Yo2

The stream of speech

ROBERT E. REMEZ
PHILIP E. RUBIN

Barnard College, Columbia University, USA
Haskins Laboratories, USA

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

The use of sinusoidal replicas of speech signals reveals that listeners can perceive speech solely from temporally coherent spectral variation of nonspeech acoustic elements. This sensitivity to coherent change in acoustic stimulation is analogous to the sensitivity to change in configurations of visual stimuli, as detailed by Johansson. The similarities and potential differences between these two kinds of perceptual functions are described.

R. E. Remez, Department of Psychology, Barnard College, Columbia University, New York, New York 10027, USA.

Studies have shown that the continuously changing stream of speech can be approximated by a kind of acoustic animation, at the theoretical heart of which is an idealization of the human vocal tract as a *resonant horn* (e.g., Chiba & Kajiyama, 1941; Stevens & House, 1955; Fant, 1956, 1960). The details of the acoustics of speech can thereby be explained by noting that the vocal horn can be constricted at different places along its length, that it may be multiply excited, and, especially, that its shape can be changed rapidly. In practical situations, such as speech synthesis, the assumption of the hornlike properties is tacit, presupposed; the synthesizer speaks by the excitation of a lumped-circuit resonator (or its digital equivalent) which is itself approximate to horns of many types, including vocal tracts.

Although the term "speech stream" is often used to refer to the acoustic products of human vocalization, speech has commonly been studied by conceiving this metaphoric stream as an imbrication of more or less isolable elements, such as steady-state or transitional formant patterns, plosive bursts, band-limited noise, and stretches of silence. In our perceptual accounts, then, the exclusive attention to perceptual effects of specific elements in the acoustic pattern has led us to undervalue the coherence of the speech stream. In contrast to theoretical characterizations of the speech stream emphasizing structural continuity (Liberman, 1970), experimenters find it quite agreeable in practice to treat the perceptually relevant acoustic structure as if it consisted of distinct elements. Within this framework, properties of change in the speech signal figure primarily as a problem; from a dynamic array, the listener must somehow extract static elements or cues, perhaps even by means of a specialized decoding device. However, research continually reveals that perceivers care very little about the momentary attributes of speech signals. Even under direct test, and in highly favorable conditions, listeners seem unable to report acoustic properties of the stimulation on which their phonetic percepts are reliably based (Best, Morrongiello & Robson, 1981; Pisoni, 1971; Mattingly, Liberman, Syrdal & Halwes, 1971).

To describe the speech stimulus in a manner appropriate for perception, we must therefore characterize the time-varying spectrum of the acoustic pattern. In doing so, we elaborate the acoustic coherence of the speech stream, and avoid reducing it to a sequence of static acoustic moments irrelevant to the operating principles of the perceptual system. Only in this fashion may we gain a clue about the "smart" perceptual processes (Runeson, 1977) applied to speech stimuli. In addition, the success of the discrete-cue approximations of speech would then be explained by observing that a sequence of cues, however it reconstitutes properties of coherent change in the speech stream, is

