657

Perception & Psychophysics
1989, 45 (3), 237-250

Discovering phonetic coherence in acoustic patterns

CATHERINE T. BEST
Wesleyan University, Middletown, Connecticut
and Haskins Laboratories, New Haven, Connecticut

MICHAEL STUDDERT-KENNEDY
University of Connecticut, Storrs, Connecticut
Yale University, New Haven, Connecticut
and Haskins Laboratories, New Haven, Connecticut

SHARON MANUEL
Massachusetts Institute of Technology, Cambridge, Massachusetts

and

JUDITH RUBIN-SPITZ
NYNEX Science and Technology, White Plains, New York

Despite spectral and temporal discontinuities in the speech signal, listeners normally report coherent phonetic patterns corresponding to the phonemes of a language that they know. What is the basis for the internal coherence of phonetic segments? According to one account, listeners achieve coherence by extracting and integrating discrete cues; according to another, coherence arises automatically from general principles of auditory form perception; according to a third, listeners perceive speech patterns as coherent because they are the acoustic consequences of coordinated articulatory gestures in a familiar language. We tested these accounts in three experiments by training listeners to hear a continuum of three-tone, modulated sine wave patterns, modeled after a minimal pair contrast between three-formant synthetic speech syllables, either as distorted speech signals carrying a phonetic contrast (speech listeners) or as distorted musical chords carrying a nonspeech auditory contrast (music listeners). The music listeners could neither integrate the sine wave patterns nor perceive their auditory coherence to arrive at consistent, categorical percepts, whereas the speech listeners judged the patterns as speech almost as reliably as the synthetic syllables on which they were modeled. The outcome is consistent with the hypothesis that listeners perceive the phonetic coherence of a speech signal by recognizing acoustic patterns that reflect the coordinated articulatory gestures from which they arose.

To master their native language, children must learn not only to listen, but to speak. In the speech signal, they must discover information that not only distinguishes among the words they hear, but also specifies how the words are to be spoken. This dual function of the speech signal has been largely disregarded in research on speech perception. Researchers have generally accepted the linguist's description of speech as a sequence of syllables or phonemes, compounded from "bundles of features," and have then looked in the signal for the "information-bearing elements" or "cues" that correspond to the linguist's abstract descriptors, without considering whether or how these cues might specify the articulatory gestures that give rise to them. The strategy has been successful to the extent that we now have detailed lists of cues—pitch contours, formant patterns, silent gaps, patches of band-limited noise, and so on—that may be mimicked by terminal analog synthesis to render intelligible speech.

Such synthesis typically proceeds, however, without appeal to general principles of either auditory or articulatory organization, or of their interrelationship. Even if an experimenter follows certain "rules for synthesis," the rules are rarely more than a summary of previous experimenters' prescriptions for copying spectrograms within the constraints of a particular synthesizing device (but see Mattingly, 1981). The criterion for a successful copy is simply a listener's judgment as to whether or not the synthesized pattern renders an acceptable phonetic form (i.e., an acceptable acoustic-articulatory pattern of sound) in the language under study.

What is the basis for listeners' phonetic percepts? What do they listen for in the signal? From the facts of speech synthesis, we might suppose that they listen for discrete

Preparation of this paper was supported in part by National Institutes of Health Grant HD-01994 to Haskins Laboratories. We thank Len Katz for statistical advice, and Michael Dorman, Peter Jusczyk, and Bruno Repp for useful comments on the texts of earlier versions. Address correspondence to Catherine T. Best, Haskins Laboratories, 270 Crown St., New Haven, CT 06511 or Department of Psychology, Wesleyan University, Middletown, CT 06457.

Copyright 1989 Psychonomic Society, Inc.

acoustic cues. However, the notion of *cue extraction* poses a logical puzzle of definition, as noted by Bailey and Summerfield (1980). To establish that a particular piece of acoustic "stuff" deserves the status of a cue, researchers commonly use speech synthesis to set all other portions of the array at values that will ensure perceptual ambiguity. They then manipulate the potential cue, so that particular settings resolve the ambiguity. However, when the ambiguity is resolved—that is, when listeners consistently identify the pattern as an instance of a particular phonetic category—which is the cue? Is it the element that was manipulated or is it the context? The context without the cue is ambiguous, and the cue without its context is typically heard as nonspeech (e.g., Mattingly, Liberman, Syrdal, & Halwes, 1971). We have no grounds for preferring one to the other as an effective or necessary component of the pattern.

If neither cue nor context (itself composed of an indefinite number of other cues) can independently and unambiguously specify the speech sounds we hear, the functional unit of speech perception must be the entire acoustic pattern that the acoustic cues compose. What is this pattern; and why do the diverse "cues" that compose it cohere perceptually?[1]

According to one account, listeners extract discrete cues, but judge them only in relation to each other, so that the phonetic segment is a result of their perceptual integration (e.g., Cutting, 1976; Jusczyk, Smith, & Murphy, 1981; Pastore, 1981; Schouten, 1980). The reason why isolated cues are often heard as nonspeech is that perceptual categorization depends on the relations among cues, and these relations are destroyed when a cue is removed from context. A variant of this view treats the supposed cues as independent "features" to which listeners assign weights on the basis of their representation in the signal. Listeners then sum or multiply the weights and compare the integrated outcome with a stored "prototype" to arrive at a probabilistic estimate of the percept (Oden & Massaro, 1978). We refer to the mechanism proposed by these accounts as *cue integration*.

According to another account, cues have no functional role in determining the sound pattern of speech. Rather, the pattern coheres according to Gestalt principles analogous to those in visual form perception, such as proximity, similarity, good continuation, and closure (Bregman, 1981). Thus, the melodic coherence of vowel sequences, essential to prosody, may be maintained across consonantal constrictions by the smooth contour of their fundamental frequencies (Bregman, 1981). The harmonics of a vowel formant may cohere by virtue of temporal proximity, that is, of their simultaneous onsets and offsets (Darwin, 1984). Temporal proximity may also account for coherence of the spectrally diverse cues to voicing in consonant-vowel (CV) syllables, discussed below. Good continuation and spectral similarity may be at work in a CV syllable when a stop consonant release burst effectively conveys information about place of articulation only if it is spectrally continuous with the following formant transition (Dorman, Studdert-Kennedy, & Raphael, 1977). Finally, formant transitions may perform not only segmental functions, but also syllabic functions, by eliminating from the signal abrupt discontinuities that might excite an unwanted increase in neural firing (Delgutte, 1982). The transitions would thus assure syllabic coherence and, incidentally, correct perception of the temporal order of syllabic components in rapid speech (Cole & Scott, 1974; Dorman, Cutting, & Raphael, 1975). If this account is correct, phonetic forms emerge from the signal by virtue of their *auditory coherence*. Notice that this account, unlike that based on cues, has nothing to say about the units of linguistic information that the speech signal conveys. Principles of auditory coherence are presumed to apply not only to phonetic segments, but to every other unit of linguistic analysis, from the feature to the prosodic contour.

A final account invokes a principle that we call *phonetic coherence*. The basis for perceptual coherence, according to this account, is said to be the coordinated pattern of articulatory gestures that produced the signal. The principle is implicit in the well-known explanation offered over 20 years ago by Abramson and Lisker (1965) for the spectral and temporal diversity of covarying cues to voicing distinctions in many languages: release burst intensity, degree of aspiration, and first formant (F1) onset frequency. They proposed that all these cues arise from the relative timing of laryngeal and supralaryngeal gestures in stop-vowel syllables:

> Laryngeal vibration provides the periodic or quasi-periodic carrier that we call voicing. Voicing yields harmonic excitation of a low frequency band during closure, and of the full formant pattern after release of the stop. Should the onset of voicing be delayed until some time after the release, however, there will be an interval between release and voicing onset when the relatively unimpeded air rushing through the glottis will provide the turbulent excitation of a voiceless carrier, commonly called aspiration. This aspiration is accompanied by considerable attenuation of the first formant, an effect presumably to be ascribed to the presence of the tracheal tube below the open glottis. Finally, the intensity of the burst, that is, the transient shock excitation of the oral cavity upon release of the stop, may vary depending on the pressures developed behind the stop closure where such pressures will in turn be affected by the phasing of laryngeal closure. Thus it seems reasonable to us to suppose that all these acoustic features [cues], despite their physical dissimilarities, can be ascribed ultimately to actions of the laryngeal mechanisms. (Abramson & Lisker, 1965, pp. 1-2)

Variations in voice onset time underlie voicing distinctions in many, if not all, languages, and this elegant account of the articulatory origin of the diverse cues to voicing has been widely accepted. What is important here however, is the general principle that the model proposes the speech signal coheres not because of (perhaps even in spite of) its auditory properties, but because coordinated patterns of gesture (i.e., of phonetically functional articulatory actions, such as lip closure, velum lowering

tongue raising, etc.) give rise to coordinated patterns of spectral and temporal change. By adopting the articulatory gesture, and its acoustic correlates, as its linguistic primitive, this account proposes an objectively observable unit common to both production and perception. Gestures thus form both the patterns of information that listeners listen for in synthetic speech and the patterns that children must discover in natural speech if they are to learn how to talk. Thus, the phonetic coherence account is the only account discussed here to offer a direct, concrete basis for the perception-production link and, hence, for imitation. This gestural account is compatible with any level of abstract linguistic unit, from the phoneme (Fowler & Smith, 1986; Studdert-Kennedy, 1987) to the word (Browman & Goldstein, 1986).

