SPEECH PERCEPTION

Robert E. Remez
Department of Psychology and Program in Neuroscience and Behavior
Barnard College, Columbia University

1. AUDITORY PERCEPTUAL ORGANIZATION AND PERCEPTUAL ANALYSIS OF SPEECH

Perceptual organization pertains to the starting point for perception, and consists of functions that resolve and bind contours within a sensory field. In the proposals that introduced the idea of organization (von Ehrenfels, 1890; Wertheimer, 1923), both visual and auditory sensitivity were considered, with the goal of identifying the autochthonous attributes of perception that establish coherence and promote perceptual analysis. One might imagine that the scientific motive in the Gestalt reaction to structuralism and nativism faded in influence long ago. To be certain, there is little theoretical or empirical work that is conducted today using introspection as the method, or toy pianos and tuning forks as equipment. Still, the remnants of this dispute can be seen in contemporary approaches to organization and analysis, especially within the auditory system, and especially regarding the perception of speech.

It might seem surprising for the study of a domain-general function to focus on a specific kind of object — a phonetic segment or a syllable composed of segments — but the reasons for empirical attention to speech are both opportunistic and practical. Speech is produced by a natural albeit complex sound source that is well modeled (Stevens, 1999). Using speech to evaluate emerging ideas about organization permits a realistic test of claims that have developed with frank indifference to natural objects and events. Indeed, a disproportionate majority of empirical investigations of auditory perceptual organization exploited arbitrary test patterns created with audio-frequency oscillators, noise generators and their digital counterparts. Such patterns have typically been designed according to formal idealizations, without regard to the mechanics of natural sound production. The ensuing empirical and theoretical rationale has resolutely relied on nominal conceptualizations of similarity and simplicity (Hochberg, 1974). If speech provides an opportunity for a realistic test of the idealizations of auditory organization, the use of speech empirically also offers a chance for practical application. Specifically, the potential for real-time implementation of accounts of perceptual organization can be seen in assistive devices worn by individuals whose hearing is impaired, and in other technology that is controllable by speech. One drawback of present computational models is poor performance in finding and following an acoustic stream of speech. Whatever theoretical disputes might inflect discussions among scientists, this practical difficulty with computational implementations urges modesty in our commitments to present understanding.

Two divergent claims presently compete in describing auditory perceptual organization of speech. One results from direct examination of spoken sound samples (Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez & Thomas, 2013). The other represents an application to speech of Gestalt-derived criteria confirmed in studies using arbitrarily composed test patterns (Bregman, 1990; Darwin, 2008). Although the description given here is concerned chiefly with aspects of organization that are verified empirically with speech samples, the contrasting claims of the Gestalt-derived conceptualization are mentioned throughout, in order to highlight the principles and practices that distinguish the claims.

1.1. Modulation sensitivity or piecemeal cue capture

All accounts of perception begin by designating primitive properties of extrinsic energy to which receptors respond. From such receptor states, a notion of the receptor bitmap can be entertained, at least conceptually, as a prelude to the cascading functions acting on a sensory sample to resolve contours. The challenge of organization (Bregman & Pinker, 1978) is to parse an auditory field into streams, each sensory stream properly assigned to a source of sound. To the extent that this function succeeds, analysis of an attended stream can include all of the properties that

To appear in the *Oxford Handbook of Computational Perceptual Organization* (S. Gepshtein, L. Maloney and M. Singh, Eds.).
Please address correspondence to R. E. Remez, Department of Psychology, Barnard College, 3009 Broadway, New York, New York 10027-6598 U. S. A. Email: remez@columbia.edu.
This research was sponsored by a grant from the National Institute on Deafness and Other Communication Disorders (DC000308.

issued from a source, permitting perceptual analysis to operate on an intact sample of a stream free of intrusive or absent effects. In the auditory instance, the resolution of the concurrent components of a stream spans frequency; a single vocal source can be 5 kHz wide. The organizational challenge to follow an evolving sound source in time in the case of speech can span 2 s or more. Because a spoken utterance readily exceeds the persistence of the ephemeral auditory trace, organizational functions arguably hand off an incompletely resolved portion of an incident contour as it is bound.

Despite outstanding progress in modeling the auditory periphery physiologically and computationally, such efforts have not yielded a description of the primitive sensory elements that evoke the organization of an auditory field into contours. Indeed, they cannot, for reasons that are conceptual, not empirical (Gallistel, 2007). Neither the acoustic spectrum nor the auditory system itself creates units analogous to blocks of data in a digital communication line. For this reason, it is implausible that the first step in auditory organization begins in a sensory sample composed of elementary, commutable, dissociable stimuli. Far likelier is a starting point that detects coordinate spectrotemporal variation within an auditory sample (Elliott & Theunissen, 2009). This notion is consistent with the findings of organizational studies of speech (Remez et al., 1994; Remez, Rubin, Pisoni, & Carrell, 1981). In those empirical projects, synthetic speech consisting of a small number of time-varying sinusoids was constructed to exhibit the spectrotemporal properties of a natural sample, while discarding the natural acoustic constituents of vocally produced sound. The result created a spectrum unlike speech in detail, lacking broadband resonances, harmonics, short-term aperiodic transients, aspirate spectral bands, and the band-limited noise of consonantal friction, for example, in /s/ or /f/. A perceiver is left rather little to grasp beyond continuous frequency variation in three or four concurrently varying tones. Moreover, the tone complexes that compose sine-wave speech utterly lack the subjective quality of speech, evoking an impression instead of mistuned whistles with indistinct musical qualities.

