2 Perceptual Organization of Speech

ROBERT E. REMEZ

How does a perceiver resolve the linguistic properties of an utterance? This question has motivated many investigations within the study of speech perception and a great variety of explanations. In a retrospective summary 15 years ago, Klatt (1989) reviewed a large sample of theoretical descriptions of the perceiver's ability to project the sensory effects of speech, exhibiting inexhaustible variety, into a finite and small number of linguistically defined attributes, whether features, phones, phonemes, syllables, or words. Although he noted many distinctions among the accounts, with few exceptions they exhibited a common feature. Each presumed that perception begins with a speech signal, well-composed and fit to analyze. This common premise shared by otherwise divergent explanations of perception obliges the models to admit severe and unintended constraints on their applicability. To exist within the limits set by this simplifying assumption, the models are restricted to a domain in which speech is the only sound; moreover, only a single talker ever speaks at once. Although this designation is easily met in laboratory samples, it is safe to say that it is rare *in vivo*. Moreover, in their exclusive devotion to the perception of speech the models are tacitly modular (Fodor, 1983), whether or not they acknowledge it.

Despite the consequences of this dedication of perceptual models to speech and speech alone, there has been a plausible and convenient way to persist in invoking the simplifying assumption. This fundamental premise survives intact if a preliminary process of perceptual organization finds a speech signal, follows its patterned variation amid the effects of other sound sources, and delivers it whole and ready to analyze for linguistic properties. The indifference to the conditions imposed by the common perspective reflects an apparent consensus that perceptual organization of speech is simple, automatic, and accomplished by generic means. However, despite the rapidly established perceptual coherence of the constituents of a speech signal, the perceptual organization of speech cannot be reduced to the available and well-established principles of auditory perceptual organization.

In D.B.Pisoni and R.E. Remez (Eds.) Handbook of Speech Perception. Oxford: Blackwell, 2005.

2.1 Perceptual Organization and the Gestalt Legacy

2.1.1 *A generic auditory model of organization*

The dominant contemporary account of auditory perceptual organization is Auditory Scene Analysis (Bregman, 1990). This theory of the resolution of auditory sensation into streams, each issuing from a distinct source, developed empirically in the past 30 years, though its intellectual roots run deep. The Gestalt psychologist Wertheimer (1923/1938) established the basic premises of the account in a legendary article, the contents of which are roughly known to all students of introductory psychology. In visible and audible examples, Wertheimer described the coalescence of elementary figures into groups and contours, arguing that sensory experience is organized in patterns, and is not registered as a mere spatter of individual receptor states. By considering a series of hypothetical cases, and without knowing the sensory physiology that would not be described for decades (Mountcastle, 1998), he justified organizing principles of *similarity, proximity, closure, symmetry, common fate, continuity, set,* and *habit.* Hindsight suggests that Wertheimer framed the problem astutely, given our contemporary understanding of the functions of the sensory periphery that integrate the action of visual and auditory receptors (Hochberg, 1974).

Setting the indefinitely elastic principle of habit aside, the simple Gestalt-derived criteria of grouping are arguably reducible to two functions: (1) to compose an inventory of sensory elements; and (2) to create contours or groups on the principle that like binds to like. Whether groups occur due to the spectral composition of auditory elements, their common on- or offset, proximity in frequency, symmetry of rate of change in an auditory dimension, harmonic relationship, or the interpolation of brief gaps, and so on, each is readily understood as a case in which similarity among a set of auditory sensory elements promotes grouping. A group composed according to these functions forms a sensory contour or perceptual stream. It is a small but necessary extrapolation to assert that an auditory contour consists of elements originating from a single source of sound, and therefore that perceptual organization parses sensory experience into concurrent streams each issuing from a different sound producing event (Bregman & Pinker, 1978).

In a series of ongoing experiments, researchers adopted Wertheimer's auditory conjectures, and calibrated the resolution of auditory streams by virtue of the principles and their corollaries. For example, Bregman and Campbell (1971) reported that auditory streams formed when a sequence of 100 ms tones differing in frequency was presented to listeners. According to a procedure that has become standard, the series of brief tones was presented repetitively to listeners, who were asked to report the order of tones in the series. Instead of hearing a sequence of high and low pitches, though, listeners grouped tones into two streams each composed of similar elements, one of high pitch and another of low (see Figure 2.1). Critically, the perception of the order of elements was veridical within streams, but perception of the intercalation order across the streams was erroneous. In another example, Bregman, Ahad, and Van Loon (2001) reported

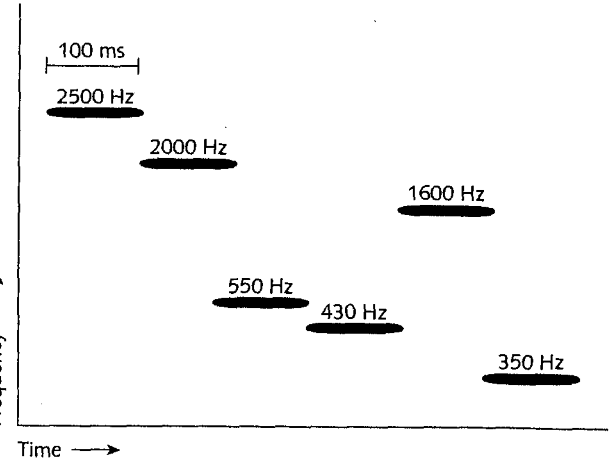

Figure 2.1 This sequence of tones presented to listeners by Bregman & Campbell (1971) was reported as two segregated streams, one of high and another of low tones. Critically, the intercalation of the high and low streams (that is, the sequence: high, high, low, low, high, low) was poorly resolved.

that a sequence of 65 ms bursts of band-limited noise were grouped together or split into separate perceptual streams as a function of the similarity in center frequency of the noise bursts. A sizable literature of empirical tests of this kind spans 40 years, and calibrates the sensory conditions of grouping by one or another variant of similarity. A compilation of the literature is offered by Bregman (1990), and the theoretical yield of this research is summarized by Darwin (1997).

Typically, studies of auditory perceptual organization have reported that listeners are sensitive to quite subtle properties in the formation of auditory groups. It is useful to consider an exemplary case, for the detailed findings of auditory amalgamation and segregation define the characteristics of the model and ultimately determine its applicability to speech. In a study of concurrent grouping of harmonically related tones by virtue of coincident onset, a variant of similarity in a temporal dimension, Dannenbring and Bregman (1978) reported that synchronized tones were grouped together, but a discrepancy as brief as 35 ms in lead or lag in one component was sufficient to disrupt coherence with other sensory constituents, and to split it into a separate stream. There are many similar cases documenting the exquisite sensitivity of the auditory sensory channel in segregating streams on the basis of slight departures from similarity: in *frequency* (Bregman & Campbell, 1971), in *frequency change* (Bregman & Doehring, 1984), in *fundamental frequency* (Steiger & Bregman, 1982), in *common modulation* (Bregman, Abramson, Doehring, & Darwin, 1985), in *spectrum* (Dannenbring & Bregman, 1976; Warren, Obusek, Farmer, & Warren, 1969), due to brief *interruptions* (Miller & Licklider, 1950), in *common onset/offset* (Bregman & Pinker, 1978), in *frequency continuity* (Bregman & Dannenbring, 1973, 1977), in *melody* and *meter* (Jones & Boltz, 1989); these are reviewed by Bregman (1990) and by Remez, Rubin, Berns, Pardo, and Lang (1994).