To summarize, each of these accounts offers a view, implicit or explicit, of (1) the information (i.e., the linguistic structure) that a speaker encodes in the signal, and (2) the mechanism by which a listener recovers that structure. Both the simple cue extraction and the cue integration accounts propose that the information is a collection, or sequence, of abstract linguistic elements—features, phonemes, or perhaps syllables—and that the recovery mechanism entails the simple extraction, or the extraction and integration, of discrete cues to those elements. The auditory coherence account, the least linguistically oriented, is neutral on the nature of the linguistic information, but proposes that listeners perceive the sound patterns of speech (whatever they may be) according to general principles of auditory form perception. Finally, the phonetic coherence account proposes that the linguistic structure of the signal is articulatory, a pattern of gestures, and that listeners recover this structure because it is implicit in the acoustic signal to which it gives rise. Whether the articulatory gestures are grouped and segregated so as to specify abstract units at an intermediate phonological level (phonemes, syllables) or only at the level of lexical items (morphemes, words) is a separate issue, not considered here.

The following three experiments were designed to test these accounts of speech perception. First, they bring further experimental evidence to bear on the arguments presented above concerning the role of cues in speech perception: They ask whether listeners can better learn to identify, and discriminate between, contrasting acoustic patterns by focusing attention on a discrete acoustic cue, or by focusing on the entire acoustic pattern of which the cue is a part. Second, if attention to the entire pattern yields superior performance, the experiments are so designed that we can ask further whether the contrasting patterns emerge according to principles of cue integration or auditory form perception, or from listeners' directing attention to their potential phonetic coherence.

We compared the perceptual effects of an attentional focus on a phonetic contrast, /r/ versus /l/, with the effects of attention to a discrete acoustic cue signaling that contrast, both in and out of context. Our stimulus materials were a continuum of sine wave speech syllables. Sine wave speech can be heard either as distorted, but recognizable, speech, or as sounds unrelated to speech (e.g., distorted musical chords or bird-like chirps) (Best, Morrongiello, & Robson, 1981; Remez, Rubin, Pisoni, & Carrell, 1980). This dissociation allowed us to compare perceptual responses to the same signal under nonspeech (auditory) and speech (phonetic) modes of attention.

EXPERIMENT 1

Experiment 1 investigated which of the views outlined above best accounts for the discrete perceptual categories in a minimal pair speech contrast: simple cue extraction, cue integration, auditory coherence, or phonetic coherence.

To test these possibilities, we developed a sine wave syllable continuum based on the time-varying formant frequencies characteristic of American English /ra/ versus /la/. The members of this continuum differed only in the direction and rate of the third formant (F3) transition, which varied systematically in approximately equal steps. Two additional series were developed for control comparisons: synthetic full-formant versions of the syllable continuum and frequency-modulated single tones corresponding to the F3 elements of the sine wave syllable series. Listeners participated in one of two conditions: speech bias or music bias. Pretest instructions and perceptual training tasks were designed to focus the speech listeners' attention on hearing the sine wave syllables as distorted versions of /ra/ and /la/, and the music listeners' attention on hearing them as carrying a binary contrast (steady vs. rising) on the transition of the F3 tone, the highest tone in a distorted three-tone musical chord.

If the speech categories depend on extraction of a single cue (F3 transition), both groups of listeners should perform identically in categorizing both the sine wave continuum and the full-formant continuum: the 50% crossover points on their identification functions (i.e., their category boundaries) and the slopes of these functions should be the same. Moreover, both groups should categorize the isolated single tones (F3) exactly as they categorize the sine wave and full-formant syllables.

If the speech categories depend on the extraction and integration of acoustic cues, music listeners, trained to attend to the sine wave F3 transition, should be able to extract and integrate that transition no less consistently than speech listeners. They may assign a different weight to the transition in the overall structure, and thus discover a somewhat different perceptual organization with a different category boundary from that of the speech listeners, but the consistency (i.e., the slope) of the sine wave categorization functions should be roughly the same for the two groups.

Similarly, if the speech categories reflect the operation of general Gestalt principles of auditory form perception, music listeners should be able to discover at least some consistent pattern (either the same as or different from that of the speech listeners) in the sine wave stimuli. Thus,

according to the auditory coherence account, we would again expect the slopes, if not the category boundaries, to be essentially the same for both groups.

Finally, if the speech categories depend on phonetic coherence, the two groups should differ in their judgments of the sine wave syllables. The speech listeners should recognize the phonetic coherence of the sine wave syllables and should categorize them as they categorize the full-formant versions, with perhaps somewhat shallower slopes (i.e., greater response variability) for the sine waves, due to the phonetically impoverished and unfamiliar patterns of the sine wave "dialect." By contrast, the music listeners (unable to suppress the perceptual influence of the lower tones, or perhaps to reject the automatically given auditory coherence of the sine wave patterns) should have difficulty in separating the F3 tone perceptually from its chordlike frame and, therefore, in categorizing the sine wave syllables on the basis of the F3 binary contrast. Their difficulties should be evident in flat sine wave syllable slopes, or at least in significantly shallower slopes for these patterns than the speech listeners show. Thus, the main test of the competing hypotheses lies in the effects of the instruction condition on the slopes of the sine wave syllable functions.

Method

Subjects. Twenty-four subjects were tested in a between-groups design, with 12 subjects each in the speech bias group (3 males, 9 females) and the music bias group (7 males, 5 females) (Experiment 1a). In addition, 5 of the music subjects returned for a second session under speech bias conditions, permitting a partial within-group test (Experiment 1b). It was not possible to conduct a full within-group study, with condition orders counterbalanced, since our own and other researchers' experience with sine wave speech indicates that once subjects have perceived the stimuli as speech, it is extremely uncommon for them to be able to revert to hearing them again as nonspeech.

All subjects were young adults with normal hearing and with negative personal and family histories of language and speech disorders. Each was paid $4 per test session. Based on their answers to post-test questionnaires (see Procedure), one female subject was eliminated from the speech group for failure to hear any of the sine wave syllables as speechlike, and one female was eliminated from the music group because she heard the sine wave syllables as "sounding like r," thus reducing the number in each group to 11. Their elimination was necessary because evaluation of the hypotheses depended on consistent group differences in hearing the sine wave syllables as either speech or nonspeech.

Stimuli. The full-formant /ra/–/la/ series was developed first, using the OVE-IIIc serial resonance synthesizer. The continuum contained 10 items, which differed from each other only in the onset frequency, and in the duration of the initial steady-state portion, of the F3 transition (see Figure 1). These F3 properties were varied in nearly equal steps (slightly constrained by the step-size limitations of the synthesizer). Each stimulus was 330 msec in duration. The series was designed to be biased toward perception of more /la/ than /ra/ tokens. This was done so that the phonetic category boundary should fall neither at the perceived shift from steady to rising F3 transitions nor at the continuum midpoint, since these physical properties were two likely foci for a psychoacoustically based category distinction.

The /l/ biasing was accomplished by using a rapid F1 transition, a long-duration steady state at the onset of F1, and a steady-state F2, in the portion of the stimuli that was kept constant throughout the series, and by including one stimulus (the first on the continuum) with a slightly falling F3 transition. These formant characteristics are associated with natural tokens of /l/; natural /r/ typically has short steady states at the onsets of the formant transitions, a relatively slow F1 transition, a slightly rising F2 transition, and a clearly rising F3 transition (MacKain, Best, & Strange, 1981). In all stimuli F0 began at 119 Hz and fell steadily to 100 Hz by the end of the stimulus. In the constant portion of the stimuli, F1 onset began at 349 Hz, remained there for 75 msec, then rose linearly to 673 Hz by 100 msec, where it remained to the end of the syllable. The steady-state frequency of F2 was 1297 Hz. The constant portion of F3 was a steady-state 2870 Hz in the final part of the stimulus, beginning at 125 msec into the syllable. The F3 onset frequencies for the 10 stimulus items were 3019 (at the /la/ end of the continuum), 2870 (a flat F3), 2729, 2576, 2431, 2278, 2119, 1972, 1821, and 1670 Hz (at the /ra/ end of the continuum). In the 1st stimulus (/la/), the steady-state onset ended and the F3 transition began at 65 msec into the stimulus. In the 10th stimulus (/ra/), this breakpoint occurred at 20 msec into the stimulus. The temporal position of the breakpoint was varied systematically in 5-msec steps for the intervening stimuli.

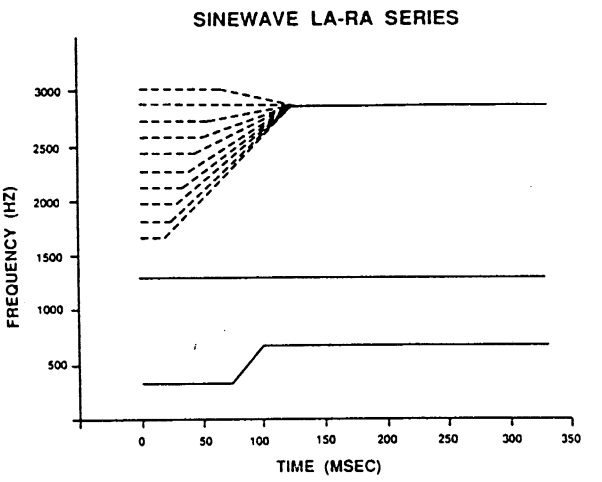

Figure 1. Schematic diagram of the center frequencies of the three formants in each of the sine wave syllables and each of the full-formant syllables, in the 10-item stimulus continua.

