PERCEPTUAL ORGANIZATION OF SPEECH IN ONE AND SEVERAL MODALITIES: COMMON FUNCTIONS, COMMON RESOURCES

Robert E. Remez

Department of Psychology
Barnard College
3009 Broadway
New York, New York 10027-6598

ABSTRACT

In order to understand speech the perceiver meets two challenges: 1) to Þnd a speech signal within ongoing sensory activity, and 2) to project its properties into linguistic phonetic attributes. These functions have customarily been designated as perceptual organization and perceptual analysis. The case of multimodal perceptual organization is revealing to consider because the perceiver Þnds sensory ingredients spanning modalities. Contemporary accounts offer alternative conceptualizations of these functions based largely on the study of single modalities. A Gestalt-derived account hypothesizes that perceptual organization precedes analysis, grouping sensory elements into perceptual streams by a variety of similarity criteria. An account deriving from probabilistic functionalism describes analysis occurring within modalities preceding a stage of organization that binds the derived features. These alternatives and their hybrids appear implausible on empirical and theoretical grounds for accommodating multimodal perceptual organization. Additionally, our studies using sinewave replicas of utterances reveal that the customary models are untenable accounts of unimodal no less than multimodal perceptual organization. A third way, justiÞed by our results, describes auditory perceptual organization of sinewave sentences as a speciÞc instance of the general susceptibility to coherent sensory variation. This account potentially allows a single description of uni- and multimodal perceptual organization.

1. CONTRASTING APPROACHES TO PERCEPTUAL ORGANIZATION

Attempts to explain the perception of speech exhibit a common feature despite their differences. SpeciÞcally, it has regularly been assumed that the analysis of linguistic properties simply begins with a speech signal. By presupposing a raw signal, neatly isolated within an organized Þeld of concurrent sensations, such accounts of perception tacitly restrict the application of phonetic analysis to the sensory properties of a single stream of speech. Admittedly, this gambit relieves the necessity of explaining many subsidiary processes that contribute to perceptual analysis, though it is reasonable only if the explanations of perceptual organization are satisfactory. Our recent attention to organizational matters has exposed the inadequacies of two familiar accounts, and instead proposes an alternative description of the perceptual organization of speech [7]. Although our work has aimed to describe speech perception by ear alone, the formulation that we derive from this evidence is compatible with the observations of multimodal speech perception. Accordingly, the goal of this brief note is to review competing conceptualizations of perceptual organization, to identify the challenge to these views inherent in multimodal perception of speech, and to present some of the evidence that unimodal and multimodal speech perception is organized by similar principles.

To observe that perceptual analysis and perceptual organization are contingent has not always seemed like a recommendation to organize Þrst, analyze second. One contemporary approach to this topic [11; cf. 12] depicts the contingency of organization and analysis as a feature binding problem, which describes the aggregation of the reports of analyzers as object descriptions. This approach recalls the spirit of Brunswik's probabilistic functionalism, in which the perceptual apprehension of objects and events is described as beginning with unaggregated sensory elements, and as culminating with the determination of the likeliest distal cause. Such an account is plausible if the acoustic cues can be listed in a table of probabilities, for this actuarial approach to perception requires the memorization of correspondence between typical acoustic elements and typical phonetic features. The model of Massaro [4] is a variant of this approach, in which feature binding is achieved by comparison of a sensory array to prototypes of items in the distal set.

Although the occasional slip of the ear may recommend this explanation, as if it were a mistaken binding of veridically analyzed consonant or vowel features, this ordering—analyze then organize—has not been pursued consistently in speech research, and it is easy to see why. None of the acoustic elements that compose a speech signal is unique to speech. Instead, it seems as though the phonetic value of an element of a speech signal depends on its conÞguration, and even within a speech stream the same acoustic element changes its phonetic valence in different contexts[3, 5, 10]. Under such conditions, the organization of the auditory world into perceptual streams must precede phonetic analysis, and in this respect the traditional formulation of Wertheimer [13] has been prominent.

