889

Psychological Review
1994, Vol. 101, No. 1, 129-156

0033-295X/94/$3.00

On the Perceptual Organization of Speech

Robert E. Remez, Philip E. Rubin, Stefanie M. Berns, Jennifer S. Pardo, and Jessica M. Lang

A general account of auditory perceptual organization has developed in the past 2 decades. It relies on primitive devices akin to the Gestalt principles of organization to assign sensory elements to probable groupings and invokes secondary schematic processes to confirm or to repair the possible organization. Although this conceptualization is intended to apply universally, the variety and arrangement of acoustic constituents of speech violate Gestalt principles at numerous junctures, cohering perceptually, nonetheless. The authors report 3 experiments on organization in phonetic perception, using sine wave synthesis to evade the Gestalt rules and the schematic processes alike. These findings falsify a general auditory account, showing that phonetic perceptual organization is achieved by specific sensitivity to the acoustic modulations characteristic of speech signals.

How does the listener perceive properties of objects from sounds that strike the ears? The answer to this fundamental question about auditory perception is commonly split in two. One familiar part pertains to auditory perceptual analysis in which the specific sensory details of an incident waveform provide information about a particular object that produces sound. Recent instances of this approach include studies of the perception of shape and surface texture (through echolocation, e.g., Habersetzer & Vogler, 1983; Schmidt, 1988); of the range, velocity, direction, and elevation of sound sources (Butler, Levy, & Neff, 1980; Riquimaroux, Gaioni, & Suga, 1991); of the distinctive timbres of musical instruments (Grey & Gordon, 1978); of the perceptual differentiation of individuals by their hand-clapping sounds (Repp, 1987) or walking sounds (Li, Logan, & Pastore, 1991); of the auditory discrimination of everyday mechanical events (Warren & Verbrugge, 1984); and of the recognition of the linguistic attributes of utterances (Liberman & Studdert-Kennedy, 1978; Stevens & Blumstein, 1981). Despite the great variety of object properties and sensory elements that concern these instances of auditory perceptual analysis, they exploit a common albeit implicit premise. Namely, each instance presumes that the listener sorts the acoustic elements available at the ear into separate streams, each relevant to the perception of attributes of a particular object or event. With this presumption, the first part of the answer rests tacitly on a second, for perceptual analysis requires the achievement of auditory perceptual organization. By virtue of the function of perceptual organization, the sensory details of an incident waveform are resolved into coherent streams, each specific to its source in the world.

Current explanations of auditory perceptual organization customarily allude to the perils of conversation at a cocktail party (Cherry, 1953). Listening in such hazardous circumstances, a perceiver must isolate the speech signal of a conversational partner from an acoustic background consisting of other voices and of various nonvocal party sounds. This disentangling of concurrent acoustic streams, clearly necessary if a listener is to perceive a talker's utterance, is the exemplary case of auditory perceptual organization. Recent reports on the details of perceptual organization include work on auditory models of the segregation of simultaneous and successive acoustic components (Weintraub, 1987), perceptual experiments on the separation of speech signals from extraneous sounds (Darwin & Gardner, 1987) and of the disentangling of simultaneous talkers (Brokx & Nooteboom, 1982; Summerfield & Assmann, 1989), studies of the formation of groups of musical sounds (Deutsch, 1982; Jones, 1976; Jones & Boltz, 1989; C. Palmer & Krumhansl, 1987), and psychophysical studies of the formation of perceptual streams from spectrally stationary and transient acoustic elements (Bregman, 1978b, 1981, 1987, 1990; McAdams, 1989). Among the themes that emerge from this research is the reliance on principles of organization deriving from Gestalt psychology (Koffka, 1935; von Ehrenfels, 1890; Wertheimer, 1923/1938).

There now exists an abundance of reports about the grouping of displays of simple acoustic elements. A growing collection of findings seeks to extend some of these principles to less arbitrary sound patterns, including speech signals (Bregman, 1990; Cole & Scott, 1973; Dorman, Cutting, & Raphael, 1975). This development is both necessary and welcome, for no explanation of speech perception provides the account of perceptual organization that it implicitly invokes. Conversely, the definitive test of any account of auditory organization is provided by natural

Robert E. Remez and Jennifer S. Pardo, Department of Psychology, Barnard College; Philip E. Rubin, Haskins Laboratories, New Haven, Connecticut; Stefanie M. Berns, Department of Neuropsychology, Queens College of the City University of New York; Jessica M. Lang, Department of Psychology, Columbia University.

This research was supported by National Institutes of Health Grant DC00308 to Robert E. Remez and Grant HD01994 to Haskins Laboratories.

To the individuals who encouraged our work gently, who argued with us avidly about its interpretation, or who supplied corrections to earlier versions of this article unsparingly, we offer our great thanks: Peter Bailey, Albert Bregman, Lynn Cooper, Peter Eimas, Carol Fowler, Julian Hochberg, Mari Riess Jones, Carol Krumhansl, Alvin Liberman, Ignatius Mattingly, Lynne Nygaard, David Pisoni, and Bruno Repp.

Correspondence concerning this article should be addressed to Robert E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, New York 10027-6598. Electronic mail may be sent to remez@paradise.barnard.columbia.edu.

cases, yet few areas of psychoacoustic research are sufficiently articulated theoretically and empirically, with the exception of speech, to permit a test of this kind. However justified this commingling of enterprises is, we doubt that the future is bright for an account of perceptual organization of speech based on Gestalt principles. Our misgivings derive from a review of the perceptual evidence and an examination of the acoustic characteristics of speech signals. Thus motivated, we performed perceptual tests of the axioms of a general auditory account, the results of which are reported here. These findings oppose an exclusive or primary role for the Gestalt principles of form in the perceptual organization of speech signals. Moreover, our tests have shown that the organizational resources supplementary to the Gestalt principles are unlikely to include schematic representations of general knowledge, despite the claim of the most prominent general auditory view. Instead, it seems that the perceptual organization of speech depends on sensitivity to time-varying acoustic patterns specific to phonologically governed vocal sources of sound. In presenting the case here, we consider the main approach to auditory perceptual organization, the acoustic and perceptual concerns that mitigate its potential, and our new findings that put its Gestalt-based views to a strong test.

Organization of Visual and Auditory Forms

By depicting a few simple, characteristic cases, Wertheimer (1923/1938) first exposed the principles of perceptual organization, thereby rationalizing his observation that elements of stimulation were perceived in unmistakable alignment, combination, and separation. The visual, rhythmic, and pitch impressions that he described differed in orderly fashion from the pictures and acoustic patterns driving perception, and no explanation by appeal to experience was readily available for many of the phenomena in his report. By means of elegant drawings now known to every student of introductory psychology, he identified the organizing principles of *proximity, similarity, common fate, set, continuity, symmetry,* and *closure.* After including a single concession to experience in a proof of the action of *habit,* he offered the principles as a collection from which to draw in explaining the fact that perception is organized and is not a piecemeal summation of sensory elements.

Explicit auditory instances of organizational principles were again offered by Julesz and Hirsh (1972), who sought a common rubric for visual and auditory modalities. Although their essay concluded that the dissimilarities of vision and hearing outweighed the shared attributes, their view of auditory perceptual organization brought an influential information-processing rationale to subsequent studies. These authors were circumspect in contending that the Gestalt principles alone might be inadequate to explain perceptual integrity of speech signals because of the acoustic complexity of speech or because of the potential influence of the listener's linguistic knowledge. They mentioned two problems:

> The harmonics of the voice are equally separated on a linear scale of frequency, but certain groups of them get reinforced by resonant properties of the vocal tract. These "formants" do not stand out as separate figures but turn out instead to be the bases for identifying the spoken vowels. (p. 300)

Also, "whether the structural features in spoken sound patterns show this [perceptual coherence] by virtue of properties of the stimulus configuration or of the language habits of the listeners is not clear" (p. 305).

In the 20 years since the appearance of the article by Julesz and Hirsh (1972), a large body of evidence has been gathered about its hypotheses. Although the extension of Gestalt organizational principles to speech has occurred infrequently, many studies of the organization of elementary auditory forms have produced evidence of grouping tendencies of the kind described therein. We draw from this empirically rich base to characterize the current view of auditory perceptual organization and to gauge its suitability for speech.

Gestalt Principles in Auditory Organization

Using rapid, repetitive presentation to exaggerate auditory grouping, Bregman and colleagues have elaborated many of the speculations of Julesz and Hirsh (1972) empirically. This collection of findings demonstrates the dimensions and parameters of the perceptual disposition to form groups. For example, a principle of proximity, here set in the frequency domain, was offered to explain the formation of groups observed by Bregman and Campbell (1971). From a repeating sequence of six 100-ms tones with frequencies of 2.5 kHz, 2 kHz, 550 Hz, 1.6 kHz, 450 Hz, and 350 Hz, listeners perceived a pattern forming two concurrent groups: one of the three low tones (550 Hz, 450 Hz, and 350 Hz) and another of the three high tones (2.5 kHz, 2 kHz, and 1.6 kHz). The independence of the auditory groups was determined from the relative ease of identifying the order of elements within a high- or low-pitched stream, in contrast with the relative difficulty of identifying the intercalation order of the series of elements (high, high, low, high, low, low). In other tests, variation in the duration of the elements and the rate of repetition of the sequence affected the proximity along the dimension of frequency at which grouping occurred (Bregman, 1978a; Handel, Weaver, & Lawson, 1983; Miller & Heise, 1950; Tougas & Bregman, 1985; Van Noorden, 1975), although the formation of groups within narrow frequency ranges persisted. Generally, the faster the rate of presentation, the smaller the frequency difference between sets of elements heard as separate tone groups.

The principle of similarity also applies to the formation of auditory groups. The evidence comes from studies (Bregman & Doehring, 1984; Steiger & Bregman, 1981) in which simple harmonic relations or similar frequency excursions among a set of tones promoted the formation of perceptual groups. In an extension of these studies, grouping of dichotically presented tones was observed when harmonic relations occurred among them (Steiger & Bregman, 1982). Again, rapidly repeating tones were grouped because they shared a common fundamental frequency, and even small departures from this harmonic relation blocked dichotic fusion. Last, spectral similarity appears to play a significant role in perceptual organization, as observed by Dannenbring and Bregman (1976). The strongest instance of grouping that they observed across parametric variations of temporal and spectral factors was the segregation of periodic from band-limited noise elements, indicating that separate noise groups and tone groups had formed on several similarity

criteria (cf. Ciocca & Bregman, 1987; Warren, Obusek, Farmer, & Warren, 1969).

Likewise, the principle of common fate was translated into the auditory domain by Bregman and Pinker (1978) in a test of grouping by relative onsets of tone elements. When brief (147 ms) tones were synchronously onset and offset, they were grouped together; when tones were offset by 58 ms or more, they were grouped into separate perceptual streams. Similar effects were noted by Dannenbring and Bregman (1978), who tested group formation with three-tone complexes varying in the onset or offset lead or lag of the mid-frequency component. Brief tone pulses (137 ms) were set to the frequencies of 500 Hz, 1 kHz, and 2 kHz, and asynchronies as brief as 35 ms were effective in disrupting auditory groups of harmonically related tones, in cases both of onset lead and offset lag. The fusion of elements that onset together had been noted by Hirsh (1959), reporting a critical asynchrony of 20 ms required to distinguish periodic and aperiodic targets temporally (see also Rosen & Howell, 1987). In other observations of the principle of common fate, Bregman, Abramson, Doehring, and Darwin (1985), Bregman, Levitan, and Liao (1990), and McAdams (1989) noted effects of amplitude modulation or frequency modulation on the grouping of tones; the modulation frequency and phase shared by simultaneous components defined the groups (see Hall & Grose, 1990; Hall, Haggard, & Fernandes, 1984; McFadden, 1986; Yost, Sheft, & Opie, 1989). These examples approximate common fate: An acoustic transformation imposed on different elements induced a perceptual group.

Organization by set has also been reported, in cases of musical experience, by Jones (1976; see also Jones & Boltz, 1989), although in some instances of this kind, perceptual organization also shows evidence of a symmetry principle. In one respect, this principle of form may stand apart from the others we are considering, for melodic sets occur readily without repetitive presentation; here, a prior portion of a musical display appears to induce an implicit expectation about the latter portion of a musical display, in the dimensions both of melodic pitch and of the temporal attributes of a melody. Appropriately, such devices noted by Jones are explained by resort to the allocation of attention, in this case to the hierarchical metrical and harmonic relations evident among melodic elements, in contrast with the distinctly sensory and perceptual emphasis found in the exposition of the other Gestalt principles (however, see French-St. George & Bregman, 1989).

Both principles of continuity and closure apply to the grouping of simple tone glides reported by Bregman and Dannenbring (1973, 1977). Group formation here occurred for tones that were continuous in their frequency contours despite interruptions by silence (up to 20 ms) or by noise bursts (up to 500 ms; see also Miller & Licklider, 1950; Sasaki, 1980; Vicario, 1960; Warren, 1984). Continuity and closure also operate in the amplitude domain, in which the more abrupt the amplitude rise time, the likelier the occurrence of grouping (Dannenbring, 1976).

Overall, the evidence is strong that repetitive presentation of a variety of ambiguous acoustic displays promotes dissociation of the elements into segregated perceptual streams, delineating a tendency for perception to depart from the details of physical stimulation. The principles demonstrated by Wertheimer (1923/1938) have proven remarkably useful in guiding the elaboration of auditory cases. Moreover, in addition to raising the technical issues pertinent to characterizing the auditory modality, research on the primitive principles of auditory organization, like the visual studies that served as precedents, also framed basic questions about the perception of objects and events. With the proposal that perceptual organization parses the acoustic field, achieving an "auditory scene analysis" (Bregman & Pinker, 1978; cf. Bregman, 1990), these processes were held to indicate distal sources of sound and not simply to identify proximal sound streams. In this extended conceptualization, the consequences of perceptual organization are not restricted to a sensory domain of auditory patterns. Although basic processes initiate grouping on extremely fine-grain auditory criteria, each stream that results is attributed perceptually to a different source of sound in the domain of objects.

Proving this account is difficult. An emphasis on exposing the perceptual phenomena of grouping has produced an ample collection of individual instances, although no formalization or unified theory of auditory organization exists (nor of visual organization: Hochberg, 1974, appraises the prospects). Consequently, a test of the current hypotheses about Gestalt factors in auditory perceptual organization must follow the empirical precedents closely, or test the axioms of the theoretical conceptualization, given the absence of a computational generalization of the devices of grouping to explore parametrically. Critically, there are few studies that examine perceptual organization of known sources of sound, hence little to establish the plausibility of the singular claim that perceptual coherence in the domain of objects rests on auditory forms created according to Wertheimer's (1923/1938) brief list. It must be conceded, too, that the emphasis placed on Gestalt principles in contemporary auditory accounts hardly forecloses the possibility that collateral (or alternative) resources no less fundamental than auditory grouping principles promote organization of some classes of sound sources. Despite these reservations, the remarkable consistency of the findings of research on auditory grouping permits a test of the approach to organization based on Gestalt principles. In evaluating the adequacy of this view for speech, we have used the empirical precedents as an implicit theory of the coherence of auditory elements; if the technical details of the test have happened to be subtle ones, the nature of the critical evidence is utterly obvious: To accommodate the case of speech, the theory of auditory scene analysis must identify vocal sources of sound no less readily than it groups the patterns made by audio-frequency oscillators and noise generators.

Perceptual Organization of Speech

In asking whether this general account of auditory perceptual organization is adequate for speech, we emphasize four points. First, we specify principles of coherence consistent with the Gestalt-based auditory view. In this step, we raise a crucial consideration that is often overlooked in the general account and designate the properties definitive of a single perceptual stream according to the Gestalt principles. Next, we consider whether this rendition of perceptual coherence presents a useful acoustic description of the speech signal, chiefly on theoretical grounds. Then, we review perceptual evidence suggesting that primitive

auditory mechanisms impose no more than weak constraints on the grouping of elements in phonetic perception. Last, we consider a specific model of organization, auditory scene analysis (Bregman, 1990), which uses a schematic component to reconcile the simplicity of the primitive analysis with the complexity of the speech signal.

Clearly, the standards for success in accounting for speech differ substantially from those applicable to the treatment of auditory forms. When an abstract acoustic display owes its structure solely to an experimenter's imagination, it is hardly relevant to ask whether the perceptual grouping of elements that ensues is true to the properties of objects and events: There are none to perceive. In contrast, acoustic displays of speech commonly derive from a particular talker producing a particular utterance, and we can test whether attributes of the displayed speech are apparent to the perceiver (see Pisoni, 1987). The perceiver's experience of speakers and vocal events counts as a test of the principles derived from ideal objectless acoustic episodes, and a general auditory account is justified only to the extent that its principles agree with the actual perceptual organization of the speech signal.

Auditory Coherence

To express the view of perceptual coherence implicit in the general auditory account, we need only paraphrase our preceding section on auditory organization, directing attention to the properties internal to streams rather than the properties that differentiate them (Remez, 1987). The derivation answers the question, What makes the perceiver hear a single source of sound? rather than, What makes the perceiver separate one source of sound from another? Roughly, acoustic elements that are proximate in frequency are attributed to the same source; the rate at which the components succeed each other influences their cohesiveness: The slower the procession, the greater the disposition to cohere. Acoustic elements must also be similar in frequency changes to be grouped together, not only in the extent of change but also in the temporal properties.

A more subtle form of similarity also promotes grouping of simultaneous components, namely, when harmonic relations occur among them. Similarity in spectrum also appears to warrant cohesion of components, such that aperiodic (noisy) acoustic elements are grouped together and periodic elements are grouped together. Acoustic elements that occur concurrently must exhibit common onsets and offsets and must show common (frequency or amplitude) modulation to coalesce. Continuity of changes in frequency or in spectrum is also required for elements to form a single perceptual group. In general, small temporal or spectral departures from these various similarities, continuities, common modulations, and proximities result in the loss of coherence and the splitting of auditory elements into separate groups. If the ordinary perceptual cohesion of the elements in a speech signal is attributable to domain-independent auditory principles of organization, we should find no violations of these criteria in the acoustic properties of speech.