The sine wave syllable continuum was generated with a multiple sine wave synthesizer program developed for the DEC PDP-11/45 computer at Haskins Laboratories. The frequency characteristics of the sine wave syllables mimicked those from the full-formant continuum, except that each formant was now represented as a single, time-varying tone rather than as the wider band of harmonics found in the formants of natural and synthetic speech. In the sine wave syllables, there was no tone to represent the original F0 contour (see Figure 1). The isolated F3 tone continuum was made up of the F3 tones from the sine wave syllable continuum, presented without the tones corresponding to F1 and F2.

Four additional stimulus series were developed for the perceptual training sequences that were presented in each condition before the categorization test for the sine wave syllables. These series were designed to focus the listener's attention either on the phonetic properties of the sine wave syllables, or on their properties as three-note chords differing only in the onset characteristics of the highest note. There were two speech bias training series, one based on the endpoint /ra/ stimulus of the sine wave syllable continuum, and the other based on the sine wave /la/ syllable with

the flat F3 (the second stimulus in the continuum). Each speech training series contained 11 stimuli that provided a gradual, stepwise change from the full-formant syllable to the corresponding sine wave syllable. The first item of each series was the pure full-formant version of the syllable; the last item was the pure sine wave version of the syllable. The nine intervening stimuli were produced by mixing the exactly synchronized, matching sine wave and full-formant syllables in inversely varied proportions (i.e., the relative amplitude of the full-formant stimulus was reduced in equal steps, while the amplitude of the sine wave syllable was correspondingly increased). The two music bias training series also consisted of 11 items each, but the transformations progressed from the isolated endpoint F3 tone at the /ra/ end of the continuum to the corresponding endpoint sine wave syllable, and from the isolated flat F3 tone at the /la/ end of the continuum to the corresponding sine wave syllable.

Procedure. The subjects were tested in groups of 2 to 5 in a sound-attenuated experimental room. The stimuli were presented to them at a comfortable listening level (75 dB SPL) over TDH-39 headsets.

All subjects first completed the experimental task, consisting of the appropriate perceptual training sequence for the condition randomly assigned to their group (speech bias or music bias), followed by the categorization test with the sine wave syllables. This task was administered first so that the subjects' performance with the sine wave syllables could not be influenced by exposure to the full-formant and isolated F3 sine wave continua. The speech subjects were instructed that they would be tested on their ability to categorize computer-distorted versions of the syllables /la/ and /ra/, whereas the music subjects were instructed that they would be asked to categorize computer-distorted chords according to whether or not there was a rising frequency glide at the onset of the highest tone in the chords.

The subjects were then told that they would first receive some perceptual training to aid in focusing their attention on the identities of the distorted syllables (speech group) or on the steady-state versus upgliding properties of the highest notes in the distorted chords (music group). Each training sequence proceeded in five steps. The speech subjects first heard the pair of full-formant clear-case stimuli (/la/ and /ra/) repeated five times, whereas the music subjects heard five repetitions of the clear-case F3 tones (flat and rising onsets), with 1-sec interstimulus intervals (ISIs) and 3-sec intertrial intervals (ITIs). Next, the subjects completed a 10-item practice test to categorize a randomized sequence of these clear-case syllables or F3 tones, presented individually with 3-sec ISIs, by entering their choices on an answer sheet. Third, they listened to, but did not explicitly categorize, the gradual 11-step transformation, beginning with the pair of full-formant syllables or the pair of F3 tones, and ending with the clear-case pair of computer-distorted sine wave stimuli. This transformation was then played in reverse order. The forward and reverse transformation was played three times, with ISIs of 3 sec, and interblock intervals (IBIs) of 6 sec. Fourth, subjects heard the pair of clear-case sine wave stimuli presented five times with 1-sec ISIs and 3-sec ITIs. Finally, they were given a randomized 20-item practice sequence of the clear-case sine wave stimuli presented one at a time with 3-sec ISIs, and they wrote their choices on answer sheets. For all practice trials, the correct answers were printed on the answer sheets, but were covered by a strip of paper; the subjects uncovered each correct answer only after writing down their own response.

Subjects in both conditions then took a categorization test with the complete series of sine wave syllables. The test contained 20 blocks of the 10 items in the sine wave syllable continuum, randomized within each block. The stimuli were presented individually, with 3-sec ISIs and 6-sec IBIs. Subjects in the speech group circled "la" or "ra" on their answer sheets to indicate the category identity of each item in the test. Music subjects circled "steady" or "upglide" to indicate whether the highest tone in the distorted chord had a flat frequency trajectory or a rising glissando at the onset.

After they had finished the sine wave syllable test, subjects in both groups completed categorization tests with the two control series, the full-formant syllables and the isolated F3 tones. Each control test contained 20 randomized blocks of the 10 items in a given series, with 3-sec ISIs and 6-sec IBIs as before. On the full-formant syllable test, all subjects circled "la" or "ra" to indicate category assignments, and on the F3 tone test, they circled "steady" or "upglide."

At the end of the test session, each subject answered a questionnaire about what the sine wave syllables had sounded like to them, whether they had been able to maintain the perceptual focus intended by their group's instructions, and whether they had made any judgments on the basis of the opposing group's perceptual set, that is, whether the speech listeners had categorized the stimuli on the basis of musical (or nonspeech) properties, and whether the music listeners had heard any as syllables.

Results

Experiment 1a: Between-groups comparison. The categorization data were tabulated, for each subject on each continuum, as the percentage of times that each item was categorized as "la" in the full-formant syllable test and in the speech condition of the sine wave syllable test, or categorized as "steady" in the F3 tone test and in the music condition of the sine wave syllable test.

Figure 2 displays the averaged results for the two groups. The category boundaries (50% crossovers) fall at somewhat different points on the three continua (highest for the sine wave syllables, lowest for the F3 tones), but are essentially the same for the two groups. The slopes of the functions clearly differ across continua (steepest for the full-formant syllables, shallowest for the sine wave syllables), and are roughly the same for the two groups on the F3 tones and full-formant syllables. The groups differ markedly on the slopes of their sine wave syllable functions, with the slope for the music listeners being the shallower.

To test the significance of these effects, the data of each subject were first submitted to a probit analysis (Finney, 1971) to determine the mean (category boundary) and slope (reciprocal of the standard deviation) of the best fitting ogive curve by the method of least squares.[2] Table 1 lists the mean category boundaries and slopes, together with their standard errors, for each group on each continuum. High slope values indicate steep slopes; low values indicate shallow slopes. Note that the computed mean category boundaries and slopes, based on individual probit analyses, are not expected to correspond exactly with those read from the group functions of Figure 2 (e.g., the relatively steep mean slope value of 2.30, listed in Table 1 for the speech listeners on the full formant syllables, was due to very steep slopes given by 3 out of the 11 listeners, but is not apparent in the group function of Figure 2). On the sine wave syllables, the music listeners were less consistent than the speech listeners in their category boundaries (higher standard error), but more consistent in their (low) slope values (lower standard error).

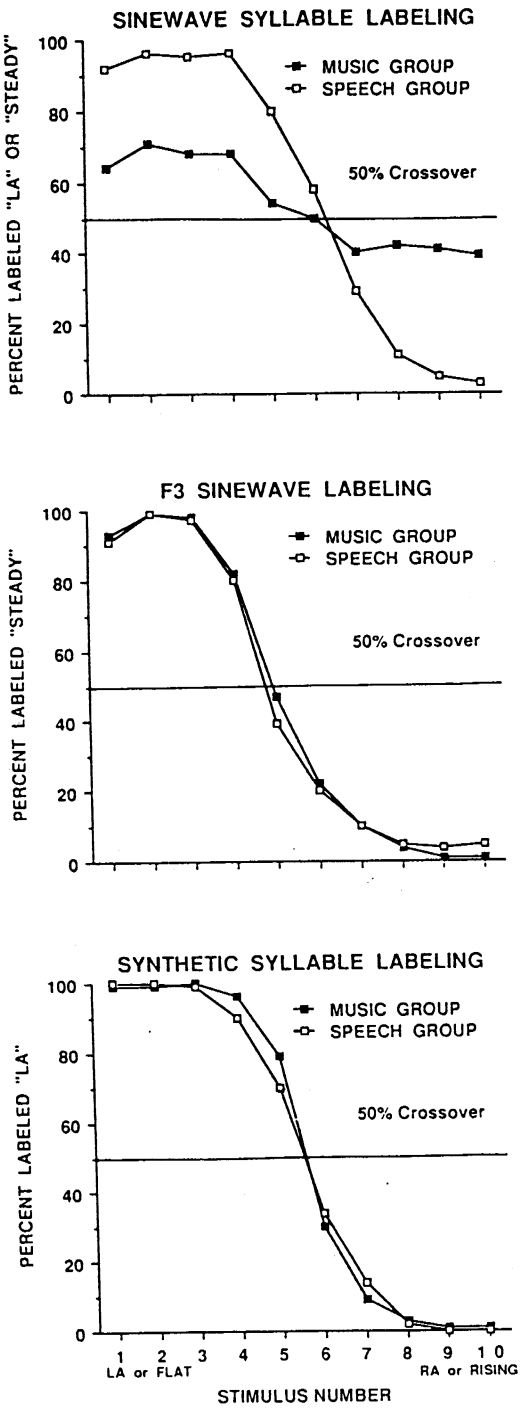

Figure 2. Labeling functions for the speech and music listeners on each of the three stimulus continua in Experiment 1a.