The cases considered by Wertheimer are familiar to every student of introductory psychology as the organizational principles of proximity, similarity, common fate, set, continuity, symmetry, closure and habit. Essentially, these terms name the dimensions along which plane shapes or of tone sequences seem to compose groups. Perceptual analysis of objects occurs once the Þeld of stimulation is organized by the application of these principles, according to the clariÞcation of this viewpoint described by Julesz & Hirsh [2]. An explicit multistage model, *auditory scene analysis* [1] offers the closest thing to a standard account of organization and analysis in this vein, and has been widely in_uential in the cognitive sciences. Its organizational functions begin by applying principles derived from those of Wertheimer to an acoustic array, forming groups of like

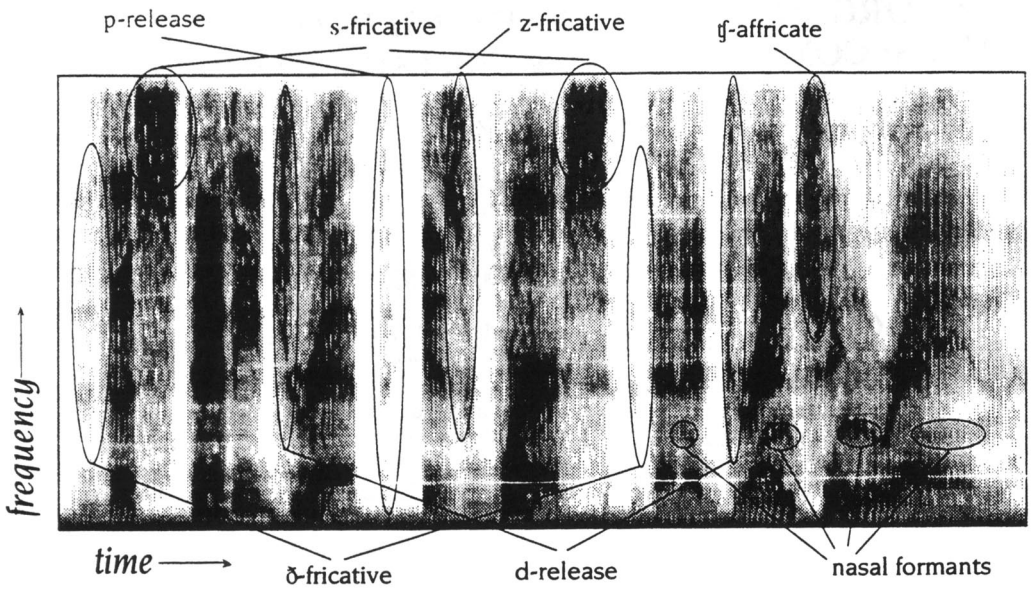

Figure 1: A spectrogram of the sentence, "The steady drip is worse than a drenching rain," analyzed into its acoustic constituents subject to perceptual organization [7]. See text.

elements, each group segregated from the others. Perceptual analysis applies separately to each segregated group of elements, or stream. It is unfortunate for theories of speech perception that would assert a standard account of perceptual organization that auditory scene analysis generally gives incorrect descriptions of speech signals.

2. GESTALT-BASED ORGANIZATION AND PHONETIC ORGANIZATION

Based on a review of the spectrotemporal criteria for stream formation given in auditory scene analysis, we recently considered a sentence produced in a quiet background, and characterized it from the point of view of a Gestalt-based [7]. The results were not encouraging of the standard account. Basically, the acoustic constituents of an unexceptional utterance, "The steady drip is worse than a drenching rain," exhibit sufÞcient variety and discontinuity to fracture into separate streams of like elements (see Figure 1). Each of the oral formants onset and offset, or rose and fell in amplitude and frequency asynchronously, at different rates and to different extents, acoustic properties that lead to segregation into three separate streams according to Gestalt-based criteria. Nasal formants appeared and disappeared rapidly and discontinuously in the spectrum, constituting a fourth stream. Release bursts of voiceless stops differed as well from voiceless affricate releases and from voiced stop releases, and voiced friction differed in spectrum from voiceless friction, constituting the Þfth, sixth, seventh, eighth and ninth streams. The spectra of fricatives also differed with articulatory place, promoting segregation of linguo-dental friction from apical friction, composing the tenth stream. Clearly, application of the standard principles of grouping fracture a speech signal into multiple streams instead of preserve its coherence.