2.1.2 Gestalt principles of organization applied to speech

Because explanations of speech perception have depended on an unspecified account of perceptual organization, it has been natural to take Auditory Scene Analysis as a theory of first resort for understanding the perceptual solution to the cocktail party problem (Cherry, 1953), specifically, of attending to a single stream of speech amid other sound sources. However, this premise was largely unsupported by direct evidence. The crucial empirical cases that had formed the model had rarely included natural sources of sound, neither the instruments of the orchestra (though, see Iverson, 1995) which are well modeled physically (Rossing, 1990), nor ordinary mechanical sources (Gaver, 1993), nor the sounds of speech, with several provocative exceptions. It is instructive to consider some of the cases in which tests of perceptual organization using speech sounds appeared to confirm the applicability to speech of the general auditory account of perceptual organization.

In one case establishing grouping by similarity, a repeating series of syllables of the form CV–V–CV–V was observed to split into distinct streams of like syllables, one of CVs and another of Vs, much as Gestalt principles propose (Lackner & Goldstein, 1974). Critically, this perceptual organization precluded the perceptual resolution of the relative order of the syllables across streams, analogous to the index of grouping used by Bregman and Campbell (1971). In another case calibrating grouping by continuity, a series of vowels formed a single perceptual stream only when formant frequency transitions leading into and out of the vowel nuclei were present (Dorman, Cutting, & Raphael, 1975). Without smooth transitions, the spectral discontinuity at the juncture between successive steady-state vowels exceeded the tolerance for grouping by closure – that is, the interpolation of gaps – and the perceptual coherence of the vowel series was lost. In another case examining organization by the common fate, or similarity in change of a set of elements, a harmonic component of a steady-state vowel close to the center frequency of a formant was advanced or delayed in onset relative to the rest of the harmonics composing the synthetic vowel (Darwin & Sutherland, 1984). At a lead or lag of 32 ms, consistent with findings deriving from arbitrary patterns, the offset harmonic segregated into a different stream than the synchronous harmonics composing the vowel. In consequence, when the leading or lagging harmonic split, the height of the vowel was perceived to be different, as if the perceptual estimate of the center frequency of the first formant had depended on the grouping. In each of these instances, the findings with speech sounds were well explained by the precedents of prior tests using arbitrary patterns of sound created with oscillators and noise generators.

These outcomes should have seemed too good to be true. It was as if an account defined largely through tests of ideal notions of the resolution of similarity in simple auditory sequences proved to be adequate to accommodate the diverse acoustic constituents and spectral patterns of natural sound. With hindsight, we can see that accepting this conclusion does require one credulous assumption: namely, that tests using arbitrary trains of syllables, meticulously phased harmonic components, and sustained steady-state vowels adequately express the ordinary complexity of speech, and the perceiver's ordinary sensitivity. In short, a sufficient test of organization by the generic principles of Auditory Scene Analysis is obliged to incorporate the kind of variability that has defined the technical description of speech perception. And a closer approximation to the conditions of ordinary listening must motivate the empirical tests. By satisfying these constraints, a set of functions rather different from the generic auditory model can be seen at work in the perceptual organization of speech.

2.2 The Plausibility of the Generic Account of Perceptual Organization

2.2.1 *A brief review of the acoustic properties of speech*

One challenge of perceptual organization facing a listener is simple to state: To find and follow a speech stream. This would be an easy matter were the acoustic constituents of a speech signal or their auditory sensory correlates unique to speech; or if the speech signal were more or less stationary in its spectrum; or if the acoustic elements and the auditory impressions they evoke were similar, moment by moment. None of these is true, however, which inherently undermines the plausibility of any attempt to formalize perceptual organization of speech as a task of determining successive or simultaneous similarities in auditory experience. First, none of the multitude of naturally produced vocal sounds composing a speech signal is unique to speech. Arguably, the physical models of speech production succeed so well because they exploit an analogy between vocal sound and acoustic resonance (Fant, 1960; Stevens & House, 1961). Second, one signature aspect of speech is the presence of multiple acoustic maxima and minima in the spectrum, and the variation over time in the frequencies at which the acoustic energy is concentrated (Stevens & Blumstein, 1981). This frequency variation of the formant centers is interrupted at stop closures, creating an acoustic spectrum that is both nonstationary and discontinuous. Third, the complex pattern of articulation by which talkers produce consonant holds and approximations creates heterogeneous acoustic effects consisting of hisses, whistles, clicks, buzzes, and hums (Stevens, 1998). The resulting acoustic pattern of speech consists of a nonstationary, discontinuous series of periodic and aperiodic elements none of which in detail is unique to a vocal source.

The diversity of acoustic constituents of speech is readily resolved as a coherent stream, perceptually, though the means by which this occurs challenges the potential of the generic auditory account. Although some computational implementations of Gestalt grouping have disentangled spoken sources of simple nonstationary spectra (Parsons, 1976; Summerfield, 1992), these have occurred for a signal free of discontinuities, as occurs in the production of sustained, slowly changing vowels. Slow and sustained change in the spectrum, though, is hardly typical of ordinary speech which is characterized by consonant closures that impose rapid spectral changes and episodes of silence of varying duration. To resolve a signal despite silent discontinuities requires grouping by closure to extrapolate across brief silent gaps. To invoke generic auditory properties in providing this function would oppose present evidence, though. For example, in an empirical attempt to discover the standard for grouping by closure (Neff, Jesteadt & Brown, 1982) the temporal threshold for gap detection was found to diverge from the tolerance of discontinuity in grouping. It is unlikely, then, that a generic mechanism of extrapolation across gaps is responsible for the establishment of perceptual continuity, whether in auditory form or in the perception of speech.

Evidence from tests of auditory form suggest that harmonic relations – that is, sharing a fundamental frequency – and amplitude comodulation – that is, pulsing at a common rate – promote grouping albeit weakly (Bregman, Levitan, & Liao, 1990), and these two characteristics are manifest by oral and nasal resonances and by voiced frication. This might be the likeliest principle to explain the coherence of voiced speech by generic auditory means, for an appeal to similarity in frequency variation among the formants is unlikely to explain their coherence. Indeed, the pattern of frequency variation of the first formant typically differs from that of the second and neither the first nor second resembles the third, due to the different articulatory causes of each (Fant, 1960). To greatly simplify a complex relation, the center frequency of the first formant often varies with the opening and closing of the jaw, while the frequency of the second formant varies with the advancement and retraction of the tongue, and the frequency of the third formant alternates in its articulatory correlate. Accordingly, different patterns of frequency variation are observed in each resonance due to the relative independence of the control of these articulators (see Figure 2.2). Even were generic auditory functions to bind the comodulated formants into a single stream, without additional principles of perceptual organization, a generic Gestalt-derived parsing mechanism that aims to compose perceptual streams of similar auditory elements would fail; indeed, it would fracture the acoustically diverse components of a single speech signal into streams of similar elements, one of hisses, another of buzzes, a third of clicks, and so on, deriving an incoherent profusion of streams despite the common origin of the acoustic elements in phonologically governed sound production (Darwin & Gardner, 1986; Lackner & Goldstein, 1974; Remez et al., 1994). Apart from this consideration, in principle, a small empirical literature exists on which to base an adequate account of the perceptual organization of speech.