The Speech Stream

Acoustic characterizations of speech are plentiful. To abstract these for the moment, note that the spectrum of a speech signal exhibits prominent features: multiple broadband resonances that vary in center frequency over time, harmonic structure, and frequent interruptions of this structure by silence and aperiodic constituents (Fant, 1962; Stevens & Blumstein, 1981; Zue & Schwartz, 1980). Grossly, the spectrum of a speech signal is nonstationary, or time varying, as the spectrogram of the sentence "Why lie when you know I'm your lawyer?" shown in Figure 1 attests. Some typical acoustic attributes to note in Figure 1a are the continuity of the lowest frequency resonance (the first formant) despite intermittent energy (with gaps exceeding 75 ms) in resonances of higher frequency (the second, third, and nasal formants), the marked dissimilarity in frequency trajectory of the first formant and those of higher frequency formants, and the lack of temporal coincidence of large changes in resonant frequencies. Nonetheless, the resonances are excited in common by a pulse train produced by the larynx. This imposes harmonic relations and common amplitude modulation across the formants (although the formant center frequencies are not

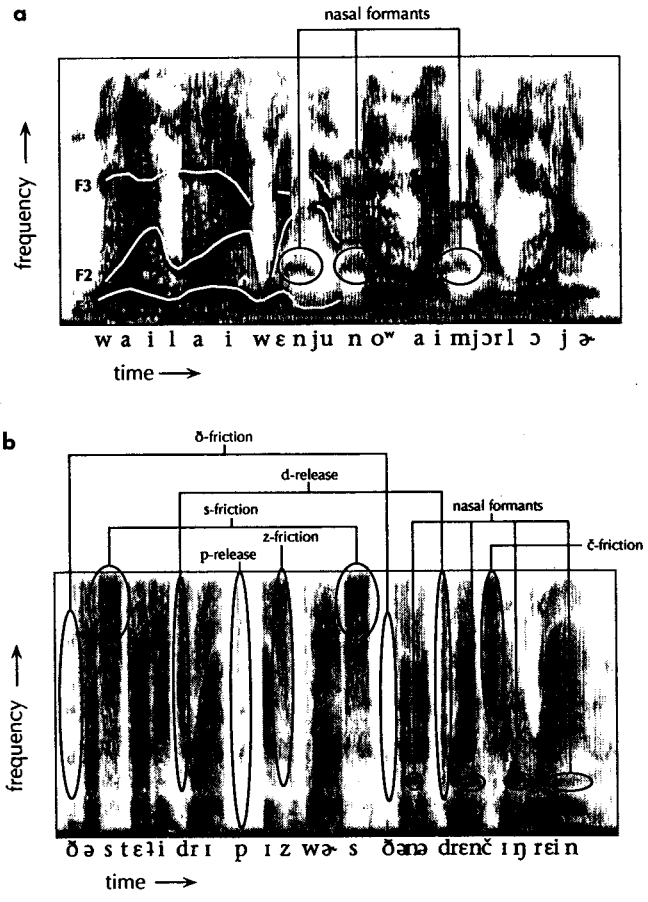

Figure 1. A spectrogram of the acoustic energy, by time and frequency, of two sentences: (a) "Why lie when you know I'm your lawyer?" consisting primarily of liquid and nasal consonants and vowels; three lowest frequency formant patterns are marked in the first half of the sentence, and nasal formants are circled. (b) "The steady drip is worse than a drenching rain," comprising more varied acoustic elements: silences, friction, broadband pops, and both rapidly and slowly changing formant configurations. (See text for explanation.)

harmonically related[1]). The attributes of harmonicity and common modulation may offer the only basis for grouping the phonated formants as a single coherent stream because no other Gestalt criterion (continuity, proximity, symmetry, or similarity) is appropriate (Bregman, 1990; Bregman et al., 1985). In the absence of common pulsation and harmonic structure, we should find this signal fracturing into multiple perceptually segregated streams when primitive auditory principles are applied: the first formant forming a single continuous stream; the second formant splitting from the first as an intermittent stream with highly varying frequency; and the nasal and third formants segregating from the others, each varying rather less in frequency than the second formant.

The acoustic composition of the sentence depicted in Figure 1b, "The steady drip is worse than a drenching rain," is considerably more varied. Although the spectrum exhibits many of the characteristics of voiced speech illustrated in Figure 1a—including asynchronous frequency variation among the formants, which the general auditory account holds to be destructive of grouping—it also exhibits numerous intermittent episodes of silence and noise and four brief appearances of a nasal formant. By criteria of spectral and frequency similarity, the sentence should split into 10 streams: 3 that correspond to the three oral formants; a fourth, which is composed of the four intermittent and discontinuous occurrences of nasal resonance (in *thaN, dreNchiNG, and raiN*); a fifth, which is composed of the noise associated with the voiced fricatives (in *THe* and *THan*); a sixth, which is composed of the noise manifest by the two unvoiced fricatives (in *Steady* and *worSe*); a seventh, which is composed of the noise at the release of the affricate (in *drenCHing*); an eighth, which is composed of two consonant release bursts (in *Drip* and *Drenching*); a ninth, which contains the pulsed noise of the voiced fricative (in the word *iS*); and a tenth, which is composed of the consonantal release (in *driP*). If the principle of common fate is applied to portions of the spectrum that are comodulated by the pulsing of the larynx, then the oral formants, the nasal formants, and perhaps the voiced fricatives are grouped as 1 stream, leaving 4 remaining aperiodic streams associated respectively with the acoustic correlates of unvoiced fricatives, affricates, and consonant releases. Neither continuity, nor closure, nor symmetry, nor any obvious principle of "goodness" can accomplish the reduction of this spectrum to the single vocal source that the listener perceives. These commonplace examples suggest that the principles of the primitive auditory analysis fall short of explaining the perceptual coherence of a single speech stream.

Were spoken communication to consist solely of phonated or whispered synchronously changing oral resonances, the Gestalt principles would succeed easily.[2] This acoustic dictate is satisfied by sequences of vowels or vowellike sounds, and some partial models of auditory organization fare well when applied to vowels alone (Assmann & Summerfield, 1990; Parsons, 1976; see McAdams, 1989, for a related example). However, the irreducible acoustic diversity of the two speech samples of Figure 1, to say nothing of the acoustic–phonetic diversity across languages, is sufficient to nourish our skepticism about the plausibility of the conventional auditory primitives. It is difficult to imagine that the elements of a speech signal would exhibit their self-evident coherence if the Gestalt-based auditory principles were the sole resource promoting organization. The crucial considerations that we have raised hypothetically have occasionally been expressed in empirical studies, and a review of the evidence offers a few surprises.

Perceptual Phenomena

Some perceptual experiments with speech have hinted at the inadequacy of an account of organization based solely on primitive auditory analysis deriving from Gestalt principles. (Experiments with visual patterns have also hinted at the inadequacy of the Gestalt rules for perceptual organization in the visual modality [e.g., Rock & Brosgole, 1964].) To make the test precise, these studies have often cataloged the perceptual phenomena that result when a speech signal has been altered acoustically. Given the clear precedents of grouping by temporal, spectral, and comodulatory similarity, this research has found that listeners report multiple sources of spoken sound when the acoustic conditions favor the fission of elements into separate streams, much as the auditory account warrants. Oddly, these perceptual streams, which should be segregated from each other according to the account given by primitive auditory analysis, are combined nonetheless to produce phonetic impressions. This literature supports a claim that is devastating to the present accounts of domain-independent auditory analysis: Namely, whatever mechanism is responsible for Gestalt-based grouping may be irrelevant to the disposition to integrate acoustic elements phonetically (Liberman & Mattingly, 1989; Mattingly & Liberman, 1990). We consider the evidence, with special emphasis on a recent remodeling of the auditory account that aims to protect primitive auditory analysis from falsification by studies of phonetic perception.

In one of the first studies of the perceptual organization of speech, Broadbent and Ladefoged (1957) investigated the effect of common pulsing on the formant bands of a sentence. They made a two-formant synthetic replica of a sentence but presented the first formant to one ear and the second formant to the other. Despite the different location of the two signals, a rather obvious violation of spatial similarity, listeners heard a single voice when the formants were excited at the same fundamental frequency. When two different fundamentals were used, listeners reported hearing two voices, as if fusion was lost; this occurred even when the differently pulsed formants were both presented to the same ear. Viewed in hindsight as a test of audi-

[1] The parameters of a vocal resonance are defined by Fant (1956). The center frequency and amplitude of a formant correspond to the peak of the function drawn to enclose the harmonics. The formant bandwidth is the half-power (−3 dB from the peak amplitude) frequency difference between the upper and lower skirts of the function.

[2] One may imagine an easy success of primitive auditory analysis in identifying a vocal source were speech to consist solely of fricative trains or of nasal murmurs. However, such a language would have little to exploit phonologically save reduplication of a few basic acoustic ingredients. Not surprisingly, this is the only device that an account based on the perceptual segregation of repetitive acoustic elements permits. Even so, the acoustic elements of an imaginary language consisting solely of labial, alveolar, and velar release bursts might lack sufficient similarity to be grouped as a single stream when produced in series (cf. Dorman, Studdert-Kennedy, & Raphael, 1977; Stevens & Blumstein, 1981).

tory organization, the findings seem wholly consistent with the Gestalt-based approach elaborated in subsequent research, except for a single curious fact. When each formant had a different fundamental, this created an impression of two voices saying the same utterance, rather than two unspeechlike buzzes varying in pitch. Listeners evidently combined the information from each resonance to form phonetic impressions despite the concurrent perception that each resonance issued from a different vocal source. In other words, the differently excited resonances were phonetically coherent, although they also were split into two separate perceptual streams in the listener's impression of the auditory scene.

Other experiments on the perception of consonant properties in brief utterances have likewise found that a common fundamental is not a prerequisite for the phonetic coherence of formant bands (Cutting, 1976; Darwin, 1981; Gardner, Gaskill, & Darwin, 1989; Repp, Milburn, & Ashkenas, 1983). Whalen and Liberman (1987) reported that even a tone with an average frequency no less than four octaves above the fundamental frequency of phonation was integrated with a formant complex corresponding to a syllable, despite the spectral dissimilarity of tone and resonant components. In these cases, subjects reported both the segregation of sound sources and the fusion of the acoustic components sufficient for the perception of subtle phonetic properties caused by their combination.

If the primitive auditory principles make the right prediction about the fusion of formants only when they share pulsed excitation, it is reasonable to ask whether harmonic properties of the fine acoustic grain ever promote auditory coherence. In this regard, Darwin and Gardner (1986) investigated the effect of mistuning a single harmonic coinciding with the peak of the first formant in a steady-state synthetic vowel. When a harmonic of the fundamental frequency in the region of 500 Hz was mistuned by 8%, this had the perceptual effect of segregating it from the rest of the vowel spectrum. Although the short-term spectrum envelope retained a formant center-frequency characteristic of /ɛ/, the consequence of stream segregation was to alter the identity of the vowel, which became more /ɪ/-like. This acoustic display was heard as two sources: (a) The mistuned harmonic was heard as a single tone, and (b) the formant complex was heard as a vowel. In contrast with observations by Broadbent and Ladefoged (1957), Cutting (1976), Repp et al. (1983), and Whalen and Liberman (1987), a listener in the study by Darwin and Gardner did not combine segregated streams to produce phonetic impressions. The streams that formed along similarity criteria remained segregated, as the general auditory account warrants. In this instance, the harmonic coherence of the components was definitive of the perceptual streams, with the coherence of the vowel formants overlaid on it; it remains to be seen whether such principles operate in conditions more nearly approximate to ordinary speech sounds, in which formant frequency and fundamental frequency are time varying. Related findings also implicate harmonic phase coherence in establishing and maintaining a perceptual stream (A. R. Palmer, Winter, Gardner, & Darwin, 1987; see Bregman et al., 1990, for a nonphonetic parallel[3]).

In addition to studies of harmonic effects on coherence, researchers have implicitly examined temporal and spatial aspects of common fate in the components of a speech signal.[4] For example, Liberman, Isenberg, and Rakerd (1981) performed a study using synthetic speech in which most of the acoustic components of a syllable were presented to one ear, and the remainder, consisting of brief second and third formant frequency transitions, were presented to the other. The listener, who heard the syllable as /sɑ/, /spɑ/, or /stɑ/, integrated several temporally offset components to perceive the phonetic properties. First, the fricative noise, which led the periodic components by 90 ms in one instance and by 190 ms in another, was integrated with the formants. Second, three steady-state formants were integrated with each other. The first formant onset led the second and third by 40 ms, although all were excited at the same frequency. Last, two brief formant frequency transitions were presented to the ear opposite the noise and steady-state formants, timed to offset just as the steady-state formants in the opposite ear onset. Liberman et al. reported that subjects experienced the phonetic fusion of these components, evidenced by the perception of the three-way consonantal contrast that depended on the combination of the dichotic signals. Nevertheless, the effects of the multiple temporal and spatial disjunctions were also observed: The formant transitions that affected the place of articulation

[3] In a test that opposed common amplitude modulation with a harmonic relationship between two tones, Bregman et al. (1990) found that comodulation was a stronger determinant of coherence. Nonetheless, in estimating the relevance of their finding to the coherence of a speech signal, they wrote, "When a human voice is registered in the auditory system, it is probable that the system can detect both a pulsing within each basilar membrane filter and a regular harmonic series. There is reason to believe, given the present experiments, that the auditory system makes use of both kinds of information. It is not appropriate, however, to conclude from these experiments that the pulsation-based mechanism is more important in the grouping of spectral regions than is one based on harmonicity. Although this may have been true in our experiments, in which there were very few partials to establish the harmonic series, it may not be true in the human voice, which is rich in harmonics" (p. 73; cf. Summerfield & Culling, 1992).

[4] In fact, much perceptual research on phonetic contrasts has used acoustic techniques that count on coherence despite the principles of similarity and common fate. For one example, consider the original study on the acoustic cues to consonant voicing, which manipulated the lead or lag of the first formant relative to the second and third (Liberman, Delattre, & Cooper, 1958). On the one hand, extremely voiced consonants were produced by starting the first formant 100 ms in advance of the second and third. To produce impressions of voicelessness, on the other hand, these investigators "cut back," or delayed, the onset of the first formant 50 ms relative to the second and third. For another example, recall the classic observation on the perception of the distinctive place features of fricative consonants (Harris, 1958). This study used a technique that presupposed the cohesion of dissimilar and temporally offset components. The method of this study contrasted the perceptual effect of the center frequencies of the noise bands of natural fricatives with the frequency transitions of the voiced oral formants that followed. The findings revealed many instances in which the fricative place feature was determined perceptually by the periodic oral formant transitions or perceived place relied on the successive albeit discontinuous occurrence of both fricative and oral formant transitions. These perceptual contingencies likewise presuppose coherence of the acoustic signal elements despite their dissimilarity of spectrum and onset characteristics. More extensive surveys of the acoustic correlates of phonetic contrasts can be found in Fant (1962), Pisoni (1978), and Zue and Schwartz (1980).

of the stop consonant, /p/ or /t/, were also heard simultaneously as sources of sound separate from the syllable with which they were self-evidently integrated. These other sources of sound had the auditory quality of chirps (see also Nygaard, 1993; Nygaard & Eimas, 1990).

A comparison of the report by Liberman et al. (1981) with the study of fundamental frequency by Broadbent and Ladefoged (1957) is illuminating. In each case, an acoustic manipulation that should have disrupted coherence did so: The listener heard the display as if it was issued from multiple distal sources of sound. Paradoxically, this organization did not disrupt the combination of perceptual streams to yield phonetic information. This divergence of phonetic and auditory organization seems neither a specific effect of differences in fundamental frequency nor of a particular temporal or spatial dissimilarity. In the views of Mattingly and Liberman (1990) and Bregman (1990), this topic of *duplex perception* is singularly instructive in accounting for phonetic perceptual organization, and we return to it later.

In contrast with the findings that undermine the principle of temporal coincidence in the formation of phonetic streams are several studies on the coherence of individual harmonics in a vowel (Darwin, 1981; Darwin, Pattison, & Gardner, 1989; Darwin & Sutherland, 1984; cf. Gardner & Darwin, 1986). These studies featured an analogous approach to the harmonic mistuning cases and assessed the effect of a leading or lagging single harmonic among the otherwise stationary components of a vowel. Recall that in the case of the mistuned harmonic, this tone was split into a perceptual stream of its own, separate from the vowel impression. Consequently, the center frequency of the first formant changed, thereby altering the perception of a vowel by making an /ɛ/ spectrum effectively more /ɪ/-like. In the case of temporal misalignment of a single harmonic component, the leading or lagging harmonic was segregated from the other harmonics with misalignment as brief as 32 ms, again, changing the impression of the vowel. This is a second case in which a Gestalt-based auditory principle anticipates an outcome of an experiment using speech. Two sources of sound were perceived and remained segregated for phonetic identification. Again, it is uncertain whether grouping by temporal coincidence requires a stationary spectrum. If it does, the role of this principle in the organization of natural speech may truly be marginal.

Overall, except for the case of the harmonic coherence in an isolated synthetic steady-state vowel, it appears that a lack of temporal coincidence of the components of a speech signal does not cause their segregation. Conversely, is simple temporal coincidence sufficient cause to group harmonic components together? Evidence about this complementary case, of comodulation of spectral components without fusion, is mentioned by Liberman and Studdert-Kennedy (1978). They reported a phenomenon in which a synthetic sentence was accompanied by arbitrary, surplus resonance bands that coincided temporally and often in center frequency with the formants of the sentence; all components were pulsed at the same fundamental frequency. To the listener, the synthetic pattern seemed to consist of a sentence against a background of chirps and bleats; neither coincidence in onset and offset, nor similarity of frequency excursions, nor comodulation at a common fundamental were adequate to cause the listener to fuse the pattern into a single stream. In this case, the criteria for fusion established by precedents of primitive auditory analysis were satisfied by the pattern of false and true formants, although the extraneous resonances neither contributed to the perceptual analysis of the phonetic properties of the sentence nor were they perceptually coherent with the vocal source.