Such principles will parse an acoustic world into streams according to sources only when the elements common to a sound source are physically similar to each other. The principles fail to organize speech because the acoustic constituents are heterogeneous, including whistles, clicks, hisses, buzzes and hums. The problem of organizing speech signals can be deÞned as one of detecting coherence despite the dissimilarity and discontinuity of the constituents, and framed in this way it is possible to see how a characterization of perceptual organization for the listener is applicable to the multimodal circumstance in which the heterogeneous sensory elements span senses.

Our proposal, at Þrst approximation, was that speech signals are organized according to principles outside the Gestalt-based set. Before we recommended this alternative, though, we had to rule out a role in speech organization for the schema-driven error handler that auditory scene analysis uses to survive mistakes imposed by the basic level Gestalt-based process. The schematic device leaves organization to the moderating effects of learning or effortful attention, thereby to form perceptual streams that conform to typical sensory manifestations of some sound sources that the Gestalt processor misses.

3. PERCEPTUAL ORGANIZATION OF SINEWAVE SENTENCES

Our experiments took three forms. In each test, the acoustic test materials were tonal analogs of speech [8]. In this kind of copy synthesis, time-varying sinusoids replicate the estimated amplitude and frequency changes of oral, nasal and fricative formants. The resulting tone complexes evoke the perceptual incoherence warranted by Gestalt rules, and naive listeners simply report hearing several simultaneous tones when sinewave sentence replicas are presented to them. However, an instruction to transcribe a synthesized sentence was often sufÞcient to allow listeners to group the tones phonetically, forming a speech stream despite the violation of grouping principles and the durable impression of unspeechlike timbre. This Þnding encouraged a claim that phonetic organization occurred neither

through Gestalt-based nor schematic resources. First, while Gestalt-based organization split the tone complex into its individual components, as it should have, phonetic properties were apparent at the same time, as if two concurrent organizations were available to the listener. This established the likelihood that something other than Gestalt rules were responsible for phonetic coherence. Second, the great physical and psychoacoustic difference between the acoustic products of natural vocalization and the pure-tone replicas argued that sinewave replicas of speech would fail to satisfy a schematic representation of the typical acoustic correlates of phonetic segments.

Two kinds of dichotic listening test conÞrmed this premise. First, we arrayed the tones across the ears, to determine whether phonetic perception of sinewave sentences required the components to originate from the same location. Had listeners failed to identify the words when Ear 1 heard analogs of the Þrst and third formant and Ear 2 the analog of the second formant, we would know that spatial similarity, a Gestalt principle of organization, was responsible for establishing the coherence of the tones. In fact, listeners fused the tones across the ears despite the spatial discrepancy, reporting the sentences [7]. The crucial evidence here was that dichotic performance exceeded the combination of each ear's contribution, estimated in two control conditions. Once again, the anomalous spectra of sinewave sentences block an explanation of perceptual organization that appeals to schemas representing the likely acoustic manifestations of phonetic features.

Further evidence resolving the non-Gestalt principles in the perceptual organization of speech come from a study of dichotic competitive presentation of sinewave sentences. The format of the test is illustrated in Figure 2. It shows the components of a sinewave replica distributed across the ears. The listener must integrate acoustic elements composing the sentence despite spatial and other dissimilarity. A sentence replica lacking its second formant analog is presented to one ear, and the second formant tone by itself is presented to the opposite ear. Crucially, a foil of the second formant tone is presented to the same ear as