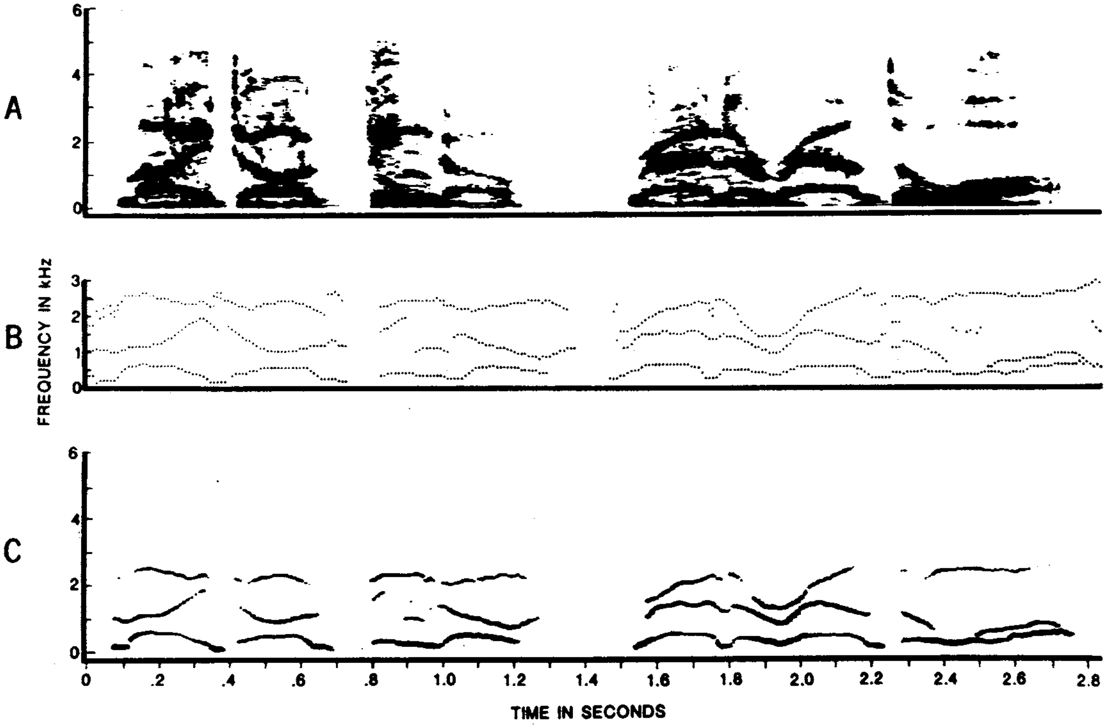

Fig. 1. (A) The English sentence, "My dog Bingo ran around the wall." The signal was produced by a natural talker and was quantified by a spectrum analyzer employing a bank of acoustic filters. (B) The same sentence is represented here as a sequence of center-frequency values of the first three formants. The method of Linear Prediction was used to prepare this series of spectral sections at 10 msec intervals. (C) The same sentence is realized here in a signal composed of three time-varying sinusoids. The pattern of relatively continuous change of the natural signal is retained, while the broadband formants and harmonic structure of the natural speech signal are eliminated.

perceptible because it conserves the important properties of acoustic *variation,* rather than because it conserves appropriate short-time spectra or discrete acoustic elements.

RECENT EVIDENCE

Although similar notions have been expressed from time to time in speech research (e.g., Liberman, 1970), several recent experiments especially promote these speculations about the importance of spectrum variation in speech perception (Remez, Rubin & Carrell, 1981; Remez, Rubin, Pisoni & Carrell, 1981). In these studies, listeners perceived phonetic segments from acoustic stimuli consisting solely of two or three sinusoids. Frequency and amplitude variations of the sinusoids imitated the changes of the vocal resonances found in natural speech signals (see Fig. 1). Specifically, each sinusoidal tone was matched in amplitude and frequency to one of the formants, or resonances, of a natural utterance that served as a model. Matching values were determined for successive 10 msec sections of the natural utterances, and these values were used to control a sinusoidal synthesizer.

None of the acoustic cues typically believed to underlie phonetic perception was present in the sinusoidal patterns: neither formant transitions nor steady state formants, because there were no broadband vocal resonances in this three-tone signal; nor fundamental frequency, because the three-tones were not harmonically related; nor band-limited noise, because the signal had only three periodically unrelated components. Thus, the short-time spectra of the signals did not satisfy the amplitude and frequency requirements of the spectral templates that are sometimes claimed to be useful in analyzing the acoustic pattern into phonetic units (Stevens & Blumsten, 1981). Despite this absence of vocally producible constituents, the overall pattern of frequency and amplitude variation imitated natural acoustic patterns. Listeners who found these sinewave replicas of speech to be intelligible evidently disregarded the inappropriate momentary acoustic structure, and were untroubled by the lack of traditional acoustic *cues*. Rather, they must have attended to the coherence of the time-variation of the tones, which betrays the vocal origin of the signal, and, at the same time, specifies an impossible sounding voice.

Unlike vocal resonances that share the same laryngeal source, each sinusoid is acoustically independent, and listeners readily reported this distinctness. Accordingly, it was rare that naive listeners spontaneously attended to phoretic information in this grossly unnatural phonetic carrier. Listeners, told nothing in advance about the three-tone signals, heard them simply as three simultaneous tones, modulated asynchronously as if the three-part counterpoint. However, the simple instruction to listen for a sentence enabled almost 70% of naive listeners to detect a sizeable chunk of the message, if not its entirety. In other words, listeners made use of phonetic information that was exclusively time-varying in nature, in the absence of short-time spectra characteristic of vocalization.

