A MORE INTIMATE COCKTAIL PARTY PHENOMENON: THE PERCEPTUAL ORGANIZATION OF TONAL ANALOGS OF SPEECH

Robert E. Remez

Barnard College, New York, New York U.S.A.

Perceptual organization of auditory patterns is often explained by appeal to Gestalt grouping principles. Despite the evidence for such principles in the grouping of simple acoustic displays, the time-varying nature of speech spectra eludes a similar account. Studies with sinewave analogs of speech show that auditory grouping is not required for phonetic perception, nor are principles of the Gestalt variety sufficient to explain the perceptual integration of oral, nasal, and fricative formants. We have identified a grouping principle keyed to speechlike spectral change, in tests using dichotic sinusoidal components which are grouped both in phonetic and auditory modes. Our findings warrant extending the simple characterizations available within the framework of the Gestalt laws.

1. PRINCIPLES OF PERCEPTUAL ORGANIZATION

In 1923, Max Wertheimer [16] reported the results of an inquiry into perceptual organization. His quest was to show that perception of an ambiguous display was organized, and not a simple summary of the elements of stimulation. Our recent concern has been the validity of these Gestalt principles in accounting for the perceptual organization of speech, and is occasioned by the fact that Wertheimer's principles are still very much with us in auditory form. They are often cited as a preliminary step in auditory recognition of objects and events. Supporting evidence comes from many studies of arbitrary acoustic displays, though we have recently submitted the principles to test using speech signals, or replicas presenting spectro-temporal attributes common to speech. These new studies are not encouraging about the descriptive or theoretical adequacy of the Gestalt account when it is applied to the case of speech.

Wertheimer exposed the principles using ambiguous plane shapes and brief tone sequences, deriving a collection of perceptual devices for grouping the figural elements of stimulation: proximity, similarity, common fate, set, continuity, symmetry, closure, and habit. Explicitly auditory instances of many organizational principles were again offered by Julesz & Hirsch [8] who sought common principles for perception in visual and auditory modalities. Although they concluded that the dissimilarities of vision and hearing outweighed the shared attributes, their review brought an influential information-processing rationale to subsequent studies. Julesz & Hirsch themselves contended that the Gestalt organizational principles alone might prove inadequate to explain the perceptual integrity of the speech signal, due both to its acoustic complexity, and to the contribution to perception of the listener's extensive knowledge of speech and language. In the 25 years since this article appeared, a large body of evidence has been gathered about its numerous hypotheses. These studies

support the detailed claims that an auditory scene is organized or analyzed perceptually according to principles of proximity [2], similarity [3, 15, 6], common fate [1, 4] and closure [10], operating in the domains of frequency, amplitude, and spectrum. The application of the grouping principles is held to promote the formation of separate auditory streams, which, once sorted, are passed along for more detailed perceptual analysis about the objects and events which gave rise to the stimulation.

2. PERCEPTUAL ORGANIZATION OF SPEECH SIGNALS

Our question is simple: Do the diverse components of a single speech signal cohere perceptually through the application of Gestalt grouping principles? Pertaining to speech, this question is typically framed about the isolation of a single voice against an acoustic background of other talkers, clinking glasses, popping champagne corks, and whirring air conditioning systems; in short, the familiar "cocktail party phenomenon" [5]. But our present focus on perceptual organization requires a more intimate setting. After all, the listener must perceive that a single speech signal produced in a quiet room is the product of a single vocal source of sound. Does streaming account for that?

Current formulations of auditory perceptual organization warrant the fracture of a speech signal into perceptually incoherent streams, rather than accounting for the fusion of the diverse acoustic components into the single ongoing perceptual event which the listener hears. This outcome is due to the reliance of the grouping principles on durable similarities or coordinate changes occurring among the elements of the incident acoustic pattern. This is the clear and unavoidable consequence of grouping acoustic elements by physical similarity, physical continuity and common, coordinate transient characteristics. The fracture of speech into incoherent streams then follows from the acoustic nature of the speech signal, which is ordinarily replete with failures of similarity, continuity and common fate. These familiar acoustic attributes are observed when the frequency changes of one formant center do not match frequency changes of another formant in direction, degree, or duration; when onsets and offsets fail to occur in synchrony; and when episodes of nasal and fricative formants occur, lacking frequency continuity and spectral similarity with the oral resonances. None of these types of lapse is particularly exotic, as any spectrogram will reveal [7].

Despite this acoustically diverse collection of elements, the listener's perception is typically of a single stream of consonants and vowels, and not of disjoint simultaneous streams, each comprising a single kind of auditory element. Disintegrated impressions can occur when a brief snippet of a speech signal is presented in a rapidly repeating train [9]. But, the specific conditions required to elicit such impressions serve to underscore the difference between the perception of speech and the segregation of auditory components through streaming.