A Matter of Method

The perceptual studies that we have reviewed here have shared a sensible assumption: that valid grouping principles must describe the organization of ordinary speech. Some of these tests have examined the perceptual organization of a spoken syllable, others a word or sentence. However, the procedures used in tests of auditory form perception have differed greatly from those in tests of speech. In studies of auditory form, a train of brief repeating tones and noises has typically been used to force or to exaggerate the expression of particular grouping tendencies. In studies of speech, a single utterance, whether natural or synthetic, is commonly used without rapid repetition. This is not surprising because rapid repetition of acoustically identical syllables actually destroys perceptual stability. Phonetic impressions of speech sounds are transformed by repetition, sometimes in ways that are unquestionably weird (Warren & Warren, 1970). The obvious contrast between the deterioration of phonetic identity accompanying the rapid repetitious presentation of speech and the stable organization of auditory forms brought about by rapid repetition again betrays the profound difference between phonetic attributes and auditory forms.

Setting that difference aside for the moment, we must consider whether the fact that speech studies have eschewed the technique of repetitive display is itself responsible for the shortage of convincing evidence of primitive auditory analysis in the perception of speech. Fortunately, a few tests with speech sounds have used the typical method of studies of auditory form, using acoustic displays of rapidly repeating syllables (Cole & Scott, 1973; Dorman et al., 1975; Lackner & Goldstein, 1974). These studies focused on perceptual organization rather than on the deterioration of phonetic properties (however, see Goldstein & Lackner, 1973). Not surprisingly, the adoption of this method with speech signals produced perceptual phenomena that paralleled the instances of abstract auditory form. In such conditions, perceptually segregated streams of auditory elements formed from the acoustic constituents of the speech signal, much as they had with tones and noise, according to primitive criteria of similarity, proximity, common fate, and continuity. In essence, rapid repetition proved to be effective in forcing perceptual segregation of acoustic elements to occur with speech sounds.

A study by Lackner and Goldstein (1974) illustrates this point. Following the stream segregation method of Bregman and Campbell (1971), Lackner and Goldstein asked listeners to transcribe the order of relative succession of four syllables, together lasting 200 ms, repeated continuously for 32 s. One test sequence comprising consonant–vowel and isolated-vowel syllables in alternation ([bi ɑ gi u]) was especially difficult for listeners to identify. Because the isolated-vowel syllables formed one perceptual stream and the consonant–vowel syllables

formed another, their relative temporal properties were perceptually inaccessible and were inaccurately transcribed. Although the result parallels that obtained with intercalated tones of different frequency (Bregman & Campbell, 1971), the theoretical problem triggered by this finding was noted clearly by Lackner and Goldstein, who remarked that the disrupted perception of temporal order in the four-syllable series, brought about by the perceptual segregation of the two concurrent streams, was uncharacteristic of the perception of ordinary utterances. In ordinary listening, the utterance as a whole composes the perceptual stream and does not disintegrate into separate trains of similarly structured syllables (e.g., one of Vs, another of CVs, yet another of CCVs, and so on) whatever rapid repetition evokes.

The finding is useful in two ways. First, it shows that the method in question, of repetitious presentation of a few brief sounds, is effective in producing segregated perceptual streams of acoustic elements whether they otherwise compose a speech signal or an arbitrary auditory form. It therefore becomes an empirical question whether the characterization of grouping derived from studies of tones and noise is descriptively adequate for the formation of streams of vocal acoustic products. This question may not be particularly urgent to answer because the second useful aspect of the finding was that it indicated that the primitive principles of organization opposed the perceptual coherence of a speech signal, creating fractures of continuity that do not occur in ordinary listening. Consequently, the problem of the organization of speech can be recast this way: If primitive auditory processes would split a speech signal into separate streams of acoustically similar elements, then the absence of such fractures under ordinary listening requires explanation. Lackner and Goldstein (1974) suspected that integration at a coarse acoustic grain prevents perceptual segregation of signal elements due to primitive auditory mechanisms.

Primitive Auditory Grouping in Auditory Scene Analysis

One prominent account of auditory perceptual organization has aimed to accommodate the special organizational problems of speech. In exposing an account of auditory scene analysis, Bregman (1990) plainly acknowledged that speech signals are more cohesive than are tone patterns and discussed the acoustic and perceptual considerations relevant to utterances. Obviously, an account of speech invoking domain-independent grouping principles would hold great appeal in its simplicity and autonomy but only if it works. Although it fails on its own, Bregman preserved the Gestalt-derived component as the first stage of auditory scene analysis, claiming that it succeeds some of the time, although its action is not error free. To explain the organization of speech when grouping principles alone cannot be responsible, Bregman hypothesized a supplementary resource, cast in the form of integrative schemas. In auditory scene analysis, a combination of segregation by grouping and integration by schemas allows the primitive principles to err without harm and undoes their mistakes by resort to knowledge of familiar aspects of speech. This dual-process model keeps speech perception within a general auditory perspective, although it stands apart from classic conceptualizations of auditory perceptual organization. In Bregman's account of 1990, the primitive auditory analysis biases the formation of one or another sound stream, although the schemas actually build the streams and impose segregation. In contrast with a classic account exploiting the Gestalt principles alone, the dual-process explanation places the burden on the schematic component to confirm the suggestions of the primitive analysis, or to correct its erroneous groupings, as presumably occurs often in the case of speech. In authorizing this role of knowledge, auditory scene analysis relinquished the claim that the primitive auditory component is adequate to motivate perceptual organization on its own.[5]

In simply ascribing a portion of perceptual organization to Gestalt grouping and reserving the remainder to non-Gestalt processes, auditory scene analysis takes a large step toward descriptive adequacy and appears to permit a general auditory account of organization to accommodate speech. However, consider the manner in which its two components are rationalized. The primitive functions are easy to characterize, as we have done here by referring to research on the manifestations of one or another Gestalt-based grouping rule. In contrast, it is difficult to specify a precise role for the knowledge-based schematic component because of the strategy recommended by Bregman (1990) for distinguishing primitive and schematic effects: Instances in which the grouping rules are correct in assigning elements to streams are to be taken to indicate the action of primitive auditory mechanisms without further proof; instances in which perceptual organization occurs despite violation of grouping rules indicate the action of schemas. It is important to recognize the snare that this premise creates for studying perceptual organization (Repp, 1992). Although perfectly clear, this rule of thumb staunchly (and tendentiously) would reserve a portion of auditory organization for the primitive component in any eventuality, even if an alternative or collateral process was responsible for organization. Bregman's strategy almost immunizes auditory scene analysis against counterevidence, deflecting tests of the necessity of either component of the model. The primitive grouping rules, demoted from assigning streams

[5] The impetus for an auditory model to incorporate phonetic perception comes from the specific challenge of duplex perception studies. In the past decade, Liberman and colleagues (Liberman et al., 1981; Whalen & Liberman, 1987) have noted many instances in which an element of an acoustic display serves to determine the phonetic identity of a consonant while it is also split into a separate stream and perceived as an auditory form. The double role of an acoustic element violated a key assumption of Gestalt-based organization, specifically, the principle of exclusive allocation: An element cannot participate in two organizations at once, as in the Rubin vase, in which case the critical contour does not serve at the same instant as the self-occluding edge of a vase and of a face. Bregman (1990) discussed the principle at length. Its negation, which presently applies within auditory scene analysis, allows a signal element to participate in streams of phonetic elements and auditory forms simultaneously, as duplex perception requires. The schematic component was thereby rationalized, and the demotion of the primitive auditory analysis was assured, by the necessity to allow two organizations of the same inflow at the same moment. It should be clear, though, that this debate over the mechanism of bistability, whether auditory scene analysis admits simultaneous and divergent organizations, is quite independent from our brief against the theory. We argue that its description of phonetic organization is false, whatever occurs concurrently.

to recommending streams, became a permanent part of perceptual organization, whereas the schematic process that ensured its permanence was cast as no more (nor less) than a catchall, deriving its functions ad hoc from the failures of Gestalt-based grouping rules. Falsifications of the primitive component are assigned to the schematic component to handle, with neither principle nor limit.

Nonetheless, we propose to have identified a way to test auditory scene analysis, both the Gestalt-based component and the schematic part responsible for saving perceptual organization from the errors of primitive auditory analysis. Let us set aside the question for a moment whether non-Gestalt perceptual organization must be based in schemas describing typical acoustic manifestations of speech, as Bregman (1990) designated. A more reasonable test of the action of Gestalt versus non-Gestalt functions in perceptual organization would conclude that Gestalt rules are warranted if fusion truly fails to occur when the rules are violated, and non-Gestalt principles are necessary and sufficient when perceptual fusion occurs and all of the Gestalt rules mandate segregation. In the perception of speech, clear counterevidence to the necessity of primitive auditory analysis would consist of the integration and phonetic perception of a signal when no Gestalt principle warrants fusion of its acoustic constituents. This would indicate a role in phonetic perception of organizational resources other than those described by Gestalt grouping rules, a susceptibility exclusive of primitive auditory analysis that is sufficient for phonetic perceptual organization.

Would such a finding warrant abandoning the Gestalt-based grouping rules? Evidence that non-Gestalt resources are sufficient, combined with evidence that Gestalt-based resources are neither necessary nor sufficient, would require a rather different account of perceptual organization than provided in primitive mechanisms or in auditory scene analysis. In contrast with Bregman's (1990) recommendation, it would mean that an error-prone and unnecessary Gestalt means of grouping, if active, contributes redundantly to sufficient non-Gestalt resources. A finding that non-Gestalt contributions to perceptual organization are primitive, rather than schematic, would add further misgivings about auditory scene analysis. Admittedly, it is not always possible to distinguish primitive effects from those occurring later, with access to knowledge, exploiting the perceiver's ability to guess accurately about familiar events. However, Bregman stipulated the psychological origins and actions of schemas, permitting us to determine the plausibility of auditory scene analysis and the evidence relevant to evaluating it (Bregman, 1990, chapters 6 & 7). Some of the key attributes are (a) schemas act after primitive auditory analysis has occurred; (b) information becomes represented schematically through a course of training or learning; (c) schemas must undo the action of the auditory analysis in building perceptual streams; and (d) schemas describe typical properties of the elements of perception, whether consonants, vowels, syllables, or other constituents of more extended utterances. With respect to the perceptual organization of speech, a suitable test of this conceptualization of schemas depends on four questions: (a) Can the organization of speech be deferred until a secondary stage of schema-based organization? (b) Are speech schemas learned? (c) Can subjects be trained to undo the effects of primitive auditory analysis and then to assign the elements of an auditory representation to arbitrary streams? and (d) Does the perceptual organization of speech require an acoustic signal to contain typical physical manifestations of each consonant and vowel?

Can the organization of speech be deferred until a secondary stage of schema-based organization? The answer depends on the time available to inspect an auditory trace in a schema-driven way once a mistaken organization is detected. The evidence is discouraging, for speech signals do not survive long in auditory storage. For instance, Pisoni (1973) observed that little of the sensory trace of a syllable remained after 200 ms, on the basis of which he rationalized the rapid time course of phonetic coding. Other estimates of auditory durability approach 100 ms or less (Cudahy & Leshowitz, 1974; Elliot, 1967; cf. Turvey, 1978, on estimating sensory duration). Schematic processes would accordingly have insubstantial remnants of an auditory trace to exploit were an error of primitive analysis to affect as much as a single syllable. Accordingly, a mechanism responsible for organizing consonants would have to work well within the 200-ms limit, without waiting for auditory analysis to err or forfeit the relevant stimulation. If a parallel arrangement of phonetic and auditory mechanisms is necessitated by the volatility of the auditory trace, this contrasts with the sequential system architecture assumed in auditory scene analysis for primitive analysis and schematic repair. The first definitive property of its non-Gestalt component therefore appears to be implausible.

Are schemas learned? Because of the secondary nature of the hypothetical schemas, Bregman (1990) alleged that the ability to organize stimulation by non-Gestalt means is acquired through experience, in contrast with the primitive auditory mechanisms that are an unlearned part of sensory processes. Schemas accommodate the failure of primitive auditory analysis in his account, and experience trains the perceiver to encode familiar failures in the secondary component of perceptual organization. This prediction is opposed clearly by evidence of non-Gestalt principles of organization observed in phonetic perception of 3- and 4-month-old infants (Eimas & Miller, 1992). These young subjects integrated spatially and spectrally dissimilar signal components to distinguish consonants, although we can be sure that an intricate course of training was not required for them to do so. Therefore, the second premise of a schema-based non-Gestalt component seems to be false.

Is a listener able to resist primitive auditory analysis and to learn an arbitrary stream assignment through familiarity with a class of acoustic signals? Despite the centrality of the claim that non-Gestalt effects in organization are revisions imposed on Gestalt-based assignments, no evidence exists that subjects are able to unparse two streams of tones segregated by a primitive principle, nor to learn to determine the intercalation of tones that are otherwise segregated by primitive grouping, nor to reassign tones to streams prescribed by an arbitrary rule. Such evidence would establish the plausibility of the claim that schemas overrule the organizations suggested by the action of the primitive auditory analysis. In the absence of evidence, the assertion that the non-Gestalt component of auditory scene analysis edits the output of the Gestalt component is unwarranted.

Does a non-Gestalt resource apply knowledge of typical acoustic properties of speech to correct the errors of primitive

auditory analysis? This question is among those considered in the experiments and interpretations reported here. Our aim was to assess perceptual organization when the auditory grouping rules are evaded and when the acoustic manifestations of speech are anything but typical or familiar. In this fashion, we examined both the classic version and the revised account, auditory scene analysis, of Gestalt-based auditory organization.

A Pointed Conclusion

This critical review of the evidence pertinent to the organization of speech allows several generalizations. First and easiest to defend is the claim that an account of auditory perceptual organization by Gestalt principles is also a hypothesis about the nature of perceptual coherence, although the method used to buttress the Gestalt-based account has charted only the loss of coherence in arbitrary acoustic patterns. We have offered a rough assessment of the adequacy of this account by referring to research on auditory form and the perception of speech. Second is the basic observation that speech signals ordinarily are perceptually coherent, although only some portions of the signal satisfy Gestalt criteria for fusion and other portions of perceptually coherent signals clearly satisfy the criteria for segregation. This may be taken as a mild embarrassment for the classic auditory approach, which explains perceptual coherence only remotely in the simultaneous fusion of formant resonances excited by a common pulsing source. Third, and potentially most troubling for a general auditory view, is an acknowledgment that the preponderance of perceptual evidence about the organization of speech signals is simply not accommodated by the principles of auditory grouping as they are presently elaborated, although the incorporation of a schema-based catchall in the account deflects some of the critique until a comprehensive test of the premises of auditory scene analysis is performed. Of course, much of the existing counterevidence was obtained for purposes other than a direct test of the assumptions of a domain-independent auditory account. This circumstance, combined with developments in digital synthesis, led us to perform a series of experiments to raise the theoretical issues directly and to convert them to an empirical setting. Our findings warrant an alternative to the auditory account that we have been considering, for we observed that phonetic organization occurs even when the simple principles of Gestalt form are consistently evaded by synthetic displays, and that the listener's perceptual impressions of such signals cleanly distinguish the perception of sound sources from the perception of phonetic messages. Therefore, this research contributes a double falsification of the premise that auditory scene analysis parses the auditory world into segregated streams of elements issuing from each sound source. In our view, the perceptual phenomena occasioned by the sine-wave replication of speech signals implicate the action of organizational principles well outside the set described by Wertheimer (1923/1938) and its contemporary auditory extrapolation.

Auditory Perceptual Objects and Sine-Wave Utterances

Although a general auditory account of organization has difficulty explaining the coherence of speech signals in all their acoustic diversity, it may be supposed on the evidence reviewed so far that the coherence of the voiced formants within a signal presents a less challenging problem than the whole signal does. Although it is true that the formants change frequency asynchronously and that their levels of energy rise, change, and fall asynchronously, and that formant patterns abound with temporal and spectral discontinuities, these attributes notwithstanding, the vocal resonances in voiced speech are excited by common pulsing.[6] Voiced formants are related by a fundamental frequency common to the harmonics of which formants are composed. To salvage the mechanism of grouping by Gestalt principles, we might adopt a conceptualization that the primitive auditory principles of organization apply to clear cases and that phonated formants are grouped into a single perceptual stream (see Carrell & Opie, 1992). The acoustically heterogeneous nature of the constituents of a speech signal would warrant recourse to superordinate perceptual processes to repair the errors committed by the primitive analysis. In this following stage, the scene analysis would be completed, and the linguistic analysis would be completed by collecting the auditory effects of release bursts, aspirations, fricatives, affricates, and the like and by placing them back in the speech stream at the appropriate junctures (as proposed by Bregman, 1990, pp. 545–546).

In contrast with the model sketched by Bregman (1990), we propose that the perceptual functions of auditory scene analysis and phonetic analysis are simply not contingent and that the detection of phonetic coherence is autonomous from the identification of sound sources (Liberman & Mattingly, 1989; Mattingly & Liberman, 1990). Our new evidence confirms the precedents indicating the divergent perception of speech and other acoustic sources and exposes an aspect of the human susceptibility to the phonetic properties of spoken language at the level of perceptual organization. The perceptual coherence of formants, whether pulsed, shocked, aspirated, or fricated, therefore depends on the perceiver's sensitivity to the spectral variation characteristic of vocal sources and not to their fine-grain comodulation or harmonic relatedness. Our rejection of this general auditory model of scene analysis is based on perceptual studies of sine-wave replicas of utterances, which provide evidence that the principles of auditory form may be irrelevant to the perception of speech.