2.2.2 *A few clues*

There is a passage in Schubert's Symphony No. 8 in B minor (D. 759, the "Unfinished," measures 13–26 of the first movement) in which the parts played by oboe and clarinet, a unison melody, fuse so thoroughly that no trace of oboe or clarinet quality remains. This instance in which two sources of sound are treated perceptually as one led Broadbent and Ladefoged (1957) to attempt a study that offered a clue about the nature of perceptual organization of speech. Beginning with a synthetic sentence composed of two formants, they created two single formant patterns, one of the first formant and the other of the second, each excited at the

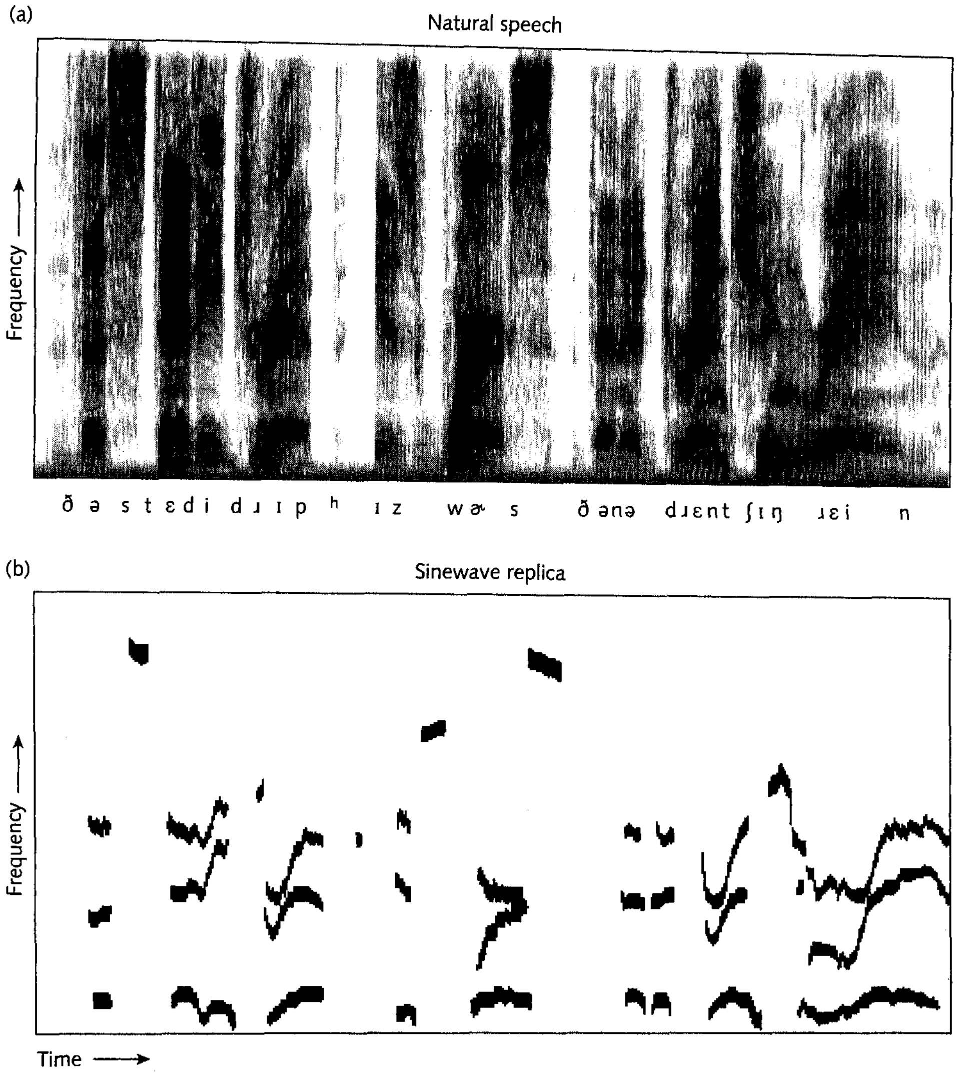

Figure 2.2 A comparison of natural and sinewave versions of the sentence, "The steady drip is worse than a drenching rain." (a) natural speech; (b) sinewave replica.

same fundamental frequency. Concurrently, the two formants evoked an impression of an English sentence; singly, each evoked an impression of an unintelligible buzz.

In one test condition, the formants were presented dichotically, in analogy to an oboe and a clarinet playing in unison. This resulted in perception of a single voice speaking the sentence, as if two spatially distinct sources had combined. Despite the dissimilarities in spatial locus of the components, this outcome is consistent with a generic auditory account of organization on grounds of harmonicity and amplitude comodulation. However, when each formant was rung on a different fundamental, subjects no longer reported a single voice, as if fusion failed to occur because neither harmonicity nor amplitude comodulation existed to oppose the spatial dissimilarity of the components. It is remarkable, nonetheless, that in view of these multiple lapses of similarity, subjects accurately reported the sentence, "What did you say before that?" although in this condition it seemed to be spoken by two talkers, one at each ear, each speaking at a different pitch. In other words, listeners reported divergent perceptual states: (1) the splitting of the auditory streams due to dissimilar pitch; and, (2) the combining of auditory streams to form speech. Although a generic Gestalt-derived account can explain a portion of the results, it cannot explain the combination of spatially and spectrally dissimilar formant patterns to compose a single speech stream.

In fine detail, research on perception in a speech mode also broached this topic, though indirectly. This line of research aimed to calibrate the difference in the resolution of auditory form and phonetic form of speech, thereby to identify psychoacoustic and psychophysical characteristics unique to speech perception. By opposing acoustic patterns evoking speech perception with nonspeech control patterns, the perceptual effect of variation in an acoustic correlate of a phonetic contrast was compared to the corresponding effect of the same acoustic property removed from the phonetically adequate context. For instance, Mattingly, Liberman, Syrdal, and Halwes (1971) examined the discriminability of a second formant frequency transition as an isolated acoustic pattern and within a synthetic syllable in which its variation was correlated with the perception of the *place of articulation* of a stop consonant. A finding of different psychophysical effect, roughly, Weber's law for auditory form and categorical perception for phonetic form, was taken as the signature of each perceptual mode. In a variant of the method specifically pertinent to the description of perceptual organization, Rand (1974) separated the second formant frequency transition, the correlate of the place contrast, from the remainder of a synthetic syllable and arrayed the acoustic components dichotically. In consequence, the critical second formant frequency transition presented to one ear was resolved as an auditory form while it also contributed to the phonetic contrast it evoked in apparent combination with the formant pattern presented to the other ear. In other words, with no change in the acoustic conditions, a listener could resolve the properties of the auditory form of the formant frequency transition or the phonetic contrast it evoked. The dichotic presentation permitted two perceptual organizations of the same element concurrently, due to the spatial and temporal disparity that blocked fusion on generic auditory principles, and due to the phonetic potential of the fused components. This phenomenon of concurrent auditory and phonetic effects of a single acoustic element was described as *duplex perception* (Liberman, Isenberg, & Rakerd, 1981; Nygaard, 1993; Whalen & Liberman, 1996) and it has been described as an effect of a peremptory aspect of phonetic organization and analysis.[1] No matter how the evidence ultimately adjudicates the psychophysical claims, it is instructive to note that the generic auditory functions of perceptual organization only succeed in rationalizing the split of the dichotic components into separate streams, and fail to provide a principle by which the combination of elements occurs.