Were we to suspect that the common vocal excitation in the formants ordinarily holds them together perceptually—a kind of common fate—we would nonetheless have a hard time explaining phonetic perception of tonal analogs of speech. Here, the co-modulation of formant centers is eliminated by the use of digital synthesis to compose a collection of linear emitters which convey the momentary acoustic maxima. The familiar timbres of consonants and vowels are not evoked by such resonance-free and grossly unnatural short-term spectra [14], yet phonetic perception occurs nonetheless. The listener's ability to transcribe these odd replicas of speech depends on the perceptual disposition to treat three- and four- tone analogs as coherent despite their violation of grouping principles and unfamiliar timbre [11, 12, 13].

3. TESTS WITH SINEWAVE REPLICAS OF SPEECH

In tests to determine the kind of organization occurring in tone analogs of speech we have tried to distinguish perceptual organization, in which simultaneous dissimilar components are actually integrated, from a low-level peripheral fusion of sinusoids, due, perhaps, to auditory coupling of the first and second formant tones. The latter possibility is a likely mechanical consequence of some models of basilar function [17], and if true eliminates much of the interest in the case of sinewave analogs.

To determine the likelihood that sinusoidal components in a tonal analog are organized due to peripheral interactions at transduction, our tests used dichotic presentation requiring the listener to integrate a single tonal component presented to one ear—corresponding to the second formant—with the remainder of the replica presented to the other ear. Were transcription to deteriorate in this dichotic presentation, relative to the binaural case, we would conclude that (i) perceptual organization is a trivial consequence of auditory transmission of tonal components; and, (ii) disjunctive azimuth precludes active perceptual organization. Were transmission to survive dichotic presentation of essential acoustic ingredients, we would conclude that (iii) organization is not attributable to passive conduction of tonal components; and, (iv) failures in similarity, continuity, common fate—not to forget azimuth—are insufficient to prevent phonetic organization from occurring.

Our first test compared binaural and dichotic presentation of sinewave replicas. The second test assessed the selective power of phonetic organization by requiring the listener to combine the appropriate dichotically presented components despite the presence of a competing speechlike tone.

The acoustic materials which we used in these tests were sinusoidal replicas of utterances of sentences. In the basic dichotic test, one ear received the tones corresponding to the first, third and fourth formant centers, while the other ear received only the tone corresponding to the second formant center. In one control, the dichotic condition was compared with the binaural presentation of the full tone set. In another, the intelligibility of the patterns presented to each ear in the dichotic case was assessed in binaural tests; one test examined the information available from Tones 1, 3, and 4, lacking the second formant tone; the other assessed the phonetic effects of of Tone 2 alone.

Second, we ran a test of organization with interfering acoustic material, to identify the acoustic criteria of the organizational principles. In this test, the phonetically coherent tones were presented dichotically, along with a foil in the ear opposite Tone 2. This distractor tone either exhibited speechlike variation or constant frequency, though neither distractor satisfied the acoustic criteria for grouping with other concurrent tones. Although neither distractor was phonetically coherent, we expected only the tone with speechlike properties to compete organizationally with the true second formant tone.

4. THE FINDINGS

Can listeners integrate tonal components presented dichotically? The first test compared transcription accuracy for ten sentences in four conditions, and Figure 1 portrays the outcomes. Integration occurred despite violations of Gestalt grouping principles, for the dichotic performance surpassed the combination of each ear's contribution from a partial signal. Note, also that there is a clear performance decrement with dichotic organization relative to the binaural case, perhaps reflecting attentional load differences.

Our question in the second test series derived from the first: Is phonetic perception driven by an organizational principle that works by acoustic similarity? If so, then listeners should be indifferent to the presence of a tone

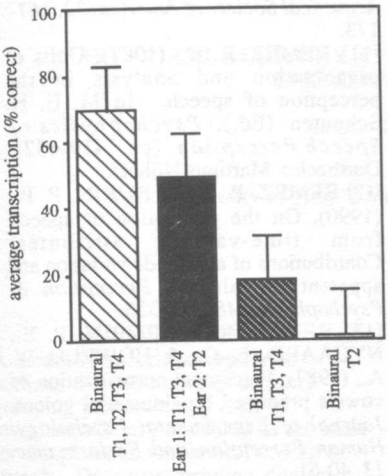

Figure 1. Group results for the perceptual organization test.