Sine-Wave Replicas of Utterances

Our recent studies have examined a simple perceptual question: Does phonetic perception depend on the occurrence of specific familiar acoustic elements within the speech signal? The customary answer to this question is found in the literature

[6] Actually, the formants are excited during voiced speech by a mixture of pulsing and noise, and the spectral contour of noise excitation will often differ from the spectral contour of harmonic excitation. Therefore, the precise mixture of noise and harmonics will vary across the formants, rendering it unlikely that each formant exhibits the identical proportion of periodic and aperiodic excitation. Whether different mixtures of noise and buzz within 100-Hz bands separated by 1 kHz violate a principle of grouping by similar spectra, or by comodulation, is an empirical question, although one of slight urgency from our perspective.

on the elementary acoustic particles that cue distinctions between consonants or vowels. Whether the particles are isolated spectral elements—the formant frequency transitions, momentary aperiodicities, and low-frequency murmurs that are correlated with articulatory place, voicing, and manner (Delattre, Liberman, & Cooper, 1955; Zue & Schwartz, 1980)—momentary spectral envelopes (Stevens & Blumstein, 1981), or sequences of spectral envelopes (Klatt, 1979), the perceiver is portrayed as a trader, busily tallying the acoustic cues and exchanging them for phonetic segments. The computational networks that stand as the metaphor for the process of cue exchange embrace this trading assumption (Elman & McClelland, 1986; Massaro, 1987), substituting an automatic mechanism for an explicitly psychological account of the trades.

An alternative to the detection of discrete cues is suggested in a few lines of research (Bailey & Summerfield, 1980; Best, Studdert-Kennedy, Manuel, & Rubin-Spitz, 1989; Jenkins, Strange, & Edman, 1983; Kewley-Port, 1983; Liberman, 1970; Remez, Rubin, Pisoni, & Carrell, 1981). This small literature expresses common themes: that there does not seem to be a core of invariant acoustic cues on which to base a mechanism of phonetic recognition; neither does the variability of the acoustic properties of speech signals implicate a normal set of acoustic tendencies around which cue variation occurs. Instead, these studies have emphasized the perceptual effects of spectral variation independent of the acoustic and auditory elements composing the signal and have suggested that this property of stimulation is no less responsible for phonetic impressions than the specific psychoacoustic effects of the cues. If the acoustic cue is not an analytic entity drawn from a finite stock of vocal effects, as this view urges, then any search for typical acoustic features of one or another phone must be forever frustrated. The relevant description of acoustic variation pertains to the phonetic effects of spectro-temporal patterns. Empirically, it has recently been possible to separate this perceptual susceptibility to time-varying patterns of speech signals from the more familiar psychoacoustic response to particular minimally contrasting speech cues, by using an acoustic ruse, the replication of utterances using a few time-varying sinusoids. Such a signal is obviously not speech, but it conveys phonetic information in its spectro-temporal attributes.

Sine-wave replication uses a variant of digital synthesis in which artificial acoustic spectra are specified, and the corresponding waveforms are calculated for conversion to acoustic patterns (Rubin, 1980). In contrast with the familiar formant synthesizer, by which one can attempt a faithful replication of the elementary spectral details of an utterance (Klatt, 1980), we use a sine-wave synthesizer to impose the coarse-grain properties of the spectrum of natural utterances on ideal (and unmistakably nonspeech) acoustic vehicles. The tone analogs of speech that result from sine-wave replication include three or four time-varying sinusoids, each of which reproduces the center frequency and amplitude of a vocal resonance in a natural utterance. A sine-wave replica of an utterance therefore retains the overall configuration of the speech spectra on which it is modeled, although it lacks the regular pulsing, aperiodicities, and broadband formant structure characteristic of natural speech. Figures 2 and 3 portray some of the prominent acoustic differences between natural speech, synthetic speech, and sine-wave replicas of speech signals. Figure 2 shows a signal produced with a software cascade-type speech synthesizer, which imitates the short-term spectra and spectro-temporal variation of the sentence in Figure 1b and the sine-wave replica, which substitutes a time-varying sinusoid for the oral, nasal, and fricative resonances associated with phonetic distinctions. Figure 3 illustrates short-term spectral differences among natural speech, synthetic speech, and a sine-wave replica. Because sinusoidal sentences present a pattern of spectro-temporal variation without the spectral details characteristic of vocal acoustic products, our perceptual tests can be taken to reveal the effects of auditory variation independent of the momentary impressions of vocal sound.

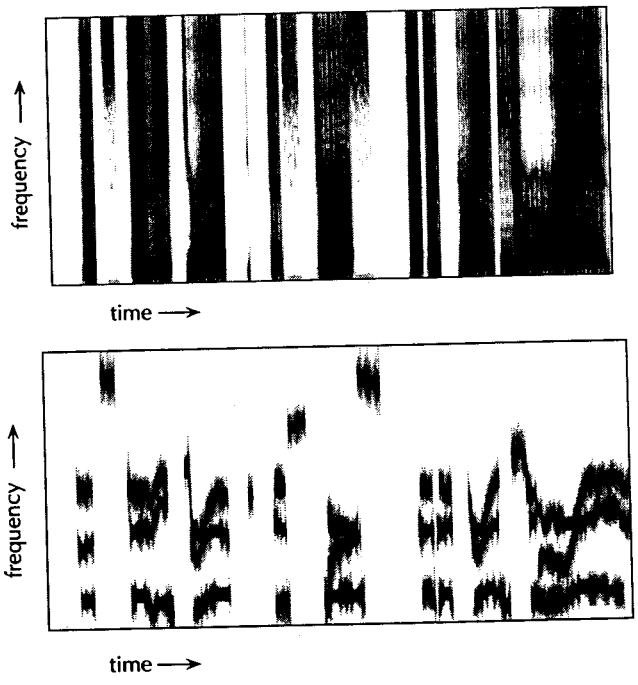

Figure 2. Two acoustic realizations of the sentence, "The steady drip is worse than a drenching rain." Top panel: Synthetic speech version, in which the details of the natural spectrum are meticulously preserved. Bottom panel: Sinusoidal version, in which a small set of time-varying sinusoids is used to replicate the coarse-grain spectro-temporal properties of the signal, eliminating the regular pulsing, harmonic series, aperiodic components, and broadband resonances. Whereas synthetic speech is clearly speechlike in timbre, sinusoidal sentences are both intelligible and utterly nonvocal in timbre.

The perceptual effects of sine-wave sentences are, in a word, unusual. To naive listeners, phonetic properties are not apparent, and when challenged to describe the three- or four-tone signals, they variously report hearing contrapuntal tones, radio interference, modern music, electronic sounds, equipment failure, and the tweeting of birds (Remez et al., 1981). The simple instruction to attend to a synthesized sentence is sufficient to permit an otherwise naive listener to hear the phonetic structure of the natural utterance on which the sine-wave replica was modeled. Even when phonetic impressions are evoked by the tone complexes, the timbre of a natural voice is not (Remez et al., 1981; Remez & Rubin, in press). Our studies of the percep-

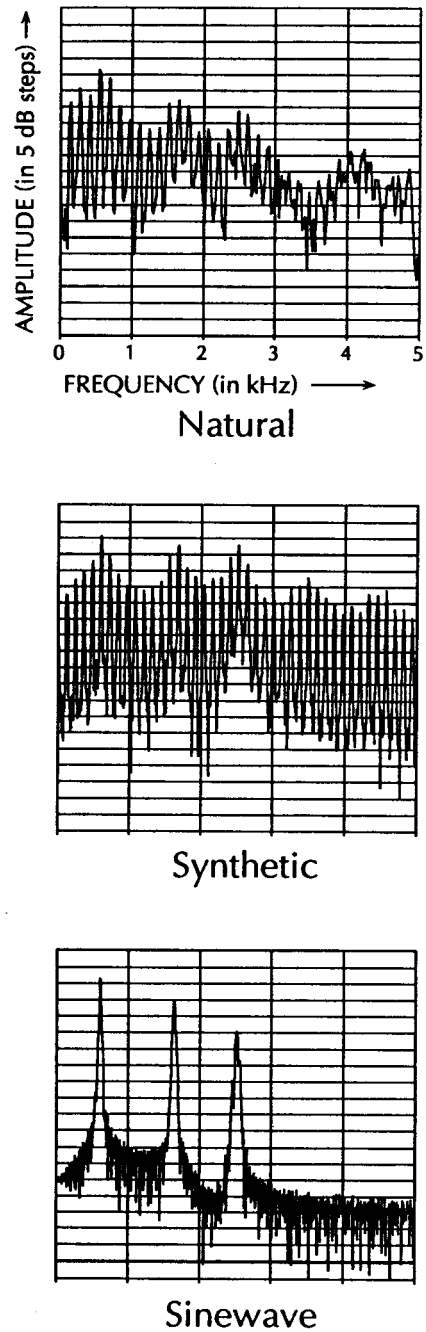

Figure 3. Short-term spectral sections of natural, synthetic, and sinusoidal sentences, illustrating the preservation of natural harmonic structure and broadband resonances in the synthetic sample. The sinewave sample preserves the resonance center frequencies and amplitudes of the three lowest resonances, eliminating the speechlike spectral fine grain from the signal.

tion of phonetic attributes of such signals show that the frequency variation of the tones provides much of the phonetic information (Remez & Rubin, 1983, 1984, 1990, in press) as long as the perceiver treats the three or four simultaneous tones as if they were a complex acoustic effect of a phonologically governed act. Indeed, our test subjects proved sensitive to subtle differences of vocal tract scale in identifying sinusoidal vowels, indicating that they treat a tonal replica as if it was issued from a specific talker (Remez, Rubin, Nygaard, & Howell, 1987). Listeners do not solve the transcription task by afterthought, good guessing, or analogical reasoning, or at least they do not appear to resort to cognitive compensation in a manner distinct from listening to ordinary speech.

It is easy to see that the perception of phonetic properties from sine-wave replicas is an anomaly from the perspective of current accounts of speech perception (see Klatt, 1989). First, the acoustic cues on which segmental distinctions are said to rely are absent from tone complexes, which lack the fine-grain acoustic structure typical of vocal products. By these accounts, the raw acoustic properties of sine-wave replicas are inappropriate for speech and should not induce phonetic perception. Second, the psychoacoustic properties of individual moments within a sine-wave replica differ drastically from impressions of natural speech. In the absence of natural cues, we might expect the perception of speech to result from elementary psychoacoustic effects, however unnaturally they are elicited (see Remez et al., 1981, Note 15; Remez et al., 1987, Experiment 2); yet, the impressions of consonants and vowels do not accompany impressions of vocal timbre in the sine-wave replicas, confirming that the sensory effects of the tones are rather distinct from their phonetic values. Third, the hypothetically modular system architecture that reserves a separate and specialized mechanism for speech describes only the phonetic requirements for engaging the speech module (Liberman & Mattingly, 1989; Mattingly & Liberman, 1990). Although the module is said to attempt a phonetic interpretation of all auditory stimulation before passing nonphonetic inflow to other systems, the account offers no more than a hint that four time-varying sinusoids should be effective in evoking phonetic impressions while leaving a sensory residue perceptible as several simultaneous whistles.

Perceptual Organization of Sine-Wave Sentences

If the reports of intelligible patterns of sinusoidal elements oppose perceptual theories that rest on a notion of typical acoustic cues, then the fact that the tones cohere and evoke phonetic impressions completely resists the account of auditory organization built on Gestalt principles of form. The argument is straightforward. First, recall that the general auditory principles poorly match the definitive phenomena of speech perception. Namely, those principles proscribe the fusion of continuous and discontinuous, periodic and aperiodic acoustic components despite the ordinary occurrence of this kind of heterogeneous grouping in speech signals. Next, note that the single encouraging point within the overall explanatory failure is the instance of glottally comodulated vocal resonances, which appear to be grouped together according to common fate. The plausibility of this explanation of the grouping of simultaneous voiced formants, notwithstanding the counterevidence of Liberman and Studdert-Kennedy (1978), musters allegiance to auditory principles in the case of speech signals. Therefore, any convincing evidence that the grouping of formants need not arise by co-

modulatory common fate diminishes the importance of this instance and exposes the Gestalt principles as unimportant in the organization of speech signals. Moreover, the perception of speech from such atypical auditory effects suggests that non-Gestalt resources describe speech abstractly and do not merely summarize typical auditory correlates of phonetic classes.

This evidence exists, we suggest, in studies of the perception of sentences conveyed by time-varying sinusoids, the components of which are neither harmonically related nor comodulated (Remez & Rubin, 1984). Rather, the sinusoidal components of a sentence exhibit all of the acoustic properties that promote perceptual segregation of each component from the others, by general principles. In a way, the incoherence that the auditory rules predict actually occurs, and listeners do hear several simultaneous tones changing in pitch and loudness as a sentence proceeds; however, to convey linguistic attributes, the simultaneously varying tones must cohere phonetically (Remez et al., 1981). The cohesion of the formant analogs shows that simple common fate is not required for grouping, not even when concurrent acoustic events lack proximity, similarity, continuity, and symmetry. To explain phonetic coherence of the sinusoids warrants a supplementary principle.

Although this argument about a proper test of the general auditory account is unequivocal, the evidence from perceptual studies of sinusoidal replicas that we have marshaled is unfortunately subject to an obvious doubt. For a listener to perceive linguistic attributes from sine-wave signals, the tones must be organized as a single ongoing perceptual stream and not as a polyphonic collection of independent elements. Although it seems unlikely that general auditory processes are responsible for the coherence that phonetic perception requires, all tones share binaural presentation and, therefore, perceived location, which may promote fusion by spatial similarity. Furthermore, recent studies of the physiology of auditory transduction suggest that other auditory interactions may dispose the perceiver to fuse the tone analogs. Wholly apart from Gestalt considerations, studies of the cochlea show that its active and passive functions are capable of producing *combination tones* (Brown & Kemp, 1984; Kim, Molnar, & Matthews, 1980), the frequencies of which depend on the components of the incident acoustic wave. Such distortion of cochlear origin, whether because of overlapping basilar membrane effects of the individual components of a replicated sentence or because of physiological properties of hair cells, may provide a potential, if remote, source of auditory information to support the perceptual organization of tone complexes in the absence of regular and fine-grain acoustic comodulation.

A combination tone brought about by the concurrent frequency variation of the first and second formant analogs, for example, could lead the perceiver to treat those tones as part of the same perceptual stream. Certainly, mere existence of cochlear distortion does not guarantee perceptual fusion of the distorted signal elements. Perceivers might not notice combination tones or might segregate them from a phonetic stream by general auditory or phonetic principles of organization. In any case, it is critical to determine whether the grouping of formant analogs is attributable to spatial similarity alone, or to mechanical or physiological distortion of the signals, or whether the fusion of tones in a sinusoidal sentence is an instance in which diverse acoustic products of a vocal source cohere perceptually. Were cochlear effects responsible for the fusion of formant analogs, and ultimately for the creation of phonetic impressions, we would have found an explanation of perceptual organization of sinusoidal sentences in the mechanics of transduction. Although such an explanation would obviously not derive from Gestalt principles, in its automatic and low-level nature it would minimize the force of our claim, namely, to have discovered a phenomenon of perceptual organization that falsifies primitive auditory analysis and resists secondary appeals to schema-driven repairs.

Present Studies

At this juncture, it seems as though the phonetic organization of sine-wave analogs of formants may be strong evidence against the generality of a Gestalt-based account of perceptual organization, although only if phonetic grouping of sine-wave sentences is not attributable to a property of cochlear physiology. A definitive test to determine the role of perceptual mechanisms in the sine-wave phenomenon is required to choose from these alternative interpretations, and we sought to achieve this in a series of experiments. Experiment 1 determines whether the organization of sinusoidal constituents in a sentence replica results from spatial similarity or cochlear distortion of the first and second formant analogs. Experiment 2 assesses the cognitive contributions to the measures obtained in Experiment 1. The results of these studies relieve the principal objections to considering a sine-wave sentence a true instance of perceptual organization. Experiment 3 obtains a clear index of the phenomena of organization pertinent to speech signals. By presenting sine-wave sentences dichotically, in which the first and second formant analogs occur concurrently but at different ears, we minimize the cochlear effects and pose a stricter test of perceptual organization in the same step. If perceivers nevertheless group tones occurring in different locations, lacking comodulation, proximity, similarity, continuity, and symmetry and bearing only abstract resemblance to speech, then neither Gestalt principles of form, specific cochlear distortion, nor schematic legerdemain can be held responsible for phonetic perceptual organization.

Experiment 1

Is Fusion of Sinusoidal Components Attributable to Cochlear Distortion?

Here, we exploited a finding by Remez et al. (1981) that phonetic perception of sinusoidal sentences required the concurrent availability of the tones analogous to the first and second formants. Listeners in that study simply transcribed the sentence carried in the tone patterns, and the experimenters controlled which combinations of the three tones composing the sentence were available in each condition. No listener was able to identify the spoken message given single tones to transcribe, and phonetic perception was possible only when the pair of tones replicating the first and second formants were presented together. Other pairs (second and third formant analogs or first and third formant analogs) were barely transcribed. These re-

sults clearly coincide with decades of speech research on the critical importance of the two lowest frequency oral resonances as vehicles of information for distinctive phonetic properties, although among other things this outcome suggests that organization of sinusoidal constituents for phonetic purposes may require the simultaneous availability of both components. Of course, if the mechanism of transduction itself drives the fusion of first and second formant tones, no appeal to principles of organization is necessary.[7]

To rule out a cochlear account of organization directly requires that we assess grouping when the first and second formant analogs occur at different ears, thereby eliminating the possibility of the induction of fusion due to peripheral interactions of the two tones critical for phonetic perception. The tests that follow this premise used sine-wave sentence replicas in four different conditions of presentation. First, to establish baseline performance, all constituents of a sentence replica were presented binaurally. This listening condition had elicited good transcription performance in our earlier research and served as one kind of control. The second condition was the critical test of peripheral induction of grouping: One ear received a sentence replica lacking its second formant analog while the other ear received the second formant analog belonging to the sentence. In this dichotic presentation, the coherence of the first and second formant tones, on which phonetic perception depends, could not arise through cochlear interactions of the critical sinusoidal constituents. Were fusion of dichotic tones to occur nonetheless, we would find that transcription performance in the dichotic condition exceeds the sum of the transcriptions of the partial signals available to each ear alone. Therefore, in the third and fourth test conditions, we assessed the transcribability of the partial replicas available to each of the ears to estimate the best performance in the dichotic condition were fusion to fail.