Two-factor (instruction group × stimulus continuum) analyses of variance were then carried out on the category boundaries and on the slopes. In these analyses, instruction group was a between-groups variable and stimulus continuum a within-groups variable. The category boundary analysis yielded no significant effects and no significant interaction. The slope analysis yielded a significant effect of stimulus continuum [$F(2,40) = 15.21$, $p < .001$], an effect just short of significance for instruction condition [$F(1,20) = 3.59$, $p < .07$], and no significant interaction [$F(2,40) = 0.67$, $p > .10$]. Scheffé tests of the stimulus continuum effect showed that slopes were significantly steeper for the full-formant syllables than for the F3 tones [$F(1,20) = 8.90$, $p < .05$], and steeper for the F3 tones than for the sine wave syllables [$F(1,20) = 14.19$, $p < .02$].

The lack of an interaction in the slope analysis was unexpected, given the functions illustrated in Figure 2. Presumably, the fact that the music group had a shallower mean slope than the speech group on all three continua gave rise to the marginal effect of instruction condition, but the interaction failed to reach significance due to the large component contributed to the error variance by performance on the full-formant syllables (see Table 1). Nonetheless, since a test of the slope difference between instruction groups on the sine wave syllables was a key to distinguishing among the competing hypotheses, we carried out a planned comparison by means of a simple t test on these data (for which the error variance was relatively low: see Table 1). The result was highly significant [$t(20) = 4.18$, $p < .0005$].

Finally, as a test for category formation on the sine wave syllable continuum, we carried out for each group a one-factor analysis of variance with repeated measures on stimulus items. For the speech listeners there was a significant effect of stimulus item [$F(8,80) = 110.18$, $p < .0001$]; Scheffé tests between all possible pairs yielded significant differences between all items in the group of Stimuli 2 through 5 and all items in the group of Stimuli 7 through 10 ($p < .05$), but none between items within these groups, indicating the presence of two distinct categories. For the music group there was a significant effect of stimulus item [$F(8,80) = 5.88$, $p < .0001$]; however, Scheffé tests between all possible item pairs yielded significant differences only between Item 2 and Items 7 and 10, indicating no consistent categorization.

Experiment 1b: Within-groups comparison. The data for the 5 subjects who served in both instruction conditions were treated in the same way as the data of Experi-

Table 1
Experiment 1a: Mean Category Boundaries and Slopes, Determined from Individual Probit Analyses on Stimulus Items 2 Through 10, for Speech and Music Listeners

	Category Boundaries			Slopes		
	Sine Wave	F3	Full-Formant	Sine Wave	F3	Full-Formant
	Speech Listeners (n = 11)					
M	6.24	5.13	5.56	0.80	1.03	2.30
SE	0.17	0.26	0.27	0.15	0.18	0.54
	Music Listeners (n = 11)					
M	6.69	5.18	5.67	0.14	0.96	1.70
SE	1.16	0.24	0.20	0.05	0.10	0.35

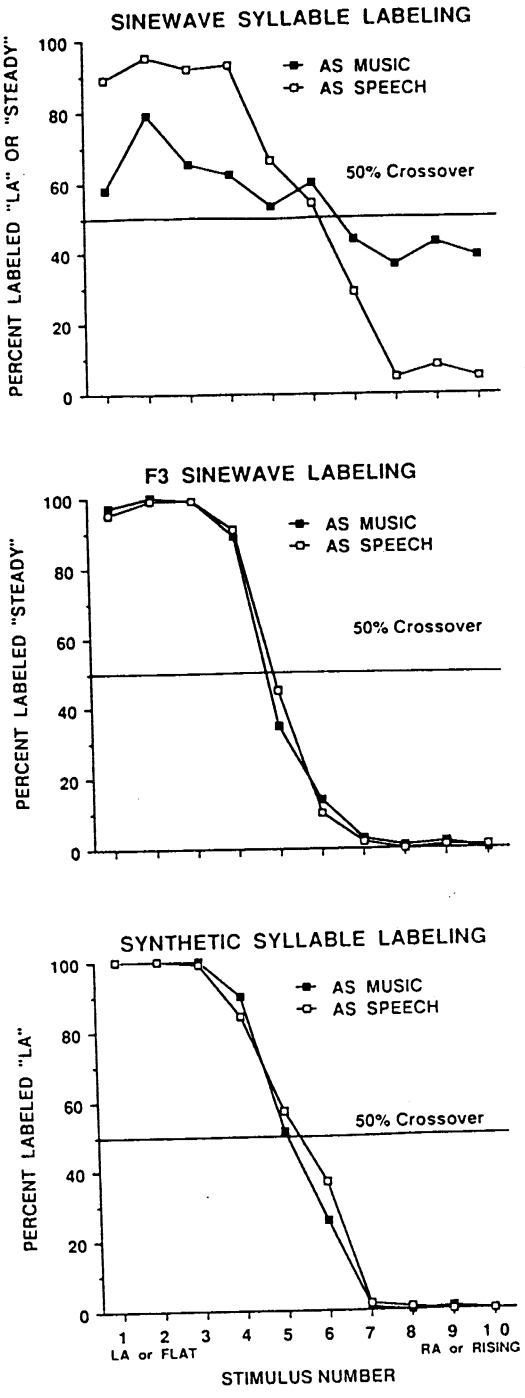

Figure 3. Labeling functions for the speech and music listeners on each of the three stimulus continua in Experiment 1b.

ment 1a.[3] Figure 3 displays the group functions, and Table 2 lists the mean category boundaries and slopes, with their standard errors. The general pattern of results is similar to that of the between-groups comparison. On the sine wave syllables, the music condition yields a higher standard error than the speech condition for the category boundaries, and a lower standard error for the slopes.

A two-factor analysis of variance with repeated measures on both factors (instruction condition and stimulus continuum) was carried out on the category boundaries and on the slopes. The category boundary analysis yielded a significant effect of stimulus continuum [$F(2,8) = 9.70$, $p < .01$], but no effect of instruction condition and no significant interaction. None of the three possible pair-by-pair category boundary comparisons between stimulus continua was significant on post hoc Scheffé tests. The slope analysis yielded a significant effect of stimulus condition [$F(2,8) = 6.08, p < .02$], but no effect of instruction condition and no significant interaction. None of three pair-by-pair slope comparisons between stimulus continua was significant according to Scheffé tests. Again, despite the lack of an interaction, we carried out the planned t test (for matched pairs) to compare the slopes of the sine wave syllable functions between the two instruction conditions. The result was highly significant [$t(4) = 6.02, p < .004$].

Tests for category formation on the sine wave syllable continuum yielded a significant effect of stimulus item for the speech condition [$F(8,32) = 24.27, p < .0001$], with Stimuli 2 through 5 and 7 through 10 again falling into distinct categories, according to Scheffé tests ($p < .05$ for all between-category comparisons, but not significant for any within-category comparisons). For the music condition there was a significant effect of stimulus items, but none of the Scheffé tests between item pairs was significant, indicating the absence of clear-cut categories.

Discussion

The significant effect of stimulus continuum in the within-group comparison of category boundaries disconfirmed the prediction of the simple cue extraction hypothesis that boundaries would be identical on the three continua. There were no grounds in either experiment for rejecting the null hypothesis of equal category boundaries on the sine wave and full-formant continua.

The most decisive results came from the slope analyses. The two groups performed identically on the two control continua (F3 tones and full-formant syllables), but differed

Table 2
Experiment 1b: Mean Category Boundaries and Slopes, Determined from Individual Probit Analyses on Stimulus Items 2 Through 10, for Listeners Who Served in Both Speech and Music Instruction Conditions ($n=5$)

	Category Boundaries			Slopes		
	Sine Wave	F3	Full-Formant	Sine Wave	F3	Full-Formant
	Speech Condition					
M	5.97	5.00	5.31	0.71	1.20	1.54
SE	0.43	0.11	0.44	0.14	0.17	0.20
	Music Condition					
M	7.15	4.75	5.48	0.14	1.34	1.79
SE	0.68	0.19	0.21	0.07	0.20	0.64

on the sine wave syllable continuum. The significantly shallower slopes for the music than for the speech listeners, on both between-groups and within-group comparisons, suggest that attention to phonetic properties of the syllables facilitated categorization, whereas attention to purely auditory properties hindered it. This outcome is predicted only by the phonetic coherence hypothesis.

