Sound-producing sources as objects of perception: Rate normalization and nonspeech perception

724

Carol A. Fowler
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511 and Dartmouth College, Hanover, New Hampshire 03755

(Received 17 July 1989; accepted for publication 17 May 1990)

In a variety of experiments and paradigms, researchers have attempted to determine whether or not speech perception is specialized by comparing perception of speech syllables to perception of nonspeech analogs. While nonspeech analogs appear optimal as comparisons to speech because they are acoustically similar without being recognized as speechlike, it is argued that the comparison they offer is confounded and uninterpretable. Two experiments are designed to show that, in auditory perception generally where acoustic signals are causal consequences of mechanical events, perceptual experiences are of the mechanical events themselves, not of the acoustic signal. This has two consequences. One is that there is a confounding in comparisons of speech with sine wave analogs that, whereas the one perceived as speech also has a definite causal source, the other, perceived as nonspeech, has an indeterminate or ambiguous source. A second is that response patterns in classification tasks such as those used in the literature comparing speech to nonspeech will reflect properties of the perceived sound-producing event; they will not provide a clear window on auditory system processes used to recover event properties. Experiment 3 is designed to show that perception of many acoustic-signal-producing events can appear to be special by the logic of speech–sine wave comparisons—even events that cannot plausibly be supposed to involve a specialization.

PACS numbers: 43.71.An, 43.66.Ba, 43.66.Lj

INTRODUCTION

In the literature on speech perception, one way that researchers have investigated whether aspects of speech perception are or are not special to speech is to compare perception of speech with perception of nonspeech analogs (e.g., Diehl and Walsh, 1989; Liberman *et al.*, 1961; Pisoni *et al.*, 1983; and see the reviews in the Symposium on Perception of Speech versus Non-speech in the Ninth International Congress on Phonetic Sciences, 1979). In a similar, slightly more elegant procedure, investigators compare perception of the same signal, selected to be sufficiently ambiguous that listeners hear it as speech under one attentional set but as nonspeech under another (Best *et al.*, 1981, 1989). Finally, in "duplex" perception (e.g., Mann and Liberman, 1983), they compare perception of the same acoustic fragment, perceived simultaneously as part of a syllable and as a nonspeech pitch glide.

The logic of these comparisons is, I believe, as follows. Generally, specialized perceptual processes applied to an acoustic signal will, in virtue of their specialization, yield a different perceptual outcome than would be yielded by general auditory system processes. Accordingly, one way to determine whether specialized or general perceptual processes underlie speech percepts is to compare perception of speech signals with perception of other signals that are as similar to the speech acoustically as possible, but that are not processed as speech. The importance of making the speech and nonspeech signals as similar acoustically as possible is in helping to guarantee that, in the absence of specialized processes for speech, the same general auditory-system processes will be applied to both sets of signals. In turn, the processes applied to the signal will determine the listeners' response patterns in a perception test thereby revealing themselves as similar or different to investigators.

In comparisons of speech with nonspeech analogs, results have been mixed. In some instances (e.g., Liberman *et al.*, 1961), findings are dissimilar for speech and nonspeech sounds; in others (e.g., Pisoni *et al.*, 1983), they are similar; Finally, Diehl and Walsh (1989) find similarity in respect to one comparison (frequency transition duration) and a difference in respect to another (amplitude rise time). In the other designs, response patterns to speech and nonspeech sounds are generally different. In short, a survey of the collection of findings does not yield a consistent answer to the question they pose whether speech perception is special.

There may be several reasons for the divergent outcomes. One possibility is that some processes applied to speech signals are special while others are not (e.g., Eimas, 1985; Miller and Eimas, 1983). While this is plausible in general, it is unlikely to explain the pattern of findings in the literature. In the literature, the aspects of perception found either to be special to speech or else general to auditory perception do not seem to form natural classes. (For example, it is not that perception of consonantal signals is special, while perception of vocalic ones are not, or that perception of formant transitions is special while perception of other acoustic consequences of speaking are not.) An alternative possibility is that some nonspeech acoustic analogs are inappropriate in either of two ways. They may be so speechlike as to be processed as speech [as Diehl and Walsh (1989) speculate regarding signals used by Pisoni *et al.* (1983)], giving rise to

 0001-4966/90/091236-14$00.80

spurious evidence against a specialization. Alternatively, nonspeech analogs may be insufficiently similar to the speech sounds they were meant to mimic so that different general auditory processes apply to them, giving rise to spurious evidence for a specialization. This alternative will not suffice to explain extant findings either. Similar response patterns are found to speech signals and to rather dissimilar nonspeech signals (Diehl and Walsh, 1989), while different response patterns are found to the same signals depending on whether they are heard as speech or as nonspeech (Best *et al.*, 1981; 1989).

I would like to suggest another reason for the inconsistent outcomes. It is that the comparisons of speech and nonspeech leave out of consideration the function that perception of acoustic signals serves the perceiver and the influence that function has on response patterns in these experiments. The function of perception is to acquaint perceivers with the environment via the casual effects that environments have on stimulation available at the sense organs. For their part, perceivers cannot avoid attempts at recovery of environmental properties from stimulation; that is what their perceptual systems have evolved to do, specialized or not and confronted or not with stimulation that specifies a real source in the environment. A consequence is that response patterns in tests of classification of speech and nonspeech signals do not particularly index the character of perceptual processes; they are strongly affected by what the signals are perceived *as*. When investigators use sine wave signals or any nonspeech signals that do not specify a particular sound-producing source in the environment, they make it difficult or impossible to determine what causal sources will be ascribed to the signals, and hence to identify that source of variation in listeners' responses. They do not prevent, because (I argue) they cannot, attempts *by* perceivers at source recover.

In the remainder of the Introduction, I will briefly unpack this argument and apply it to a particular series of experiments in the literature. Following that, experiments 1–3 are designed to demonstrate the effect that perception of environmental source properties can have on response patterns and to reveal the consequent uninterpretability of speech/nonspeech comparisons in terms of the similarity or dissimilarity of auditory-system processes applied to the different signals.

A. The environment-directed character of perception

In early research at Haskins Laboratories, speech researchers obtained findings that, in their view, called for a radical, counter-intuitive theory of speech perception (see Liberman *et al.*, 1967 for a summary of that work). Their findings were that certain dimensions of a listener's speech percept correspond poorly with the dimensions of the acoustic signal found to support it, but correspond more obviously with the vocal tract actions giving rise to the signal. A well-known example is the case of synthetic, two-formant /di/ and /du/. In these syllables, the /d/ percept is supported by second formant transitions, which are high in the spectrum and rising in /di/ and lower in the spectrum and falling for /du/. Isolated from their contexts, the transitions sound distinct and, indeed, appear to be perceived literally—the one sounding like a high-pitched rising glide and the other like a lower pitched fall. The latter finding, considered unsurprising by Haskins' researchers, suggested to them that acoustic signals interpreted as nonspeech give rise to acoustic perceptual objects. However, the findings on /di/ and /du/, suggested (among other findings obtained by Haskins' researchers) that perceptual objects of speech are articulatory. Due to coarticulation, the transitions following consonantal release are context sensitive; however, both /d/'s are produced by the same constriction gesture of the tongue blade against the alveolar ridge of the palate. The motor theory of speech perception was devised to explain both the speech-specific nature of the percept and its motor character as well. According to the theory, the listener's own speech motor system is recruited in perception to help disentangle effects that coarticulation has on the acoustic speech signal (see, e.g., Liberman *et al.*, 1967 and Liberman and Mattingly, 1985 for additional details).

Looked at in a larger context, however, findings suggesting that listeners recover the signal-producing source in speech perception do not call for a motor theory of speech perception. Rather, they suggest that speech perception serves a function identical to that severed by perceptual systems generally.

