in speech context, and concluded that speech perception cannot be accounted for on the same basis as their experimental result. We face a similar question. Can we account for the discrimination function for the speech stimuli on a strictly psychoacoustic basis? To do so requires us either to point out resemblances to the corresponding nonspeech functions or to propose some convincing explanation for the differences. We recall first that the forward speech functions have characteristic peaks and troughs; these peaks and troughs occur consistently for all Ss and are obvious in the pooled data of Fig. 9. The same peaks and troughs, much less pronounced, appear in the speech function for the backward condition. Nothing corresponding to these peaks and troughs occurs for the nonspeech stimuli, except that the nonspeech functions, like the speech functions, have peaks near or at 0. As we shall see shortly, this is probably a coincidence, and there is no obvious parallel in the nonspeech function for the other peak of the forward speech function or for its troughs. Furthermore, we note that performance is consistently better for nonspeech stimuli in backward condition than in forward condition, while for speech stimuli there is no corresponding consistent improvement (Fig. 10).

As we have seen, the perception of the speech stimuli tends to be categorical: the peaks are found near phonetic boundaries while the troughs correspond to zones inside these boundaries. It has been noted before (Liberman, 1957) that there is an obvious articulatory reference for such perception. When there is articulatory continuity, as in several tokens of [t], each with somewhat different second-formant transitions, such as might, in a human speaker, have resulted from different varieties of apical closure, the listener finds it difficult or impossible to discriminate. When, on the other hand, the difference in the formant transition, though physically no greater, is at a point in the continuum such that it could only have resulted from one sound having been made with labial closure and the other by apical closure, there is a discontinuity in articulation and the listener discriminates quite readily. Because of the particular vowel used in the stimuli, this point of discontinuity happened to fall at stimulus 0. For a vowel with a higher (or lower) second-formant steady state, the boundary would have been lower (or higher) relative to this steady state.

The articulatory basis for the fact that initial transitions result in better phonetic separation than final transitions is less clear, but a study

by Ohman (1966) suggests a possible answer. He found that consonants tend to be coarticulated much more with a following vowel than with a preceding vowel. In production of V_1CV_2 syllables, the character of the transition from V_1 to C depends not merely on V_1 and C but quite considerably on V_2, whereas the transition from C to V_2 is only slightly affected by V_1. Thus, an initial transition (CV) is apt to be a better consonantal cue than a final transition (VC). And, in fact, in natural speech final stops are often followed by a release, consisting of a burst (itself a supplementary cue to point of articulation), and low-amplitude transitions toward [ə]; unreleased stops, on the other hand, are notoriously ambiguous. The stops in the backward speech stimuli used in this experiment were, of course, unreleased.

In previous experiments comparing perception of speech and nonspeech, the nonspeech results were interpreted as representing the discrimination of an acoustic variable before the acquisition by the Ss of this articulatory knowledge. Differences between the discrimination of speech as opposed to nonspeech could then be assigned to "acquired distinctiveness" or "acquired similarity." The results of the [to]–[do] and *rapid–rabid* experiments were taken as evidence of acquired distinctiveness. A more conservative and, we now think, more proper view would have taken the results of those experiments, just as we take the results of our own present experiment, to be evidence for the existence of a speech mode that differs in interesting ways from the auditory mode. Questions about the role of learning in the development of the speech mode stand apart from questions about its existence and are answered by experiments different from those of the kind we have been considering here. Thus, to see the effects of experience we should look to the cross-language studies of Lisker and Abramson (1970; also Abramson & Lisker, 1970) on the perception of the distinction between voiced and voiceless stops. These studies have shown that peaks in discrimination similar to those of our experiment are present or absent depending on the linguistic background of the listener. It does not follow, however, that the peaks are simply a consequence of differential reinforcement or of the mediational processes usually associated with the concepts of acquired distinctiveness and acquired similarity. In that connection we should take note of other results obtained by the same investigators which show that the location of the voiced–voiceless boundaries is very much the same in a number of unrelated languages. When we consider, in addition, that the voicing distinction is universal, or very nearly so, we see that learning does not, in any case, exert its effect in the arbitrary way that Lane (1965), for example, or Quine (1960: 85–90) suppose. The biologically given constraints are important and must surely be

of the greatest interest to anyone who is concerned to understand the development of consonant perception and the peaks that characterize consonant discrimination. This view is strengthened by the findings of recent experiments on infants by Moffitt (1969) and Eimas *et al.* (1970) which show that consonant discrimination is present at a very early age. In the study by Eimas *et al.* it was found that 1-month-old infants discriminate synthetic [ba] and [pa]. Of even greater interest is the fact that, given a fixed physical difference in the relevant acoustic cue, these infants discriminate better across a phonetic boundary than within a phonetic category. Thus, like our adult Ss, they show a discontinuity in discrimination of the voiced–voiceless distinction just as our adult Ss do for the place distinction. It is most likely that the infants' perception of the voicing distinction was, like so many deeply biological processes, not entirely uninfluenced by their experience. If they had been reared in a soundless environment, they would conceivably not have been able to discriminate [ba] from [pa] as they did. Indeed, it is possible that the experience of having heard speech was a necessary condition for the performance that Eimas *et al.* found. But it is hardly conceivable that the effects were produced at the age of 1 month by the simple processes of differential reinforcement or by the more complex mediational mechanisms implied by the concepts of acquired distinctiveness and acquired similarity.