2.2.3 Organization by coordinate variation

A classic understanding of the perception of speech derives from study of the acoustic correlates of phonetic contrasts and the physical and articulatory means by which they are produced (reviewed by Raphael, this volume; also, see Fant, 1960; Liberman, Ingemann, Lisker, Delattre, & Cooper, 1959; Stevens & House, 1961). In addition to calibrating the perceptual response to natural samples of speech, researchers also used acoustic signals produced synthetically in detailed psychoacoustic studies of phonetic identification and differentiation. In typical terminal analog speech synthesis, the short-term spectra characteristic of the natural samples are preserved, lending the synthesis a combination of natural vocal timbre and intelligibility (Sawusch, this volume). Acoustic analysis of speech and synthesis that allows parametric variation of speech acoustics have been important for understanding the normative aspects of perception, that is, the relation between the typical or likely auditory form of speech sounds encountered by listeners and the perceptual analysis of phonetic properties (Diehl, Molis, & Castleman, 2001; Lindblom, 1996; Massaro, 1994; Stevens, 1998).

However, a focus on natural samples and on synthetic idealizations of natural speech discounts the adaptability and versatility of speech perception, and draws attention from the properties of speech that are relevant to understanding perceptual organization. Because grossly distorted speech remains intelligible (for example, Licklider, 1946; Miller, 1946) when many of the typical acoustic correlates are absent, it is difficult to sustain the hypothesis that finding and following a speech stream crucially depends on meticulous registration of the brief and numerous acoustic correlates of phonetic contrasts described in classic studies. But, if the natural acoustic products of vocalization do not determine the perceptual organization and analysis of speech, what does?

An alternative to this conceptualization was prompted by the empirical use of a technique that combines digital analysis of speech spectra and digital synthesis of time-varying sinusoids (Remez, Rubin, Pisoni, & Carrell, 1981). This research has revealed the perceptual effectiveness of acoustic patterns that exhibit the gross spectrotemporal characteristics of speech without incorporating the fine acoustic structure of vocally produced sound. Perceptual research with these acoustic materials (and their relatives – noise band vocoded speech: Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; acoustic chimeras: Smith, Delgutte, & Oxenham, 2002; see, also, Remez, Yang, Piorkowski, Wissig, Batchelder, & Nam, 2002) has permitted an estimate of a listener's sensitivity to the time-varying patterns of speech spectra independent of the sensory elements that compose them.

The premise of sinewave replication is simple, though in practice it is as laborious as other forms of copy synthesis. Three or four tones, each approximating the center frequency and amplitude of an oral, nasal, or fricative resonance, are created to imitate the coarse grain attributes of a speech sample. Lacking the momentary aperiodicities, harmonic spectra, broadband formants, and regular pulsing of natural and most synthetic speech, a sinewave replica of an utterance differs acoustically and qualitatively from speech while remaining intelligible. A spectrogram of a sinewave sentence is shown in the bottom panel of Figure 2.2;

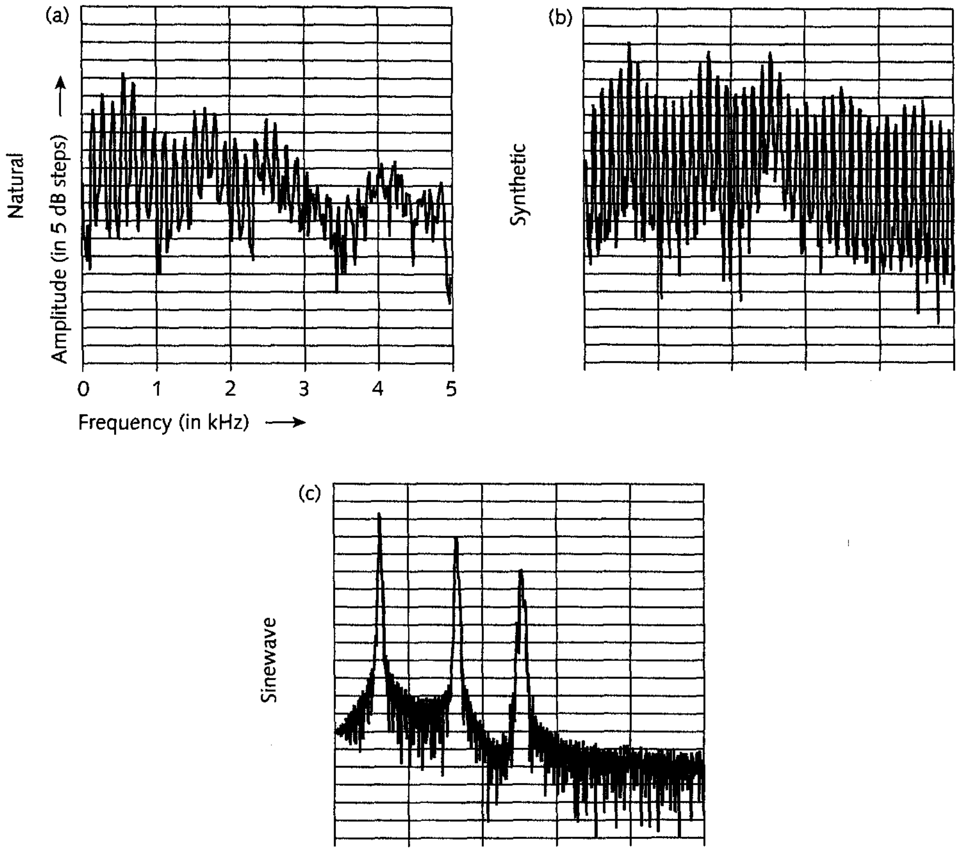

Figure 2.3 A comparison of the short-term spectrum of (a) natural speech; (b) terminal analog synthetic speech; and (c) sinewave replica. Note the broadband resonances and harmonic spectra in natural and synthetic speech, in contrast to the sparse, nonharmonic spectrum of the three tones.

a comparison of short-term spectra of natural speech and both synthetic and sinewave imitations is shown in Figure 2.3.

It is significant that three or four tones reproducing a natural formant pattern evoke an experience in a naive listener of several concurrent whistles changing in pitch and loudness, and do not automatically elicit an impression of speech. In other words, the immediate experience of the listener is accurately predicted by a generic auditory account, because acoustic elements that change frequency at different rates to different extents, onsetting and offsetting at different moments in different frequency ranges, are dissimilar along many dimensions that specify separate perceptual streams according to Gestalt principles. However, once instructed that the tones compose synthetic speech, a listener readily reports linguistic properties as if hearing the original natural utterance on which the sinewave replica was modeled. To be precise, intelligibility of sinewave speech is variable, and performance under different listening and instructional conditions has varied between 50% and 85% correct (Liebenthal, Binder, Piorkowski, & Remez, 2003; Remez et al., 1994). Within this range of performance levels, these acoustic conditions pose a crucial test of a Gestalt-derived account of perceptual organization, for a perceiver must integrate the tones in order to compose a single coherent speech stream, thereby resolving the linguistic properties of the signal. Several tests support this claim of true integration preliminary to analysis.

In direct assessments, the intelligibility of sinewave replicas of speech exceeded intelligibility predicted from the presentation of individual tones (Remez, Rubin, Nygaard, & Howell, 1987; Remez et al., 1981; Remez et al., 1994). This superadditive performance is evidence of integration, and it persisted even when the tones came from separate spatial sources, violating similarity in location (Remez et al., 1994; cf. Broadbent & Ladefoged, 1957). In combining the individual tones into a single time-varying coherent stream, however, this complex organization necessary for phonetic analysis does not exclude an auditory organization as independently resolvable streams of tones (Remez & Rubin, 1984, 1993). In fact, the perceiver's resolution of the pitch contour associated with the frequency pattern of tonal constituents is acute whether or not the fusion of the tones supporting phonetic perception occurs (Remez, Pardo, Piorkowski, & Rubin, 2001). On this evidence rests the claim that sinewave replicas are *bistable*, exhibiting two simultaneous and exclusive organizations.