Nonetheless, as the stimulus item analysis indicated, the identification function of the music listeners on the sine wave syllables was not flat. Although they gave no evidence of reliable category formation, the music listeners' functions sloped in the same direction as those of the speech listeners in both experiments. We were therefore concerned that the music listeners might have been influenced by factors other than attentional mode. Specifically, the music group might have been disadvantaged by lack of practice with the crucial acoustic features of the category exemplars (i.e., steady vs. rising F3 tones) before the categorization test on the sine wave syllables. In addition, the words "steady" and "upglide" may have been more arbitrary as labels for the sine wave syllable endpoints in the music condition than were the "la" and "ra" labels in the speech condition.

EXPERIMENT 2

To meet the foregoing objections, we performed a second experiment in which we had each group first complete the control test that would constitute practice for the sine wave syllable categorizations (the F3 categorization test for the music listeners; the full-formant syllable test for the speech listeners) before completing the sine wave syllable test itself. We also provided the music listeners with nonverbal symbols of the endpoint F3 trajectories (—— vs. /‾‾) to use as category identifiers in the sine wave syllable test, rather than the perhaps arbitrary verbal labels used in Experiment 1.

Method

Subjects. Thirty young adults were tested. Of these, 13 were tested in the music bias condition (6 males, 7 females) and 17 were tested in the speech bias condition (4 males, 13 females). On the basis of their answers on the posttest questionnaires (see Experiment 1), 6 subjects were eliminated from the analyses, leaving a total of 12 subjects in each group. One female was withdrawn from the music group because she began to hear words or names in the sine wave syllables, and 5 subjects (1 male, 4 females) were withdrawn from the speech group for failing to hear the sine wave speech patterns as syllables. It may be of interest that of the latter participants, one was not a native speaker of English and another was dyslexic. All remaining subjects lacked any personal or familial history of language and speech problems, were monolingual English speakers, and had normal hearing. Each received $4 participation in the test session.

Stimuli. The stimuli designed for Experiment 1 were used again in this experiment.

Procedure. The procedures were identical to those described in Experiment 1, except in the following respects: The music bias group used nonverbal labels (—— vs. /‾‾) rather than the words "steady" and "upglide" to identify the items in the sine wave syllable and isolated F3 tone continua, and they completed the F3 categorization test before the training sequence and test with the sine wave syllable continuum. They did not take a categorization test with the full-formant syllables. The speech bias group, on the other hand, completed the categorization test with full-formant syllables before the training and test with the sine wave syllables. They again used "la" and "ra" as their category labels. They did not take a test with the isolated F3 tones.

Results

The data were treated as in the previous experiment. Figure 4 displays the group functions, and Table 3 lists the mean category boundaries and slopes, with their standard errors. For the speech listeners the category boundary is somewhat higher, and the slope shallower, on the sine wave syllables than on the full-formant syllables and the F3 tones. For music listeners the projected category boundary falls outside the continuum for the sine wave syllables, because most of the stimuli were labeled as "——" (steady) across the whole series; their F3 category boundary falls well below the continuum midpoint. The sine wave syllable slope for the music listeners is close to zero, very much shallower than on the F3 tone series. The two groups differ strikingly in both category boundary and slope on the sine wave syllable continuum. Once again, the pattern of standard errors indicates that the music listeners were highly variable in their sine wave syllable boundaries, but highly consistent in their low slope values.

To test the significance of the category boundary and slope variations, we carried out t tests for correlated samples on the sine wave syllable and control continua for each group, and t tests for independent samples, comparing the groups on the sine wave syllable continuum. For the speech listeners, category boundaries on sine wave and full formant syllables did not differ, but slopes differed significantly, being shallower for the sine wave syllables [$t(11) = 3.45, p < .05$]. For the music listeners, despite the large difference in means, category boundaries did not differ significantly on sine wave syllables and F3 tones, presumably due to the very high variability of their data. However, their slopes on the two continua differed significantly, being shallower for the sine wave syllables [$t(11) = 3.84, p < .0034$]. The two groups did not differ in their category boundaries on the sine wave syllable continuum (presumably again due to the high variability of the music listeners), but they did differ significantly in their slopes [$t(22) = 5.62, p < .0001$].

Tests for category formation on the sine wave syllables, analogous to those of Experiment 1, showed that Stimuli 2 through 5 and 6 through 10 fell into distinct categories for the speech listeners ($p < .05$ for all between-category comparisons; not significant for any within-category comparisons). For the music listeners, there was no effect of stimulus item, indicating that the slope of their mean function did not differ significantly from zero.

Discussion

The results replicate and strengthen those of Experiment 1. The shallower slopes for the speech listeners on sine wave than on full-formant syllables indicate, not un-

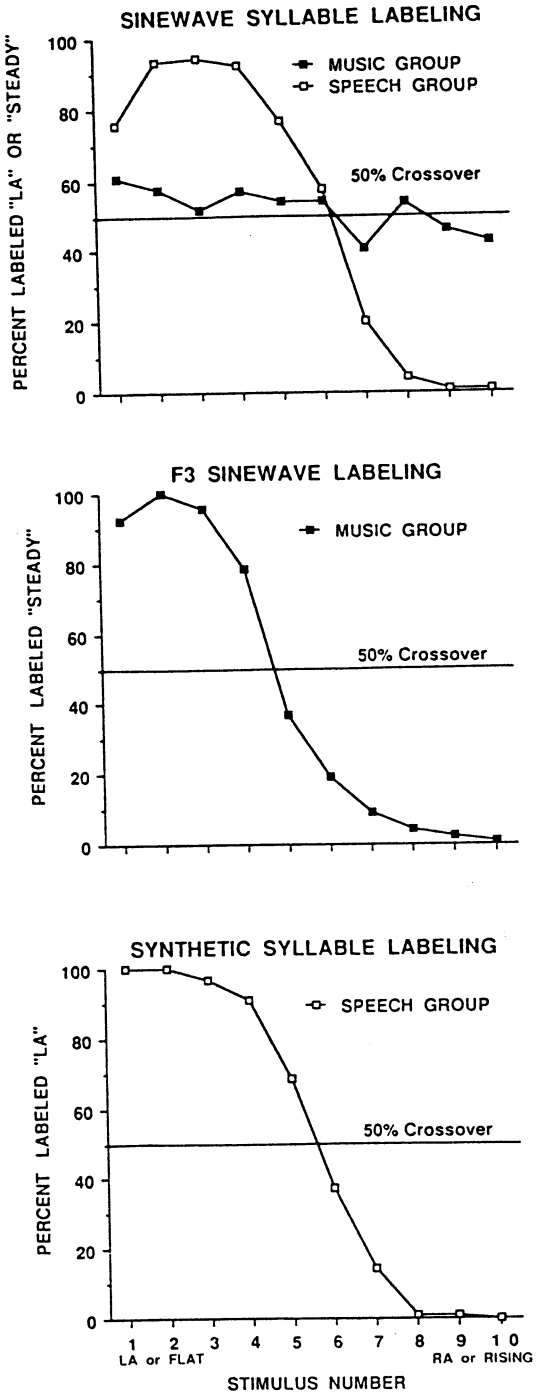

Figure 4. Labeling functions for the speech and music listeners on each of the three stimulus continua in Experiment 2.

expectedly, that their responses were less consistent for the phonetically impoverished and unfamiliar syllables of the sine wave dialect than for standard synthetic speech. Otherwise, they treated the two forms of speech identically, as the phonetic coherence hypothesis predicts. The music listeners, on the other hand, despite their practice with the F3 tones before hearing the sine wave syllables, performed even less consistently than in Experiment 1. They categorized isolated F3 tones with fair consistency, but they were quite unable to exploit this supposed cue when they heard it in the context of the two lower tones.

Nonetheless, the possibility remained that the group differences might be attributable to the use of labels per se. People have had so much more experience with naming words and syllables than with labeling nonspeech sounds, particularly with labeling slight differences in the onset properties of single notes within a chord, that this experiential difference alone might account for the relatively poor performance of the music listeners on the sine wave syllable task. Moreover, the memory demands of the task, which requires listeners to remember the category exemplars in order to label individual items from the continuum, may be much greater for the music listeners than for the speech listeners.

EXPERIMENT 3

A categorization task that does not require labels and that provides the subject with category exemplars on each trial would circumvent the difficulties noted above, as would a discrimination task. We therefore conducted a third experiment, using an AXB categorization procedure that provides clear-case exemplars on each trial, but does not require category labels (e.g., Bailey, Summerfield, & Dorman, 1977; Best et al., 1981), as well as an AXB discrimination task that places relatively low demands on short-term memory (Best et al., 1981).

Method

Subjects. Thirty-four young adults were tested in Experiment 3. The speech bias condition was run on 12 subjects (6 males, 6 females), and the music bias condition on 22 subjects (13 males, 9 females). All had normal hearing and negative personal and family histories of language or speech difficulties, and each was paid $4 for participation. Subsequently, on the basis of posttest questionnaire answers, 1 female speech subject was eliminated because she did not hear the sine wave syllables as speech. Nine subjects were eliminated from the music condition: 8 of these had begun to hear the sine wave syllables as words or syllables (6 males, 2 females), and the remaining subject (a female) failed to perceive the full-

Table 3
Experiment 2: Mean Category Boundaries and Slopes, Determined from Individual Probit Analyses on Stimulus Items 2 Through 10, for Speech and Music Listeners

	Category Boundaries			Slopes		
	Sine Wave	F3	Full-Formant	Sine Wave	F3	Full-Formant
	Speech Listeners ($n=12$)					
M	5.90	—	5.66	0.61	—	2.27
SE	0.24	—	0.33	0.10	—	0.46
	Music Listeners ($n=12$)					
M	10.81	3.97	—	0.04	1.25	—
SE	7.74	0.20	—	0.02	0.31	—

formant series as /la/ and /ra/. The final samples were therefore unequal, with *n*s of 11 and 13 for the speech and music groups, respectively.