Perceptual systems provide the means—indeed, the only means—by which animals can come to know the environment in which they participate as actors. In its public aspect,[1] we know generally how perception serves that function. Consider vision, haptics, and, finally, speech perception by ear. Objects and events in the environment causally structure light; generally, different objects and events structure light differently. Precisely, because structure in the light is caused by properties of objects and events and because the structure is largely specific to its causal source, it can serve as information for its source to a sensitive perceiver. Evidently, that is what it does do (see, e.g., Gibson, 1966, 1979; Warren, 1981). While persons on the street can describe their surrounding environment adequately, without a course on optics, they cannot describe the patterning in the reflected light that supports their perceptual experience and knowledge.

Consider the effect that has on observers' response patterns in a perception experiment. Imagine that observers are asked to judge the distances from themselves to a set of stakes in a field. Clearly the major sources of variation in the observers' responses are the real distances to the stakes from the observers. If the exponent relating judged to real distance is 1, we learn nothing from the response patterns about visual perceptual processes except that they are transparent to the judged environmental variation. Even if the exponent is not 1, we have not necessarily learned something about visual system processes from the response patterns; at least some departures from linearity can be ascribed to the decisional processes by which subjects map their percepts onto the permissible responses in the experiment (e.g., Wagner and Baird, 1981).

Haptic perception is no different. Perceivers handling an object can describe the object, but not the mosaic of skin deformation and joint-angle changes that handling the ob-

ject causes. Response patterning in a haptics experiment will reveal what the perceiver knows about the object; it will not provide a window on perceptual processes.

It is crucial to the perceiver that perception work in this way—that is, that it recover real world objects and events from the stimulation they cause at the sense organs, rather than yielding an experience of the stimulation itself. In this way, the perceptual systems together yield knowledge of the world in which the perceiver participates as an actor. While it matters not at all whether a perceiver collides with a symmetrically expanding optic array, it matters a great deal whether or not he/she collides with the moving automobile that causes the optic array to expand. The same consideration applies to auditory systems. The world of events that causes disturbances in the air needs to be perceived, not the disturbed air itself; if auditory perception is like visual and haptic perception, perceptual objects will be mechanical acoustic-signal-causing events, not acoustic signals themselves. (That we hear sounds as emanating from a place in the environment is perhaps a hint that auditory perception—even of sine waves— is environment-directed.)

Considered in the context of the foregoing discussion, findings that gave rise to the motor theory should not have been explained by appeal to a specialization of the brain in which the listener's own speech motor system gave the speech percept its correspondence with vocal-tract actions. Nor, however, should theorists, adopting a view that speech perception is not special, work to show that perceptual objects of speech are acoustic (or "acoustic/auditory" as in Diehl and Kluender, 1989). If speech perception is not special with respect to the perceptual objects it renders, then perceptual objects should be the events in the environment that cause the acoustic speech signals, not the acoustic signals *per se*. Those events are linguistically significant vocal-tract actions (e.g., Fowler, 1986).

Consider the importance of these conclusions for interpreting comparisons of speech perception with perception of sine wave analogs in terms of specialized or general auditory-system processes. They suggest that response patterns on the identification tests used in this research domain will be strongly affected by listeners' attempts to use acoustic signals as (I argue) their evolutionary heritage requires—as information for the ostensible causes of the signals in the environment. Similar response patterns may be found to speech and nonspeech stimuli because, in respect to the property under test, the perceived speech and nonspeech events are similar; by the same token, different response patterns may occur because, *in respect to the property under test*, the perceived events are different. Neither outcome may be interpreted safely as evidence that auditory processes applied to the signals are the same or different.

The experiments I report below were designed to address these points. They were provoked by a series of experiments on rate normalization for speech in the literature that I review next.

B. Rate normalization or durational contrast?

The series of studies begins with an experiment (Miller and Liberman, 1979) designed to investigate listeners' use of after-coming acoustic information (in this case, ostensibly for speech rate) to scale their classifications of a consonant as a stop or glide. More specifically, Miller and Liberman (1979) reported an effect of after-coming acoustic structure on phonetic classification as "b" or "w" of members of a continuum of formant transitions varying in duration (and in amplitude rise time). In general, on consonant–vowel (CV) syllables, continuum members with short-duration transitions were identified as "b," appropriate to the rapid gesture associated with bilabial-stop release, and longer-duration members were identified as "w" appropriate to the glide's slower rate of release. In addition, however, Miller and Liberman found that longer- as compared to shorter-duration steady-state formants for a following /a/ vowel were associated with an increased probability of classifying continuum members as "b" rather than "w." Subsequent experiments in the series showed that other ways of increasing the duration of the speech signal following the consonantal transitions had smaller effects on listeners' classifications, or even effects opposite to that of increasing steady-state vowel duration.

Miller and Liberman (1979) interpreted their findings in the following way: /b/ and /w/ share place of articulation and differ saliently in the velocity of their opening gestures; consequently, explosive formant transitions for /b/ are spectrally similar to those for /w/, but, as compared to those for /w/, they are compressed in time. Accordingly, transitions for a /w/ spoken at a fast rate of speech may closely resemble those for a slow-rate /b/. As compared to a short vowel, a long vowel signals to a listener that a syllable including the vowel has been produced at a slow rate of speech. Listeners use variation in vowel duration in a CV syllable as information for speaking rate, and their classifications of a consonant as a relatively rapidly released stop (/b/) or a relatively slowly released glide (/w/) reflect normalization for rate. Other ways of increasing syllable duration that do not signal a decrease in speaking rate do not give rise to rate scaling in /b/–/w/ classifications.

Since publication of Miller and Liberman's research, two studies have challenged the conclusion that listeners literally normalize for speech rate in classifying a syllable-initial bilabial consonant as "b" or "w" even though what listeners do instead has that consequence (Pisoni *et al.*, 1983; Diehl and Walsh, 1989). Both studies focus their challenge on a conclusion that Miller and Liberman do not in fact draw explicitly, but that Pisoni *et al.* consider to be implied in their interpretation and that Diehl and Walsh consider to be entailed by it.[2] It is that normalization for speaking rate is a speech-specific process using mechanisms and/or processes special to speech.

To test the conclusion, both sets of investigators obtained bipolar classifications ("abrupt onset"–"gradual onset") of members of nonspeech acoustic continua designed to be similar to the /ba/–/wa/ syllables of Miller and Liberman. In the research by Pisoni *et al.*, sine waves traced the center frequencies of the three synthetic speech formants of the /ba/–/wa/ continua. Listeners' classifications along the sine wave continua replicated the findings of Miller and Liberman (1979). Increasing the steady-state duration of the

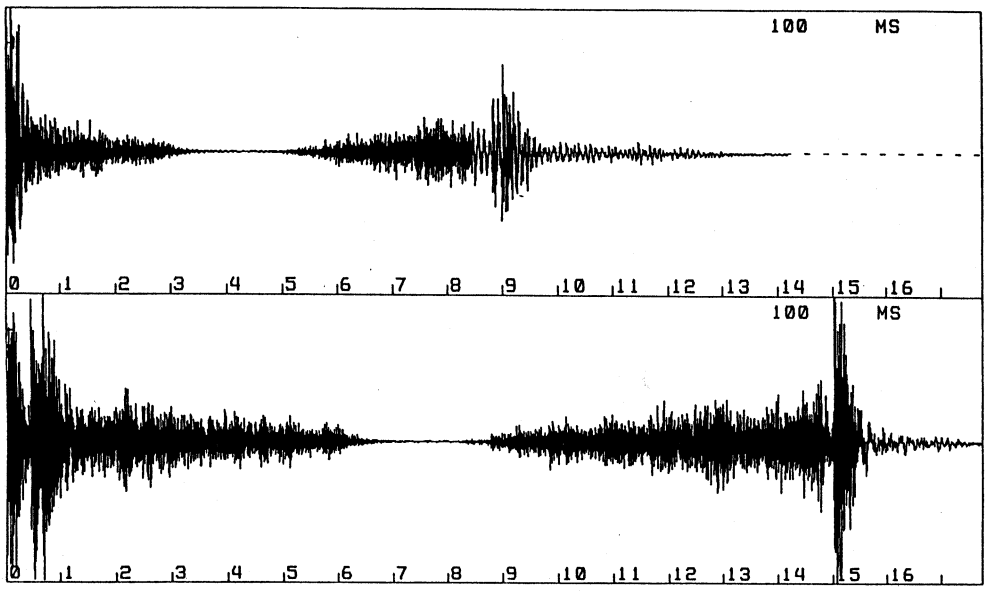

FIG. 4. Phase 2 sounds of the steel ball rolling up then down the steel-surfaced portion of the ramp after having first rolled down the 10-deg (top) and 50-deg (bottom) sandpaper-covered portions of the ramps. Divisions mark 100-ms intervals.

ment 1, one phase 2 sound from each of the two bent tracks was selected, input to the computer, and stored. Table I gives the durations of the sounds used in the tests.