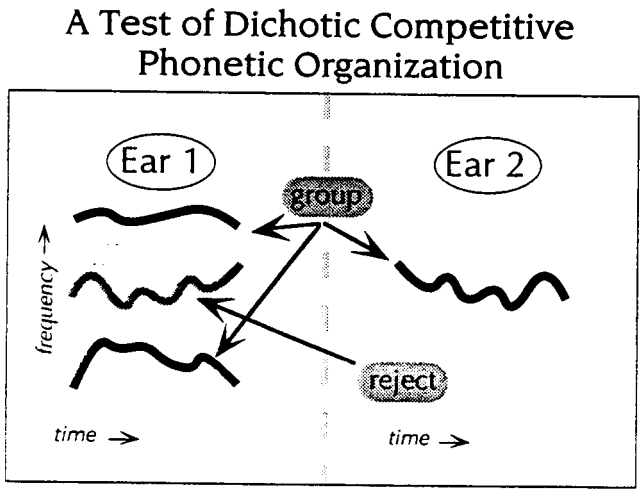

Figure 2: A schematic description of dichotically presented sinewave sentence, with an extraneous tone in the frequency region of the second formant. Dark lines represent tonal analogs of the formants of a natural utterance; gray line represents an extraneous tone added to the signal.

the sinewave pattern lacking its second formant tone. Here, the subject must reject a spatially coherent though phonetically incoherent element in the presentation and fuse the dichotically presented true second formant analog of the sentence. In the test, we varied the likeness of the second formant foil tone to speech, on the hypothesis that the principle of phonetic organization includes a time-varying Þlter that passes speechlike variation in the coarse spectrotemporal grain. Although Gestalt rules would split the second formant foil tone into its own stream, apart from any of the other formant analogs, an organizer keyed to speechlike spectrotemporal properties should group it with speech, which we can see by the effect on transcription of the dichotically fused tones of the sinewave sentence; unspeechlike tones, even those that are nonstationary, should be blocked, and should not interfere with dichotic fusion of the phonetically coherent tones.

We varied the speechlikeness of the second formant foil tone by imposing a frequency strain on a temporally re_ected version of the true tone analog of the second formant. At one extreme, the foil exhibited the natural range of frequency variation. To produce less speechlike spectrotemporal properties in other conditions, variation around the mean frequency was reduced 33%, 67%, or completely, at which extreme the foil became a constant frequency tone at the average frequency of the true second formant. Performance was compared to the condition in which the foil was a dithering tone, nonstationary but also nonphonetic, in which 200 ms tone segments, one 10% greater and the other 10% lower in frequency than the mean of the second formant, alternated; and with performance when there was no competing tone. The results of this test are shown in Figure 3. It is apparent that the more speechlike the foil tone, the more it competed with organization of the dichotically presented formant analogs of the sentence. Clearly, too, the dithering tone and the constant frequency tone interfered minimally with phonetic organization, as shown in the transcription performance.

4. UNIMODAL AND MULTIMODAL PERCEPTUAL ORGANIZATION

The results of our investigations indicate that when the perceiver listens to speech, the superÞcial sensory form of the signal may matter far less for organization than the pattern of spectrotemporal change, which must be consistent with phonetically governed sound production. However, the speechlike variation that drives organization does not apparently evoke a phonetic feature analysis, for our studies have shown that a single tone varying in a phonetic manner—exactly the kind of element that a phonetic organizer recruits—is not analyzable phonetically even when a listener is given ample time and rehearings. Evidently, organization does not depend on a success of symbolic analysis, and is distinct therefore from varieties of pandemonium in contemporary models.

The problem that led to this line of research was the unmistakable heterogeneity and discontinuity of the acoustic elements of speech. Organization for the listener is a function that establishes unity among the constituents of an auditory sensory register despite dissimilarities that violate the Gestalt rules. Disparity of the elements undergoing organization is self-evident in the multimodal case, requiring a principle of