Even if the processes by which these states occurred were strictly parallel, the bistable occurrence of auditory and phonetic perceptual organization is not amenable to further simplification. A sinewave replica of speech allows two organizations, much as the celebrated cases of visual bistability do: the duck-rabbit figure, Woodworth's equivocal staircase, Rubin's vase, and Necker's cube. Unlike the visual cases of alternating stability, the bistability that occurs in the perception of sinewave speech is simultaneous. A conservative description of these findings is that an organization of the auditory properties of sinewave signals occurs according to Gestalt-derived principles that promote integration or segregation; and, that phonetic perceptual analysis is incompatible with that organization. However, the concurrent variation of the tones satisfies a non-Gestalt principle of coordinate auditory variation despite local dissimilarities, and these promote integration of the components into a single stream. This organization is susceptible to phonetic analysis.

2.3 The Perceptual Organization of Speech

2.3.1 *Characteristics of the perceptual coherence of speech*

While much remains to discover about perceptual organization dependent on complex coordinate variation, research on the psychoacoustics and perception of speech from a variety of laboratories permits a rough sketch of the parameters. The portrait of perceptual organization offered here gathers evidence from different research programs that aimed to address a range of perceptual questions, for there is no unified attempt at present to understand the organization of perceptual streams that approach the complexity of speech. Overall, these results expose the perceptual organization of speech as fast, unlearned, nonsymbolic, keyed to complex patterns of sensory variation, indifferent to auditory quality, and requiring attention whether elicited or exerted.

The evidence that perceptual organization of speech is *fast* rests on long-established findings that the auditory trace of speech fades rapidly. Although estimates vary with the task used to calibrate the durability of unelaborated auditory sensation, all of the measures reflect the urgency with which the fading trace is recoded into a more stable phonetic form (Howell & Darwin, 1977; Pisoni & Tash, 1974). It is unlikely that much of the auditory form of speech persists beyond a tenth of a second, and it has decayed beyond access by 400 ms. The sensory integration required for perceptual organization is tied to this pace. Contrary to this notion of perceptual organization as exceedingly rapid, an extended version of Auditory Scene Analysis (Bregman, 1990) proposes a resort to a cognitive mechanism occurring well after primitive grouping takes place, to function as a supplement to the Gestalt-based mechanism. Such knowledge-based mechanisms are also featured as a method to resolve difficult grouping in recent artifactual approaches to perceptual organization (for example, Cooke & Ellis, 2001). However, the formal or practical advantages that this method achieves come at a clear cost, namely, to reject boundary conditions that subscribe to the natural auditory limits of perceptual organization.

The propensity to organize an auditory pattern by virtue of complex coordinate variation is apparently *unlearned,* or nearly so. In tests with infant listeners, 14-week-old subjects exhibited the pattern of adult sensitivity to dichotically arrayed components of synthetic syllables (Eimas & Miller, 1992; cf. Whalen & Liberman, 1987). In this case, the pattern of perceptual effects evident in infants was contingent on the integration of sensory elements despite detailed failures of auditory similarity on which Gestalt grouping depends. Perhaps it is an exaggeration to claim that this organizational function is strictly unlearned, for even the youngest subject in the sample had been encountering airborne sound for three months, and undeniably had an opportunity to refine its sensitivity through learning. However, the development of sensitivity to complex auditory patterns cannot plausibly result from a history of meticulous trial and error in listeners of such tender age, nor is it likely to reflect specific knowledge of the auditory effects that typify American English phonetic expression. It is far likelier that this sensitivity represents the emergence of an organizational component of listening that must be present for speech perception to develop, and 14-week-olds still have several months ahead of them before the phonetic properties of speech become conspicuous (Houston, this volume; Jusczyk, 1997).

Research on sinewave replicas of speech has shown that the perceptual organization of speech is *nonsymbolic* and *keyed to patterns of sensory variation.* The evidence is provided by tests (Remez et al., 1994; Remez, 2001) that used tone analogs of sentences in which a sinewave replicating the second formant was presented to one ear while tone analogs of the first, third, and fricative formants were presented to the other ear. In such conditions, much as Broadbent and

Ladefoged had found, perceptual fusion readily occurs despite the violation of spatial dissimilarity and the absence of other attributes to promote Gestalt-based grouping. To sharpen the test, an intrusive tone was presented in the same ear with the tone analogs of the first, third, and fricative tones. This single tone presented by itself does not evoke phonetic impressions, and is perceived as an auditory form without symbolic properties: it merely changes in pitch and loudness without phonetic properties. In order to resolve the speech stream under such conditions, a listener must reject the intrusive tone, despite its spatial similarity to the first, third, and fricative tones of the sentence, and appropriate the tone analog of the second formant to form the speech stream despite its spatial displacement from the tones with which it combines. Control tests established that a tone analog of the second formant fails to evoke an impression of phonetic properties. Performance of listeners in a transcription task, a rough estimate of phonetic coherence, was good if the intrusive tone did not vary in a speechlike manner. That is, an intrusive tone of constant frequency or of arbitrary frequency variation had no effect on the perceptual organization of speech. When the intrusive tone exhibited the pattern of a temporally reversed second formant – exhibiting the tempo and range of frequency variation appropriate for a second formant, without supplying the proper variation that would combine with other tones to form an intelligible stream – performance suffered. It was as if the criterion for integration of a tone was specific to its speechlike variation under conditions in which it was nonetheless unintelligible.

Since the advent of the telephone, it has been obvious that a listener's ability to find and follow a speech stream is *indifferent to auditory quality*. The lack of spectral fidelity in early forms of speech technology made speech sound phony, literally, yet it was readily recognized that this lapse of natural quality did not compromise the usefulness of speech as a communication channel (Fletcher, 1929). This fact indicates clearly that the functions of perceptual organization hardly aim to collect aspects of sensory stimulation that have the precise auditory quality of natural speech. Indeed, Liberman and Cooper (1972) argued that early synthesis techniques evoked phonetic perception because the perceiver cheerfully forgave departures from natural quality that were often extreme. In techniques such as speech chimeras (Smith et al., 2002) and sinewave replication, the acoustic properties of intelligible signals lie beyond the productive capability of a human vocal tract, and the impossibility of such spectra as vocal sound does not evidently block the perceptual organization of the sound as speech. The variation of a spectral envelope can be taken by listeners to be speechlike despite acoustic details that give rise to impressions of gross unnaturalness. Findings of this sort contribute a powerful argument against psychoacoustic explanations of speech perception generally, and perceptual organization specifically.

Ordinary subjective experience of speech suggests that perceptual organization is automatic, for speech seems to pop right out of a nearby commotion. Despite this impression that perceptual organization of speech is unbidden, findings with sinewave replicas of utterances show that the perceptual organization of speech *requires attention*, and is not an automatic consequence of a class of sensory effects. This feature differs from the automatically engaged process proposed in strict modular terms by Liberman and Mattingly (1985). With sinewave signals, most subjects fail to resolve the phonetic properties of sinewave words and sentences unless they are asked specifically to listen for speech (Remez et al., 1981; cf. Liebenthal et al., 2003), indicating that the auditory forms alone do not evoke speech perception. Critically, a listener who is asked to attend to arbitrary tone patterns as if listening to speech fails to report phonetic impressions, indicating that signal structure as well as phonetic attention are required for the organization and analysis of speech. The prospect that generic auditory perceptual organization is similar to speech perception in requiring attention has been raised in recent studies of arbitrary patterns (Carlyon, Cusack, Foxton, & Robertson, 2001). Of course, a natural vocal signal exhibits the phenomenal quality of speech, and this is evidently sufficient to elicit a productive form of attention for perceptual organization to ensue.