II. EXPERIMENT 1

In the first experiment, two groups of listeners were presented with sounds from one or the other sets of tracks. As shown in Table I, each sound consisted of one of the phase 1 sounds followed by the phase 2 sound from either the 50- or the 10-deg "end point" tracks. Listeners were asked to judge the slope of the ramp section of each track. If listeners use acoustic signals as information for their source in the environment, predictions are that responses in the upsloping condition will pattern similarly to those in the research of Miller and Liberman, because a long-duration phase 2 implies a short-duration phase 1; responses in the flat-ramp condition should pattern oppositely.[5] Predictions based on durational contrast[3] are that long durations of phase 2 will promote classifications in the short-duration category regardless of the physical event giving rise to the signal.

A. Method

1. Subjects

Subjects were 21 students at Dartmouth College, who received course credit for their participation. They reported normal hearing. Data from one subject were eliminated from the experiment when he reported having confused the response labels during the first part of the test.

TABLE I. Durations (ms) of phase 1 and phase 2 acoustic signals used in experiment 1. "Up" refers to the upsloping track and "flat" to the flat track.

	50	40	30	20	10
Phase 1	320	358	385	472	547
Phase 2 (up)	1776				1421
Phase 2 (flat)	524				904

2. Materials

Four continua were constructed by splicing different phase 2 sounds onto the five phase 1 sounds from the different sloping ramps (see Fig. 1). The two continua for one group of ten subjects were created by splicing phase 2 sounds from the bent-(upsloping) end point tracks (the 1776-ms sound recorded from the upsloping ramp following a run down a 50-deg slope and the analogous 1421-ms sound recorded from the upsloping track following a run down the 10-deg slope) onto the end of all phase 1 sounds. In this pair of continua, the phase 2 sounds differed in duration, and a long-duration sound in that phase was associated in the original recordings with a short-duration (steep-slope) phase 1 sound. For a second group of subjects, another pair of continua was constructed by splicing a pair of phase 2 sounds from the flat tracks onto all phase 1 sounds. Here, the phase 2 sounds also differed in duration, but now the longer-duration phase 2 sound (904 ms) has been associated in the originally recorded events with the shallowest slope in phase 1 and the shorter sound (524 ms) with the steepest slope.

Test orders were devised in which each of the ten sounds (five from each of two continua) occurred ten times each in random order with the constraint that each continuum member occur once in each successive group of ten trials. There were 3 s between trials and 6 s after each 20 trials.

3. Procedure

Subjects were tested in groups of one to three. They were shown the steel ball and the end point tracks, but were not given a live demonstration of the steel ball rolling along the tracks. Next subjects listened to the recorded sounds of the ball rolling down the 50- and 10-deg sandpaper-covered ramps (that is, just the phase 1 end-point sounds). They

heard the pair of sounds in alternation five times. They were told that the first sound of each pair was that of the steel ball rolling down the steep ramp, and the second was the sound of the steel ball rolling down the shallow ramp. Their task would be to classify the ramp as steep or shallow on each trial by writing down an identifying pair of letters on their answer sheet (writing "ST" for the steeper ramp and "FL" for the shallower or flatter ramp).

Listeners were told that they would not just hear the phase 1 sound, but they would also hear the ball rolling along the whole track. Subjects in the group of ten subjects who listened to sounds from the upsloping tracks were told that they would hear the sound of the ball rolling down the sandpaper-covered ramp, and then rolling up the steel-surfaced track until it turned and rolled back down to the bend where it was stopped. Subjects, in the group of ten subjects, who listened to the sounds from the flat tracks were told that they would hear the sound of the ball rolling down the sandpaper-covered ramp, and then along the steel-surfaced track until it rolled off the end. However, the task assigned to subjects in both groups was to classify the slope of the sandpaper-covered ramp as steep or shallow. The experimenter did not explain how the second phase of the event might help them to classify the first phase sounds; nothing was said about the differences in duration of the phase 2 events as a function of the slope of the sandpaper-covered ramp. However, subjects were reminded that the ball's behavior on the second part of the track would depend on its history. They were also told to focus their attention on the first-phase sound, and to choose between the classifications "steep" or "shallow," guessing if necessary.

B. Results and discussion

Figure 5 displays the outcomes for the two groups of subjects. Displayed are the percentages of "steep" responses as a function of ramp slope separately for the continua with the longer and shorter phase 2 sounds. Figure 5 (top) presents the results using the upsloping track and Fig. 5 (bottom) those with the flat track.

Response frequencies slope upward only weakly as a function of ramp slope. The sandpaper-covered surface of the phase 1 ramps lowered the intensity of the rolling sound considerably; the manipulation was intended, in part, to encourage attention to the more intense phase 2 sound. Evidently the manipulation was stronger than optimal. Even so, subjects were able to make the steep/shallow distinction. Trend tests on each of the four continua reveal significant linear increases in percentage of "steep" responses with phase 1 slope in all cases [upsloping track, long phase 2 sound: $F(1,9) = 7.91$, $p = 0.02$; short phase 2 sound: $F(1,9) = 6.74$, $p = 0.03$; flat track, long-duration phase 2 sound: $F[1,9] = 8.15$, $p = 0.02$; short: $F(1,9) = 7.31$, $p = 0.02$].

An analysis of variance with factors group (unsloping or flat ramp) and duration (long, short) of the phase 2 sound was performed on the total number of "steep" judgements summed across continuum members. This yielded no main effects (both Fs < 1), but a highly significant interaction [$F(1,18) = 23.61$, $p = 0.001$]. All four simple effects

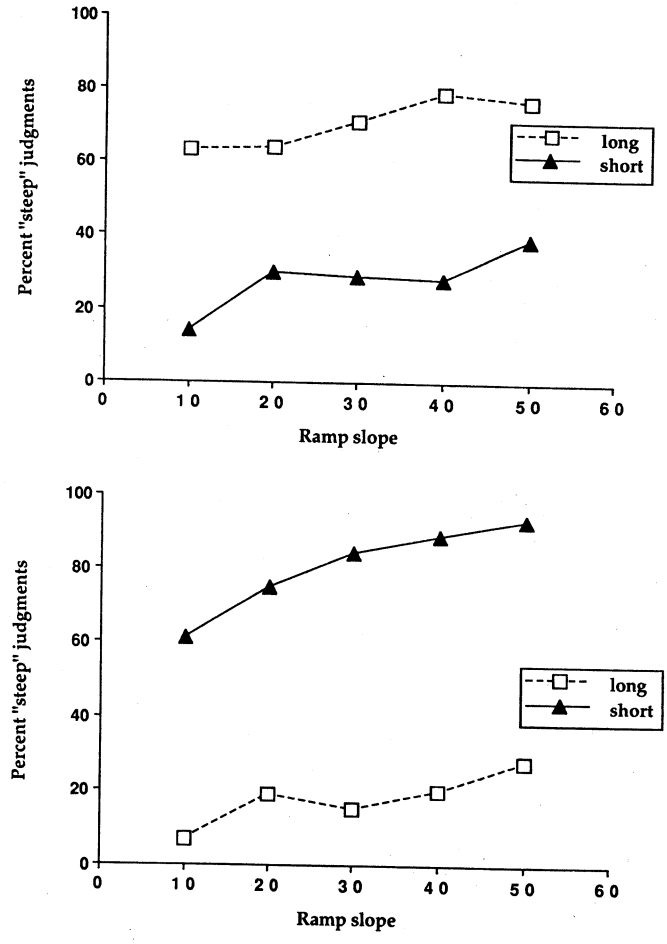

FIG. 5. Results of experiment 1. (top) Percent "steep" judgments as a function of the sandpaper-covered ramp's slope for phase 1 sounds followed by the long (50 deg) and short (10 deg) duration phase 2 sounds from the upsloping ramps (bottom) Percent "steep" judgments as a function of the sandpaper-covered ramp's slope for phase 1 sounds followed by the long (10 deg) and short (50 deg) duration phase 2 sounds from the flat tracks.

are significant; most importantly, the effect of duration is significant for each group [upsloping track: $F(1,18) = 7.66$, $p = 0.01$; flat track: $F(1,18) = 16.84$, $p = 0.001$]. That is, both groups show a significant shift in their probabilities of labeling a ramp as steep or shallow as a function of the duration of the phase 2 sound, but the shifts are opposite in direction.