A listener's first impression of sine-wave speech is dominated by the auditory form of the tones. Lacking the disposition to treat the tones as a single complex contour, perceivers hear no phonetic qualities, and the physiological response to the signal differs from states observed with speech (Liebenthal, Binder, Piorkowski, & Remez, 2003). However, the mere instruction to listen for synthetically produced speech is sufficient for a perceiver to notice coherence in the coordinate variation of the asynchronously changing tones, resulting in an impression of linguistic properties equal to the natural sample on which the sine-wave pattern was modeled. It is not unusual for intelligibility to be excellent (see Remez, Ferro, Wissig, & Landau, 2008; cf. Feng, Xu, Zhou, Yang, & Yin, 2012).

1.1.1. Sine-wave speech. Research with sine-wave speech provided a distinct portrait of auditory organization while also falsifying a long-term assumption about perception of speech. Whether concerned with organization or analysis of speech, researchers had been committed to a normative view of the causes of perception. To be clear, research since the invention of synthetic speech consistently sought to describe the elementary acoustic properties of speech that were typically correlated with linguistic properties, expressing a version of essentialism that has been common in nativist and empiricist accounts, alike (Pastore, 1971). In this approach, an unidentified sensory state is bound and analyzed by virtue of its similarity to a memorized standard. The standards, whether describing prototypes or characteristic attributes, are typical of the sensory presentation of an object or an event; some variants of this conceptualization propose that standards are established by exposure, others by inheritance. Through comparison to a group of standards, an as-yet-unidentified sensory sample was said to be categorized. Many formally intricate explanations rest on this simple idea: Perception is possible when stimulation is similar to a familiar memorized sensory state.

Because sine-wave speech evokes an impression of linguistic properties – and, the personal character of the one who spoke them — without relying on sensory details that are typical of vocally produced sound, it defeats the premise that perception depends on typical momentary acoustic properties of speech and their auditory sensory effects. In its simplest formula, it also means that acoustic or auditory essentialism is false. The perception of speech entails something wholly different than the categorization of sounds as such, that is, categorization by appeal to typical timbre, typical pitch, typical spectrum shape, typical combinations of acoustic or auditory moments, etc. Perceptual organization of a speech stream must entail sensitivity to modulation independent of the composition of the carrier, and cannot depend on the piecemeal registration of acoustic signal elements and their auditory sensory effects.

In addition to posing a radical challenge to the normative assumption that perception is caused by the detection of typical sensory properties, the intelligibility of sine-wave speech also undermines the classic accounts attributing the causes of speech perception to speech cues. The notion of the speech cue that has dominated the technical explanation of speech perception for decades (Liberman & Cooper, 1970; Raphael, 2005) holds that the spectrum of speech is an acoustic composite of whistles, clicks, hisses, buzzes and hums. According to this account and its contemporary proponents (Holt & Lotto, 2010), each of these elements has a value that it contributes to an act of

categorization, which occurs by tallying the values of the cues in the present sample, by comparing these with base-rate incidence and by concluding with the best linguistic segmental match. Sine-wave speech lacks all of these cues, yet it remains effective in causing the perception of speech. One explanation for its effectiveness (Remez, 2005) is that the perceptually causal property that individual acoustic moments possess depends on the modulation characteristic of their aggregation. Indeed, studies of acoustic chimeras (Smith, Delgutte, & Oxenham, 2002) indicate that the carrier and the shape of the spectrum can truly be arbitrary as long as the properties of the modulation remain available in a sensory sample.

A contrasting conceptualization is given in the Gestalt-derived generic auditory account. According to this view (Bregman, 1990; Darwin, 2008) the primitive constituents of perceptual organization are the acoustic moments registered as such in the auditory periphery. This unorganized auditory field coalesces into streams through the operation of similarity principles inherent in the Gestalt grouping dispositions. A sizable body of research now demonstrates grouping by similarity in: frequency; change in frequency; similarity in fundamental frequency; closure across interruptions; temporal coincidence; spectral similarity; frequency continuity; harmonicity; and common modulation. Research has also sought to calibrate the relative binding strength of each kind of similarity (summarized in Bregman, 1990; Remez et al., 1994). Although such studies collectively compose a consistent description, the account overall suffers from the lack of development with natural sound sources, and fails utterly to describe the empirical findings with the acoustically heterogeneous speech spectrum.

1.1.2. Speech and the Gestalt principles. Relatively early in the collection of evidence to buttress a Gestalt account of auditory perceptual organization, it was acknowledged that speech would not fit the emerging description (see Julesz & Hirsh, 1972; Lackner & Goldstein, 1974). For example, the hypothetical disposition to form an auditory stream composed of similar sensory elements must fracture a single speech spectrum into multiple streams. Each of the vocal resonances changes asynchronously in frequency and to different extents during production, and by their differences should split into separate streams. Likewise, nasal resonances, which form and resolve abruptly, should compose a fourth stream by dissimilarity and discontinuity in frequency with the oral resonances. When consonantal holds are released, a variety of bursts is produced, some aspirate and some unaspirated, and these should each split into separate streams of similar bursts. Because high-, mid-, and low-frequency bursts differ in frequency whether these are aspirate or not, each should also split into a separate stream distinct from the sustained resonances. Voiced fricatives are accompanied by periodic excitation, while voiceless fricatives are more nearly aperiodic, and should split into separate streams on that difference alone. Yet, some labiodental and linguodental fricatives are extremely broad-band in the frequency distribution of noise; these differ from apical and laminar fricatives which are far narrower. A glottal aspirate has the structure of an oral resonance, but because of its aperiodic spectrum must split by dissimilarity from a pulsed resonance. Each of the acoustic moments differing in spectrum shape and periodicity should compose a stream distinct from the other elements (See Figure 1.) Studies of grouping of arbitrary test patterns by Gestalt criteria show extremely tight tolerance, temporally, as well; their application to speech would even split similar acoustic elements into separate streams were they to onset or offset asynchronously by as little as 50 ms. Overall, an application of Gestalt-based similarity criteria is destructive of the coherence of speech produced by a single individual in a quiet, echoless enclosure.