Stimuli. The stimuli were the same as in the first two experiments.

Procedure. The subjects were tested under the same listening conditions as before. All subjects completed two tests on each of the three stimulus continua: an AXB categorization test and an AXB discrimination test. They completed the categorization test before the discrimination test for each continuum, in the order (1) sine wave syllables, (2) isolated F3 tones, and (3) full-formant syllables. As in Experiment 1, the sine wave syllable test was presented first to prevent any possible influence of exposure to the other stimulus series on performance with the sine wave syllables. The sine wave syllable test was preceded by the appropriate instruction and training set for the condition randomly assigned to each subject.

On each trial of the AXB categorization tests, three stimuli were presented. The first and third stimuli were constant throughout the test: the endpoint /ra/ or rising F3 item from the appropriate continuum, and the second /la/ or flat F3 item. The middle stimulus, X, varied randomly among the 10 items of the stimulus series. The subject's task on each trial was to indicate whether X belonged in the same perceptual category as A (first) or B (third). Each AXB categorization test contained 10 blocked randomizations of the trials for the 10 items in the continuum, with 1.5-sec ISIs, 3.5-sec ITIs, and 5-sec IBIs.

In the AXB discrimination tests, three stimuli were also presented on each trial. However, the first and third stimuli (A and B, respectively) were always three steps apart on the appropriate stimulus continuum and varied from trial to trial, whereas the middle stimulus (X) always matched either A or B. The subject's task on the discrimination tests was to indicate whether X was identical to A or to B. Each discrimination test contained five randomizations of the 28 possible AXB configurations, blocked in groups of 14 trials, with 1.5-sec ISIs, 3.5-sec ITIs, and 5-sec IBIs.

Results

Categorization. The data were treated as in the previous experiments. Figure 5 displays the group categorization functions, and Table 4 lists the mean category boundaries and slopes, with their standard errors. Category boundaries appear to differ across continua, being lowest on the F3 tones for both groups. The slopes on the sine wave syllables are somewhat steeper than in the previous experiments, but they are still shallowest on the sine wave and steepest on the full-formant syllables for both groups. The music listeners again give a shallower slope than the speech listeners on the sine wave syllables, and the pattern of standard errors for this continuum replicates that of the previous experiments.

Two-factor (instruction condition × stimulus continuum) analyses of variance, with repeated measures on stimulus continuum, were carried out on category boundaries and slopes. For the category boundaries there was a significant effect of stimulus continuum [$F(2,44) = 7.16$, $p < .002$] and a significant interaction between stimulus continuum and instruction condition [$F(2,44) = 3.41$, $p < .04$], but no effect of instruction condition. Scheffé tests on the three pair-by-pair boundary comparisons gave no significant differences for either speech listeners or music listeners.

Analysis of the slope data yielded a significant effect of stimulus condition [$F(2,44) = 9.81$, $p < .0003$], but no effect of instruction condition and no significant interaction. Scheffé tests showed that the sine wave syllable slopes were significantly lower than the slopes for either the F3 tones [$F(1,22) = 10.51$, $p < .05$] or the full-formant syllables [$F(1,22) = 12.56$, $p < .05$], but that the slopes for the full-formant syllables and F3 tones

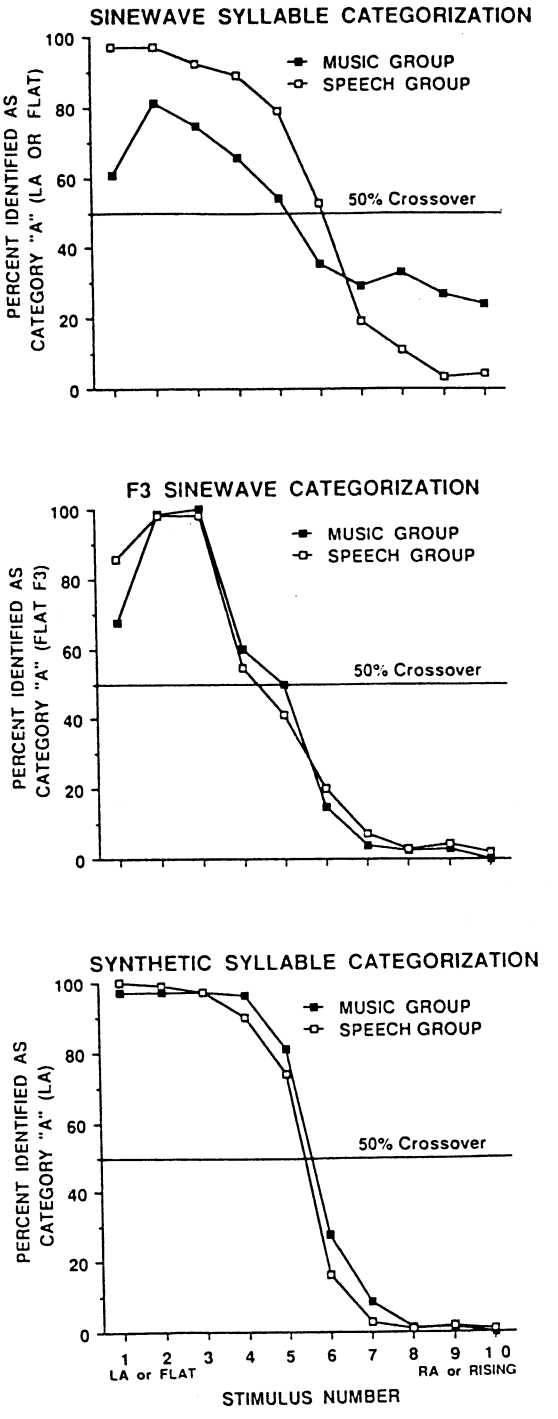

Figure 5. AXB categorization functions for the speech and music listeners on each of the three stimulus continua in Experiment 3.

Table 4
Experiment 3: Mean Category Boundaries and Slopes, Determined from Individual Probit Analyses on Stimulus Items 2 Through 10, for Speech ($n=11$) and Music ($n=13$) Listeners

	Category Boundaries			Slopes		
	Sine Wave	F3	Full-Formant	Sine Wave	F3	Full-Formant
	Speech Listeners ($n=11$)					
M	6.03	4.82	5.37	0.83	1.24	2.75
SE	0.19	0.27	0.17	0.10	0.32	0.90
	Music Listeners ($n=13$)					
M	5.19	4.84	5.65	0.26	0.98	1.62
SE	0.30	0.16	0.15	0.06	0.10	0.32

did not differ. Although there was no significant interaction, we carried out the planned t test to assess the difference between speech and music listeners' slopes on the sine wave syllables, predicted by the phonetic coherence hypothesis. The difference was again significant [$t(22) = 4.99$, $p < .0001$].

Tests for category formation on the sine wave syllable continuum, analogous to those of the previous experiments, showed a significant effect of stimulus item [$F(8,80) = 131.43$, $p < .0001$] for the speech listeners, with Stimuli 2 through 5 forming one category and Stimuli 6 through 10 another, according to Scheffé tests ($p < .05$ for all between-category comparisons; not significant for any within-category comparison). For the music listeners, there was also a significant effect of stimulus item [$F(8,96) = 16.58$, $p < .0001$]; Scheffé tests indicated that Stimuli 1 through 5 fell into one category, Stimuli 7 through 10 into another ($p < .05$ for all between-category comparisons; not significant for any within-category comparison).

Discrimination. Each subject's percent correct performance was computed for each stimulus pair in each AXB discrimination test. Figure 6 displays mean performance for the two groups on the three continua. The full-formant syllables yield a standard speech continuum pattern for both groups: performance peaks on discrimination pairs that straddle or abut the category boundary. We see a similar, but somewhat flattened, function for the speech listeners on the sine wave syllables. By contrast, the music listeners show no systematic peaks on the sine wave syllables: performance declines across the continuum, as it does for both groups on the F3 tones, although the latter elicit a generally higher level of discrimination than do the other continua.