2.3.2 *Generic auditory organization and speech perception*

The intelligibility of sinewave replicas of utterances, of noise-band vocoded speech, and of speech chimeras reveals that a perceiver can find and follow a speech signal lacking the multiple detailed similarities among acoustic and auditory constituents on which Gestalt-based generic functions operate. These findings show that perceptual organization of speech can occur solely by virtue of attention to the complex coordinate variation of an acoustic pattern. Of course, the use of such exotic acoustic signals for the proof creates some uncertainty that ordinary speech perception is satisfactorily characterized by tests using these acoustic oddities. An argument of Remez et al. (1994) for considering these tests to be a useful index of the perception of commonplace speech signals begins by noting that phonetic perception of sinewave replicas of utterances depends on a simple instruction to listen to the tones as speech. Because the disposition to hear sinewave words and sentences appears readily, without arduous or lengthy training, this prompt adaptation to phonetic organization and analysis suggests that the ordinary cognitive resources of speech perception are operating for sinewave speech. Although some form of short-term perceptual learning might be involved, the swiftness of the appearance of adequate perceptual function is evidence that any special induction to accommodate sinewave signals is a marginal component of perception.

Despite all, natural speech consists of large stretches of glottal pulsing, which create amplitude comodulation over time and harmonic relations among concurrent portions of the spectrum. This has led to a reasonable proposal (Barker & Cooke, 1999) that generic auditory grouping functions, although not necessary for the perceptual organization of speech, contribute to perceptual organization when speech spectra satisfy the Gestalt criteria. A critical empirical test was provided by Carrell and Opie (1992) and in detail it offers an index of the plausibility of the claim. In the test, the intelligibility of sinewave sentences was compared in two acoustic conditions: (1) three-tone time varying sinusoids; and (2) three-tone time varying sinusoids on which a regular amplitude pulse was imposed. Although the tone patterns in the first condition were not susceptible to Gestalt-based grouping, because they failed to exhibit similarity in each of the relevant dimensions that we have discussed, the pulsed tone patterns in the second condition exhibited amplitude comodulation and harmonicity in its complex spectra (Bregman et al., 1990). All other things being equal, the perceptual organization attributable to complex coordinate variation should have been reinforced by perceptual organization attributable to similarity that triggers generic auditory grouping. Indeed, Carrell and Opie found that pulsed sentences were more intelligible than smoothly varying sinusoids, as if the spectral components once bound more securely were more successfully analyzed.

The assertion offered by Barker and Cooke (1999) about this phenomenon is that generic auditory functions can reinforce the grouping of speech signals, although the evidence on close examination does not yet warrant an endorsement of a hybrid model of perceptual organization. Carrell and Opie (1992) had used a range of pulse rates and conditions in their study, and reported that the intelligibility gain attributable to pulsing a sinewave sentence was restricted to a pulse rate in the range of 50–100 Hz. No benefit of pulsing was observed for a pulse rate of 200 Hz. While this topic certainly merits additional study, the available evidence supports a conclusion that a hybrid model of perceptual organization is restricted to speech signals produced by low bass voices, and whatever benefit is seen for such speech does not extend to tenors, to say nothing of altos and sopranos. Most generously, we might conclude that the relation of primitive Gestalt-based generic auditory grouping and the more abstract organization by sensitivity to coordinate variation cannot be defined without stronger evidence, and that it is premature to conclude that the Gestalt set plays a prominent or even a secondary role in the perceptual organization of speech.

2.4 Implications of Perceptual Organization for Theories of Speech Perception

2.4.1 *The nature of speech cues*

What causes the perception of speech? A classic answer takes a linguistically significant contrast – voicing, for instance – and provides an inventory of acoustic correlates of a careful articulation of the contrast (for example, Lisker, 1978). A perceptual account that reverses the method would depict a meticulous listener collecting individual acoustic correlates as they land and assembling them in a stream, thereby to tally the strength with which a constellation of cues indicates the likely occurrence of a linguistic constituent. Klatt's (1989) retrospective survey of perceptual accounts describes many approaches that treat the acoustic signal as a straightforward composite of acoustic correlates. The function of perceptual organization, usually omitted in such accounts, establishes the perceiver's compliance with the acoustic products of a specific source of sound, and in the case of speech, it is the function that finds and tracks the acoustic products of vocalization. However, it is clear from evidence of several sorts – tolerance of distortion, effectiveness of impossible signals, forgiveness of departures from natural timbre – that the organizational component of perception which yields a speech stream fit to analyze cannot collect acoustic cues piecemeal, as this simple view describes. The functions of perceptual organization act, instead, as if attuned to a complex form of regular if unpredictable spectrotemporal variation within which the specific acoustic and auditory elements matter far less than the overall configuration they compose.

The evolving portrait of speech perception that includes organization and analysis recasts the cue as the property of perception that gives speech its phenomenality, though not its phonetic effect. If the transformation of natural speech to chimera, to noise-band vocoded signal, and to sinewave replica is phonetically conservative, preserving the fine details of subphonemic variation while varying to the extremes of timbre or auditory quality, then it is apparent that the competent listener derives phonetic impressions from the properties that these different kinds of signal share, and derives qualitative impressions from their unique attributes. The shared attribute, for want of a more precise description, is a complex modulation of spectrum envelopes, although the basis for the similar effect of the infinitely sharp peaks of sinewave speech and the far coarser spectra of chimerical and noise-band vocoded speech has still to be explained. None of these manifests the cues present in natural speech despite the success of listeners in understanding the message. The conclusion supported by these findings is clear: phonetic perception does not require speech cues. Instead, the organizational component of speech perception operates on a spectrotemporal grain that is requisite both for finding and following a speech signal and for analyzing its linguistic properties. The speech cues that seemed formerly to bear the burden of stimulating phonetic analyzers into action appear in hindsight to provide little more than auditory quality subordinate to the phonetic stream.

An additional source of evidence is encountered in the phenomenal experience of perceivers who listen to speech via electrocochlear prostheses (Goh, Pisoni, Kirk, & Remez, 2001). Intelligibility of speech perceived via a cochlear implant is often excellent, rivaling that of normal hearing, and recent studies with infant and juvenile subjects (Svirsky, Robbins, Kirk, Pisoni, & Miyamoto, 2000) suggest that this form of sensory substitution is effective even at the earliest stages of language development (see Pisoni, this volume). The mechanism of acoustic transduction at the auditory periphery is anomalous, it goes without saying, and the phenomenal experience of listeners using this appliance to initiate neural activity differs hugely from ordinary auditory experience of natural speech. Despite the absence of veridical perceptual experience of the raw qualities of natural speech, electrocochlear prostheses are effective in the self-regulation of speech production by their users, and are effective perceptually despite the abject deficit in delivering speech cues. What brings about the perception of speech, then? Without the acoustic moments, there is no stream of speech, but the stream itself plays a causal role beyond that which has been attributed to momentary cues since the beginning of technical study of speech.