Method

Acoustic test materials. Four versions were made of each of 10 sentences, using the technique of sinusoidal replication of natural utterances. The natural signals from which the sinusoidal patterns derived were produced by a male speaker (Robert E. Remez) recording a single utterance of each sentence in a sound-attenuating chamber, using a condenser microphone, low-noise audiotape, and half-track tape recorder. These recorded signals were then low-pass filtered at 4.5 kHz and sampled at 10 kHz with 12-bit amplitude resolution and stored on a VAX-based computer system. Spectra were estimated from the sampled data by the method of linear prediction (Markel & Gray, 1976) at intervals of 10 ms throughout each utterance. After correcting erroneous values by checking the linear prediction spectra against traditional spectrograms, the frequency and amplitude values of the estimated formants were used to control the output of a sine-wave synthesizer (Rubin, 1980), a software device that calculates the waveforms of signals generated by multiple independent audio-frequency oscillators. Three or four sinusoids were used to replicate each sentence, one sinusoid for each of the three lowest frequency formants and a fourth for a fricative formant when appropriate. A graphic representation of a sentence replica is shown in the top panel of Figure 4.

The listening tests were prepared on the VAX and output to tape by digital-to-analog conversion. Listening sessions were conducted by means of playback of the test tapes, through a power amplifier, using matched headsets. The average listening level was attenuated to 60 dB SPL (sound pressure level) re 0.0002 dynes/cm^2.

Four tests were composed from the basic stock of 10 sinusoidal sentence replicas. In the binaural test, complete sentence replicas were delivered on both channels to the headphones. In the dichotic test, neither channel received the complete sine-wave sentence replica: One channel contained a single tone following the frequency and amplitude pattern of the second formant; the other channel contained the analogs of the first, third, and fricative formants. In the Tone 2 missing test, both channels contained the sentence patterns with the analogs of the second formant deleted. Last, in the Tone 2 alone test, the analog of the second formant was delivered on both channels.

Procedure. Each of the four tests was conducted with a different group of listeners. A test session consisted of three blocks in all of which listeners were simply asked to transcribe the synthesized sentences. First, a warm-up block occurred, consisting of a sequence of nine binaurally presented sinusoidal sentences differing from the main test sentences to accustom the listeners to the unusual timbre of these synthetic signals. The first three sentences in this set were transcribed for the listeners by the experimenter in advance to facilitate the perceptual adjustment to replicated utterances, to give examples of transcriptions, and to provide indirect feedback on the initial three trials. After a short pause, one of the four test conditions was presented, consisting of 10 sinusoidal sentences, none of which duplicated any of the items of the warm-up set. The dichotic condition was counterbalanced across listeners for the ear receiving the second formant analog and for the ear receiving the first, third, and fricative formant analogs. Following the presentation of one of the four test conditions, the warm-up sequence was presented once more. The transcriptions of this repeated set served to identify subjects who were unable to transcribe sinusoidal patterns even under highly favorable circumstances. Individuals who failed to transcribe the three sentences in the repeated set that had been transcribed by the experimenters in advance in the warm-up presentation were not included in the data set.

Every trial had the same structure. A sinusoidal pattern was presented eight times, each separated by 3 s from the next and each block of eight separated from the next by 10 s. Listeners were instructed to write while the sentences were repeating and were encouraged to report whatever words or fragments they heard.

Subjects. Ninety-nine listeners were tested, each assigned to one of the four test conditions. After eliminating those who failed to follow instructions or who did not transcribe the sentences in the retest, 15 remained in each of the four conditions. Some listeners were paid for participating, whereas others received course credit in introductory psychology in exchange for participation. Each subject reported normal hearing, and none had participated in any other study using sinusoidal sentence replicas. Testing occurred in groups of 6 or fewer listeners.

Results and Discussion

The transcriptions provided by the subjects were scored for the number of syllables correctly identified. An average percent-

[7] Another study reported by Remez et al. (1981) opposed the interpretation that transduction in the cochlea, or by any other irresistible sensory process, is simply responsible for the listener's ability to hear the tones as speech. By comparing the performance of subjects instructed to listen for a sentence with that of subjects asked more obliquely to identify the sounds in the headphones, it was determined that phonetic attributes were heard only when listeners were instructed to attend to a strangely synthesized sentence, with few exceptions. The phonetic grouping of the tonal analogs of formants was not an obligatory consequence of auditory sensations evoked by the tones, although this would surely be the case were phonetic grouping to depend on automatic auditory processes. (This issue is elaborated in the General Discussion section.)

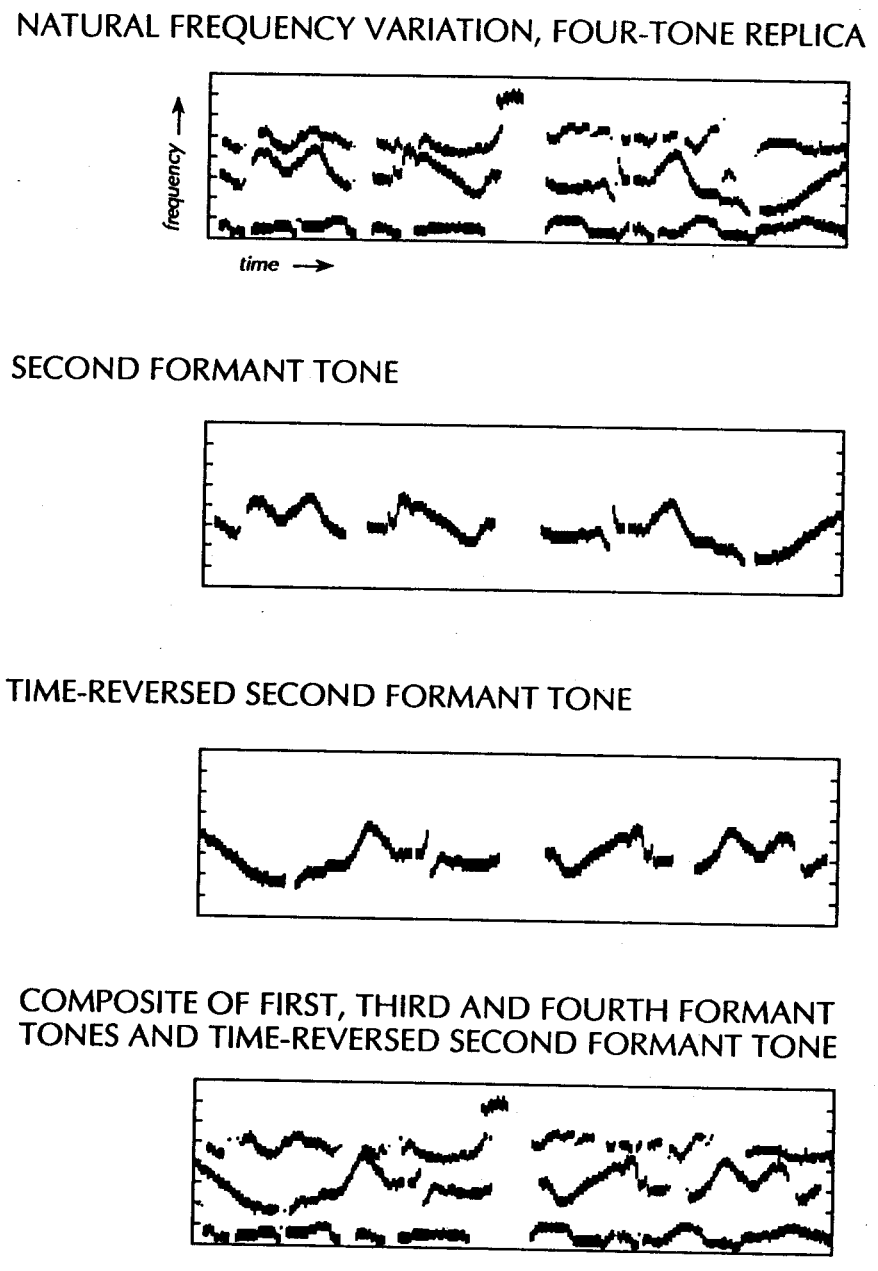

Figure 4. Schematic representation of sinusoidal sentence replicas. Top panel: Four sinusoids are used to replicate the sentence, "The beauty of the view stunned the young boy"; there is one tone for each of the three lowest formants, including bursts, aspirations, and nasals and a fourth tone for the fricative formant. Second panel: The second formant tone pattern isolated from the sentence replica. Third panel: The pattern of the second formant tone time reversed (reflected temporally). Bottom panel: The pattern formed by the tonal replica lacking a second formant analog combined with a temporally reflected second formant tone.

age correct score was derived for each subject, on which the statistical analyses of group performance were conducted. Although this measure may underestimate the phonetic effectiveness of sinusoidal vehicles, this conservative technique has proven to be a reliable correlate of speech perception with sinusoidal replicas (e.g., Remez & Rubin, 1990, in press). In the present conditions, transcription performance was observed to vary across the four listening conditions; the mean correct transcription performance for each group of 15 subjects is shown in Figure 5.

Performance differed significantly across the four tests, as revealed by a one-way analysis of variance, $F(3, 56) = 69.9$, $p < .001$. A Scheffé test estimated the smallest significant difference for pairwise comparisons of the group means (13.4, $\alpha = .05$), and this value is represented in Figure 5 in the height of each T bar. Binaural presentation of the sinusoidal replicas led to the

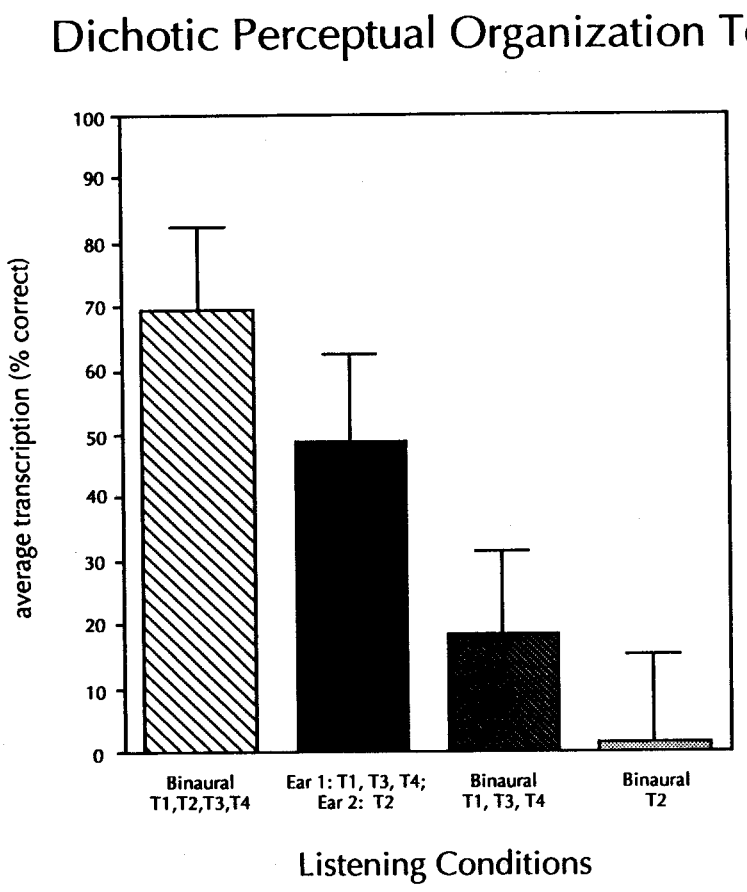

Figure 5. Group performance in the four listening conditions of Experiment 1: binaural presentation of sentence replicas (light stripes); dichotic presentation in which the second formant analog was presented to one ear and the remaining tones to the other (dark filled bar); binaural presentation of tone analogs of the first, third, and fricative formants (dark stripes); and binaural presentation of the second formant analog (light filled bar). T = tone.

best transcriptions (light stripes: 69% correct), and the dichotic presentation was transcribed well (dark filled bar: 49% correct), although neither of the partial tone complexes were well transcribed (heavy stripes: Tones 1, 3, and 4 at 18% correct; light filled bar: Tone 2 alone at 2% correct). The performance in the dichotic condition was better than would have been expected on the basis of the estimates of the contribution of each ear alone, and binaural performance was significantly better than dichotic performance.

The dichotic condition is critical for testing the hypothesis that fusion of sinusoidal components in a sentence replica is induced peripherally. Were peripheral effects responsible for grouping the sinusoids, a dichotically presented set of components would not fuse, and dichotic performance would therefore be no better than the sum of the transcription performance in the two conditions in which the partial signals were evaluated. Here, we observed that dichotic transcription was more than twice as good as the sum of the mean transcription achieved in the two tests using incomplete tone ensembles, showing that listeners were indeed able to obtain phonetic information from the dichotic signal that neither ear supplied alone. This test offers a clear indication that fusion of auditory elements is possible without satisfying the grouping criteria derived from the Gestalt principles of form. Neither does it seem that auditory distortion of cochlear origin is the only cause for combining sinusoidal analogs of formants.

The performance level difference between the dichotic and binaural sentences is important to note, too, for it reveals a cost of discrepancy in location despite the phonetic coherence of the components. Although this difference may stem from a violation of an auditory grouping principle—for example, that similarity in location promotes fusion—we cannot be sure that the performance difference between these two instances of phonetic fusion is not simply attributable to the unique problem posed by the dichotic task, namely, to sample, hold, and combine different signals across the ears. All other things being equal, these two conditions differ in difficulty. The dichotic condition requires the subject to integrate sensory events at different ears, and this imposition may draw resources away from the analysis of the auditory details or from the analysis of the combined pattern achieved through perceptual organization. On the basis of classic studies of the sensory processes accompanying phonetic

perception (e.g., Darwin & Baddeley, 1974; Liberman, Mattingly, & Turvey, 1972; Pisoni, 1973), we must expect an unanalyzed auditory representation of a speech signal to decay rapidly and any delay or compromise in the creation of phonetic impressions to be costly. Of course, these results also reflect the performance of completely naive listeners unpracticed in the laboratory task of organizing dichotic inflow, whereas ordinary listening is more nearly approximated by the condition of binaural presentation. If the ordinary experience of our subjects prepares them better to perform in the binaural than in the dichotic condition, this disadvantage hampers but does not prevent subjects in the dichotic test from organizing the components as if hearing a speech signal. Whether performance improves with additional exposure to dichotic sentences, or remains 20% poorer than a binaural condition, is left for further study.

To summarize, sinusoidal formant analogs were organized phonetically in dichotic presentations that preclude an account of fusion by appeal to peripheral induction. Because listeners fused the pieces of formantlike variation even when the components arrived in different ears, performance in the dichotic case exceeded what we expected if each ear contributed independently to phonetic perception. This outcome is anticipated by our critique of the standard auditory treatment of perceptual organization. Having ruled out a lower level account of fusion in this test, we suggest that this outcome is a manifestation of a kind of perceptual organization surpassing the simple forms permitted by Gestalt principles. In this way, the listener amalgamates the concurrent tone variation available at each ear to form a single pattern consistent in an abstract fashion with the time-varying acoustic properties of vocalization. The coherent ensemble of tones is thereby analyzable into specific phonetic attributes.

However likely this conclusion may appear from the results of Experiment 1, an alternative to our account is possible. The facts of the dichotic test can be interpreted differently with reference to the literature on the restoration of phonetic and phonemic attributes from interrupted and masked speech signals (Blank, 1979; Samuel, 1981; Sherman, 1973; Warren, 1970). To sketch such an account of Experiment 1, consider that the presence of the second formant analog in the ear opposite the first, third, and fourth may trigger the use of cognitive resources to complete the phonetic attributes of the utterance. In that case, no fusion of dichotically arrayed components occurs; rather, partial phonetic information is available at one ear in an incomplete acoustic pattern, but the missing attributes are not restorable without the concurrent distracting acoustic material from the opposite ear. Experiment 2 contained two tests to assess the likelihood that restoration occurs when sinusoidal sentences are presented dichotically.

Experiment 2

Is Perception of Dichotic Sine-Wave Sentences Based on Phonemic Restoration?

Many studies of phonemic restoration show that speech perception depends on perceptual analysis and cognitive structure alike. In the typical instance of restoration, an utterance of a familiar word is altered through acoustic editing in which a brief moment of the speech signal is deleted and replaced with an acoustic foil lacking a speechlike spectrum. Although the listener notices the prominent intrusion of the sound replacing the speech, the utterance is often perceived as if the speech signal had been perfectly intact (Warren, 1984). The linguistic representation is so completely restored that, in fact, listeners seem unable to distinguish similar versions of a word: one missing a portion of its signal and another merely containing intrusive noise (Samuel, 1981). Of course, there are many circumstances in which listeners have great difficulty attending to auditory correlates of speech (e.g., Mattingly, Liberman, Syrdal, & Halwes, 1971).

One of the signals used in Experiment 1 could be considered to be an incomplete although potentially restorable pattern: This is the tone ensemble containing the first, third, and fourth tones, lacking the second formant analog. Although it is consistent with precedents for an incomplete phonetic display to be perceived poorly because of a hole in the signal, as we observed when this pattern was presented to both ears in the third condition, it is also consistent with the findings of restoration for the perceiver to fill in the missing phonetic detail once a foil occupies the acoustic hole (Warren, 1984). Did we observe the effects of restoration of the linguistic information due to the presence of a foil in the opposite ear? Alternatively, did the dichotic condition measure perceptual organization by a non-Gestalt principle of phonetic coherence? For the results of Experiment 1 to count as counterevidence to a cochlear account of sinusoidal fusion, the latter must have occurred.

No direct parallel to the case of Experiment 1 exists in the studies of restoration (cf. Bashford & Warren, 1979), neither in the sentence-length duration of the missing acoustic material nor in the specific exclusion of a second formant and the dichotic placement of the foil. Therefore, we cannot say definitively whether natural signals treated analogously to the dichotic test of Experiment 1 would evoke restoration. It is a simple matter, nonetheless, to perform a specific test pertinent to the interpretation of Experiment 1, forgoing a more general study of phoneme restoration. This is accomplished by composing a dichotic test to determine whether a phonetically incompatible substitute for a second formant tone acts as a foil in triggering restoration when it is presented in the ear opposite to the first, third, and fricative tones. We know from Experiment 1 that the performance level of a dichotically arrayed complete sentence replica approaches 50% correct. If this success is attributable to cognitive restoration rather than to perceptual fusion, then a tone varying within the frequency band of the second formant will act as a foil, even if it fails to replicate the frequency and amplitude values of the second formant of the sentence, and restoration will occur. Conversely, if performance in the dichotic condition of Experiment 1 is the consequence of perceptual fusion, then we should observe poor performance when dichotically arrayed sinusoids fail to replicate the time-varying spectral peaks of a natural utterance.