A three-factor (instruction condition × stimulus continuum × discrimination pair) analysis of variance, with repeated measures on stimulus continua and discrimination pairs, was carried out. A significant stimulus continuum effect [$F(2,44) = 51.99$, $p < .00001$] indicated that discrimination performance was highest overall for the isolated F3 tones and lowest for the sine wave syllables. A significant interaction between instruction condition and stimulus continuum [$F(2,44) = 5.56$, $p < .01$],

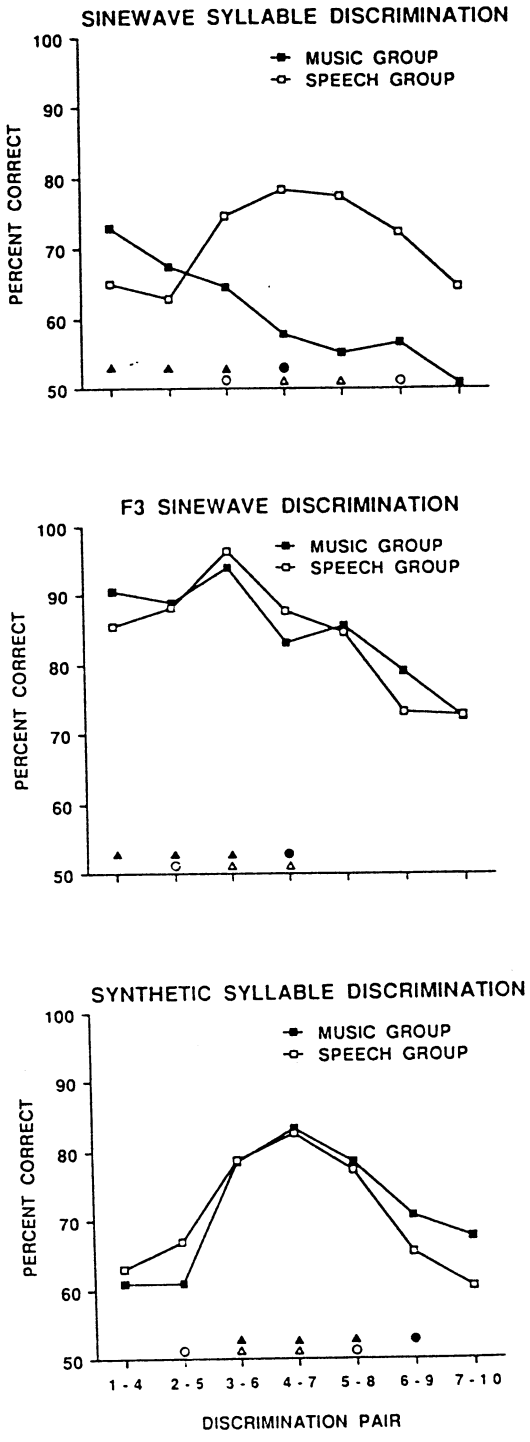

Figure 6. Discrimination functions for the speech and music listeners on each of the three stimulus continua in Experiment 3. The markers above the abscissa indicate the status of each stimulus pair with respect to categorization judgments. Triangles indicate pairs that straddle the category boundary for a given group. Circles indicate pairs in which one item is at or near the boundary. Filled markers represent the music listeners' data, open markers the speech listeners' data.

combined with Scheffé tests, indicated that the two groups differed in performance only on the sine wave syllable test, on which the speech listeners outperformed the music listeners [$F(1,22) = 13.88, p < .05$]. Further Scheffé tests showed that performance on the sine wave and full-formant syllable tests did not differ for the speech listeners, but did differ significantly for the music listeners [$F1,12) = 75.58, p < .001$].

The effect of discrimination pair was significant [$F(6,132) = 16.29, p < .00001$], indicating that overall performance level was not uniform throughout the stimulus continua, but rather showed higher performance on some discrimination pairs than on others. A significant interaction between discrimination pair and stimulus continuum [$F(12,264) = 5.29, p < .00001$] indicated further that the pattern of the discrimination function differed among the three stimulus continua. A three-way interaction of instruction group with discrimination pair and stimulus continuum [$F(12,264) = 3.01, p = .005$] evidently arose because, according to a Scheffé test of the instruction × discrimination pair interaction [$F(6,132) = 6.01, p < .01$], only the speech group showed a peak in performance level near the category boundary on the sine wave syllable test.

Discussion

The influence of attentional mode on perception of the sine wave syllables demonstrated in Experiments 1 and 2 was replicated in Experiment 3. Evidently the difference in the way speech and music listeners categorize these syllables reflects a true difference in perceptual response, and is not simply a function of differences in ability to assign labels to speech versus nonspeech stimuli, or in the influence of short-term memory for the speech versus nonspeech category exemplars. Although removal of the requirement for overt labeling, as well as presentation of category exemplars for comparison with the target item on each trial, certainly permitted the music listeners to form somewhat more consistent sine wave syllable categories than were observed in the previous experiments, these listeners were still less consistent than the speech listeners, and gave no evidence of a peak at their sine wave syllable category boundary in the discrimination test. Evidently their categories, such as they are, are less robust and more dependent on experimental conditions than are those of the speech listeners. Thus, support for the phonetic coherence hypothesis, over the alternative psychoacoustic hypotheses, was considerably strengthened by Experiment 3.

GENERAL DISCUSSION

The results of these three experiments are inconsistent with the claim that speech perception entails the simple extraction, or the extraction and integration, of discrete information-bearing elements or cues. All listeners could correctly classify, within psychophysical limits, the transitions on the isolated F3 tones as steady or rising. However, music listeners, biased to listen for the transition in the context of the lower F1 and F2 tones, could not then reliably recover the target pattern. They also could not either integrate the F3 cue with other cues in the F1–F2 array or apprehend the auditory coherence of the total pattern so as to arrive at a unitary, distinctive percept for each category. By contrast, listeners biased to hear the sine wave patterns as speech were evidently immune to whatever psychoacoustic interactions blocked consistent judgments of the patterns by music listeners, in that the former classified the sine wave syllables only somewhat less consistently than they classified the full-formant syllables. These results agree with those of several other studies of sine wave speech in arguing for a specialized mode of speech perception (e.g., Best et al., 1981; Tomiak, Mullenix, & Sawusch, 1987; Williams, 1987).

How are we to characterize this mode? What did the speech listeners in these experiments do that the music listeners did not? Consider, first, the music listeners' performance with the sine wave syllables. In Experiment 1, and particularly in Experiment 3, where labeling was not required so that listeners could compare whole signals without attempting to isolate distinctive cues, the music group's categorization function sloped in the "correct" direction. At least some of these listeners grasped certain contrastive properties of the signals, even though, according to the posttest questionnaire, they did not perceive them as speech. One suspects that, with sufficiently prolonged training under suitable experimental conditions (e.g., those provided by AXB categorization, as used in Experiment 3), these listeners might even come to render judgments of the sine wave syllables no less consistent than those of the speech listeners. However, if they did so, it would remain notable that they require extensive training, whereas the speech listeners require very little. Moreover, even if extensive training aided the music listeners, would they then be perceiving the patterns as speech? The answer would surely be yes, if they could tell us the names of the sounds they had heard, that is, if they had discovered the articulatory patterns implied by the signals. However, if they could not tell us the names, the answer would be no. Their condition might, in fact, be much like that of nonhuman animals trained to distinguish between speech sound categories (Kluender, Diehl, & Killeen, 1987). Alternatively, they might be categorizing the patterns on adventitous nonspeech properties, rather as a color-blind individual might correctly classify two objects of different colors on the basis of their differences in brightness rather than of their differences in hue.

Consider here another class of listener well known to have difficulty with the English /r/-/l/ distinction: monolingual speakers of Japanese. Figure 7 displays a group categorization function for 7 such speakers attempting to classify full-formant patterns on an /r/-/l/ continuum similar to the continuum of the present experiments (MacKain, Best, & Strange, 1981). The function is remarkably like that of the music listeners attempting to

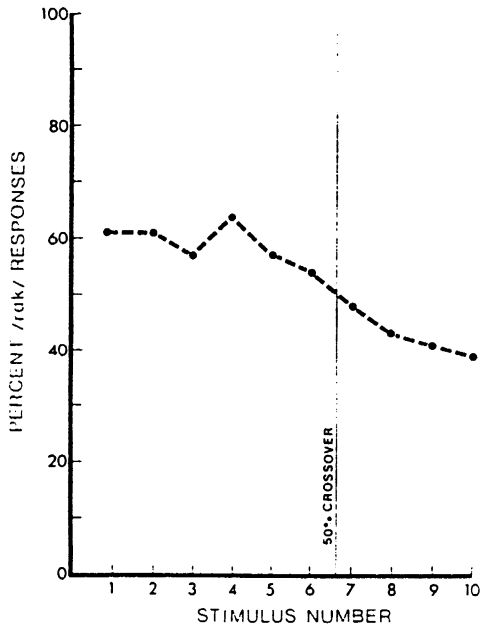

Figure 7. Labeling of a rock–lock continuum by Japanese adults without extensive English conversation experience. (Adapted from "Categorical Perception of English /r/ and /l/ by Japanese Bilinguals" by K. S. MacKain, C. T. Best, and W. Strange, 1981, *Applied Psycholinguistics*, 2, p. 378. Copyright 1981 by Cambridge University Press. Reprinted by permission.)

classify the corresponding sine wave syllables: some of these Japanese listeners also seem to have captured certain contrastive auditory properties of the signals. However, unlike the music listeners, who were diverted from hearing the sounds as speech by being trained to attend to an acoustic cue rather than to the whole pattern of which it was a part, these Japanese listeners were diverted because they had not discovered the relevant properties of contrastive sounds spoken in an unfamiliar language. However, we know that, with sufficient exposure to the English /r/–/l/ contrast in natural contexts where it serves a phonological function, Japanese listeners can come to hear the contrast correctly and categorically (MacKain et al., 1981). What have they then learned or discovered? What, more generally, has any second language learner—or, indeed, any child learning a first language—discovered when the auditory patterns of a target language drop into phonological place? Presumably, they have learned to do more than classify auditory patterns consistently. They also have discovered the correct basis for classification, namely, the articulatory structures that the patterns specify.