2.4.2 *A constraint on normative descriptions of speech perception*

The application of powerful statistical techniques to problems in cognitive psychology has engendered a variety of normative, incidence based accounts of perception. Since the 1980s, a technology of parallel computation based loosely on an idealization of the neuron has driven the creation of a proliferation of devices that perform intelligent acts. The exact modeling of neurophysiology is rare in this enterprise, though probabilistic models attired as neural nets enjoy a hopeful if unearned appearance of naturalness that older, algorithmic explanations of cognitive processes unquestionably lack. Used as a theory of human cognitive function, it is more truthful to say that neural nets characterize the human actor as an office full of clerks at an insurance company, endlessly tallying the incidence of different states in one domain (perhaps age and zip code, or the bitmap of the momentary auditory effect of a noise burst in the spectrum) and associating them (perhaps, in a nonlinear projection) with those in another domain (perhaps, the risk of major surgery, or the place of articulation of a consonant).

In the perception of speech and language, the ability of perceivers to differentiate levels of linguistic structure has been attributed to a sensitivity to inhomogeneities in distributions of specific instances of sounds, words, and phrases. Although a dispute has taken shape about the exact dimensions of the domain within which sensitivity to distributions can be useful (for instance, Peña, Bonatti, Nespor, & Mehler, 2002; *contra* Seidenberg, MacDonald, & Saffran, 2002), there is confident agreement that a distributional analysis of a stream of speech is performed in order to derive a linguistic phonetic segmental sequence. Indeed, this is claimed as one key component of language acquisition in early childhood (Saffran, Aslin, & Newport, 1996). The presumption of this assertion obliges a listener to establish and maintain in memory a distribution of auditory tokens projectable into phonetic types. This is surely false. The rapid decay of an auditory trace of speech leaves it uniquely unfit for functions lasting longer than 100 ms, and for this reason it is simply implausible that stable perceptual categories rest on durable representations of auditory exemplars of speech samples. Moreover, the notion of perceptual organization presented in this essay argues that a speech stream is not usefully represented as a series of individual cues, neither for purposes of perceptual organization nor analysis. Indeed, in order to determine that a particular acoustic moment is a cue in fact, a perceptual function already sensitive to coordinate variation must apply. Whether or not a person other than a researcher compiling entries in the *Dictionary of American Regional English* can become sensitive to distributions of linguistic properties as such, it is exceedingly unlikely that the perceptual resolution of linguistic properties in utterances is much influenced by representations of the statistical properties of speech sounds. Indeed, the clerks are free to tally what they will, but perception must act first to provide the instances.

2.4.3 *Multisensory perceptual organization*

Fifty years ago, Sumby and Pollack (1954) conducted a pioneering study of the perception of speech presented in noise in which listeners could also see the talkers whose words they aimed to recognize. The point of the study was to calibrate the level at which the speech signal would become so faint in the noise that to sustain adequate performance attention would switch from an inaudible acoustic signal to the visible face of the talker. In fact, the visual channel contributed to intelligibility at all levels of performance, indicating that the perception of speech is ineluctably multisensory. But, how does the perceiver determine the audible and visible composition of a speech stream? This problem (reviewed by Bernstein, this volume, and by Rosenblum, this volume) is a general form of the listener's specific problem of perceptual organization, understood as a function that follows the speechlike coordinate variation of a sensory sample of an utterance. To assign auditory effects to the proper source, the perceptual organization of speech must capture the complex sound pattern of a phonologically governed vocal source, sensing the spectrotemporal variation that transcends the simple similarities on which the Gestalt-derived principles rest. It is obvious that Gestalt principles couched in auditory dimensions would fail to merge auditory attributes with visual attributes. Because auditory and visual dimensions are simply incommensurate, it is not obvious that any notion of similarity would hold the key to audio-visual combination. The single property that the two senses share, localization in azimuth and range, is violated freely without harming audiovisual combination, and therefore cannot be requisite for multisensory perceptual organization.

The phenomenon of multimodal perceptual organization confounds straightforward explanation in yet another instructive way. Audiovisual speech perception can be fine under conditions in which the audible and visible components are useless separately for conveying the linguistic properties of the message (Rosen, Fourcin, & Moore, 1981). In addition, neither spatial alignment nor temporal alignment of the audible and visible components must be veridical for multimodal perceptual organization to deliver a coherent stream fit to analyze (see Bertelson, Vroomen, & de Gelder, 1997; Conrey & Pisoni, 2003; Munhall, Gribble, Sacco, & Ward, 1996). Under such discrepant conditions, audiovisual integration occurs despite the perceiver's evident awareness of the spatial and temporal misalignment, indicating a divergence in the perceptual organization of events and the perception of speech. In consequence, it is difficult to conceive of an account of such phenomena by means of perceptual organization based on tests of similar sensory details applied separately in each modality. Instead, it is tempting to speculate that an account of perceptual organization of speech can ultimately be characterized in dimensions that are removed from any specific sensory modality, yet is expressed in parameters appropriate to the sensory samples available at any moment.

2.5 Conclusion

Perceptual organization is the critical function by which a listener resolves the sensory samples into streams specific to worldly objects and events. In the perceptual organization of speech, the auditory correlates of speech are resolved into a coherent stream fit to analyze for its linguistic and indexical properties. Although many contemporary accounts of speech perception are silent about perceptual organization, it is unlikely that the generic auditory functions of perceptual grouping provide adequate means to find and follow the complex properties of speech. It is possible to propose a rough outline of an adequate account of the perceptual organization of speech by drawing on relevant findings from different research projects spanning a variety of aims. The evidence from these projects suggests that the critical organizational functions that operate for speech are: fast, unlearned, nonsymbolic, keyed to complex patterns of coordinate sensory variation, indifferent to sensory quality, and requiring attention whether elicited or exerted. Research on other sources of complex natural sound has the potential to reveal whether these functions are unique to speech or are drawn from a common stock of resources of unimodal and multimodal perceptual organization.

ACKNOWLEDGMENTS

In conducting some of the research described here and in composing this essay, the author is grateful for the sympathetic understanding of: Peter Bailey, Robert Krauss, Jennifer Pardo, Rebecca Piorkowski, David Pisoni, Philip Rubin, Ann Senghas, Michael Studdert-Kennedy, and Stephanie Wissig. This work is supported by a grant to Barnard College from the National Institute on Deafness and Other Communication Disorders (DC00308).

NOTE

[1] It is notable that the literature on duplex perception contains meager direct evidence that the auditory and phonetic properties of the duplex acoustic test items are available simultaneously. The empirical evaluation of auditory and phonetic form employed sequential measures, sometimes separated by a week, that assessed the perception of auditory form in one test and phonetic form in another. Evidence is provided that phonetic perception is distinct from a generic auditory process, but the literature is silent on the criteria of perceptual organization required for phonetic analysis.

REFERENCES

Barker, J. & Cooke, M. (1999). Is the sine-wave cocktail party worth attending? *Speech Communication*, 27, 159–74.

Bertelson, P., Vroomen, J., & de Gelder, B. (1997). Auditory-visual interaction in voice localization and in bimodal speech recognition: The effects of desynchronization. In C. Benoît & R. Campbell (eds.), *Proceedings of the Workshop on Audio-visual Speech Processing: Cognitive and Computational Approaches* (pp. 97–100). Rhodes, Greece: ESCA.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Bregman, A. S. & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequence of tones. *Journal of Experimental Psychology*, 89, 244–9.

Bregman, A. S. & Dannenbring, G. L. (1973). The effect of continuity on auditory stream segregation. *Perception & Psychophysics*, 13, 308–12.

Bregman, A. S. & Dannenbring, G. L. (1977). Auditory continuity and amplitude edges. *Canadian Journal of Psychology*, 31, 151–8.

Bregman, A. S. & Doehring, P. (1984). Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics*, 36, 251–6.