Method

Acoustic test materials. Tone patterns for two dichotic restoration tests were constructed from the stock of 10 sentences used in Experi-

ment 1. In the first test, Flipped T2, the parameter table for the second formant tone was temporally reflected, or flipped, and the resulting signals produced by the sine-wave synthesizer placed the three temporally veridical analogs of the first, third, and fricative formants on one channel and the temporally reversed second formant tone on the other. A graphic depiction of a flipped second formant analog is shown in Figure 4, along with the composite formed by combining the reversed second formant analog with temporally veridical first, third, and fricative formant analogs. In the second test, Flipped T1, T3, and T4, the temporally veridical analog of the second formant was placed on one channel and the temporally reversed analogs of the first, third, and fricative formants on the other. Conditions for delivering the sounds to listeners were the same as described for Experiment 1.

Procedure. The two tests were conducted with different listeners in three blocks: a warm-up, a test block, and a retest of the warm-up sentences. The first and third blocks were used, once again, to facilitate the perceptual adjustment to replicated utterances, to give examples of transcriptions, to provide indirect feedback on the initial three warm-up trials, and to determine post hoc whether a listener was susceptible to the phonetic effects of sine-wave utterances. The presentation of dichotic materials was counterbalanced over the ears across listeners. Instructions given to subjects were the same as in Experiment 1, as were the criteria for identifying susceptible listeners.

Subjects. Forty-four listeners were tested and were assigned to one of the two test conditions. Fifteen subjects remained in each condition after eliminating those listeners who did not transcribe the sentences in the retest that had been identified by the experimenters in the warm-up block. Some listeners were paid for participating, whereas others received course credit in introductory psychology. No subject reported a history of speech or hearing disorder, nor was any subject familiar with sinusoidal sentence replicas. Testing occurred in groups of 6 or fewer listeners.

Results and Discussion

As in Experiment 1, the transcriptions provided by the subjects were scored for the number of syllables correctly identified. Again, each subject contributed an average percentage correct score to the analysis of group performance. In the two tests of Experiment 2, transcription performance was observed to differ; the mean correct transcription performance for each group of 15 subjects is shown in Figure 6, along with the results of the dichotic condition of Experiment 1 for comparison.

Several hypotheses were tested on this data set. To begin, we determined that performance differed significantly between the two dichotic conditions containing time reversed components. The group mean of the test in which the second formant analog was time flipped differed significantly from the mean of the test in which the first, third, and fricative tones were time flipped, $t(28) = 6.49, p < .0005$. The performance level contrast is stark. Very little phonetic detail was available from the veridical second formant tone with the remaining components of the dichotic display reversed temporally (light filled bar: 0.4% correct). Conversely, relatively more of the phonetic attributes of the sentence were transcribed when the second formant analog was the sole component departing from the properties of the natural signal (striped bar: 17.8% correct). To assess whether this performance level difference was attributable to restoration, for in no case could it be attributed to the phonetic coherence of an arbitrary second formant tone and the veridical formant analogs, a confidence interval for the mean of this condition was determined using the *t* statistic to permit comparison with two

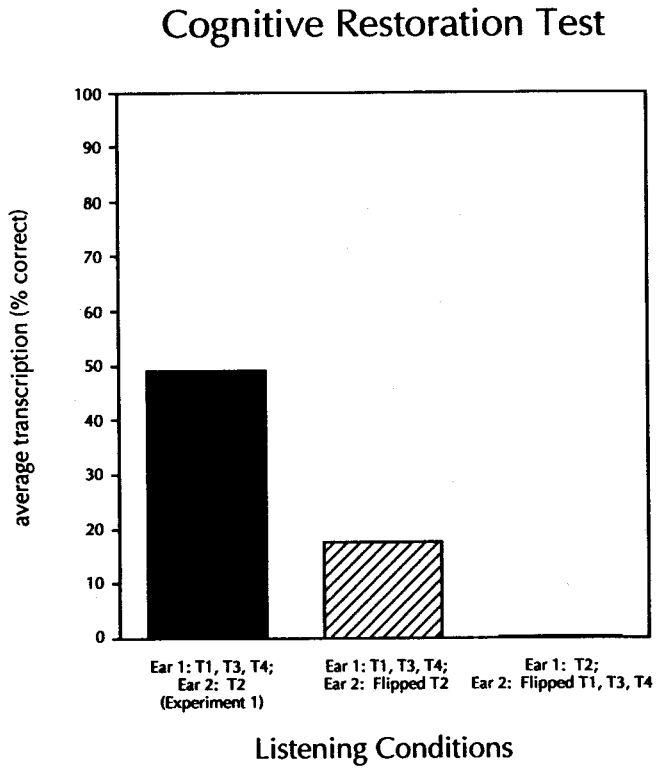

Figure 6. Group performance in Experiment 2: Dichotic presentation of tone analogs of the first, third, and fricative formants in one ear and the temporally reversed second formant in the other (light stripes) and of the temporally reversed first, third, and fricative formants in one ear and the temporally veridical second formant in the other (light filled bar, or shortest bar). The dichotic presentation of the sentence replica from Experiment 1 is shown to aid comparison (dark filled bar).T = tone.

tests of Experiment 1 ($12.2 < \mu < 23.3, \alpha = .05$). The performance on the dichotic condition of Experiment 1 fell outside this interval (49%); the group mean of this condition is depicted again in Figure 6 (dark filled bar) to facilitate the comparison. The performance on the binaural test of the tone ensemble lacking a second formant analog (18% correct; shown in Figure 5: dark stripes) fell well within the confidence interval.

To summarize the results, some phonetic information was available from dichotic patterns in which a temporally inverted single-tone analog of the second formant was presented in the ear opposite the other veridical formant analogs. However, this kind of presentation was no more effective than one simply lacking a second formant analog and was significantly less effective than a dichotically presented pattern in which the combined tones replicated a natural signal.

The two conditions of Experiment 2 complete a pair of tests that establish the phenomenon of phonetic organization in the perception of tone analogs of sentences. Experiment 1 revealed that the transcriptions of dichotically arrayed components of sentence replicas surpassed the performance expected from each channel alone, as assessed by tests presenting the partial replicas. Although this level of transcription could not be at-

tributed to the action of general auditory principles of organization, the tests of Experiment 2 were required to determine that perceptual organization, rather than restoration, actually occurred in this dichotic case. In two tests here, a temporally inverted component, speechlike in its overall variation while departing from the values of the second formant in detail, was not treated as a foil for the missing formant and did not trigger phoneme restoration. Rather than provoking the replacement of information missing because of the absence of a second formant analog, this component proved to be useless. Its presence did not affect performance, judging from the statistical similarity of this condition and the third test of Experiment 1 in which the signal contained no second formant analog.

The evidence of Experiments 1 and 2 counts against auditory scene analysis as a general theory of perceptual organization. That account fails because the cases considered in the first two experiments are not reducible to its functions and arguments. Nonetheless, although falsifying the general claim that the auditory principles supplemented by schemas offer an adequate description of the perceptual organization of speech, the first two studies provide no explicit evidence of a principle of organization superseding Gestalt grouping rules for phonetic cases. In Experiment 3, we build on the methods introduced in the prior two studies to distinguish a principle of organization applicable to speech signals.

Experiment 3

Is There a Time-Varying Principle of Phonetic Perceptual Organization?

The perceptual study in Experiment 3 used six tests to probe for effects of a principle of *grouping by phonetically governed acoustic variation.* We aimed to expose the principle by devising a condition to impede the organization of sinusoidal replicas. Assume that the listener who is given a sine-wave sentence to transcribe perceives the tones as a single phonetic source because of their speechlike variation in frequency (Remez & Rubin, 1990); the perceiver tacitly apprehends that the tone ensemble conveys resonance changes of a natural utterance. A corollary of this conclusion is that phonetic perceptual organization should be uniquely disruptable by certain kinds of concurrent patterns. For instance, an extraneous sinusoid that varies in a speechlike manner should compete for organization with the proper components of a sentence replica. However, no such competition should be observed with an unspeechlike extraneous tone. By any general auditory account of organization, neither kind of tone is enough like the constituents in a replicated sentence to be grouped with it.

In Experiment 3, we tested the susceptibility of perceptual organization to disruption by speechlike spectral variation. In the competitive organization task that we used, a sine-wave replica of a sentence was presented along with a surplus tone. In each case, the surplus tone coincided roughly with the sinusoidal analog of the second formant: either as a constant-frequency tone, the attributes of which are completely unspeechlike, or as a temporally inverted analog of the second formant. In this challenge, transcription performance reflects the listener's success in rejecting the surplus tone while retaining the components belonging to a sinusoidal replica.

This kind of test is not completely new and actually extends the studies of the resistance of speech to interference by a babble mask, a composite of several speech signals, which is used in lieu of the more familiar envelope-shaped noise (Carhart, Johnson, & Goodman, 1975; Carhart & Tillman, 1970; Kalikow, Stevens, & Elliott, 1977; Young, Parker, & Carhart, 1975). The requirement to detect a spoken message in the presence of a chattering background lends a touch of realism to measures of intelligibility, some of which have concerned the practical improvement of hearing aids, although the psychological findings interest us here. Those studies generally found that spoken masks impede speech more effectively than broadband noise masks do; also, the interfering effect of babble is correlated with the distinctness of the components. For example, an extraneous channel composed of fewer than 3 voices interfered more with the perception of the target speech than did a babble composed of 16 or more voices (Carhart et al., 1975). Whether these observations can be attributed to acoustic or to phonetic distinctness—that is, whether the effects are due to masking, to phonetic interference in immediate memory, or to competition in perceptual organization—remains an open question, although the competitors that we used left the interpretation less ambiguous.

Experiment 3 differed from these precedents in aiming the acoustic interference at a specific constituent of the signal, the second formant analog. In that respect, the unspeechlike constant-frequency tone matched the center of the frequency band within which the second formant analog varied in each test sentence. The speechlike competing tone was a close match to the second formant analog in its spectro-temporal attributes. It is reasonable to suppose that neither of these tones interferes perceptually with the sine-wave utterance at or above the level of the phonetic segment. The constant-frequency tone exhibits no speechlike attributes, and the temporally inverted analog of the second formant, by itself, is not likely to evoke phonetic impressions, as we have shown here and elsewhere (Remez et al., 1981). A temporally reversed tone is also unlikely to cohere with the first, third, and fourth tones of the sentence replica, as Experiment 2 revealed. Although interference of a speechlike competitor is unlikely to be phonetic in nature, arising in immediate memory, it is probable that interference results from competition in perceptual organization, at least in part. On the presumption that listeners seek the acoustic correlates of vocal sound production, we therefore expected to see this speechlike tone interfere more effectively with phonetic perceptual organization than with a constant-frequency competitor, as reflected in transcription performance.

Method

Acoustic test materials. Tone patterns for six tests were constructed from the stock of 10 sentences used in Experiments 1 and 2. The first three conditions were binaural competitive tests in which a sine-wave replica had no competing tone, a single-tone constant-frequency competitor, or a single-tone speechlike competitor. In the No Competitor condition, a sinusoidal replica of an utterance was used, consisting of three or four tones in which the pattern replicated the changes in resonant frequencies and amplitudes of a natural utterance. This test served to mark the best performance attainable from which to estimate the effects of competition on organization and perception. In the Constant-

Frequency Competitor condition, a surplus tone was set to the average frequency of the second formant and was added to each sentence replica, onsetting and offsetting with the second formant analog. Because of differences in phonetic composition across the sentences, the value of this constant-frequency tone ranged from 1,095 Hz to 1,640 Hz, with a mean value of 1,311 Hz. The amplitude of this constant-frequency competitor was ramped on and off in 20 ms and was set to a constant level roughly equal to the second formant analog at syllable nuclei. In the Speechlike Competitor condition, the temporally reversed frequency and amplitude values of the second formant were used to create a tone varying in frequency and amplitude as if it was a second formant, but it was not combinable with other tones to make a replica of the natural utterance. (A depiction of veridical and time-flipped second formant analogs is shown in Figure 4, with the pattern replicating the natural signal and the composite formed by combining the reversed second formant analog with temporally veridical first, third, and fricative formant analogs.) With both competitors, transcription performance reflected the apprehension of the tones replicating the utterance and the rejection of the competing tone, whether speechlike or not.

Dichotic versions of the three tests were also used in which components of the replicated sentences occurred at different ears. The dichotic No Competitor test arrayed the first, third, and fricative tone in one ear and the second formant tone in the other, as in the dichotic condition of Experiment 1. On the results of that test we expected fusion to occur, permitting perception of phonetic properties that derive from the ensemble of formant analogs. The dichotic versions of the two competitive organization tests placed the second formant analog in one ear and the first, third, fricative, and competitor tones in the other. In the Constant-Frequency Competitor condition, the competing tone, presented on the same channel as the first, third, and fricative tones, exhibited the unchanging mean frequency of the second formant, attenuated roughly to the amplitude of the second formant. The Speechlike Competitor condition used a tone exhibiting the temporally reflected values of the second formant. All other test materials and listening conditions were the same as described for Experiment 1.

Procedure. Six tests were conducted, each with different listeners. A test session included three blocks: a set of warm-up sentences, a test block, and a retest of the warm-up sentences. The first block was used to facilitate the perceptual adjustment to replicated utterances, to give practice in transcribing sentences, and to provide indirect feedback on the initial three warm-up trials. The presentation of dichotic materials was counterbalanced over the ears across listeners. Instructions to subjects were the same as in Experiment 1.

Subjects. Two hundred seventy-five listeners were tested, each assigned to one of the six test conditions. Forty-two subjects remained in each condition after those listeners who did not transcribe sentences in the retest had been eliminated from consideration. Some listeners received course credit in introductory psychology for participating, whereas others were paid for their time. No subject reported a history of speech or hearing disorder, and all were naive to sinusoidal sentence replicas. Testing occurred in groups of 6 or fewer.

Results and Discussion

The transcriptions of the test sentences were scored for the number of syllables correctly identified. Each subject contributed an average percentage correct score to the analysis of group performance. The six mean performance levels are shown in Figure 7, which plots binaural (filled round bullets) and dichotic (unfilled square bullets) results with the three competitive conditions (no competitor, constant frequency, and time-flipped second formant analog) arrayed left to right in the frame.

As the figure shows, binaural transcriptions surpassed dichotic in all conditions; there was also a clear effect on performance of different competitors. In the binaural tests, performance with no competitor was significantly better than performance with either of the two competitors, which did not differ from each other. This was determined by the analysis of variance and post hoc comparison of the means within the binaural presentation, $F(2, 123) = 8.52$, $p < .001$; Scheffé's $\alpha = .05$. In the dichotic tests, only the effect of the speechlike competitor was substantial and differed significantly from the other dichotic conditions, $F(2, 123) = 3.73$, $p < .027$, Scheffé's $\alpha = .05$. No difference was observed between the dichotic tests with a constant competitor and with no competing tone. In summary, the effect of a competing tone on perceptual coherence was contingent on the spatial array of the component tones. Either competing tone interfered with binaural performance equally; only the speechlike competitor impeded dichotic performance.

The findings of Experiment 3 reveal a manifestation of organization by phonetically governed acoustic variation. This is the unequivocal result of the dichotic tests, wherein the effect of a competing tone depended on its pattern of variation. A constant-frequency tone proved ineffective for disrupting perception relative to the condition without a competing tone. It seems that the presentation of the second formant analog and its competitor in different ears permitted the perceiver to segregate the unspeechlike tone from the first, third, and fricative tones with which it shared a spatial locus and to incorporate the authentic second formant tone from the opposing ear. This segregation of the competitor proved impossible when it exhibited speechlike variation. In that instance, it was apparently grouped with the analogs of the first, third, and fricative formants, despite the failure of that grouping to replicate an utterance. Grouping on the basis of speechlike spectral variation apparently held fast until a stage in perceptual analysis at which the phonetic information contributed by the actual second formant analog, at a separate spatial locus, was less accessible. Although grouping by similarity in location is anticipated in the Gestalt approach, the pattern-contingent suspension of this rule that we observed here is new and is an effect, we claim, of the reliance on a principle of grouping sensitive to the acoustic products of vocalization.