1.2. Binding the constituents of a speech stream

Direct investigations of speech have produced a description, albeit incomplete, of perceptual organization. Whether natural or synthetic speech is the object, perceptual organization is effortful, fast, indifferent to sensory elements, nonsymbolic, unlearned, and keyed to coordinate variation in the spectrum.

1.2.1. Effortful. It is rarely noted that perception is an active cognitive function, and often assumed tacitly that the knowledge of objects and events produced perceptually is elicited reflexively, without a perceiver's commitment (Searle, 1981). In ordinary instances of speech perception, awareness of the vocal timbre of speech is usually sufficient to draw attention to speech, concentrating cognitive resources on organization and analysis. Studies with sine-wave speech reveal that perceptual organization of the tones is bistable. Grouping the tones to form a single speech stream depends on the belief that the tones are a kind of synthetic speech. If intention is absent, the nonvocal qualities of the pitch and loudness of the tones are sufficient to promote the segregation of the tones into separate perceptual streams. Although some classic research on attention suggests that some properties of an unattended speech stream can break through to seize attention, those results are also consistent with a dynamic of drifting attention that shifts between a focal stream and unattended streams, occasionally extracting coherent impressions from the background when attention lands.

Gestalt-derived auditory organization had been described as automatic, operating without effort or attention, since Wertheimer introduced the conjecture. This premise remained untested until recently (Carlyon, Cusack, Foxton, & Robertson, 2001; though, see Remez et al., 1981). In contrast to these historical claims, even an auditory stream

of

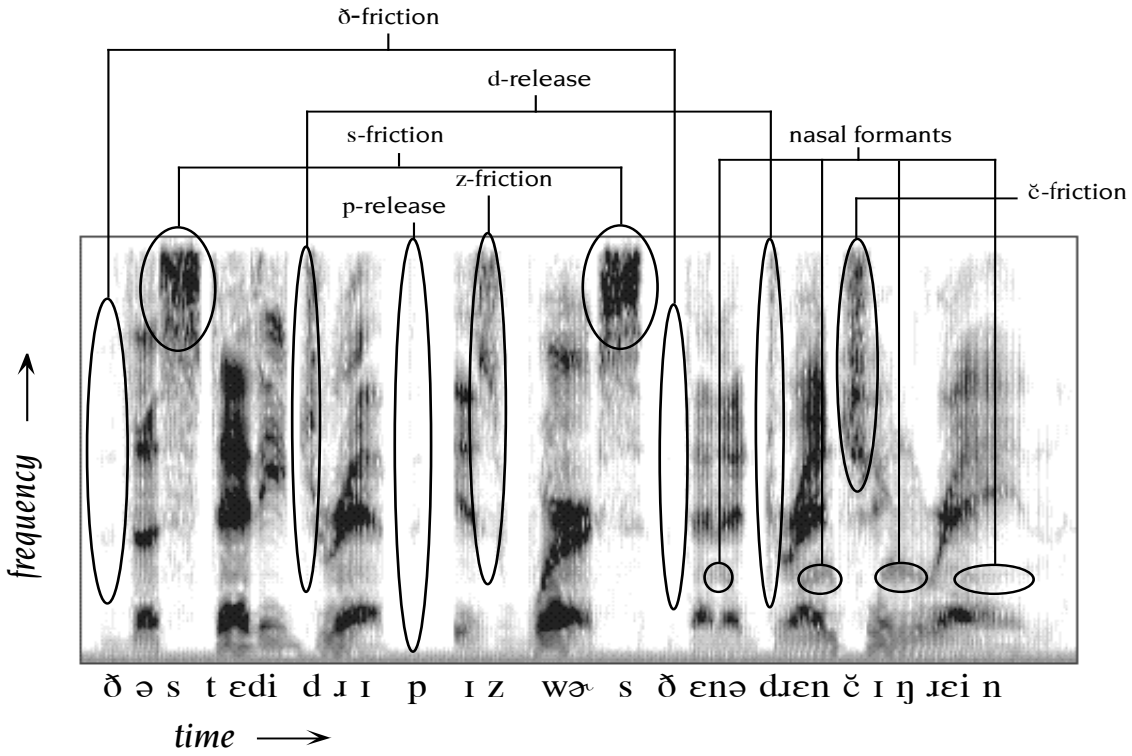

Figure 1. A spectrogram of the natural sentence, "The steady drip is worse than a drenching rain," exhibiting the acoustic diversity of the constituents of a speech stream. From Remez et al. (1994), reprinted with permission of the author and the American Psychological Association.

arbitrary composition requires attention in order to coalesce into streams, much as speech does. Streams of nonsense test patterns take time to build, as well, permitting an opportunity for short-term learning or long established knowledge to affect arbitrary stream formation. The Gestalt-derived grouping dispositions seem to be a plausible product of biological evolution, as Darwin (2008) has claimed, but not if the evidence is admitted that undermines the premise that their action is automatic. The proof that grouping is effortful blocks this claim. No matter how the Gestalt criteria of auditory grouping originate, it will be difficult to support the claim that these functions are primitive and automatic, originating in the physiological periphery, for instance.