Effects of duration of the phase 2 sound for the subjects who heard the upsloping track events are in the same direction as effects of vowel duration in the research by Miller and Liberman (1979); Pisoni *et al.* (1983); and Diehl and Walsh (1989); and of steady-state sine wave duration in research by Pisoni *et al.* and Diehl and Walsh. Were it not for the findings with the flat track, they might be ascribed to durational contrast effects. With the flat track, the effects are opposite to predictions based on durational contrast. However, both outcomes are consistent with the hypothesis that listeners use acoustic information, perhaps including durational information, that arises during the second phase of the sound-producing event, but is caused by the nature of the phase 1 event, as information for the phase 1 event. On the

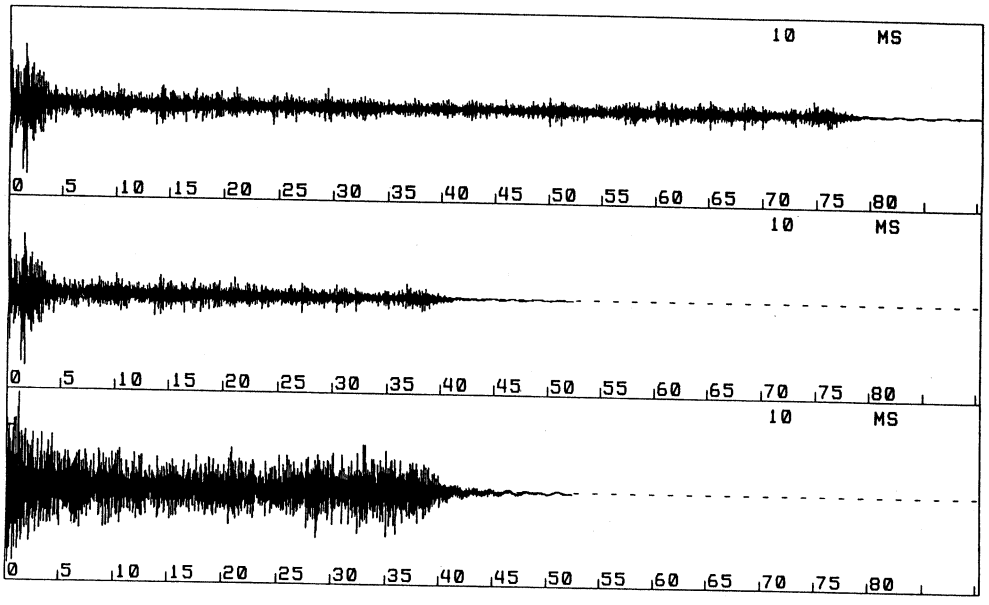

FIG. 6. Phase 2 sounds from the flat-track condition used in experiment 2. (top) From the 10-deg sloping ramp; (middle) from the 10-deg sloping ramp, duration matched to the phase 2 sound from the 50-deg ramp (bottom). Divisions mark 10-ms intervals.

upsloping track, a longer duration second phase is associated with the steeper sandpaper-covered ramp because the steeper ramp imparts more momentum to the steel ball, causing it to travel farther up the steel-surfaced track before turning around and rolling back down. On the flat track, a longer duration second phase is associated with the shallower ramp, because the shallower ramp imparts less momentum to the ball, which then rolls more slowly down the flat track before rolling off the end.

Of course, phase 2 acoustic signals do not differ only in duration. To determine whether subjects who heard flat-track events are using duration in a way opposite to predictions of a durational contrast hypothesis, it is necessary to discover whether they are, in fact, using a long duration of the phase 2 event as information for a shallow phase 1 event. That is the purpose of experiment 2.

III. EXPERIMENT 2

In this experiment, only stimuli from the flat-track condition were used. If, among other sources of information, listeners are using duration of the phase 2 event in this condition to classify the slope of the ramp producing the phase 1 sound, then: (a) phase 2 signals that differ only in duration should still yield differential effects on steepness judgments: (b) the long-duration phase 2 event should lead to more "steep" judgments if it is shortened.

A. Method

1. Subjects

Subjects were the same ten subjects who had participated in the second (flat-track) condition of experiment 1.

2. Materials

One new phase 2 sound was created by modifying the 904-ms phase 2 sound of the steel ball rolling along the flat steel-surfaced track after having run down the 10-deg sandpaper-covered ramp. Small sections of the signal were cut out, bounded by zero crossings, until the duration of the acoustic signal matched that of the ball rolling along the flat track after having run down the 50-deg ramp. Figure 6 displays waveforms of the original 10-deg phase 2 sound, the duration-modified 10-deg phase 2 sound, and the original 50-deg phase 2 sound.

Two new tests were devised. Durations of the stimuli for the tests are presented in Table II. In the first test, phase 1 sounds of experiment 1 were spliced onto (a) the 904-ms phase 2 sound from experiment 1 [Fig. 6 (top)] and (b) the duration-modified phase 2 sound [Fig. 6 (middle)]. These sounds differed in duration but were otherwise very much alike, since the duration-modified sound was created by shortening the 904-ms sound. In the second test, phase 1 sounds were spliced onto (a) the 524-ms phase 2 sound from experiment 1 [Fig. 6 (bottom)] and (b) the duration-modified phase 2 sound [Fig. 6 (middle)]. These stimuli were matched in the duration of their phase 2 sounds, but differed in every other way that a phase 2 sound differs following the steel ball's run down a 50- or a 10-deg ramp. (For example, the slower revolutions of the steel ball can be heard in the 10-deg condition.)

In both tests, each continuum member was presented in

TABLE II. Durations (ms) of phase 1 and 2 acoustic signals in experiment 2.

	50	40	30	20	10
Phase I	320	358	385	472	547
Phase 2 (test 1)					904
					524[a]
Phase 2 (test 2)	524				524[a]

[a] This is the signal depicted in the middle panel of Fig. 6 in which the 904-ms signal from the 10-deg condition has been shortened. This signal is labeled short in Fig. 7, but long-E2 in Fig. 8.

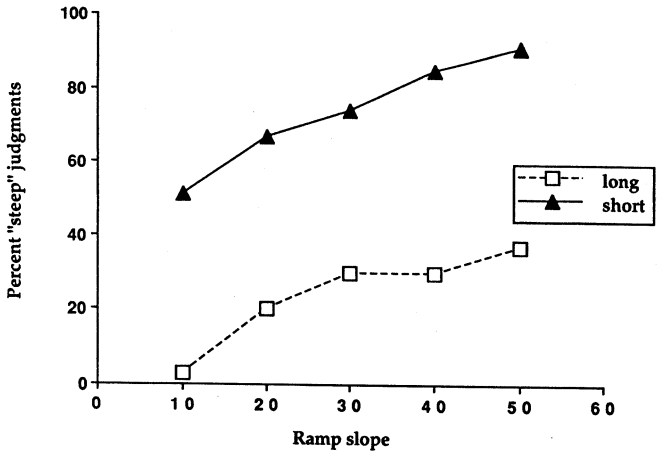

FIG. 7. Results of a test in which phase 2 sounds differ only in duration.

random order ten times subject to the same constraint as in experiment 1. As before, there were 3 s between trials and 6 s after each block of 20.