The binaural tests revealed no difference in performance between conditions of different competing tones, and phonetic perception in this instance was far from impervious to the effects of an unspeechlike competitor. This was apparently due to masking between the second formant analog and the competing tones. Neither competitor completely obliterated the second formant analogs that they opposed, for performance fell in the middle of the range in those two conditions. Here, the comparison between the binaural and dichotic tests is instructive for interpreting whether the binaural outcomes reflect masking or organization. Assume that performance on any of these six tasks is potentially affected by attentional load, by competition in perceptual organization, and by masking. Then, we may be confident that the spatial opposition of rival tones in the dichotic case eliminates two kinds of masking of the second formant analog present in the binaural case (Rand, 1974; Repp & Bentin, 1984): (a) the masking effect of the competing tones, whether constant frequency or speechlike, on the analog of the second formant and (b) the masking of the second formant analog by the first, which spreads upward in frequency. By the

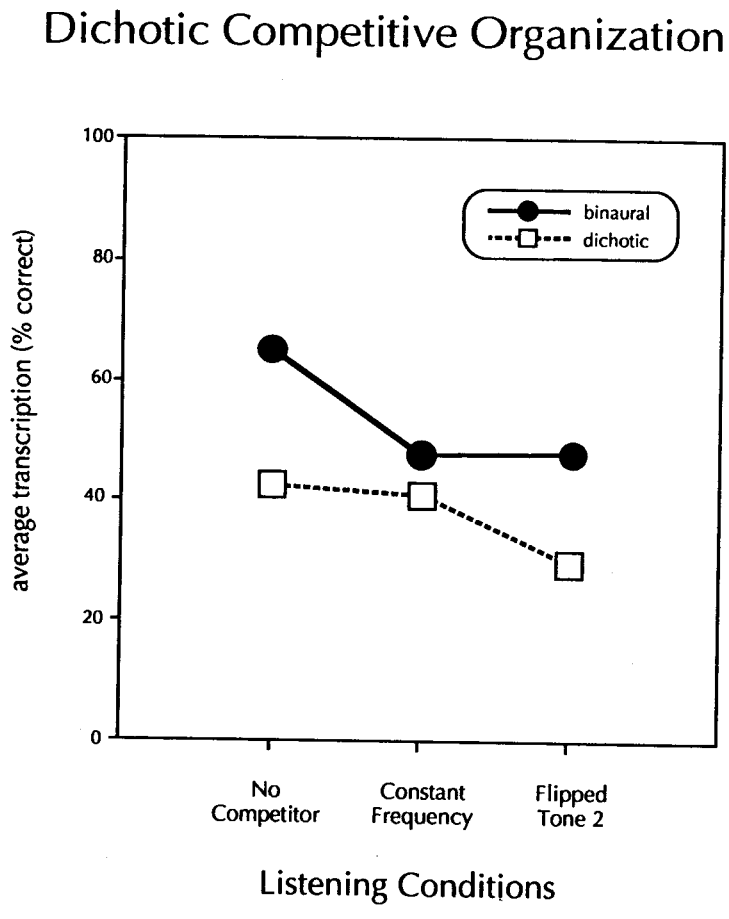

Figure 7. Group performance in Experiment 3: The binaural conditions and the dichotic conditions are shown for each competitor type (left: no competitor, center: constant-frequency competitor, and right: temporally reversed second formant analog competitor) arrayed left to right in the plot frame.

same reasoning, the competitor is dichotically freed from the masking effect of the second formant analog, rendering its spectro-temporal pattern more distinct. Dichotic presentation predictably allows the perceiver to isolate the spectro-temporal properties of the second formant analog and its opponents, although it evidently imposes a tax on performance for dividing attention between the ears. Nonetheless, these conditions permit us to observe a phonetic principle of perceptual organization by delineating the effects of peripheral masking. Although a parametric set of measures is warranted to confirm this account of the binaural case, the dichotic results offer indirect corroboration of a masking explanation of the binaural cases, while supplying direct evidence of a phonetic principle.

It should be mentioned, too, that primitive Gestalt-based auditory analysis, with or without supplementary schemas, fails to explain the outcomes of these tests. If similarity in location is presumed to be an operative principle, then the effects of the dichotic tests are impossible to handle, for grouping occurred across the ears with no competitor and in the presence of an unspeechlike competitor. If similarity in frequency variation is presumed to be an operative principle, then the grouping of tones to form sentences in binaural and dichotic conditions is inexplicable, and no premise exists for rejecting any particular tone pattern as a competitor to the tones forming the sentence replica. If a schematic component is presumed to supply information about the typical acoustic properties of speech, it is certain that sine-wave replicas fail to qualify as typical. Likewise, the rapidly fading auditory trace of the sine-wave sentence is not sufficiently durable to sustain a successful schematically driven inspection. Overall, it is difficult to see how Gestalt-based principles with schematic supplements apply to these complex ensembles. To group the auditory effects of speech production, the listener rapidly detects acoustic variation consistent with an articulating vocal tract and cannot project from proximal stimulation to distal vocal object by resorting to the impoverished similarity principles or typicality schemas offered in auditory scene analysis. On the contrary, coherence among the components of a sentence replica is established because the listener detects their plausible origin in a vocal source.

General Discussion

The listener's ability to follow a voice in a noisy acoustic field has been taken as the classic example of perceptual organization

for many years, although the kind of cocktail party in which phonetic organization occurs can be a more intimate one, indeed. Even when one talker produces speech in an anechoic field, a listener goes beyond the Gestalt rules to hear the diverse acoustic elements as one stream. What governs this function? In pursuing the answer, we transposed an argument (Bregman & Pinker, 1978) that took the grouping principles as given and supposed that they parse the auditory sensory array into sources of sound. On the evidence of the first two experiments reported here, we take the parsing of the acoustic array as given—for detecting the coherence of a single speech signal, this is not a risky assumption—and propose that the grouping principles responsible for that feat must be subtler than those entertained in Wertheimer's (1923/1938) collection.

Testing the General Auditory Account With Speech Sounds

The question motivating these studies is general: How does the listener perceive the properties of objects from their acoustic effects? Every account of auditory perceptual analysis attempts to explain how knowledge of worldly attributes results from excitation of the auditory system. However, to identify the distal correlates of a disorderly mixture of auditory elements, the acoustic inflow must be analyzed according to sources of sound in the world. The ease with which the perceiver finds perceptual streams in ordinary listening betrays the fundamental nature of this part of perception, and in psychological accounts it has seemed reasonable to suppose that this basic process can count very little on sensitivity to complex auditory properties. Rather, the simple criteria deriving from Gestalt principles of form have guided the exposition of auditory perceptual organization.

At this juncture, the archives are rich with organizational studies of arbitrary auditory forms rationalized by parameters designed by Wertheimer (1923/1938), outlined by Julesz and Hirsh (1972), and portrayed by Bregman and colleagues. Although this endeavor has produced results that are certainly consistent, as an empirical undertaking it stops short of the problem that it set for itself. The picture of auditory organization that this line of research created is largely free of ordinary objects known through auditory means. Although the promise is expressed regularly that this approach holds a key to organization according to sound sources, it has been rare for tests of grouping to use the acoustic effects of real events. The justifications may be numerous for an empirical emphasis on lab-made ensembles of tones and noises, yet the neglect of speech as a suitable prospect for the needed test is glaring. The acoustic attributes of speech signals are thoroughly familiar, and we have taken the extension of auditory organization to speech to be a convenient way of testing the general auditory approach with a natural source. The result of our review of the current conceptualizations of auditory organization is not encouraging. Speech, a commonplace sound, retains its perceptual integrity under conditions that clearly and consistently violate the hypothetical auditory criteria of grouping; the schema-based supplement fails to achieve its dual purpose of safeguarding the auditory primitives from falsification and of accommodating speech.

Speech appears to be organized by virtue of its characteristic spectro-temporal variation, which lies well beyond the simple rules and typicality schemas endorsed in the present version of auditory scene analysis. From the outset, we acknowledged that ordinary utterances comprise a variety of acoustic elements, periodic and aperiodic, continuous and discontinuous, exhibiting synchronous and asynchronous variation. From such observations, speech seemed a poor candidate for perceptual organization by rules that admit little tolerance for dissimilarity of any kind among components of a single stream. Although grouping by one or another manner of likeness has found much support in the technical literature on tones and noises, the reliance on such metrics to identify the vocal products of a single talker must fail, according to our acoustic audit of speech, stripping consonantal release bursts, voice bars, fricative formants, and nasal murmurs from the voiced formants.

Sinusoids exhibiting the frequency and amplitude variation of the formants yet lacking their fine-grain acoustic structure were treated as functional resonances, grouped together to produce phonetic perception in the absence of comodulation. The differential grouping of tones with speechlike variation, observed in our third experiment, indicates that perceptual organization can be keyed to attributes of the spectro-temporal modulation characteristic of vocal sources. It is tempting to suppose that perceptual organization specifically accommodates vocal sound, for a spoken source of sound is both unique, an anatomical resonator that exhibits graded and quantal acoustic effects of continuous articulatory action (Stevens, 1972), and ubiquitous in human auditory experience.

The prospects of a domain-independent auditory approach to perceptual organization seem no better in consequence of our attempt to lend content to its claims. There still is need of a convincing case that the Gestalt principles, whether supplemented by schemas or not, are able to approximate a perceptual stream containing all and only the natural acoustic products of a worldly source of sound. This is not to propose that oscillators, noise generators, digital filters, and anechoic chambers are unnatural constituents of a modern laboratory, only that idealized acoustic patterns existing solely in the perfect world populated by these devices fail to map the breadth of actual cases of perceptual organization. It is precisely the lack of correspondence between the domains of the laboratory and of the world that makes auditory scene analysis a portrait limited to the domain of the lab, despite an emphatic insistence on domain independence as the goal of the enterprise. Natural sources of sound in ordinary listening pose a substantial challenge, as we see here, and an account of perceptual organization that fails the world must be discarded, however comprehensively it deals with arbitrary phenomena in vitro.

On the Independence of Phonetic and Auditory Organization

Our motive to test the general auditory view stemmed from the opportunity to evaluate the application of common fate in the grouping of voiced formants. Evidence of the independence of phonetic and auditory organization is the bonus obtained by using sine-wave utterances in our attempt, regardless of how each organization is accomplished perceptually. When phonetic properties were evoked by sinusoidal ensembles, the impres-

sions of incoherent unspeechlike tones occurred, indicating that a sine-wave sentence satisfies criteria for two modes of organization simultaneously, the phonetic and the auditory. This property of sine-wave replicas shows that phonetic organization diverges from auditory scene analysis early in perception and proceeds independently (as in the case of duplex perception; see Whalen & Liberman, 1987). Presumably, phonetic principles of organization find a single speech stream, whereas auditory principles find several simultaneous whistles. Our dichotic conditions exaggerate this effect in which the components contributing phonetic information occur at different locations.

Given an unmistakable rift between auditory and phonetic functions, we must agree with Mattingly and Liberman (1990): Scene analysis and phonetic organization occur independently when the components of an acoustic display give rise simultaneously to impressions of a single phonetic and multiple auditory sources. This independence of the two organizations of sine-wave sentences hampers the generalization of our conclusions about auditory perceptual organization from the case of articulating vocal tracts to the broader class of natural sound sources. Additional tests with nonphonetic natural sources of sound are necessary to assess the potential of the Gestalt-based approach for parsing the auditory world into streams issuing from natural sound sources, as opposed to the simpler charge of organizing ambiguous auditory forms. Nonetheless, the example of speech may prove instructive. In the speech case, the Gestalt rules were proven to be inadequate to the spectro-temporal variation comprised in an ordinary utterance. A more general way to put this fact is to note that simple mechanical changes in the vocal source can create multiple acoustic effects for the perceiver (Rubin, Baer, & Mermelstein, 1981). If this formula happens to characterize the production of any nonvocal, nonphonetic sound, we may simply expect the Gestalt rules to fail for them as they do for speech. However, the simultaneous, incompatible phonetic and auditory organizations in the case of sine-wave sentences can only discourage a search for common primitives in the perception of speech and of auditory forms. We conclude, then, that the Gestalt-based auditory account of perceptual organization fails our tests on two challenges: (a) It is unable to rationalize the grouping of acoustic elements composing a speech signal, and (b) it prohibits the action of a phonetic mode of organization independent of the auditory mode.

Perceptual Organization of Speech: Natural Speech and Sine-Wave Replicas

The tests that we performed in evaluating the auditory principles of organization used exotic acoustic displays in comparison with the acoustic entities that confront the listener in ordinary conversation. However, it is plain that phonetic principles of perceptual organization apply no less to common cases than to our uncommon ones. Neither the acoustic elements nor the spectral shapes typical of speech are unique to speech. Rather, the patterned frequency variation of the spectral peaks is thought to be definitive of speech signals (Mattingly & Liberman, 1990; Remez, 1987; Stevens & Blumstein, 1981), as are the discontinuities due to articulatory gestures with quantal acoustic effects. Without the disposition to treat the signal as the product of a complex object, one with multiple sources of excitation and multiple reshapable resonant chambers, and to ignore the dissimilarities and discontinuities among acoustic constituents, the speech signal would fracture into the bits that the general account warrants.

From this perspective, sine-wave replicas probably do not enjoy a special perceptual accommodation, as if our listeners improvised a method to fuse the components of an atypical signal. Conceivably, the perceiver exploits the time-varying properties preserved from the natural utterances by sinusoidal replicas. In this way, we suppose that the perceiver conjures no special resources to handle the tone ensembles, which are abstractly phonetic by design. Overall, we conclude that the sine-wave replicas demonstrate grouping by phonetic rather than by simple auditory coherence, making use of organizational principles that ordinarily function in the perception of natural speech.

Speech Mode and Modularity

Our findings register a claim about the specific requirements for the perceptual organization of speech. With this step, we join a lively debate about the nature of speech perception that has drawn evidence from biology, engineering, linguistics, neurology, philosophy, and psychology (Mattingly & Studdert-Kennedy, 1991). The pertinent antecedent to our account of perceptual organization is a description by Mattingly and Liberman (1990) of the phonetic module, an autonomous perceptual resource independent of the functions of scene analysis and potentially preemptive of the open systems mediating the auditory recognition of objects. We grant that the findings of the three studies reported here add weight to the claim of independence of phonetic perception from other auditory functions, for in every case of sinusoidal presentation, our listeners heard multiple tones, each a distinct sound source, whereas phonetic fusion resulted in impressions of a single source of consonants and vowels (see Remez, Pardo, & Rubin, 1993). Moreover, our findings are consistent with the early separation of speech perception from auditory perception, warranted in the modular account. Does this mean that the phonetic and auditory impressions evoked simultaneously by sine-wave sentences disclose two modes of perceptual organization or two dedicated perceptual modules?

Collectively, our findings are consistent with some of the criteria of modular function defined by Fodor (1983): The processes of phonetic and auditory organization are fast and domain specific. That the action is fast we can infer, at least for phonetic processes that are pegged to occur within the time limit estimated by the decay of the auditory trace (for instance, Pisoni, 1973). That the processes are domain specific is warranted both for auditory perceptual organization and phonetic perception. The former identifies sound sources in locations in which amplitude or phase differences between the ears are deterministically projected into values of direction and elevation. The function of phonetic organization, as we have argued, is engaged by the occurrence of spectro-temporal patterns specific to linguistically governed vocalization. In this regard, the phenomena seem to be manifestations of modular activity. On other grounds, however, we see that the perception of sine-wave sentences is not consistent with three definitive properties of perceptual modules: Phonetic perception of tone replicas is not

mandatory, nor is the process informationally encapsulated, and, most obviously, we did not observe limited central access to underlying representations, for the sine-wave patterns that elicit phonetic impressions also are readily apparent as ensembles of concurrently changing tones. However, this last inconsistency with the principles of modular function only has force if the resources serving phonetic perception also produce auditory impressions of tone timbre, which seems unlikely. The violation of mandatoriness and encapsulation are more serious to consider.

The incomplete satisfaction of the criteria of modularity is due primarily to the finding that instructions greatly influence the perceptual treatment of sine-wave sentences (Remez et al., 1981). When listeners were asked simply to describe their spontaneous impressions of the tone patterns delivered through headphones, they rarely heard sentences, reporting instead a variety of mechanical, electronic, and avian attributes. However, when a second group of listeners was explicitly instructed to transcribe "the speech of a talking computer," reports of the sentence were quite accurate under the identical acoustic conditions that had elicited so few spontaneous reports of linguistic attributes. Yet a third group, which was requested to verify that the computer had produced a particular sentence, reported that they heard most of the words. Now, the hypothetical phonetic module is an input system and necessarily receives the same sensory inflow in all three conditions of instruction. If the action of the module were mandatory, we would observe its effects regardless of instructions, for the raw properties of sensory excitation must determine whether a module operates or not and whether it completes its analysis or passes. By the premise of encapsulated function, too, the phonetic module would pursue the same course regardless of instructions, which should no more influence an instance of phonetic perception than an instance of localization. If the beliefs of listeners facilitate or suppress action of the phonetic module, then speech uses a different kind of input system than Mattingly and Liberman (1990) or Fodor (1983) allow under the modular rubric.

What is the nature, after all, of phonetic perception: mode or module? This question sharpens the point on an issue that we have parried throughout our discussion of the significance of sinusoidal sentences. Namely, the success of sinusoidal vehicles in eliciting phonetic perception at all is the undoing of a straightforward acoustic description of the causal properties of speech perception. If a perceiver is satisfied by modulations of an impossibly unspeechlike carrier, then no acoustic element plays an essential role in perception, and the acoustic grit of speech must merely provide the opportunity for a listener to detect perceptually crucial properties of spectro-temporal variation. The effects of instruction appear to inflect this function by the allocation of attention. The most pressing matter left to us in describing the phonetic perception of sinusoidal replicas is to understand the bifurcation in perceptual state that occurs when these unspeechlike patterns are transformed from exclusively auditory impressions into tones with phonetic attributes. Presently, our results encourage a conceptualization of phonetic perception independent of general auditory processes, as if there were two distinct modes of perception, although our evidence of instructional influence does not corroborate any orthodox rendition of modularity.

Multimodal Coherence

What will an account of auditory perceptual organization ultimately contain? In part, this depends on tests analogous to ours from which to determine whether Gestalt-based principles, with realistic schematic assistance, can manage the acoustic complexity of natural sound sources. The suitability of these principles for organization in the perception of objects requires a search for evidence, therefore, and can neither be asserted from existing studies of ideal auditory forms nor denied, except in the case of phonetic sources, on the basis of our findings.

In addition, what will the ultimate account of organization of phonetic sources of sound contain? Although it is surely premature to propose a specific model, some of the attributes of phonetic organization are well enough resolved to enumerate: (a) Phonetic organization exploits speechlike acoustic variation independent of short-term spectra, (b) phonetic organization occurs rapidly and analytically, (c) the grain of phonetic organization is nonsymbolic and does not derive linguistic attributes, and (d) phonetic organization need not be learned. For now, these attributes suffice to guide research on the perceptual organization of speech, although in aiming for an effective account, larger considerations also apply.