1.2.2. Fast. The temporal limit on the perceptual organization of a sensory sample is set by its persistence. In order to track a contour over its temporal course, its components must be integrated before they decay and are replaced by new samples. But, a visual scene of stable objects permits rescanning its contours, much as an auditory scene composed of steady state sound sources — a fan, the cabin of a cruising airplane — likewise permits attention to return without cost. With objects in motion, though, a visual scene requires organization on the fly, and the extraction of sensory contours matched to objects at each glimpse occurs under time pressure set by the rate of sensory decay. Similarly, the pace of spectral modulation in a speech spectrum is set by the rate of articulation, which might ordinarily be 20 Hz or faster. This falls within the rate of fading of an auditory sensory sample, which might persist from 80-100 ms. To a first approximation, the resolution of auditory contours across frequency and over time occurs within this temporal window.

Corroborating evidence comes from a few sources. The ability of listeners to tolerate regular interruptions in an ongoing speech sample has this time course (Miller & Licklider, 1950). An inference about the effect of a fading auditory trace on the perceptual organization of speech is also fostered by studies of phonetic discrimination (Pisoni, 1973). An auditory trace of speech remained available for use in distinguishing instance-specific properties of syllables for about 100 ms, after which time the grain of discriminable differences coarsened to the grain of phonologically designated differences. This loss of

sensitivity to the finest details of a spoken syllable is explained as the rapid fading of auditory properties and recoding of the syllable into a durable phonetic grain, all within a tenth of a second. The former closely track the infinitely graded variation of an incident spectrum; the latter fit the utterance to a finite and small set of phonetic and phonemic contrasts known to the listener through long-established familiarity with words. These findings and many others from the classic era of phonetic psychoacoustics reveal the urgency with which speech perception occurs. But, while they established that speech perception overall is susceptible to deadline stress caused by a volatile auditory sample, these old tests stop short of a direct performance evaluation.

Recently, several projects have addressed this empirically by examining the perception of temporally perturbed speech samples. To measure perceptual organization, these studies calibrated the tolerance of temporal desynchrony of acoustic constituents of speech. The premise of the work is that the disposition to integrate desynchronized patterns of speech spectra reflects the temporal width of the window of organization. Although some measures (Greenberg & Arai, 1998; Saberi & Perrott, 1999) had indicated a span approaching 200 ms over which integration occurred, those methods had not distinguished between perceptual effects of sensory coherence and effects of unaggregated short-term spectra. To avoid overestimating the time course of organization, noiseband vocoded speech or sine-wave speech was used in subsequent tests. Neither of these kinds of spectra preserves the properties of short-term vocal timbre. Accordingly, time-critical organizational functions are spotlighted by these methods. Two kinds of temporal distortion have been evaluated: tolerance of desynchrony (Fu & Galvin, 2001; Remez, et al., 2008), and tolerance of brief temporal reversals (Remez, Thomas, Dubowski, Koinis, Porter, Paddu, Moskalenko, & Grossman, in press). Each of these converges on an estimate between 50-75 ms as the pace of the resolution of an auditory contour composing a speech stream. The convergence of these estimates and classic psychoacoustic assays of the persistence of auditory traces adds credence to the claim that organization is fast.

1.2.3. Indifference to sensory elements. An utterance is typically composed of words that are well known to talker and listener. Despite such extensive sharing of linguistic resources, the acoustic effects of each utterance are unique, a consequence of graded variation among talkers in anatomy and physiology, in dialect and idiolect, in mood and vitality, and in situation-specific expression. Finding a speech stream entails identifying a sensory contour that is patterned according to linguistic governance even if the attributes of a specific talker and the local conditions make it unpredictable in its exact composition. Traditionally, the variation in acoustic properties of speech has been treated as a normative circumstance, as if each acoustic element were drawn from a finite set; and, as if each specific acoustic element varied only in degree within its kind. These assumptions rationalized the project to identify the acoustic cues, a hypothetical set of acoustic elements that corresponded to the distinctive features of phoneme contrasts. Yet, about this work, Liberman and Cooper (1972) said that perception succeeded even under conditions of enormous acoustic distortion, encouraging the notion of speech perception as robust. That is, utterances were perfectly intelligible even when the vocal spectrum was unnatural, the burst spectrum was unnatural, the nasal spectrum was unnatural, the fricative spectrum was unnatural, and the impression of the synthesis overall was unnatural. Their conclusion was that a competent perceiver could tell which type of speech cue a distorted token expressed, thereby recognizing its value in indicating the consonant or vowel with which it was associated. A speech synthesizer might have sounded unnatural, yet the consonant and vowel sequences it produced were thought to be organized and analyzed by likeness to the idealized acoustic effects of natural sound production.

In disparity with this sensible assumption, it must be acknowledged that the elements of a speech stream apparently do not compose a closed set. The studies showing this empirically took a different approach than projects attempting to model the acoustic output of vocal resonators excited by a laryngeal source. In sine-wave speech, the acoustic character of the spectral constituents is discarded while the time-varying modulations are preserved and imposed on a distinctly non-vocal group of pure tones. None of the classical speech cues is present in this signal, yet it remains highly intelligible. A similar effect of short-term vocal characteristics and patterned modulation is reported for noise-band vocoded speech. In this technique, devised by Shannon, Zeng, Kamath, Wygonski, & Ekelid (1995) to model the spectral blur observed in the use of a cochlear implant, a virtual stack of filters is used to analyze the energy within a set of bands that span the range of a speech spectrum. An utterance is then represented as a set of changes in power within the bands that span its frequency range. Then, each band is filled with amplitude-modulated noise according to the instantaneous energy estimated moment by moment over the course of an utterance. The result is an intelligible synthetic utterance if at least four equal-size bands represent 5 kHz of the speech spectrum. The acoustic details of speech are completely eliminated, and the instantaneous acoustic maxima and minima that give speech spectrum its formant pattern are reduced to a coarse blur. Nonetheless, in the absence of acoustic details typical of vocal sound production, perceptual organization occurs and phonetic analysis succeeds.