3. Procedure

Subjects took these tests having taken the test in experiment 1 and after taking a second test in which phase 2 sounds were matched in amplitude (Effects of this manipulation were negligible and will not be described here.) Instructions were as in experiment 1.

B. Results and discussion

Results are displayed in Figs. 7 and 8. Figure 7 shows results from the condition in which phase 2 sounds differed only in duration. That is, both phase 2 sounds derive from the original 10-deg condition, but in the "short" condition, the sound has been shortened. Clearly, there remains a large effect of the phase 2 sounds. A t test comparing the summed

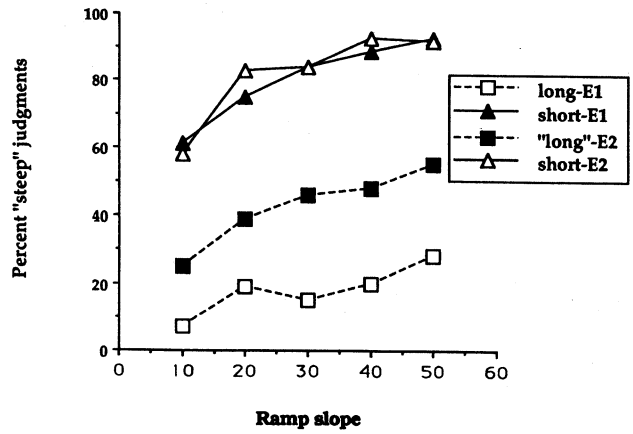

FIG. 8. Comparison of results of the original flat-track test (E1 in the legend) with a condition in which the originally long-duration (10 deg; solid line, open squares) sound has been shortened (dashed line, closed squares) to match the duration of the originally short-duration (50 deg, solid line, triangles) sound. Data represented by solid lines and triangle symbols are the original short-duration sounds, used also in experiment 2.

"steep" responses in the two duration conditions is highly significant [$t(9) = 4.04, p = 0.003$].

Figure 8 superimposes results on the original flat-track test from experiment 1 [also shown in Fig. 5 (bottom)] and results on the second test of the present experiment, in which the duration difference is eliminated. (In the figure, "long" E2 refers to the phase 2 signal depicted in the middle panel of Fig. 6 that was created by shortening the long-duration signal ("long E1" in fig. 8) in the top panel of Fig. 6. The signal was shortened to match the signal in the lowest panel of Fig. 6—the "short-E1" and "short-E2" conditions of Fig. 8). A comparison of results on these tests will show whether the absence of a duration difference between the phase 2 sounds (long-E2 versus short-E2 conditions in the figure) affects the magnitude of the influence of the phase 2 sound on the phase 1 judgments. If listeners were sensitive only to the duration of the phase 2 event, then the points represented by closed squares in Fig. 8 should have been superimposed on points represented by triangles, but Fig. 8 shows that they are not, numerically at least. If listeners are sensitive to other information in the signal, in addition to duration, then points represented by closed squares should have been, and were numerically, shifted downward from points represented by triangles toward those represented by open squares. In an analysis of variance, with factors, test (original, duration-matched), and duration of phase 2 sound (originally long or short), the effect of test was marginal [$F(1,9) = 4.75$, $p = 0.057$], with overall, more "steep" responses in the duration-matched test; the effect of duration was highly significant [$F(1,9) = 48.27$, $p < 0.001$]. Although Fig. 8 shows that the effect of test was almost entirely confined to the phase 2 sounds that were different across the two tests (that is, the 10-deg original and duration-modified sounds), the interaction between conditions was not significant [$F(1,9) = 1.86$, $p = 0.21$]. The effect of the phase 2 sound remains significant, even in the duration-matched condition, presumably because listeners can use other information for phase 1 slope in the phase 2 sounds (including, perhaps, velocity of the revolutions of the steel ball).

Together, the results of the pair of tests show that listeners do use duration of the phase 2 sound to classify the phase 1 slopes as steep or shallow (Fig. 7), although, not surprisingly, they do not use only duration (Fig. 8). Providing only a difference in duration to guide classifications, listeners still show a large effect of the phase 2 sound on phase 1 classifications. Presenting a phase 2 sound that signals a shallow phase 1 ramp, except in its duration, increases "steep" responses, albeit weakly.

There is another effect evident comparing Fig. 7 and 8 that, along with the nonsignificant interaction in Fig. 8, shows that listeners use other information besides duration to make their slope judgments. Responses to the shortened 10-deg phase 2 sounds (middle panel of Fig. 6) are labeled "short" in Fig. 7, because, as compared to the unshortened phase 2 sounded in that test, they are short. The same sounds are labeled long in Fig. 8 (dashed lines), because, in contrast to the other phase 2 sound in that test, they derived from an originally long sound. Across the two tests, there is a large difference in the average number of steep responses assigned

to that duration-modified sound. Presumably, the reasons for the difference derive from the fact that, in that phase 2 sound, duration is set in opposition to other information for phase 1 slope, rendering the sound ambiguous. As contrasted with the other phase 2 sound in the test, in which phase 2 sounds differ only in duration (Fig. 7), the duration-modified phase 2 sound is more characteristic of a steep-sloped ramp; as compared to the other phase 2 sound in the duration-matched test, it is more characteristic of a shallow-sloped ramp. Although listeners are not instructed to distribute their responses evenly between the response categories, on average, they may be unwilling to distribute their responses very unevenly. If any phase 2 sound is to count as information for the steep-sloped ramp, it must be the shortened sound in the duration-only test, but the 50-deg phase 2 sound in the duration-matched test.

IV. EXPERIMENT 3

Results of the first two experiments show that listeners are influenced by duration of phase 2 of a biphasic event in their classifications of phase 1; however, influences are different for different events. Where a steep phase 1 caused a short-duration phase 2, short duration phase 2s were associated with increases in "steep" responses; where a steep phase 1 caused a long-duration phase 2, that signal was associated with increased "steep" responses. Experiment 2 showed that duration was, indeed, an effective variable, perceptually, in the condition in which response patterns were opposite to predictions based on durational contrast. These findings show that duration *per se* of phase 2 does not affect classification responses; rather, the information that duration provides about its origin in a sound-producing event affects response patterns.

A final question addressed by this series of experiments is, whether, by the logic of the experimental comparisons of sine wave perception with speech perception, perception of rolling steel balls can be shown to be special. In the present experiment, listeners classified, as "long" or "short," the first phase of biphasic sine wave signals designed as analogs to the sounds generated on the flat tracks of experiment 1. By the logic of the research under examination here, if listeners' classifications of the sine waves have the same pattern as the steep and shallow judgements of the real steel ball sounds, then we should infer that mechanisms and processes supporting perception of the rolling sounds are those supporting perception of analogous nonspeech sounds. If the responses pattern differently, then rolling steel ball perception is supported by specialized mechanisms and processes.

Since the steel ball signals do not have narrow spectral peaks, as speech syllables do, sine wave versions of the signals cannot track any peaks. They can be analogous in some other ways, however, and, in particular in respect to the acoustic property under examination here, duration. Diehl and Walsh used syllables analogs consisting of a single sine wave tracking an inverted $F1$. These signals were analogous to the speech signals to which they were compared only in having a transitional phase and a steady state with durations matching the corresponding phases of the syllable to which it was compared. Here, I used a two-phase signal with a lower amplitude, lower frequency first phase and a higher amplitude, higher frequency second phase, with durations matching those of the flat-track sounds of experiment 1. The amplitude and frequency differences were meant to mimic corresponding differences in the flat-track signals along the sand-paper covered and metal portions of the tracks.

Sine wave analogs were made of the flat-track sounds of experiment 1, rather than the sounds from the upsloping tracks, because results in this condition patterned oppositely to a durational contrast effect. In sine wave analogs of these sounds, most likely information for causal effects of phase 1 on phase 2 (that is, for effects of rolling down a steep or shallow ramp on the subsequent rolling event on the flat track), is absent. If so, then there is no reason to expect effects of phase 2 events on phase 1 judgements, excepting durational contrast effects, if any. Thus, by the logic of the speech/nonspeech comparisons in the literature, an outcome is expected that will support an inference that rolling steel-ball perception is special.