Notice that this formulation attempts to resolve the notorious lack of correspondence between a quasi-continuous acoustic signal and its abstract linguistic predicates by positing an event with observable, physical content—the articulatory gesture—as the fundamental unit of both production and perception. We are thus dealing with patterns of movement in space and time, accessible to treatment according to general principles of motor control (Browman & Goldstein, 1986; Kelso, Tuller, & Harris, 1983; Saltzman & Kelso, 1987). The task for perceptual theory, then, is to uncover the acoustic properties that specify a particular pattern of gestures. These acoustic properties will not be simply a collection of independent cues, arrived at by (articulatorily) unconstrained manipulation of a terminal analog synthesizer, but rather sets of correlated properties that arise from coordinated patterns of gesture, tested by systematic articulatory synthesis.

We should emphasize that we are not here arguing for a "motor theory" of speech perception. We are not proposing that the *process* of arriving at a speech percept engages mechanisms outside the auditory system that humans share with many other animals. What is specific to speech, and to humans, is the final percept, a phonological structure determined by the structure of the articulatory gestures from which the signal arose (see Repp, 1987). Since (as the Japanese listeners show) this structure is inaccessible to adult humans, unless they can assimilate the sounds to the phonological categories of a language they know (see Best, McRoberts, & Sithole, 1988), we assume a fortiori that it is also inaccessible to the infant who does not yet know a language. The infant's task is to discover the phonetic coherence of phonological categories in the surrounding language by focusing attention on recurrent auditory contrasts that signal changes of meaning in that language (see Jusczyk, 1986; Studdert-Kennedy, 1986, 1987). An articulatory "representation" of the phonologically contrastive patterns is then an automatic consequence of the species-specific perceptuomotor link that underlies the child's capacity to imitate speech patterns, and thus to learn to talk.

Finally, although we have couched our experimental procedures in terms of attention, we do not mean to imply that the processes of speech perception can be engaged or disengaged at will. Although some listeners can choose to attend or not to attend to particular aspects of a speech signal, such as a speaker's "accent," or even the spectral properties of a fricative noise (Repp, 1981), it is difficult, if not impossible, to hear natural speech, spoken clearly in a language that we know, as nonspeech. Also, as we have already remarked, listeners who have once heard a particular sine wave pattern as speech find it difficult later to hear that pattern—or any other sine wave pattern modeled on speech—as entirely devoid of phonetic structure. Evidently, perceiving speechlike patterns as speech is as mandatory, automatic, and unconscious as, say, perceiving the rhythm and melody of a nonspeech auditory form.

REFERENCES

ABRAMSON, A. S., & LISKER, L. (1965). Voice onset time in stop consonants: Acoustic analysis and synthesis. In D. E. Commins (Ed.), *Proceedings of the 5th International Congress of Acoustics (A51)*. Liege: Thone.

BAILEY, P. J., & SUMMERFIELD, Q. (1980). Information in speech: Ob-

servations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 536-563.

BAILEY, P. J., SUMMERFIELD, Q., & DORMAN, M. F. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins Laboratories Status Report, SR51/52*, 1-25.

BEST, C. T., MCROBERTS, G. W., & SITHOLE, N. M. (1988). Examination of perceptual reorganization for non-native speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 345-360.

BEST, C. T., MORRONGIELLO, B., & ROBSON, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, **29**, 191-211.

BREGMAN, A. (1981). Asking the "what for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 99-118). Hillsdale, NJ: Erlbaum.

BROWMAN, C. P., & GOLDSTEIN, L. (1986). Towards an articulatory phonology. *Phonology Yearbook*, **3**, 219-252.

COLE, R. A., & SCOTT, B. (1974). Toward a theory of speech perception. *Psychological Review*, **81**, 348-374.

CUTTING, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**, 114-140.

DARWIN, C. J. (1984). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, **76**, 1636-1647.

DELGUTTE, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 131-149). New York: Elsevier.

DORMAN, M. F., CUTTING, J. E., & RAPHAEL, L. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception & Performance*, **1**, 121-129.

DORMAN, M. F., STUDDERT-KENNEDY, M., & RAPHAEL, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, **22**, 109-122.

FINNEY, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.

FOWLER, C. A., & SMITH, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123-136). Hillsdale, NJ: Erlbaum.

JUSCZYK, P. W. (1986). Toward a model of the development of speech perception. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 1-19). Hillsdale, NJ: Erlbaum.

JUSCZYK, P. W., SMITH, L. B., & MURPHY, C. (1981). The perceptual classification of speech. *Perception & Psychophysics*, **30**, 10-23.

KELSO, J. A. S., TULLER, B., & HARRIS, K. (1983). A 'dynamic pattern' perspective on the control and coordination of movement. In P. MacNeilage (Ed.), *The production of speech* (pp. 138-173). New York: Springer-Verlag.

KLUENDER, K. R., DIEHL, R. L., & KILLEEN, P. R. (1987). Japanese quail can learn phonetic categories. *Science*, **237**, 1195-1197.

MACKAIN, K. S., BEST, C. T., & STRANGE, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, **2**, 369-390.

MATTINGLY, I. G. (1981). Phonetic representation and speech synthesis by rule. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech* (pp. 415-420). Amsterdam: North-Holland.

MATTINGLY, I. G., LIBERMAN, A. M., SYRDAL, A. M., & HALWES, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131-157.

ODEN, G. C., & MASSARO, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

PASTORE, R. E. (1981). Possible psychoacoustic factors in speech perception. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 165-205). Hillsdale, NJ: Erlbaum.

REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1980). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

REPP, B. H. (1981). Two strategies in fricative discrimination. *Perception & Psychophysics*, **30**, 217-227.

REPP, B. H. (1987). The role of psychophysics in understanding speech perception. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 3-27). Boston: Martinus Nijhoff.

SALTZMAN, E., & KELSO, J. A. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, **94**, 84-106.

SCHOUTEN, M. E. H. (1980). The case against a speech mode of perception. *Acta Otolaryngologica*, **44**, 71-98.

STUDDERT-KENNEDY, M. (1986). Sources of variability in early speech development. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 58-76). Hillsdale, NJ: Erlbaum.

STUDDERT-KENNEDY, M. (1987). The phoneme as a perceptuomotor structure. In A. Allport, D. MacKay, W. Prinz, & E. Scheerer (Eds.), *Language perception and production* (pp. 67-84). London: Academic Press.

TOMIAK, G. R., MULLENIX, J. W., & SAWUSCH, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America*, **81**, 755-764.

WILLIAMS, D. R. (1987). *The role of dynamic information in the perception of coarticulated vowels*. Unpublished doctoral dissertation, University of Connecticut, Storrs.

NOTES

1. In this paper, we use the term *cohere*, and its derivatives, to refer to the effect of a perceptual process by which listeners apprehend a complex signal as a unitary pattern or configuration, rather than as a collection of discrete elements. In vision, we may add to the many examples familiar from textbook treatments of Gestalt principles, the phenomenon of face recognition, in which the identity of a face emerges as a holistic pattern, not simply as a collection of discrete features. In speech, the unitary patterns would correspond to units of linguistic function, such as phonemes, syllables, morphemes, and words.

2. Because the first stimulus on the continuum had a slightly falling transition (originally intended to help bias the full-formant series toward /l/, as noted above), listeners (particularly the music listeners on the sine wave syllables, and both groups on the F3 tones) tended to judge this stimulus with slightly less consistency than its neighbors (see Figures 2 through 5). As a result, probit analyses tended to yield lower slopes than were characteristic of the main bodies of the functions, and thus to exaggerate the slope differences between groups and conditions. We therefore omitted this stimulus from the probit analyses: all computed means and slopes in this and the following experiments are based on analyses of individual functions for Stimuli 2 through 10.

Note, furthermore, that by converting the standard deviations of the underlying distributions into their reciprocals (the slopes of the cumulative functions), we went some way toward homogenizing the group variances, as appropriate for subsequent analyses of variance. At the same time, we reduced the apparent differences between groups across stimulus continua. For example, in Table 1 the difference between the mean slopes for the two instruction conditions is not much greater on the sine wave syllables ($0.80 - 0.14 = 0.66$) than on the full-formant syllables ($2.30 - 1.70 = 0.60$). But the difference between the mean standard deviations for the two conditions is very much greater on the sine wave syllables ($1/0.14 - 1/0.80 = 5.89$) than on the full-formant syllables ($1/1.70 - 1/2.30 = 0.15$). The reader should bear this in mind when comparing slopes depicted in the figures with slope values listed in the tables.

3. One of the five music subjects from Experiment 1a, who returned to take the speech test, later turned out to be the one we rejected from that experiment because she had heard some of the sine wave syllables as sounding like r. However, her data were not appreciably different from those of other music listeners, and so we retained her in Experiment 1b. If her tendency to hear some of the sine wave syllables as r-like had facilitated her categorization of these syllables, the result would presumably have been to reduce the differences between instruction conditions. However, the results of statistical analyses on the data of Experiment 1b were unchanged when this subject's data were eliminated.

(Manuscript received April 11, 1988;
revision accepted for publication September 21, 1988.)