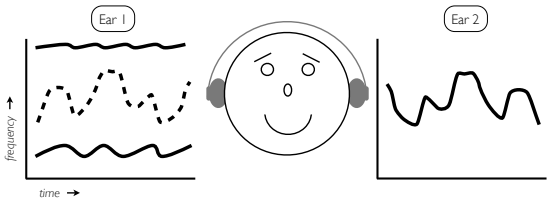

Figure 2. A schematic representation of a competitive test of perceptual organization. The tone components of a sine-wave sentence are presented dichotically, with the analog of the second formant presented to one ear and the remaining tones to the other. This is shown in the dark curves. An intrusive tone, shown here as a dashed curve, is presented at the same average frequency of the tone analog of the second formant, at the ear opposite the true tone analog. A listener must integrate the tone components of the sentence and reject the intrusive tone in order to resolve the speech stream. This paradigm was used by Remez (2001), Remez et al. (1994), and Roberts et al. (2010).

A third proof of the indifference of perceptual organization to typical acoustic effects of vocalization can be seen in studies of acoustic chimeras. In this method (Smith, Delgutte, & Oxenham, 2001) the envelope of a speech spectrum is analyzed and combined with the pattern of zero-crossings of an arbitrary signal. In a case reported by Remez (2008), the arbitrary signal was taken from a musical combination of woodwinds, brass and percussion, and the spectrum envelope taken from an unexceptional sentence of seven syllables. In the resulting bistable chimera, the linguistic properties are no less prominent that the melodic and instrumental characteristics of the musical source. (Listening examples of sine-wave, noise-band and chimerical speech are available on-line at Remez, 2008).

Overall, such studies show the relative importance of sensitivity to modulation independent of the specific details of the modulated carrier. Although an extensive survey of carrier types has yet to be performed, the most conservative conclusion consistent with these findings is that sensitivity to complex modulation promotes perceptual organization of a spectrum into a single contour; once the contour is resolved, perceptual analysis can occur. The most general conclusion consistent with these finding is that the carrier is responsible for eliciting an impression of instantaneous auditory quality; modulation is effective, causally, in the coalescing of the contour and in the resolution of linguistically governed contrasts.

The Gestalt grouping principles are largely silent about complex modulation of the kind observed in speech, although there have been many assays of amplitude comodulation of concurrent acoustic elements. One motive is the potential to discover comodulation in a neural ensemble, which would lend physiological plausibility to a portion of the Gestalt account. Unfortunately, comodulation of concurrent acoustic elements promotes binding only weakly, psychoacoustically, which significantly reduces the explanatory value of this property in the binding of elements into contours.

1.2.4. Nonsymbolic. The evidence reviewed here has shown that the perceptual organization of speech occurs by virtue of a perceiver's sensitivity to spectrotemporal variation. That is to say, isolation of a speech stream as a single contour within a busy acoustic environment depends on the perceptual resolution of the properties of modulation independent of the character of the modulated carrier. If a speech stream is unnatural, distorted or synthetic, then the coherence of the contour might also depend on the perceptual dexterity to attend to the modulation while ignoring the momentary properties of a sine-wave, noise-band or chimerical carrier. Yet, studies show clearly that attention to modulation is not governed by the likelihood that the contour is effective in evoking a distinct phonetic impression; neither is attention to modulation limited by perceptual success in deriving linguistic-phonetic features from the auditory stream. Perceptual organization of speech is distinctly nonsymbolic in nature.

The empirical paradigm that established this aspect of perceptual organization used a competitive test of integration (Remez et al, 1994; Roberts, Summers, & Bailey, 2010). Beginning with a sine-wave sentence, the tone components were arrayed across two ears. In a dichotic presentation, the tone analogs of the first, third and fourth formant were presented to one ear, and the tone analog of the second formant presented to the other ear. Intelligibility depended on fusing the dichotically arrayed tones, despite their dissimilarity in spatial location, frequency range, change in frequency, etc., because neither pattern presented to each ear was intelligible alone. In the competitive tests, a single intrusive tone that was the temporally reflected analog of the second formant was presented in the same ear as the tones replicating the first, third and fourth formant. It exhibited the same frequency and amplitude range, frequency and amplitude variation, and central tendency as the proper analog of the second formant, and was not spatially dislocated, as the true second formant analog was. In order to integrate the tones composing a synthetic utterance, a perceiver had to reject the spatially and spectrally similar intrusive tone and recruit the spatially dislocated second formant tone. (See Figure 2). When the attributes of the intrusive tone varied, its effectiveness as a perceptual lure also varied, and across a series of tests it was possible to identify the properties that drew perceptual resources for finding and following a speech stream.

The results showed that speechlike variation in frequency was the key attribute of the auditory pattern on which the binding of the contour depended. Amplitude variation alone played no role in promoting the integration of concurrent elements across the 5 kHz wide speech spectrum. Neither a constant frequency tone nor a frequency modulated tone oscillating around the average frequency at a steady rate of 5 Hz interfered with perceptual integration of the sentence. Imposing an affine strain on the competitor tone to produce a gradient of patterns ranging from constant frequency to natural frequency variation showed the corresponding graded competition in perception. When variation was natural, interference was complete; when variation was intermediate, interference was intermediate (Remez, 2001). And, when the competitor exhibited was a steady tone at the center frequency of the tone analog of the second formant, it did not interfere with perceptual organization at all.