A. Method

1. Subjects

Subjects were 11 students at Dartmouth College who participated for course credit. They were native speakers of English who reported normal hearing.

2. Materials

Stimuli consisted of two sine waves presented sequentially. First phase sounds were low-amplitude 500-Hz tones having durations of the five phase 1 sounds of experiments 1 and 2. Second phase tones were higher amplitude (ratio 1:9) 1000-Hz tones having the durations of the phase 2 sounds from the flat-track condition of experiment 1. The amplitude difference was used to mimic the abrupt increase in amplitude that accompanied shifting from the sandpaper-covered to the bare steel sections of the track. The change in frequency was meant to imitate a pitch difference between the sound of metal against sandpaper during phase 1 versus metal against metal during phase 2.

A test tape was made up using the ten stimuli (five phase 1 tones $\times$ two phase 2 tones), each presented ten times in random order. The test order was identical to that for the flat-track stimuli of experiment 1, with each sine wave analog replacing its corresponding rolling sound in the test sequence.

3. Procedure

The subjects first heard the isolated longest and shortest phase 1 tones in alternating order five times to familiarize them with the long and short sounds they would be judging. Next, they were told that they would hear those tones each followed by a louder, higher pitched tone. Their task, however, was to classify the first, lower pitched, sound in each trial as long or short, guessing if necessary.

B. Results and discussion

Data from 1 of the 11 subjects were eliminated, because here responses were random across the continua. Figure 9

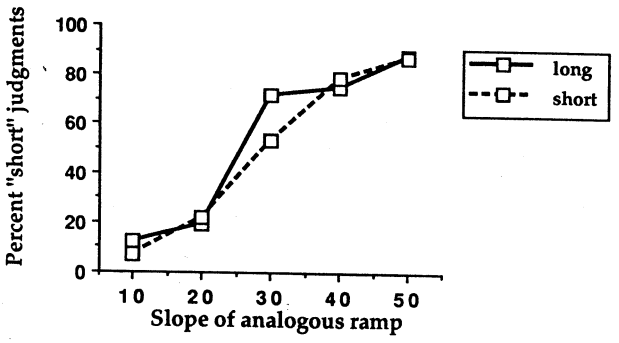

FIG. 9. Results of experiment 3. Percent "short" judgments as a function of the duration of phase 1 and phase 2 tones. Phase 1 durations vary from long to short on the abscissa. Solid triangles: short-duration phase 2; open squares: long-duration phase 2.

shows the results for the remaining ten subjects. Shifts in response percentages across the continua are considerably steeper than in experiments 1 and 2. Trend tests across the continua verified that there was a significant linear increase in percentage of "short" responses as the phase 1 sound shortened [$F(1,9) = 89.39, p < 0.001$, for the short phase 2 condition; $F(1,9) = 65.69, p < 0.0001$, for the, long phase 2 condition].

In an analysis of variance with factors, duration of phase 1 tone, and duration of phase 2 tone, the overall effect of phase 2 was marginal [$F(1,9) = 4.00$, $p = 0.077$], with overall more "short" judgments in the context of a long phase 2 tone. However, the interaction of phase 1 and phase 2 duration was significant [$F(4,36) = 6.00$, $p = 0.0008$], because, as Fig. 9 shows, the curves separate at just one, or possibly, two points. *Post hoc* tests showed a significant difference between the long and short phase 2 conditions in the third position in the continuum [$F(1,9) = 15.55$, $p = 0.003$], and a marginal difference in the longest phase 1 condition [$F(1,9) = 5.00, p = 0.052$].

To the extent that the duration of the phase 2 tone affected judgments of phase 1 durations at all, the difference was in the direction of a durational contrast effect. This is opposite to the effect of phase 2 durations in the analogous steel ball (flat-track) condition of experiment 1. There, long duration phase 2 sounds were associated with a decrease in "steep" (short) judgments. Results on summed steep or short judgments were compared across the two experiments in an analysis of variance, with factors experiment, and duration of phase 2 sound. In the analysis, the interaction of experiment and phase 2 duration was highly significant [$F(1,18) = 75.64, p < 0.0001$], because the effect of phase 2 duration was, indeed, opposite in the two experiments. Is perception of acoustic consequences of steel balls rolling along partially sandpaper-covered tracks special?

V. GENERAL DISCUSSION

Findings in the experiments favor a view that, in perceiving nonspeech sounds, listeners use structure in the acoustic signals as information for its casual source. In two similar biphasic nonspeech events, listeners used the second phases of the sounds as information for the slope of a ramp causing the first phase sound structure, presumably because attributes of the phase 2 sound were also causally affected by the ramp's slope. However, in one event, a steep slope caused a short-duration first phase and a long-duration second phase, while, in the other event, a steep slope caused short durations of both phases 1 and 2. In the first instance, results were similar, superficially, to results obtained by Miller and Liberman, by Pisoni *et al.* and by Diehl and Walsh. All of those outcomes are consistent with a hypothesis that, rather than recovering physical event properties, listeners are subject to durational contrast. In the second instance, however, results are the reverse of an outcome based on durational contrast. Across all of the experiments where the hypothesis can be evaluated unambiguously, results are consistent with the hypothesis that, perception based on acoustic signals (like perception based on optic or haptic simulation) is environment directed; listeners use structure in an acoustic signal (and in an optic or haptic array) as information for its causal source. Accordingly, their responses on perceptual tests reflect perceived source properties; similarities or differences in response patterns to different signals do not reveal similarities and differences in auditory-system properties applied to the signals. I argued earlier that findings in the remaining comparisons (including that in experiment 3 above!) are uninterpretable because influences on response systematicities by those of an apparent distal source cannot be determined. In my opinion, those comparisons should be avoided.[6]

One question to ask, then, is how are we to discover whether perception of some kinds of signals involves a specialization? It seems to me that the best evidence for identifying specialized *mechanisms* is neuropsychological. Findings on neurologically damaged individuals provides convincing evidence that there is a neurological specialization for speech. Given information from that domain, certain behavioral tests—for example, dichotic listening tests—have provided corroborative information on intact listeners.

However, I do not think that we have reliable procedures for identifying specializations based only on behavioral measures in the absence of prior hints from neuroscience. This is because we lack a convincing theory of what that is special about specializations will show up on behavioral tests involving noninvasive experimental procedures. Liberman and Mattingly (1989) propose that specializations of the brain called "modules"—including, for example, systems dealing with stereopsis, sound localization, and speech—are special in yielding special perceptual objects. Nonspecialized sensory systems yield "homomorphic" percepts—percepts, that is, whose dimensions mirror those of stimulation at the sense organ. In contrast, specializations yield "heteromorphic" percepts, whose dimensions are different from those in proximal stimulation. For example, the system dealing with stereopsis takes offset images from the two eyes and, rather than yielding an experience of offset images, yields perceived depth. The sound-localization system takes in time-of-arrival differences at the two ears and yields an experience, not of time of arrival differences, but of location in the environment of an event. The speech system

takes in formant transitions and, rather than yielding an experience of pitch glides, yields consonants.

I think that the equation of homomorphic percepts and nonspecialized processes, and heteromorphic percepts and specialized processes is mistaken, however. Every example of heteromorphy above is, at the same time, an example of a percept that is "homomorphic" with respect to some environmental property—relative distance (stereopsis), location of an event (sound localization), and consonant production (speech perception). If the arguments presented in the Introduction of the present manuscript are correct, and, if the findings of the experiments are general, successful perception is always homomorphic with respect to the environment, because perceptual systems, whether specialized to a narrow domain or general to many, have evolved to recover environmental properties from stimulation at the sense organs. Accordingly, from this perspective, research that attempts to test for a specialization for speech perception by distinguishing speech and nonspeech perception in respect to their perceptual objects is misguided, as is an approach that rejects a specialization for speech on grounds that perceptual objects are of the same sort in both domains.

If the nature of the perceptual objects yielded by a perceptual system does not index a specialization, what should the criteria be for identifying a specialization from behavioral data? To my knowledge, none has been worked out. Until they have been, perhaps decisions concerning specializations should be left to neuroscience.