The outcome of our present study also raises other doubts about the applicability of acquired distinctiveness and similarity. In the forward condition, for some distance on either side of the peaks corresponding to the phone boundaries, the speech function is well above the nonspeech function. This, therefore, we would have to attribute to acquired distinctiveness. For portions of the continuum well within phonetic boundaries, the speech function is at or near random and usually well below the nonspeech function. This we would have to attribute to acquired similarity. So far, nothing is seriously amiss, though it would be more parsimonious if it were possible to invoke only one of these processes.

In the case of the backward functions, however, our embarrassment is of a different character. The nonspeech function is higher than the speech function at almost every point. We are, therefore, compelled to invoke acquired similarity to account for the peaks as well as the troughs of the speech function. But why should there be any acquired similarity for stimuli on opposite sides of a phonetic boundary—that is, for stimuli which the listener has learned to call by different names?

Though there are surely ways out of this difficulty that yet preserve concepts like acquired distinctiveness and acquired similarity, it seems

to us preferable to conclude, rather, that we are dealing with two basically different modes of perception. One of these modes is the psychoacoustic. The results of discrimination studies in this mode require an interpretation of the kind we advanced in trying to account for the chirp and bleat data. The other mode is the speech mode. Its characteristics are the consequence of the special processor that decodes the complexly encoded speech signal and recovers the phonetic message. The results of perceptual experiments on the stop consonants do not yield to an interpretation in terms of psychoacoustic perception, with or without such modification as might have been produced by discrimination learning.

In connection with the conclusion that speech and nonspeech are processed differently, we should note that speech and nonspeech functions differ not only in their shape and level but in their reliability. The nonspeech functions vary not only from S to S but also for a single S from one session to the next. Such factors as the relative naivete, the alertness, and the motivation of the S, and the strategy he adopts for the task of discrimination, may make a very substantial difference. In informal tests, in which two of the authors served as Ss, higher levels of chirp discrimination in the forward condition were attained than for any of the Ss for whom data have been presented here. The remarkable thing about the perception of the speech-like stimuli, on the other hand, is precisely its insensitivity to all such factors. Within wide limits, the performance of a S is relatively stable and predictable, provided only he hears the synthetic stimuli as speech. Even Ss who are quite familiar with the stimuli—for example, the authors—do little better than naive Ss away from phonetic boundaries, while naive Ss do little worse than the authors near phonetic boundaries. The speech mode appears to act like some digitizing device which, accepting a signal of quite variable quality and much fine detail, converts it to a perceptual response that is coarsely but reliably quantized.

The backward speech discrimination functions at first appear to contradict what has just been said, since these functions are variable and unstable. In the backward speech test, the Ss were confronted with a confusing task. They were given speech-like stimuli which, as the identification function showed, were difficult to perceive as speech. One might have expected them, in such a situation, to discriminate speech poorly: that is, to produce a discrimination function in which the peaks corresponding to those observed in the forward condition were lower, and the troughs—near random in the forward condition—remained near random. Such an outcome, however, would have suggested that there was, after all, considerable variability in the level of speech discrimina-

tion and that, for some kinds of speech, discrimination is much less reliable than we have just suggested. What actually happens, however, is that while the peaks are indeed lower, the troughs are higher (Fig. 10). The function appears to be a combination of the forward speech function and the backward chirp function. Our interpretation is that the Ss tried to respond to the stimuli as speech. When they found this too difficult, they reverted to the nonspeech mode. But whenever they did respond to the stimuli as speech, they did so, we suspect, as reliably as in the forward condition.

This interpretation of the data bears on an important and difficult question: what conditions must be present to insure perception in the speech mode? The very fact that perceptual experimentation with very simple synthetic speech patterns has been possible shows that a high degree of naturalness is not an important factor, though it seems reasonable to suppose that, at a minimum, some representation of the first two formants may be essential. However, the Ss' response to the backward speech, where formants were present but speech cues were weak and few in number, suggests that a requirement for perception in the speech mode is that the cues for the distinctions among phonetic segments be present in sufficient strength and number to keep the perceptual machinery active. If this requirement is not met, the listener may slip into the nonspeech mode. Thus, the apparently exceptional backward speech results offer an interesting and, to us, unexpected insight into the nature of the special mode of perception which, our experiments suggest, is required for speech.

REFERENCES

Abramson, A. S., & Lisker, L. Discriminability along the voicing continuum: cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences, Prague 1967*. Prague: Academia, 1970, pp. 569–573.