Foremost in this regard is the likelihood that the auditory organization of speech is simply a single aspect of multimodal organization when the objects of perception happen to be utterances. In such circumstances, perception through auditory and visual modalities come to be aligned, as we have seen in a few studies of the past decade that have set the organization problem multimodally: The perceiver watches the talker while listening (Kuhl & Meltzoff, 1988; Rosen, Fourcin, & Moore, 1981; Sams et al., 1991; Summerfield, 1991). Despite the differences in the way each modality provides information, the now-and-then visible structures of articulation and the acoustic signal of speech combine perceptually as if organized in common to promote multimodal perceptual analysis. In some uses of this perceptual task, bimodal phonetic perception proves to be good, even when the displays are manipulated experimentally to make the auditory and visual components separately useless for conveying consonants, vowels, and words (see Bernstein, 1989). The specific points of correspondence between the modalities that are used to establish perceptual organization are simply unknown, although the tolerance for discrepancy without loss of coherence is surprisingly great (Summerfield & McGrath, 1984; see also Fowler & Dekle, 1991). Naturally, it is difficult to imagine an explanation of these phenomena by means of phonetic criteria applied serially to each modality. Instead, it is appealing to suppose that the criteria for the perceptual organization of speech—visible, audible, and even palpable—are actually specified in a general form, removed from any particular sensory modality, and are available to each when the world warrants it.

References

Assmann, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequency. *Journal of the Acoustical Society of America, 88,* 680–697.

Bailey, P. J., & Summerfield, Q. (1980). Information in speech: Observations on the perception of [s]-stop clusters. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 536–563.

Bashford, J. A., Jr., & Warren, R. M. (1979). Perceptual synthesis of deleted phonemes. In J. J. Wolf & D. H. Klatt (Eds.), *Speech communication papers* (pp. 423–426). New York: Acoustical Society of America.

Bernstein, L. E. (1989). Independent or dependent feature evaluation: A question of stimulus characteristics. *Behavioral and Brain Sciences, 12,* 756–757.

Best, C. T., Studdert-Kennedy, M., Manuel, S., & Rubin-Spitz, J. (1989). Discovering phonetic coherence in acoustic patterns. *Perception & Psychophysics, 45,* 237–250.

Blank, M. A. (1979). *Dual-mode processing of phonemes in fluent speech.* Unpublished doctoral dissertation, University of Texas, Austin.

Bregman, A. S. (1978a). Auditory streaming is cumulative. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 380–387.

Bregman, A. S. (1978b). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance, VII* (pp. 63–76). Hillsdale, NJ: Erlbaum.

Bregman, A. S. (1981). Asking the "what for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 98–118). Hillsdale, NJ: Erlbaum.

Bregman, A. S. (1987). The meaning of duplex perception: Sounds as transparent objects. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 95–111). Dordrecht, The Netherlands: Martinus Nijhoff.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics, 37,* 483–493.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89,* 244–249.

Bregman, A. S., & Dannenbring, G. L. (1973). The effect of continuity on auditory stream segregation. *Perception & Psychophysics, 13,* 308–312.

Bregman, A. S., & Dannenbring, G. L. (1977). Auditory continuity and amplitude edges. *Canadian Journal of Psychology, 31,* 151–158.

Bregman, A. S., & Doehring, P. (1984). Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics, 36,* 251–256.

Bregman, A. S., Levitan, R., & Liao, C. (1990). Fusion of auditory components: Effects of the frequency of amplitude modulation. *Perception & Psychophysics, 47,* 68–73.

Bregman, A. S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology, 32,* 19–31.

Broadbent, D. E., & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America, 29,* 708–710.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10,* 23–36.

Brown, A. M., & Kemp, D. T. (1984). Suppressibility of the 2f1—f2 stimulated acoustic emissions in gerbil and man. *Hearing Research, 13,* 29–37.

Butler, R. A., Levy, E. T., & Neff, W. D. (1980). Apparent distance of sounds recorded in echoic and anechoic chambers. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 745–750.

Carhart, R., Johnson, C., & Goodman, J. (1975). Perceptual masking of spondees by combinations of talkers. *Journal of the Acoustical Society of America, 58,* S35.

Carhart, R., & Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Archives of Otolaryngology, 91,* 273–279.

Carrell, T. D., & Opie, J. M. (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception & Psychophysics, 52,* 437–445.

Cherry, C. (1953). Some experiments on the recognition of speech with one and with two ears. *Journal of the Acoustical Society of America, 25,* 975–979.

Ciocca, V., & Bregman, A. S. (1987). Perceived continuity of gliding and steadystate tones through interrupting noise. *Perception & Psychophysics, 42,* 476–484.

Cole, R. A., & Scott, B. (1973). Perception of temporal order in speech: The role of vowel transitions. *Perception & Psychophysics, 27,* 441–449.

Cudahy, E., & Leshowitz, B. (1974). Effects of contralateral interference tone on auditory recognition. *Perception & Psychophysics, 15,* 16–20.

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

Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology, 30,* 99–114.

Dannenbring, G. L., & Bregman, A. S. (1976). Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 544–555.

Dannenbring, G. L., & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics, 24,* 369–376.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology, 33A,* 185–208.

Darwin, C. J., & Baddeley, A. D. (1974). Acoustic memory and the perception of speech. *Cognitive Psychology, 6,* 41–60.

Darwin, C. J., & Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America, 79,* 838–845.

Darwin, C. J., & Gardner, R. B. (1987). Perceptual separation of vowels from concurrent sounds. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 112–124). Dordrecht, The Netherlands: Martinus Nijhoff.

Darwin, C. J., Pattison, H., & Gardner, R. B. (1989). Vowel quality changes produced by surrounding tone sequences. *Perception & Psychophysics, 45,* 333–342.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology, 36A,* 193–208.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769–773.

Deutsch, D. (1982). Grouping mechanisms in music. In D. Deutsch (Ed.), *The psychology of music* (pp. 99–134). San Diego, CA: Academic Press.

Dorman, M. F., Cutting, J. E., & Raphael, L. J. (1975). Perception of temporal order in vowel sequences with and without formant transitions. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 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.

Ehrenfels, C. von. (1890). Ueber Gestaltqualiïtaten [On Gestalt qualities]. *Vierteljahrsschrift fur wissenschaftliche Philosophie, 14,* 249–292.

Eimas, P. D., & Miller, J. L. (1992). Organization in the perception of speech by young infants. *Psychological Science, 3,* 340–345.

Elliot, L. L. (1967). Development of auditory narrow-band frequency contours. *Journal of the Acoustical Society of America, 42,* 143–153.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech waveform. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 360–385). Hillsdale, NJ: Erlbaum.

Fant, C. G. M. (1956). On the predictability of formant levels and spectrum envelopes from formant frequencies. In M. Halle, H. Lunt, & H. MacLean (Eds.), *For Roman Jakobson* (pp. 109–120). The Hague, The Netherlands: Mouton.

Fant, C. G. M. (1962). Descriptive analysis of the acoustic aspects of speech. *Logos, 5,* 3–17.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816–828.

French-St. George, M., & Bregman, A. S. (1989). Role of predictability of sequence in auditory stream segregation. *Perception & Psychophysics, 46,* 384–386.

Gardner, R. B., & Darwin, C. J. (1986). Grouping of vowel harmonics by frequency modulation: Absence of effects on phoneme categorization. *Perception & Psychophysics, 40,* 183–187.

Gardner, R. B., Gaskill, S. A., & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America, 85,* 1329–1337.

Goldstein, L. M., & Lackner, J. R. (1973). Alteration of the phonetic coding of speech sounds during repetition. *Cognition, 2,* 279–297.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America, 61,* 1493–1500.

Habersetzer, J., & Vogler, B. (1983). Discrimination of surface-structured targets by the echolocating bat *Myotis myotis* during flight. *Journal of Comparative Physiology, 152,* 275–282.

Hall, J. W. III, & Grose, J. H. (1990). Comodulation masking release and auditory grouping. *Journal of the Acoustical Society of America, 88,* 119–125.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. *Journal of the Acoustical Society of America, 76,* 50–56.

Handel, S., Weaver, M. S., & Lawson, G. (1983). Effect of rhythmic grouping on stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 637–651.

Harris, K. S. (1958). Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech, 1,* 1–7.

Hirsh, I. J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America, 31,* 759–767.

Hochberg, J. (1974). Organization and the Gestalt tradition. In E. C. Carterette & M. P. Friedman (Eds.), *Handbook of perception: Vol. 1. Historical and philosophical roots of perception* (pp. 179–210). San Diego, CA: Academic Press.

Jenkins, J. J., Strange, W., & Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Perception & Psychophysics, 34,* 441–450.

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review, 83,* 323–335.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review, 96,* 459–491.

Julesz, B., & Hirsh, I. J. (1972). Visual and auditory perception: An essay of comparison. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 283–340). New York: McGraw-Hill.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1351.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America, 73,* 322–335.

Kim, D. O., Molnar, C. E., & Matthews, J. W. (1980). Cochlear mechanics: Nonlinear behavior in two-tone responses as reflected in cochlear-nerve-fiber responses and in ear-canal sound pressure. *Journal of the Acoustical Society of America, 67,* 1704–1721.

Klatt, D. H. (1979). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 243–288). Hillsdale, NJ: Erlbaum.

Klatt, D. H. (1980). Software for a cascade-parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971–995.

Klatt, D. H. (1989). Review of selected models of speech perception. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 169–226). Cambridge, MA: MIT Press.

Koffka, K. (1935). *Principles of Gestalt psychology.* New York: Harcourt, Brace.

Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: The Minnesota Symposium on Child Psychology* (Vol. 20, pp. 235–266). Hillsdale, NJ: Erlbaum.

Lackner, J. R., & Goldstein, L. M. (1974). Primary auditory stream segregation of repeated consonant-vowel sequences. *Journal of the Acoustical Society of America, 56,* 1651–1652.

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. *Journal of the Acoustical Society of America, 90,* 3036–3049.

Liberman, A. M. (1970). The grammars of speech and language. *Cognitive Psychology, 1,* 301–323.

Liberman, A. M., & Cooper, F. S. (1972). In search of the acoustic cues. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre* (pp. 329–338). The Hague, The Netherlands: Mouton.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1958). Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech, 1,* 153–167.

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics, 30,* 133–143.

Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243,* 489–494.

Liberman, A. M., Mattingly, I. G., & Turvey, M. T. (1972). Language codes and memory codes. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 307–334). New York: Holt, Rinehart & Winston.

Liberman, A. M., & Studdert-Kennedy, M. (1978). Phonetic perception. In R. Held, H. Leibowitz, & H.-L. Teuber (Eds.), *Handbook of sensory physiology: Vol. 8. Perception* (pp. 143–178). New York: Springer-Verlag.

Markel, J. D., & Gray, A. H., Jr. (1976). *Linear prediction of speech.* New York: Springer-Verlag.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological enquiry.* Hillsdale, NJ: Erlbaum.

Mattingly, I. G., & Liberman, A. M. (1990). Speech and other auditory modules. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Signal and sense: Local and global order in perceptual maps* (pp. 501–520). New York: Wiley.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. G.

(1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2,* 131–157.

Mattingly, I. G., & Studdert-Kennedy, M. (1991). *Modularity and the motor theory of speech perception: Proceedings of a conference to honor Alvin M.* Liberman. Hillsdale, NJ: Erlbaum.

McAdams, S. (1989). Segregation of concurrent sounds: I. Effects of frequency modulation coherence. *Journal of the Acoustical Society of America, 86,* 2148–2159.

McFadden, D. (1986). Co-modulation masking release: Effects of varying the level, duration and time delay of the cue band. *Journal of the Acoustical Society of America, 80,* 1658–1667.

Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America, 22,* 637–638.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22,* 167–173.

Nygaard, L. C. (1993). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 268–286.

Nygaard, L. C., & Eimas, P. D. (1990). A new version of duplex perception: Evidence for phonetic and nonphonetic fusion. *Journal of the Acoustical Society of America, 88,* 75–86.

Palmer, A. R., Winter, I. M., Gardner, R. B., & Darwin, C. J. (1987). Changes in the phonemic quality and neural representation of a vowel by alteration of the relative phase of harmonics near F1. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 371–376). Dordrecht, The Netherlands: Martinus Nijhoff.

Palmer, C., & Krumhansl, C. (1987). Independent temporal and pitch structures in perception of musical phrases. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 116–126.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America, 60,* 911–918.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics, 13,* 253–260.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 6. Linguistic functions in cognitive theory* (pp. 167–233). Hillsdale, NJ: Erlbaum.

Pisoni, D. B. (1987). Some measures of intelligibility and comprehension. In J. Allen, S. Hunnicut, & D. H. Klatt (Eds.), *From text to speech: The MITalk system* (pp. 151–171). Cambridge, England: Cambridge University Press.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America, 55,* 678–680.

Remez, R. E. (1987). Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 419–432). Dordrecht, The Netherlands: Martinus Nijhoff.

Remez, R. E. (in press). A guide to research on the perception of speech. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics.* San Diego, CA: Academic Press.

Remez, R. E., Pardo, J. S., & Rubin, P. E. (1993). *Making the auditory scene with speech.* Manuscript submitted for publication.

Remez, R. E., & Rubin, P. E. (1983). The stream of speech. *Scandinavian Journal of Psychology, 24,* 63–66.

Remez, R. E., & Rubin, P. E. (1984). Perception of intonation in sinusoidal sentences. *Perception & Psychophysics, 35,* 429–440.

Remez, R. E., & Rubin, P. E. (1990). On the perception of speech from time-varying attributes: Contributions of amplitude variation. *Perception & Psychophysics, 48,* 313–325.

Remez, R. E., & Rubin, P. E. (in press). Acoustic shards, perceptual glue. In P. A. Luce & J. Charles-Luce (Eds.), *Proceedings of a workshop on spoken language.* Norwood, NJ: Ablex.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 40–61.

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

Repp, B. H. (1987). The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America, 81,* 1100–1109.

Repp, B. H. (1992). Toward an emancipation of the "weaker sense." *Psychological Science, 2,* 382–386.

Repp, B. H., & Bentin, S. (1984). Parameters of spectral/temporal fusion in speech perception. *Perception & Psychophysics, 36,* 523–530.

Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. *Perception & Psychophysics, 33,* 333–337.

Riquimaroux, H., Gaioni, S. J., & Suga, N. (1991). Cortical computational maps control auditory perception. *Science, 251,* 565–568.

Rock, I., & Brosgole, L. (1964). Grouping based on phenomenal proximity. *Journal of Experimental Psychology, 67,* 531–538.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature, 291,* 150–152.

Rosen, S., & Howell, P. (1987). Is there a natural sensitivity at 20 ms in relative-tone-onset-time continua? A reanalysis of Hirsh's (1959) data. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 199–209). Dordrecht, The Netherlands: Martinus Nijhoff

Rubin, P. E. (1980). *Sinewave synthesis.* Internal memorandum, Haskins Laboratories, New Haven, Connecticut.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70,* 321–328.

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S.-T., & Simola, J. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127,* 141–145.

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General, 110,* 474–494.

Sasaki, T. (1980). Sound restoration and temporal localization of noise in speech and music sounds. *Tohuku Psychologica Folia, 39,* 79–88.

Schmidt, S. (1988). Evidence for a spectral basis of texture perception in bat sonar. *Nature, 331,* 617–619.

Sherman, G. L. (1973). *Studies of the temporal sequence of speech perception at different linguistic levels.* Unpublished doctoral dissertation, University of Wisconsin, Milwaukee.

Steiger, H., & Bregman, A. S. (1981). Capturing frequency components of glided tones: Frequency separation, orientation, and alignment. *Perception & Psychophysics, 30,* 425–435.

Steiger, H., & Bregman, A. S. (1982). Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics, 32,* 153–162.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr., & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51–66). New York: McGraw-Hill.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives in the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Summerfield, Q. (1991). Visual perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception: Proceedings of a conference to honor Alvin M. Liberman* (pp. 117–137). Hillsdale, NJ: Erlbaum.

Summerfield, Q., & Assmann, P. F. (1989). Auditory enhancement of the perception of concurrent vowels. *Perception & Psychophysics, 45,* 529–536.

Summerfield, Q., & Culling, J. F. (1992). Auditory segregation of competing voices: Absence of effects of FM or AM coherence. *Philosophical Transactions of the Royal Society of London (B), 336,* 357–366.

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36A,* 51–74.

Tougas, Y., & Bregman, A. S. (1985). The crossing of auditory streams. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 788–798.

Turvey, M. T. (1978). Visual processing and short-term memory. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Vol. 5. Human information processing* (pp. 91–142). Hillsdale, NJ: Erlbaum.

Van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences.* Unpublished doctoral dissertation, Eindhoven University of Technology, The Netherlands.

Vicario, G. (1960). L'effetto tunnel acustico [The acoustic tunnel effect]. *Rivista di Psicologia, 54,* 41–52.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 393–395.

Warren, R. M. (1984). Perceptual restoration of obliterated sounds. *Psychological Bulletin, 96,* 371–383.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. (1969). Auditory sequence: Confusion of patterns other than speech or music. *Science, 164,* 586–587.

Warren, R. M., & Warren, R. P. (1970). Auditory illusions and confusions. *Scientific American, 223,* 30–36.

Warren, W. H., Jr., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing: A case study in ecological acoustics. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 704–712.

Weintraub, M. (1987). Sound separation and auditory perceptual organization. In M. E. H. Schouten (Ed.), *Psychophysics of speech perception* (pp. 125–134). Dordrecht, The Netherlands: Martinus Nijhoff.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A sourcebook of Gestalt psychology* (pp. 71–88). London: Kegan Paul, Trench, & Trubner. (Original work published 1923)

Whalen, D. H., & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science, 237,* 169–171.

Yost, W. A., Sheft, S., & Opie, J. (1989). Modulation interference in detection and discrimination of amplitude modulation. *Journal of the Acoustical Society of America, 86,* 2138–2147.

Young, L. L., Jr., Parker, C., & Carhart, R. (1975). Effectiveness of speech and noise maskers on numbers embedded in continuous discourse. *Journal of the Acoustical Society of America, 58,* S35.

Zue, V. W., & Schwartz, R. M. (1980). Acoustic processing and phonetic analysis. In W. A. Lea (Ed.), *Trends in speech recognition* (pp. 101–124). Englewood Cliffs, NJ: Prentice Hall.

Received February 10, 1992
Revision received May 7, 1993
Accepted July 22, 1993 ■