A related important question concerns whether (and, if so, how) behavioral measures can be obtained that do shed light on covert auditory–perceptual mechanisms and processes, whether or not those mechanisms and processes involve specializations. Perhaps they can; however, if so, they must be obtained using research especially designed to induce subjects to provide public measures of private processes.[7]

In the early days of cognitive psychology, when investigation of covert processes was once again licensed following the waning influence of behaviorism, a major research goal was to find procedures that would provide public manifestations of covert processes (see, e.g., Neisser, 1967, for a review of the early work). Researchers were aware that simply providing a visible letter or letters, for example, and asking for a response would generally yield no more than a close reflection of the distal input.

One way in which information-processing psychologists hypothesized that perceptual processes can be exposed is by interfering with processing and observing behavioral consequences. To take an important sample technique in this field, consider backward masking. Early ideas about masking were that the mask could be used to halt perceptual processing of a target at various times after stimulus onset. If so, masking reveals that perception has a time course, and it provides information on the attributes of target stimuli that perceptual processes work on earlier and later.

A second procedure is to make predictions about behavior that (a) are based on known perceptual-system properties and (b) are uniquely predictions based on those properties. An example, among others, is Sakitt's work (e.g., 1976) on ionic memory relating its properties to known properties of the rod system.

I do not refer to these research domains because they have provided accurate information about perceptual processes. Indeed, research in both domains has proven highly controversial. I refer to them only to point out that, in the domain of investigating covert speech and nonspeech perceptual processes, as in the domain of visual-information processing, researchers need to devise special tools that allow them to observe traces of perceptual processes in behavior. I do not see anything in the research designs used by Pisoni *et al.* or Diehl and Walsh promoting behavioral responses that expose perceptual processes and mechanisms. There are no manipulations, such as masking, to interfere with perceptual processes. Likewise, predictions are not based on known properties of auditory processes or mechanisms, and are not unique to such properties as are hypothesized. (Predictions are not, in particular, based on information that the auditory system exhibits durational contrast to stimuli such as those used in their research; see footnote 3.)

One might argue that, because sine waves have no real distal source of a kind we can perceive, response patterns to the sine waves can only reflect auditory system processes; accordingly, response patterns to the sine waves can be used to pick out patterns in responses to speech signals that have the same origin. This is not a safe conclusion, however. Sometimes sine wave signals are heard *as* speech; sometimes they are heard *as* musical tones. Heard over a loudspeaker, they always sound as if they come from some location in the environment. Can we know that it is even possible to hear them as nothing at all, except an acoustic signal, despite their being localized as if they were caused by an event in the world? How, indeed, could the auditory system decide, in a given instance, whether to recover a distal event from a given acoustic signal, or else, to perceive the signal instead? It could not make the decision based on whether or not the signal fully specifies some environmental source, because the presence of noise in the environment and of signals arriving at the ear from long distances away must guarantee that signals with real sources in the environment do not always fully specify them. There is, I believe, the strong likelihood that response patterns to sine wave signals in classification tasks reflect the listeners' best guesses as to the nature of an environmental source of the signals.

From the foregoing perspective, the best designs in the literature comparing speech to sine waves are the Best designs (Best *et al.*, 1981, 1989) that examine perception of the same acoustic signal, heard either as speech or as nonspeech. Since there is no difference in the signals, different response patterns depending on attentional set must, it seems, be ascribed to the perceiver. This procedure has obtained different response patterns to signals, depending on whether they are heard as speech or as nonspeech. However, I do not think that an inference from the findings to a specialization for speech perception is much safer than the opposite inferences based on the designs of Pisoni *et al.*, and Diehl and Walsh. The stimuli in these experiments have to be ambiguous—perhaps as the Gestalt ratman drawing is ambiguous. In turn, this may mean that different subsets of the acoustic

structure, or more likely, in these experiments, different organizations of the same signal survive the attentional filter. These different subsets or organizations may signal distal events that differ in respect to the properties under test, and these different apparent-distal-event properties may underlie the different response patterns. There is, after all, really a pictured main in the rat-man display, and there is really a depicted rat.

To summarize, while I do not think that it is impossible to learn something about perceptual systems from behavioral data, I do not see anything in the designs of most of the research in the literature comparing speech to sine waves that actively promotes obtaining observable perceptual-system intrusions on response patterns. Listeners base their responses on their perceptual experiences, and successful perceptual experiences are of the environment. To get information as well about perceptual mechanisms or processes must require special experimental designs.[8]

[1]The investigation of perception has two aspects. One is to understand what happens from the time that a sense organ is stimulated until a percept is achieved. This is the study of covert or private aspects of perception. The second is to understand the character of events in the environment, the extent to which they causally structure media that can stimulate sense organs and the extent to which that structure is used by perceivers to recover event properties. This is the study of public aspects of perception.

[2]According to Pisoni *et al*: "Although not stated explicitly, Miller and Liberman implied that this form of perceptual normalization for speaking rate was unique to processing of speech signals and the mechanisms used in phonetic categorization" (p. 314).

According to Diehl and Walsh: "Miller and Liberman (1979) interpreted the effect as follows: a longer vowel is evidence of a slower rate of articulation, and to compensate perceptually (i.e., to normalize for rate), listeners accept a greater range of transition durations as corresponding to the stop category.

This account of the stop/glide boundary shift is speech-specific in that the normalization process it invokes depends on listeners' sensitivity to variation in articulatory rate."

For reasons that may become clear, I think that these inferences are mistaken. Listeners may normalize for speech rate because the acoustic signal provides information encouraging normalization—just as the optic array to the eye does when perceivers are able to identify, for example, various acts of walking or dancing that occur at various rates. Moreover, to the extent that rate affects utterances and acoustic signals as it does some other acoustic-signal producing event and its acoustic signals, rate normalization for speech need not be special to speech.

[3]Tests of durational contrast cannot be designed with any confidence. While the literature on contrast effects in general is large, that on durational contrast is small (but see, e.g., Behar and Bevan, 1961; Walker *et al.*, 1981). As far as I have been able to determine, it does not establish whether or not durational contrast works in a right-to-left direction in the durational range of the ba/wa stimuli to which Diehl and Walsh apply it as an explanation. In fact, Diehl and Walsh (1989) and Kluender *et al.* (1988) propose durational contrast as an account of their findings as well as for a popular durational patterning across languages (that is, long vowels before voiced as compared to voiceless obstruents) without citing a single experiment showing durational contrast for acoustic signals and without testing for it directly (that is, without obtaining duration judgments from their listeners).

[4]My original intent was to devise one type of event physically very similar to the /ba/–/wa/ speech events and another different one. Given other constraints of the research design and practical constraints on what I could implement, this did not come to pass. In the speech syllables, durational information for rate and for the stop–glide distinction combine in a particular way that I did not succeed in mimicking. In particular, rate is a global variable that affects duration globally—that is, everywhere in the syllable. In a CV, a stop or glide affects duration locally and converges with rate effects syllable initially. If listeners do normalize for rate, they do so by using information for rate in the vowel, where (in the synthetic syllables of Miller and Liberman anyway), it is unambiguous, to scale for rate in the consonantal portion of the syllable where it is confounded with variability due to consonant manner. Residual or relative duration in the consonant then serves as information for the stop or glide. In the events used experiments 1 and 2 below, there is no convergence of two independent sources of durational variation, one global and one local to phase 1 only. The only similarity to Miller and Liberman's stimuli is the abstract one that, if listeners recover physical event properties from the acoustic signals and use duration in phase 2 as information for the character of phase 1, in one set of stimuli the nature of the physical events is such that a long duration phase 2 provides information for a short-duration phase 1; in the other set of stimuli, the events are such that the relationship is reversed.