In no case did the successful lures oppose perceptual resolution of the speech stream by evoking phonetic impressions. On the contrary, when the lure prevented the establishment of a coherent speech stream, this occurred by misleading the listener about its auditory composition. Concurrent tones that failed to compete were innocuous because their frequency variation was sufficiently unspeechlike to be disregarded. Because the speech stream was resolved incompletely, or with extraneous constituents, it was not well composed for an effective analysis. Linguistic phonetic properties, an aspect of the symbolic inventory of a language, were not evoked when perceptual organization of the auditory contour was erroneous.

1.2.5. Unlearned. In a competent listener, perceptual analysis of linguistic attributes can occur once the organizational functions of perception extract a contour from sensory samples. In doing so, the conditional relation between organization and analysis is demonstrated, namely, the decomposition of speech into its linguistic attributes begins with the resolution of the speech stream as a coherent contour within ongoing sensory activity. This is one way that perceptual organization is critical in the origin of perceptual experience. A second origin takes place during early development.

In the first few months of life, infants exhibit the ability to resolve speech in its time-varying complexity. This sensitivity to the coherence of the acoustically heterogeneous speech stream (observed at 14 weeks: Eimas & Miller, 1992; at 20 weeks: Newman & Jusczyk, 1995) antedates an infant's attention to the properties of speech that distinguish the native language prosodically and phonemically. Apparently, the organizational ability to integrate the constituents of speech despite dissimilarity in the spectra or timing of the constituents appears early in infancy, as this function must if it is to permit an effective perceptual analysis leading to the establishment of language. Some evidence indicates that the perceptual resolution of a complex auditory stream also confers protection from masking (Nittrouer & Tarr, 2011), promoting attention to the structure of a speech stream.

It is implausible that three months of exposure to airborne sound is sufficient for perceptual learning to induce sensitivity to complex patterns of modulation across 5 kHz. Although a fetus can hear a bit of the acoustic scene before birth, this can hardly count as much of a head start in finding and following a speech stream. Maternal tissues and fluid-filled volumes act collectively as a low-pass acoustic filter, permitting an auditory experience for the fetus of variation in voicing and little else. It is unlikely that much of the maternal speech spectrum above 500 Hz is even available to a fetus (Griffiths, Brown, Gerhardt, Abrams, & Morris, 1994). At birth, the transition from speech transmitted in utero to airborne speech imposes a stark discontinuity in a fetus's history of exposure to speech. The perceptual ability to bind the diverse properties of the speech spectrum based on the modulations characteristic of a vocal source of sound seems to develop without meticulous training or a protracted period of trial and error. Until it is proven that the sound issuing from a vocal source admits a simple description that is easy to master on minimal exposure and fast mapping, this intricate perceptual sensitivity seems to be unlearned.

2. MULTIMODAL PERCEPTUAL ORGANIZATION AND ANALYSIS OF SPEECH

Perceptual resolution of the coordinate variation of speech spectra differs in kind from perceptual grouping of acoustic elements based on simple similarity. In apprehending a speech stream as a single complex contour, a perceiver follows heterogeneous constituents across 5 kHz in a pattern produced by the modulation of a natural source of sound. The variety of the constituents and the spectrotemporal pattern of their distribution is specific to a phonologically controlled resonator no less than the sound of an oboe reflects the action of a reed exciting a column of air enclosed by a cone 65 cm long. The principles of the generic auditory account, alternatively, are devoted solely to the properties of sensory patterns, as if the accurate partition of a jumble of sound into its sources only required sorting by constituent, indifferent to the mechanics of sound sources. The principles were believed by Wertheimer as they are today to derive from properties intrinsic to a sensory nervous system.

Research on the perceptual organization of speech has naturally focused on auditory perceptual organization, acknowledging the sufficiency of the auditory sense to establish and to maintain the perception of spoken language. Considered more broadly, recent interest in the

merging of the senses has spurred the examination of audiovisual speech perception and the corresponding function of intermodal perceptual organization. With few exceptions, these studies have found that intermodal combination is ineluctable when a perceiver can see the talker while listening. In one of the first projects, Sumby and Pollack (1954) showed that audiovisual combination occurred even when the visible component was readily identifiable in silence, or that the audible component was readily identified without the aid of vision. When it occurs, intermodal combination is complete, and a perceiver usually cannot distinguish the visible from the audible properties in the mixture. Most surprising, the results of intermodal combination do not appear to be constrained by typicality or familiarity. In a study that established this by using an open response set, McGurk and MacDonald (1976) reported impossible consonant clusters in the protocols of their listeners. Evidently, integration of auditory and visual sensory samples occurs without the intrusion of belief, or expectation of the likely phonemic effects of speech perception (see section 1.2.4).

When auditory and visual samples are combined, the similarity criteria of the Gestalt account can have little to contribute descriptively or theoretically besides noting an unbridgeable gulf between the sensory elements of a visual contour and those of an auditory contour. Despite the incommensurate dimensions of vision and hearing and the resulting dissimilarity of their sensory grain, there are properties that auditory and visual samples share nonetheless. The time course can be similar in onset, duration and offset; and the spatial coordinates can be similar, specifically, the heading, which combines azimuth and elevation. However, studies that have sought to calibrate the role of temporal and spatial similarity reveal that intermodal binding discounts dissimilarity in these attributes, as if integration were indifferent to coincidence in these properties (Bertelson, Vroomen, & deGelder, 1997; Munhall, Gribble, Sacco, & Ward, 1996). Although the principles of intermodal combination are not yet described completely, it is clear that audiovisual combination cannot be attributed to the similarity of the sensory details. Neither does temporal or spatial dissimilarity impede audiovisual binding.