AN ANALOGY WITH VISUAL EVENT PERCEPTION

The analogy is readily apparent between our experiments on phonetic perception from sinusoids, and Johansson's (e.g., 1975) experiments on the perception of locomotory and other movements from point-light displays. In both cases, it appears that the pattern of coherent change in the stimulation conveys information about the event in progress. In Johansson's case, visual displays are made by videotaping a human figure moving in the dark, illuminated only at the joints of the articulating limbs. Although it is impossible to identify the content of the dot display from the single snapshots, the *moving* dots of light convey a wide range of subtle locomotory information. It is this organized change in the constellation of lights that carries information about the walking actor, despite the absence of static information to reveal which light belongs to which joint. Our sinewave element is like a speech formant in the same way that Johansson's light spot is like a radiocarpal joint—*the value of each element is estalished only by virtue of the coherent configuration to which it belongs*.

The analogy is not perfect, though. The distal object for Johansson's subject was a walker who seemed to mean nothing by his walking. In contrast, the distal object for our subjects was a message spoken by a strange talker. Our subjects perceived a highly structured message, and Johansson's did not. But, this may merely be a superficial methodological discrepancy if the visual observer can perceive whether the person in the display is performing a tango or a fox trot; or whether the person is using body "language" or American Sign Language, and what the message is (Poizner, Bellugi & Lutes-Driscoll, 1981). In each case, then, the perceiver identifies a person (one talking, one dancing), a structured transformation (one linguistic, one terpsichoric), and a strange medium (one sinusoidal, one dotty).

There is an additional discrepancy between Johansson's paradigm and ours. Subjects seem to find the information in the moving dot displays to be more accessible than the information in the sinusoidal displays. We do not understand this very well, but the fact that so many naive subjects hear sinusoids phonetically when instructed to do so may reduce the significance of this difference between the visual and linguistic cases.

A POTENTIAL FORMAL DIFFERENCE

In view of the similarity, it seems appropriate to distinguish formally some properties of visual motion perception and speech perception. Faced with the task of describing the geometry of optical flow, Johansson writes, "the self-motion component in the proximal flow (is combined with) tremendously complex object motion flow. The result is from a mathematical point of view like a chaos or is at least mathemtically complex to absurdity" (1981; page 6). Of course, the perceiver often disentangles the various components easily despite the limitations we otherwise experience as descriptive geometers. Acoustic change in the case of speech may not prove quite so elusive to describe, though. In principle, the physical limits of the variation of speech sounds are far narrower—constrained anatomically and lingistically—than are the physical limits of optical flow, which appear to be set, after all, by mechanics. And, we have a tremendous head start of forty years of "ecological acoustics" of the vocal tract. In any event, we suggest that studying the *coherence* of the speech stimulus—describing the *stream of speech*—requires a change of emphasis that brings research on phonetic perception closer to the approach established by Johansson for studying the visual perception of events.

This research was supported by grants HD 15672 (to R. E. R.) and HD 01944 (to Haskins Laboratories) from the National Institute of Child Health and Human Development. We gratefully acknowledge the advice of Alvin Liberman, Ignatius Mattingly, David B. Pisoni, Frank Restle, and Michael Studdert-Kennedy.

REFERENCES

Best, C. T., Morrongiello, B. & Robson, R. Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception & Psychophysics*, 1981, *29*, 191–211.

Chiba, T. & Kajiyama, M. *The vowel: its nature and structure*. Tokyo: Tokyo-Kaiseikan, 1941.

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

Fant, C. G. M. *The acoustic theory of speech production*. The Hague: Mouton, 1960.

Johansson, G. Visual motion perception. *Scientific American*, 1975, *232*(6), 76–89.

Johansson, G. About visual event perception: Perception of motion, dynamics and biological events. Paper presented at the First International Conference on Event Perception, Storrs, Connecticut, USA, 1981.

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

Mattingly, I. G., Liberman, A. M., Syrdal, A. K. & Halwes, T. G. Discrimination in speech and nonspeech modes. *Cognitive Psychology*, 1971, *2*, 131–157.

Pisoni, D. B. On the nature of categorical perception of speech sounds. *Supplement to Status Report on Speech Research*, SR–27, Haskins Laboratories, New Haven, 1971.

Poizner, H., Bellugi, U. & Lutes-Driscoll, V. Perception of American Sign Language in dynamic point-light displays. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, *7*, 430–440.

Remez, R. E., Rubin, P. E. & Carrell, T. D. Phonetic perception of sinusoidal signals: Effects of amplitude variation. *Journal of the Acoustical Society of America*, 1981, *69*, S114.

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

Runeson, S. On the possibility of "smart" perceptual mechanisms. *Scandinavian Journal of Psychology*, 1977, *18*, 172–179.

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

Stevens, K. N. & House, A. S. Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 1955, *27*, 484–493.