[5]These predictions are clear based on the physical characters of the ramp events. However, they raise an issue that concerned me for a while and that a reviewer of this manuscript raised. In speech, a fast rate of talking shortens everything so that a short vowel (phase 2 of a CV) is associated with a short consonant phase 1). In this sense, the CV syllables are analogous to the flat ramp condition, not the upsloping condition; yet the predictions are that responses in the flat track condition will pattern oppositely to those in the speech syllables. The reason for this is clear if the differences among the physical events are considered. In the speech syllables, effects of rate are global; effects of the stop/glide manner are local and converge locally with effects of rate. If listeners truly normalize for rate, then they must do so by using durational information for rate in the vowel to scale for durational effects of rate in the consonant where it converges with effects of stop manner. Listeners judge, not rate, but the rate-scaled local variable, consonant manner. When a vowel is long, they ascribe more of the transition's duration to rate and less to consonant manner, increasing "b" judgments.

Neither ramp event has this kind of structure in which two variables, one local and one global, have converging influences on duration during phase 1. Only one variable affects duration of phase 1 and that is ramp slope, the judged variable. It happens that ramp slope also affects duration in phase 2 so that duration in phase 2 can be used as information about ramp slope. But the information in phase 2 is for ramp slope, the judged variable, not for another variable that ramp slope is scaled with respect to.

All that is common to the upsloping ramp events and the speech events is that, if listeners recover the physical sound producing events from the consequent acoustic signals, a long-duration phase 2 will increase the frequency of classifications in the short-duration phase 1 category.

[6]This is not to say that use of sinewave signals should be avoided in speech research. Research by Remez and his colleagues (e.g., Remez *et al.*, 1981) has been quite illuminating in showing what information in acoustic signals is crucial to recovery of phonetic properties of speech. It is only to argue that sinewaves or any signal with no definite distal source provide unrepresentative and uninterpretable instances of "nonspeech" perception.

[7]A caveat: My own interest are in public aspects of perceiving; accordingly, this discussion may not be as informed as one that someone else might be able to provide. My reason for providing speculation at all is only that, in the literature under examination here, the issue is not addressed explicitly. I think that my research indicates that it needs addressing by those investigators interested in studying perceptual mechanisms and processes based on behavioral measures.

[8]There is one more issue to address. Elsewhere, I have espoused a theory of speech perception as direct (e.g., Fowler, 1986); I still do. However, the theory is controversial (see, e.g., the commentary following Fowler, 1986), and I would not like the arguments of the present manuscript to be rejected on grounds that readers consider direct realism implausible. In fact, the arguments I raise in the present manuscript do not derive from claims that are special to the direct-realistic theory.

A direct realist theory (e.g., Shaw and Bransford, 1977) holds that perception is unmediated in both of two respects in which competing theories generally hold that it is mediated. To illustrate, I will describe the direct-realist theory in the domain in which it has been developed most extensively, namely, visual perception (e.g., Gibson, 1979). In some theoretical points of view—an information-processing view for example—perceivers are held to use stimulation at the eye to construct a mental representation which is the basis for perceptual experience. In one sense of mediation, then, perception of the environment is mediated by the representation. A direct-realist theory holds that world itself is perceived, not a mental representation of it. In a second sense of mediation, a direct-realist theory denies the perception involves hypothesis testing, inference making, or other processes that accomplish anything other than direct recovery of environment

properties from information in stimulation. Here, the claim is that, for any property of the environment that is visually *perceived* (rather than inferred or guessed at), information in reflected light specifies that property.

In the field of visual perception, theorists who reject direct realism do so on grounds that, in their view, perception is—and must be—mediated. Importantly, however, there is no disagreement that perceptual objects are environmental. Theorists may disagree as to whether the perceived world is the same as the real world, because failures of correspondence will be an occasional cost of mediation. But, however error-prone a mental representation may be held to be, it is still considered a representation of the environment, not of the light. For obvious reasons, a view that visual perceptual objects are reflected light structures has no proponents. Accordingly, a view that visible objects are environmental is not a special claim of a direct realist theory; the special claims of that theory have to do with how the perceptual objects are recovered.

In the field of speech perception, a claim that listeners hear phonetic gestures of the vocal tract is only accidentally associated in particular with a direct-realist theory (indeed, it is held as well by motor theorists for whom perception is mediated by processes of analysis-by-synthesis). As in visual perception, the "direct-realism" in the direct-realist theory of speech perception derives from its claim that recovery of events from stimulation is unmediated. Accordingly, just as the view is widely held in visual perception that perception of the environment is, instead, mediated, so could the view be held in auditory perception that perception of acoustic-signal-causing events occurs, but, in contrast to the claims I make as a direct realist, perception is mediated. From such a perspective, the arguments that I have raised in this manuscript based on the idea that perceptual objects are necessarily environmental are still valid.

Behar, I., and Bevan, W. (**1961**). "The perceived duration of auditory and visual intervals: Crossmodal comparison and interaction," Am. J. Psychol. **74**, 17–26.

Best, C., Morrongiello, B., and Robson, R. (**1981**). "Perceptual equivalence of acoustic cues in speech and nonspeech signals," Percep. Psychophys. **29**, 191–211.

Best, C., Studdert-Kennedy, M., Manuel, S., and Rubin-Spitz, J. (**1989**). "Discovering phonetic coherence in acoustic patterns, Percep. Psychophys. **45**, 237–250.

Diehl, R., and Kluender, K. (**1989**). "On the objects of speech perception," Ecolog. Psychol. **1**, 121–144.

Diehl, R., and Walsh, M. (**1989**). "An auditory basis for the stimulus-length effect in the perception of stops and glides," J. Acoust. Soc. Am. **85**, 2154–2164.

Eimas, P. (**1985**). "The equivalence of cues in the perception of speech by infants," Infant Behav. Dev. **8**, 125–138.

Fowler, C. (**1986**) "An event approach to the study of speech perception from a direct realist perspective," J. Phon. **14**, 3–28.

Fowler, C. (**1989**). "Real objects of perception," Ecolog. Psychol. **1**, 145–160.

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

Gibson, J. J. (**1979**). *The Ecological Approach to Visual Perception* (Houghton-Mifflin, Boston).

Kluender, K., Diehl, R., and Wright, B. (**1988**). "Vowel length differences before voiced and voiceless consonants: An auditory explanation," J. Phon. **16**, 153–169.

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

Liberman, A., Harris, K., Kinney, J., and Lane, H. (**1961**). "The discrimination of relative onset-time of the components of certain speech and nonspeech patterns, J. Exp. Psychol. **61**, 379–388.

Liberman, A., and Mattingly, I. (**1985**). "The motor theory revised," Congnition **21**, 1–36.

Liberman, A., and Mattingly, I. (**1989**) "A specialization for speech perception," Science **243**, 489–494.

Mann, V., and Liberman, A. (**1983**). "Some differences between phonetic and auditory modes of perception," Cognition **14**, 211–235.

Miller, J., and Eimas, P. (**1983**). "Studies on the categorization of speech by infants," Cognition **13**, 135–165.

Miller, J., and Liberman, A. (**1979**). "Some effects of later-occurring information on the perception of stop consonant and semivowel," Percep. Psychophys. **25**, 457–465.

Neisser, U. (**1967**). *Cognitive Psychology* (Appleton Century Crofts, New York).

Pisoni, D., Carrell, T., and Gans, S. (**1983**). "Perception of the duration of rapid spectrum changes in speech and nonspeech signals," Percep. Psychophys. **34**, 314–322.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.

Sakitt, B. (**1976**). "Iconic memory," Psychol. Rev. **83**, 257–276.

Shaw, R., and Bransford, J. (**1977**). "Introduction: Psychological approaches to the problem of knowledge," in *Perceiving Acting and Knowing: Toward an Ecological Psychology*, edited by R. Shaw and J. Bransford (Erlbaum, Hillsdale, NJ) pp. 1–39.

Wagner, M., and Baird, J. (**1981**). "A quantitative analysis of sequential effects with numerical stimuli," Percep. Psychophys. **29**, 359–365.

Walker, J., Irion, A., and Gordon, D. (**1981**). "Simple and contingent aftereffects of perceived duration in vision and audition," Percept. Psychophys. **29**, 475–486.

Warren, R. (**1981**). "Measurement of sensory intensity," Behav. Brain Sci. **4**, 175–189.