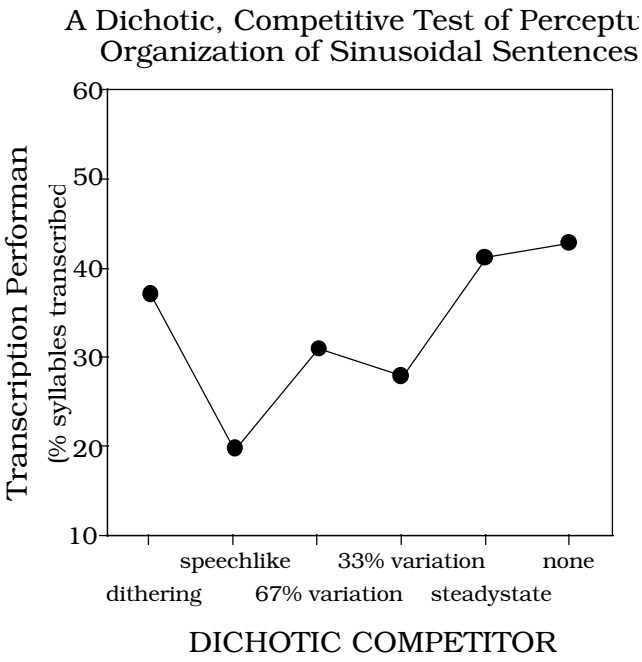

Figure 3: Group results of a test of phonetic perceptual organization using dichotic presentation of tones competing with the second-formant analog.

coherence that disregards superÞcial differences, for example, between the visual projection of the shape of the vermilion border or self-occluding edge of the lips and the auditory timbre of a vowel. In contrast to accounts of perceptual organization that assert an independence of the senses, our research yields a view of perceptual organization that is general, and reasonably extrapolated to a case in which the listener also views the talker. Recent Þndings, including studies that manipulate the intersensory discrepancies temporally [6] and spectrally [9] expose the formal equivalence of perceptual organization in the multimodal detection of correspondence and in the portrait of organization that we have offered in general form, though, admittedly, from consideration solely of the circumstances of auditory perceptual organization. Although research on auditory organization presents the problem in a rather subtler form than the multimodal presentation does, the precedent of the sinewave studies reveals a means by which to derive principles of general use from modality-speciÞc cases.

REFERENCES

1. Bregman, A. S. *Auditory Scene Analysis*. Cambridge: MIT Press, 1990.

2. Julesz, B., & Hirsh, I. J. “Visual and auditory perception: An essay of comparison.” In E. E. David & P. B. Denes (Eds.), *Human Communication: A UniÞed View* (pp. 283-340). New York: McGraw-Hill, 1972.

3. Liberman, A. M., Delattre, P. & Cooper, F. S. “The role of selected stimulus-variables in the perception of the unvoiced stop consonants.” *American Journal of Psychology* 65: 497-516, 1952.

4. Massaro, D. W. “Psychological aspects of speech perception: Implications or research and theory.” In M. A. Gernsbacher (Ed.)., *Handbook of Psycholinguistics* (pp. 219-263). New York: Academic Press, 1994.

5. Miller, J. L., & Liberman, A. M. “Some effects of later-occurring information on the perception of stop consonant and semi-vowel.” *Perception & Psychophysics* 25: 457-465, 1979.

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

7. Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. “On the perceptual organization of speech.” *Psychological Review* 101: 129-156, 1994.

8. Remez, R. E., Rubin, P. E., Pisoni, D. P., & Carrell, T. D. “Speech perception without traditional speech cues.” *Science* 212: 947-950, 1981.

9. Saldaña, H. M., Pisoni, D. B., Fellowes J. M., & Remez, R. E. “Audio-visual speech perception without speech cues: A Þrst report.” In D. G. Stork (Ed.), *Speechreading by Man and Machines: Models, Systems and Applications*. New York: Springer-Verlag. (in press).

10. Shattuck, S. R., & Klatt, D. H. “The perceptual similarity of mirror-image acoustic patterns.” *Perception & Psychophysics* 20, 470-474, 1979.

11. Triesman, A. “The perception of features and objects.” In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, Awareness and Control: A Tribute to Donald Broadbent* (pp. 5-35), 1993.

12. Triesman, A., & Gelade, G. “A feature-integration theory of attention.” *Cognitive Psychology* 12: 97-136, 1980.

13. Wertheimer, M. “Laws of organization in perceptual forms.” In W. D. Ellis (Ed.), *A Sourcebook of Gestalt Psychology* (pp. 71-88). London: Kegan, Paul, Trench & Teubner 1938. (Original work published in 1923).

Acknowledgment. The author gratefully acknowledges the advice of Stefanie Berns, Lila Braine, Jennifer Fellowes, Julian Hochberg, and Philip Rubin. This research was supported by grant DC00308 from the National Institute on Deafness and Other Communication Disorders.