2.1. Audiovisual rivalry and audiovisual agreement.

Many studies have examined intermodal integration using a variant of the McGurk effect. In this paradigm, a composite is created to present a discrepant video display and an acoustic sample. For example, in the report that launched this work (McGurk & MacDonald, 1977), a video of a talker saying the syllable [gɑ] was combined with an acoustic sample of speech of the syllable [bɑ]. When the perceiver identified the presentation as an instance of [dɑ] the integration was seen as evidence of the combination of the two sensory streams, and the consequent analysis incorporated features from both modalities. Because some phonetic features, like a voicing contrast, are thought to be more commonly confused visually than auditorily; while others, like consonantal place, are thought to be more confusable auditorily than visually, it has been possible to construct an account of integration based on the hypothetical reliability of the information obtained from each sense. This view of integration has been described as the Principle of Inverse Effectiveness (Stein & Meredith, 1993). Through the instrumental use of visual and acoustic noise and other techniques to titrate uncertainty at the moment of recognition, it has been possible to mathematize this approach to intermodal integration (Massaro & Stork, 1998).

One consequence of this practice is that there have been few studies after Sumby & Pollack that examine the integration of auditory and visual sensory streams that actually belong together. Inasmuch as this original study is acknowledged to count as counterevidence to the Principle of Inverse Effectiveness, some projects have directly concerned the perceptual effect of correspondence across the senses — as opposed to the reduction of conflicting information between the senses (Tye-Murray, Sommers, Spehar, Myerson, & Hale, 2010). The challenge that these studies offer to the conventional view is that of superadditivity. Specifically, under some conditions of presentation, neither the visual nor the auditory speech stream alone are intelligible. When this occurs, it is neither possible to fix the likelihood of identification in each modality nor the uncertainty independent of combination. Nonetheless, some reports describe cases in which visible and audible streams are not separately identifiable yet the combination is very nearly normal in intelligibility, blocking an account of integration that depends on a prior assessment of the reliability of the information available in each sense (Bernstein, Auer, & Moore, 2004; Remez, Dubowski, Ferro, & Thomas, 2013; Rosen, Fourcin & Moore, 1981). Instead, the evidence suggests that intermodal organization is preliminary to phonetic analysis.

Considering this set of studies, it seems as if the disposition to integrate depends on variation intrinsic to the sensory streams, independent of the symbolic properties that can or cannot be extracted from each sense independent of the other. Studies of the neural correlates of integration have been intriguing but limited in identifying the underlying pathways of superposition or integration (for example, Besle, Bertrand, & Giard, 2009; Calvert el al., 1997; Lange, Christian, & Schnitzler, 2013; Sams, et al., 1999) Although these studies have encouraged the hypothetical convergence of visual and auditory streams in a secondary auditory cortical area suspected in sensory analysis of speech, the form of the combination is not known. Overall, direct integration of visual and auditory

contours seems unlikely, due to the incommensurate dimensions of the sensory samples. Could multimodal integration occur in a metric common to auditory and visual contours, whatever this might be? There is little evidence — though significant conjecture (Rosenblum, 2005) — of steps that derive an amodal code for auditory and visual contours, or for the reliance on one or the other sensory constituents of audiovisual perceptual organization.

3. *EMPIRICAL AND THEORETICAL CHALLENGES.*

The study of perceptual organization began a century ago with the premise that contours in a sensory field were composed according to an ordinary notion of simplicity. The inherent acoustic complexity of speech made it an attractive test case for the generic auditory account of perceptual organization inspired by such simplicity principles. A brief but intense history of technical attention to modeling vocal sound production in language made the ensuing tests sharp. In consequence, the findings show that the generic auditory account offers no plausible explanation of the binding of discontinuous, periodic and aperiodic, spectrally distributed and asynchronously varying constituents of speech into a coherent stream. The generic auditory account explains well the formation of streams from acoustic patterns composed narrowly to test its premises, but fails with this natural sound source. Whether the Gestalt-derived description of perceptual organization would survive other tests with natural sound sources (for instance, Iverson, 1995) remains for researchers to discover.

Although the empirical case for perceptual organization by sensitivity to complex modulation is made chiefly by tests using speech, no evidence indicates that this form of sensitivity is a specialized function. So few tests of perceptual organization have been conducted with ordinary sound sources — those that exhibit resonant properties (Gaver, 1993), in contrast to the linear emitters typically used to create acoustic test patterns — it is premature to conclude that simplicity is a natural default and complexity a departure achieved only by specialization. In view of the findings that organization by similarity of elements takes time to build, and requires effort to establish and to maintain (Carlyon et al., 2001) the evidence favoring the Gestalt-derived theory of perceptual organization seems far weaker now than ever, independent of tests with speech.

Going forward, the empirical and theoretical challenges concern the generality of modulation sensitivity exhibited in the perceptual organization of speech. In the audiovisual instance, it also seems as though the characteristics of visible and audible sensory samples are inherently different (Julesz & Hirsh, 1972), standing as an ultimate challenge to a straightforward notion of similarity or modulation sensitivity to promote binding. Because it seems likely that intermodal integration is required for perceptual analysis of speech to occur, audiovisual integration must be based on a rather abstract form of modulation compatible with both senses. Considering the tolerance for temporal and spatial dissimilarity, the confluence of audible and visible streams defies all present accounts, and there is potentially much to be discovered about perceptual organization from considering that setting of the scientific question.

REFERENCES

Bernstein, L. E., Auer, E. T., Jr., & Moore, J. K. (2004). Audiovisual speech binding: Convergence or association. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of Multisensory Processing* (pp. 203-223). Cambridge, MA: MIT Press.