that is both dissimilar to the acoustic components of a speech signal and incoherent, in the sense that it could not have issued from a the same vocal source as the other tones in the presentation. We tested this with a condition requiring perceivers to reject a temporally flipped second formant tone in the pattern at one ear and to integrate the dichotically available veridical second formant tone. Neither second formant tone was similar in the Gestalt sense to the acoustic ensemble of Tone 1, Tone 3, and Tone 4, but only one was coherent in that it belonged to the tonal replica. This task of rejecting a second formant tone that had appropriate azimuth and speechlike time-varying frequencies—but which nevertheless was inappropriate for the rest of the tonal ensemble—and integrating the appropriate second formant tone presented in the other ear proved to be quite difficult for our listeners. However, when the dichotic competing tone lacked the spectro-temporal attributes of speech, exhibiting constant frequency, listeners easily rejected it, as if it were not competitive at all. Figure 2 shows performance in this test. The combined results suggest that listeners are vigilant in listening for plausible speechlike components, and are therefore misled by natural frequency variation and azimuth of the phonetically incompatible tone.

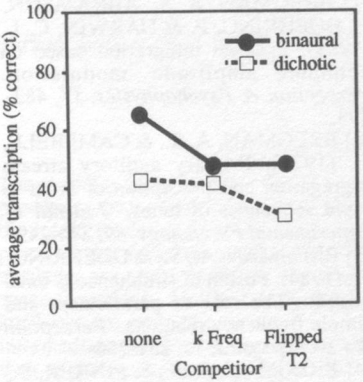

Figure 2. Group results for the test of competitive organization. Listeners heard either no distractor, or one of two possible distractors: a constant frequency tone, or a temporally reflected formant tone.

5. CONCLUSION

To summarize the outcome of our tests, the perceptual organization of speech does not rely necessarily on the acoustic properties featured in contemporary—or historical—accounts of grouping. Listeners were quite able to integrate a pattern of tones lacking similarity, continuity and common fate in the acoustic spectrum. No resort to auditory familiarity is available for explaining this finding, given the highly unnatural timbre of tonal analogs of speech. Moreover, the integrating mechanism seems keyed to speechlike signal variation in the simultaneous component tones, for listeners were less able to reject a speechlike tone that shared azimuth with the remainder of the tone ensemble in favor of a phonetically appropriate tone presented with inappropriate azimuth. Altogether, then, it seems that the complex spectrum of speech places it beyond the reach of the simple, elegant Gestalt rules. Our

search leads now toward the underlying physical and linguistic principles of perceptual organization of speech, and the perceiver's ability to exploit them in understanding an utterance.

6. REFERENCES

[1] BREGMAN, A. S., ABRAMSON, J., DOEHRING, P. & DARWIN, C. J., (1985), Spectral integration based on common amplitude modulation. *Perception & Psychophysics, 37*, 483-493.

[2] BREGMAN, A. S., & CAMPBELL, J., (1971), Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89*, 244-249.

[3] BREGMAN, A., S., & DOEHRING, P., (1984), Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics, 36*, 251-256.

[4] BREGMAN, A. S., & PINKER, S., (1978), Auditory streaming and the building of timbre. *Canadian Journal of Psychology, 32*, 19-31.

[5] CHERRY, E. C., (1953), Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America, 25*, 975-979.

[6] DANNENBRING, G. L., & BREGMAN, A. S., (1978), Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics, 24*, 369-376.

[7] FANT, C. G. M., (1962), Descriptive analysis of the acoustic aspects of speech. *Logos, 5*, 3-17.

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

[9] 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.

[10] MILLER, G. A. & LICKLIDER, J. C. R., (1950), The intelligibility of interrupted speech. *Journal of the Acoustical Society of America, 22*, 167-173.

[11] REMEZ, R. E., (1987), Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of Speech Perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

[12] REMEZ, R. E., & RUBIN, P. E., (1990), On the perception of speech from time-varying attributes: Contributions of amplitude variation and apparent naturalness. *Perception & Psychophysics, 48*, 313-325.

[13] REMEZ, R. E., RUBIN, P. E., NYGAARD, L. C., & HOWELL, W. A., (1987), Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 40-61.

[14] REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL T. D., (1981), Speech perception without traditional speech cues. *Science, 212*, 947-950.

[15] STEIGER, H., & BREGMAN, A. S., (1982), Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics, 32*, 153-162.

[16] WERTHEIMER, M., (1923), Untersuchungen zur Lehre von der Gestalt, II. *Psycholigische Forschung, 41*, 301-350.

[17] WILSON, J. P., & JOHNSTONE, J. R., (1975), Basilar membrane and middle-ear vibration in guinea pig measured by capacitative probe. *Journal of the Acoustical Society of America, 57*, 705-723.

The author gratefully acknowledges the assistance of S. M. Berns, J. M. Lang, J. S. Nutter and P. E. Rubin. This research was supported by a grant from the NIDCD (DC00308).