Brady, P. T., House, A. S., & Stevens, K. N. Perception of sounds characterized by a rapidly changing resonant frequency. *Journal of the Acoustical Society of America*, 1961, 33, 1357–1362.

Cross, D. V., & Lane, H. L. *An analysis of the relations between identification and discrimination functions for speech and nonspeech continua*. Report No. 05613-3-P. Ann Arbor: Behavior Analysis Laboratory, University of Michigan, 1964.

Eimas, P., Siqueland, E. R., Jusczyk, P., & Vigorito, J. Speech perception in early infancy. Paper presented to the Eastern Psychological Association, April, 1970.

Kimura, D. Left-right differences in the perception of melodies. *Quarterly Journal of Experimental Psychology*, 1964, 16, 335–358.

Kimura, D. Functional asymmetry of the brain in dichotic listening. *Cortex*, 1967, 3, 163–178.

Kirstein, E. Perception of second-formant transitions in non-speech patterns. *Status Report on Speech Research*, 1966, 7/8, paper 9. New York: Haskins Laboratories.

Kirstein, E., & Shankweiler, D. Selective listening for dichotically presented con-

sonants and vowels. Paper presented to the Eastern Psychological Association, Philadelphia, 1969.

LANE, H. L. The motor theory of speech perception: a critical review. *Psychological Review*, 1965, 72, 275–309.

LIBERMAN, A. M. Some results of research on speech perception. *Journal of the Acoustical Society of America*, 1957, 29, 117–123.

LIBERMAN, A. M. Some characteristics of perception in the speech mode. *Proceedings of the Association for Research in Nervous and Mental Diseases*. Baltimore: Williams & Wilkins, in press.

LIBERMAN, A. M., COOPER, F. S., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431–461.

LIBERMAN, A. M., DELATTRE, P. C., COOPER, F. S., & GERSTMAN, L. J. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, 1954, No. 8, 1–13.

LIBERMAN, A. M., HARRIS, K. S., EIMAS, P., LISKER, L., & BASTIAN, J. An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 1961, 4, 175–195. (b)

LIBERMAN, A. M., HARRIS, K. S., HOFFMAN, H. S., & GRIFFITH, B. C. The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 1957, 53, 358–368.

LIBERMAN, A. M., HARRIS, K. S., KINNEY, J. A., & LANE, H. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, 1961, 61, 379–388. (a)

LISKER, L., & ABRAMSON, A. S. A cross-language study of voicing in initial stops: acoustical measurements. *Word*, 1964, 20, 384–422.

LISKER, L., & ABRAMSON, A. S. The voicing dimension: some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences, Prague 1967*. Prague: Academia, 1970, pp. 563–567.

MATTINGLY, I. G. Experimental methods for speech synthesis by rule. *IEEE Transactions Audio and Electroacoustics*, 1968, 16, 198–202.

MATTINGLY, I. G., & LIBERMAN, A. M. The speech code and the physiology of language. In K. N. Leibovic (Ed.), *Information processing and the nervous system*. Berlin: Springer Verlag, 1969. Pp. 97–117.

MATTINGLY, I. G., LIBERMAN, A. M., SYRDAL, A. K., & HALWES, T. Discrimination of F2 transitions in speech context and in isolation. *Journal of the Acoustical Society of America*, 1969, 45, 314–315.

MOFFITT, A. R. Speech perception by 20–24-week-old infants. Paper presented to the Society for Research in Child Development, Santa Monica, Calif., March, 1969.

OHMAN, S. E. G. Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, 39, 151–168.

POPPER, R. Linguistic determinism and the perception of synthetic voiced stops. Unpublished Ph.D. thesis, U.C.L.A., 1967.

QUINE, W. v. O. *Word and object*. Cambridge, Mass.: MIT Press, 1960.

SHANKWEILER, D., & STUDDERT-KENNEDY, M. Identification of consonants and vowels presented to left and right ears. *Quarterly Journal of Experimental Psychology*, 1967, 19, 59–63.

STRANGE, W., & HALWES, T. Confidence ratings in speech research: experimental evaluation of an efficient technique for discrimination testing. *Perception and Psychophysics*, in press.

STUDDERT-KENNEDY, M., & SHANKWEILER, D. P. Hemispheric specializations for speech perception. *Journal of the Acoustical Society of America*, 1970, 48, 579–594.

STUDDERT-KENNEDY, M., LIBERMAN, A. M., HARRIS, K. S., & COOPER, F. S. Motor theory of speech perception: a reply to Lane's critical review. *Psychological Review*, 1970, 77, 234–249.

SYRDAL, A. K., MATTINGLY, I. G., LIBERMAN, A. M., & HALWES, T. Discrimination of F2 transitions in speech and nonspeech contexts. *Journal of the Acoustical Society of America*, 1970, 48, 94.

(Accepted December 11, 1970)