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). September 26-27, 1997. Rhodes, Greece: ESCA.

Besle, J., Bertrand, O., & Giard, M.-H. (2009). Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex. *Hearing Research*, 258, 143–151.

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

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

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iverson, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. Science 276 594-596.

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, 115-127.

Darwin, C. J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society B*, 363, 1011–1021.

Ehrenfels, C. von. (1890). Über Gestaltqualitäten. [On the qualities of form.] *Vierteljahrsschrift fur wissenschaftliche Philosophie*, 14, 249-292.

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

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology*, 5. 3 e1000302.

Feng, Y.-M., Xu, L., Zhou, N., Yang., G., & Yin, S.-K. (2012). Sine-wave speech recognition in a tonal language. *Journal of the Acoustical Society of America*, 131, EL133-EL138.

Fu, Q.-J., & Galvin, J. J., III. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users. *Journal of the Acoustical Society of America*, 109, 1166-1172.

Gallistel, C. R. (2007). Flawed foundations of associationism? *American Psychologist*, 62, 682-685.

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

Greenberg, S., & Arai, T. (1998). Speech intelligibility is highly tolerant of cross-channel spectral asynchrony. In P. Kuhl & L. Crum (Eds.), *Proceedings of the Joint Meeting of the Acoustical Society of America and the International Congress on Acoustics* (pp. 2677-2678). Melville, New York: Acoustical Society of America.

Griffiths, S. K., Brown, W. S., Jr., Gerhardt, K. J., Abrams, R. M., & Morris, R. J. (1994). The perception of speech sounds recorded within the uterus of a pregnant sheep. *Journal of the Acoustical Society of America*, 96, 2055-2063.

Hochberg, J. (1974). Organization and the Gestalt tradition. In E. C. Carterette and M. P. Friedman (Eds.), *Handbook of Perception, Vol. I: Historical and Philosophical Roots of Perception* (pp. 179-210). New York: Academic Press.

Holt, L. L., & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception & Psychophysics*, 72, 1218-1227.

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception & Performance*, 21, 751-763.

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.

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.

Lange, J., Christian, N., & Schnitzler, A. (2013). Audio–visual congruency alters power and coherence of oscillatory activity within and between cortical areas. *NeuroImage*, 79, 111–120.

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: Mouton.

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, 549-558.

Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration: A 240 year old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, 86, 236-244.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

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

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, 58, 351-362.

Newman, R. S., & Jusczyk, P. W. (1995). The cocktail party effect in infants. *Perception & Psychophysics*, 58, 1145-1156.

Nittrouer, S., & Tarr, E. (2011). Coherence masking protection for speech in children and adults. *Attention, Perception & Psychophysics*, 73, 2606–2623.

Pastore, N. (1971). *Selective History of Theories of Visual Perception: 1650-1950*. New York: Oxford University Press.

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

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni and R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 182-206). Oxford: Blackwell.

Remez, R. E. (2001). The interplay of phonology and perception considered from the perspective of organization. In E. V. Hume and K. A. Johnson (Eds.), *The Role of Speech Perception Phenomena in Phonology* (pp. 27-52). New York: Academic Press.

Remez, R. E. (2005). Perceptual organization of speech. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 28-50). Oxford: Blackwell.

Remez, R. E. (2008). Sine-wave speech. In E. M. Izhikovitch (Ed.), *Encyclopedia of Computational Neuroscience* (pp. 2394). (Cited as *Scholarpedia*, 3, 2394.)

Remez, R. E., Dubowski, K. R., Ferro, D. F., & Thomas, E. F. (2013). Audiovisual asynchrony tolerance in the perceptual organization of speech. Unpublished research report.

Remez, R. E., Ferro, D. F., Wissig, S. C., & Landau, C. A. (2008). Asynchrony tolerance in the perceptual organization of speech. *Psychonomic Bulletin & Review*, 15, 861-865.

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, 129-156.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-949.

Remez, R. E., & Thomas, E. F. (2013). Early recognition of speech. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4, 213-223.

Remez, R. E., Thomas, E. F., Dubowski, K. R., Koinis, S. M., Porter, N. A. C., Paddu, N. U., Moskalenko, Marina, Grossman, Y. (in press). Modulation sensitivity in the perceptual organization of speech. *Attention, Perception & Psychophysics*, 00, 000-000.

Roberts, B., Summers, R. J., & Bailey, P. J. (2010). The perceptual organization of sine-wave speech under competitive conditions. *Journal of the Acoustical Society of America*, 128, 804-817.

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

Rosenblum, L. D. (2005). Primacy of multimodal speech perception, In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78). Oxford: Blackwell.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature*, 398, 760.

Sams, M., Sulendo, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Luh, S., & Sinola, J. (1991). Seeing speech: Visual formation from lip movements modify activity in the human auditory cortex. *Neuroscience Letters*, 127, 141-145.

Searle, J. R. (1981). The intentionality of intention and action. In D. R. Norman (Ed.), *Perspectives in Cognitive Science* (pp. 207-230). Hillsdale, NJ: Ablex.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

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

Stein, B. E., & Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, Massachusetts: MIT Press.

Stevens, K. N. (1999). *Acoustic Phonetics*. Cambridge, Massachusetts: MIT Press.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the Principle of Inverse Effectiveness. *Ear & Hearing*, 31, 636-644.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt, II. Psychologische Forschung, 4, 301-350. [Translated as, "Laws of organization in perceptual forms," in W. D. Ellis (Ed.), *A Sourcebook of Gestalt Psychology* (pp. 71-88). London: Routledge & Kegan Paul, 1938.]