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

Bregman, A. S., Abramson, J., Doehring, P., & Darwin, C. J. (1985). Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, 37, 483–93.

Bregman, A. S., Ahad, P. A., & Van Loon, C. (2001). Stream segregation of narrow-band noise bursts. *Perception & Psychophysics*, 63, 790–7.

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

Broadbent, D. E. & Ladefoged, P. (1957). On the fusion of sounds reaching different sense organs. *Journal of the Acoustical Society of America*, 29, 708–10.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27, 115–27.

Carrell, T. D. & Opie, J. M., (1992). The effect of amplitude comodulation on auditory object formation in sentence perception. *Perception & Psychophysics*, 52, 437–45.

Cherry, E. (1953). Some experiments on the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, 25, 975–9.

Conrey, B. L. & Pisoni, D. B. (2003). Audiovisual asynchrony detection for speech and nonspeech signals. *Proceedings of Audio Visual Speech Processing 2003*, 25–30.

Cooke, M. & Ellis, D. P. W. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35, 141–77.

Dannenbring, G. L. & Bregman, A. S. (1976). Stream segregation and the illusion of overlap. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 544–55.

Dannenbring, G. L. & Bregman, A. S. (1978). Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics*, 24, 369–76.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, 1, 327–33.

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, 838–44.

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

Diehl, R. L., Molis, M. R., & Castleman, W. A. (2001). Adaptive design of sound systems. In E. Hume and K. Johnson (eds.), *The Role of Speech Perception in Phonology* (pp. 123–39). San Diego: 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*, 104, 121–9.

Eimas, P. & Miller, J. (1992). Organization in the perception of speech by young infants. *Psychological Science*, 3, 340–5.

Fant, C. G. M. (1960). *The Acoustic Theory of Speech Production*. The Hague: Mouton.

Fletcher, H. (1929). *Speech and Hearing*. New York: D. Van Nostrand.

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

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology*, 5, 285–313.

Goh, W. D., Pisoni, D. B., Kirk, K. I., & Remez, R. E. (2001). Audio-visual perception of sinewave speech in an adult cochlear implant user: A case study. *Ear & Hearing*, 22, 412–19.

Hochberg, J. (1974). Organization and the Gestalt tradition. In E. C. Carterette & M. P. Friedman (eds.), *Handbook of*

Perception, Vol. I: Historical and Philosophical Roots of Perception (pp. 179–210). New York: Academic Press.

Howell, P. & Darwin, C. J. (1977). Some properties of auditory memory for rapid formant transitions. *Memory & Cognition*, 5, 700–8.

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 751–63.

Jones, M. R. & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, 96, 459–91.

Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, MA: MIT Press.

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.

Lackner, J. R. & Goldstein, L. M. (1974). Primary auditory stream segregation of repeated consonant-vowel sequences. *Journal of Acoustical Society of America*, 56, 1651–2.

Liberman, A. M., Ingemann, F., Lisker, L., Delattre, P., & Cooper, F. S. (1959). Minimal rules for synthesizing speech. *Journal of the Acoustical Society of America*, 31, 1490–9.

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, 30, 133–43.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21, 1–36.

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. 329–38). The Hague: Mouton.

Licklider, J. C. R. (1946). Effects of amplitude distortion upon the intelligibility of speech. *Journal of the Acoustical Society of America*, 18, 429–34.

Liebenthal, E., Binder, J. R., Piorkowski, R. L., & Remez, R. E. (2003). Short-term reorganization of auditory analysis induced by phonetic experience. *Journal of Cognitive Neuroscience*, 15, 549–58.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, 99, 1683–92.

Lisker, L. (1978). Rapid vs. rabid: A catalog of acoustic features that may cue the distinction. *Haskins Laboratories Status Report on Speech Perception*. New Haven, CT: Haskins Laboratories, SR-54, 127–32.

Massaro, D. W. (1994). Psychological aspects of speech perception: Implications for research and theory. In M. A. Gernsbacher (ed.), *Handbook of Psycholinguistics* (pp. 219–63). San Diego: Academic Press.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. G. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, 2, 131–57.

Miller, G. A. (1946). Intelligibility of speech: Effects of distortion. In *Transmission and Reception of Sounds Under Combat Conditions* (pp. 86–108). Washington, DC: National Defense Research Committee.

Miller, G. A. & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, 22, 167–73.

Mountcastle, V. B. (1998). *Perceptual Neuroscience*. Cambridge, MA: Harvard University Press.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58, 351–62.

Neff, D. L., Jesteadt, W., & Brown, E. L. (1982). The relation between gap discrimination and auditory stream segregation. *Perception & Psychophysics*, 31, 493–501.

Nygaard, L. C. (1993). Phonetic coherence in duplex perception: Effects of acoustic differences and lexical status. *Journal of Experimental Psychology*, 19, 268–86.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America*, 60, 911–18.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, 298, 604–7.

Pisoni, D. B. & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, 15, 285–90.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, 55, 678–80.

Remez, R. E. (2001). The interplay of phonology and perception considered from the perspective of perceptual organization. In E. Hume & K. Johnson (eds.), *The Role of Speech Perception in Phonology* (pp. 27–52). San Diego: Academic Press.

Remez, R. E. & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics*, 35, 429–40.

Remez, R. E. & Rubin, P. E. (1993). On the intonation of sinusoidal sentences: Contour and pitch height. *Journal of the Acoustical Society of America*, 94, 1983–8.

Remez, R. E., Pardo, J. S., Piorkowski, R. L., & Rubin, P. E. (2001). On the bistability of sine wave analogues of speech. *Psychological Science*, 12, 24–9.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, 101, 129–56.

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, 41–60.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947–50.

Remez, R. E., Yang, C. Y., Piorkowski, R. L., Wissig, S., Batchelder, A., & Nam, H. (2002). The effect of variation in naturalness on phonetic perception identification. *Journal of the Acoustical Society of America*, 111, 2432.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. J. (1981). Voice pitch as an aid to lipreading. *Nature*, 291, 150–2.

Rossing, T. D. (1990). *The Science of Sound*. Reading, MA: Addison-Wesley.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–8.

Seidenberg, M. S., MacDonald, M. C., & Saffran, J. R. (2002). Does grammar start where statistics stop? *Science*, 298, 553–4.

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303–4.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416, 87–90.

Steiger, H. & Bregman, A. S. (1982). Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics*, 32, 153–62.

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge, MA: MIT Press.

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 on the Study of Speech* (pp. 1–38). Hillsdale, NJ: Lawrence Erlbaum.

Stevens, K. N. & House, A. S. (1961). An acoustical theory of vowel production and some of its implications. *Journal of Speech & Hearing Research*, 4, 303–20.

Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–15.

Summerfield, Q. (1992). Roles of harmonicity and coherent frequency modulation in auditory grouping. In M. E. H. Schouten (ed.), *The Auditory Processing of Speech: From Sounds to Words* (pp. 157–66). Berlin: Mouton de Gruyter.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological Science*, 11, 153–8.

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, 586–7.

Wertheimer, M. (1923). Unsuchungen zur Lehre von der Gestalt, II, *Psychologische Forschung* (pp. 301–50). [Translated as, "Laws of organization in perceptual forms," in W. D. Ellis (ed.) (1938). *A Sourcebook of Gestalt Psychology* (pp. 71–88). London: Routledge & Kegan Paul.]

Whalen, D. H. & Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, 237, 169–71.

Whalen, D. H. & Liberman, A. M. (1996). Limits on phonetic integration in duplex perception. *Perception & Psychophysics*, 58, 857–70.