Speech Recognition with Primarily Temporal Cues

Robert V. Shannon,* Fan-Gang Zeng, Vivek Kamath, John Wygonski, Michael Ekelid

Nearly perfect speech recognition was observed under conditions of greatly reduced spectral information. Temporal envelopes of speech were extracted from broad frequency bands and were used to modulate noises of the same bandwidths. This manipulation preserved temporal envelope cues in each band but restricted the listener to severely degraded information on the distribution of spectral energy. The identification of consonants, vowels, and words in simple sentences improved markedly as the number of bands increased; high speech recognition performance was obtained with only three bands of modulated noise. Thus, the presentation of a dynamic temporal pattern in only a few broad spectral regions is sufficient for the recognition of speech.

The recognition of speech has been thought to require frequency-specific (spectral) cues. Spectral energy peaks in speech (formants), for example, reflect the resonant properties of the vocal tract and thus provide acoustic information on the production of the speech sound. However, efforts to identify acoustic cues that convey phoneme identity reliably under various listening conditions and with various talkers have met with only limited success (*1*). Studies that used amplitude compression (*2*) and spectral reduction (*3*) have demonstrated the robustness of speech recognition under these conditions. However, these manipulations resulted in stimuli that were still highly complex in their temporal-spectral characteristics. Even total removal of spectral cues from speech resulted in stimuli that carried a surprising amount of information on consonant identity (*4*). Work on prosthetic electrical stimulation of the auditory system by cochlear implants has refocused attention on amplitude and temporal cues, which are the principal cues transmitted by these prostheses (*5*). In our study, we preserved amplitude and temporal cues while systematically varying the amount of spectral information. This combination not only allowed us to parametrically assess the role of spectral detail in speech recognition independently of temporal cues, but also simulates the stimulation pattern of a cochlear implant (*6*).

Spectral information was removed from speech by replacement of the frequency-specific information in a broad frequency region with a band-limited noise (Fig. 1). The acoustic signal was divided into several frequency bands and the amplitude envelope was extracted from each band by half-wave rectification and low-pass filtering. Low-pass filters with cutoff frequencies of 16, 50, 160, and 500 Hz were used for envelope extraction to evaluate the effect of reducing the bandwidth of temporal envelope information. The envelope signal was used to modulate white noise, which was then spectrally limited by the same bandpass filter used for the original analysis band (*7*). Thus, temporal and amplitude cues were preserved in each spectral band, but the spectral detail within each band was removed. All bands were then summed and presented to the listeners through headphones. One, two, three, or four band processors were used, each with envelope information low-pass–filtered at 16, 50, 160, or 500 Hz, for a total of 16 conditions.

Eight normal hearing listeners (*8*) listened to 16 medial consonants (a/C/a), eight vowels (h/V/d), and simple sentences in each of the signal conditions (*9*). Consonants and vowels were presented in random order to each listener. The listeners were instructed to identify the presented stimulus by selecting it from the complete set of 16 consonants or 8 vowels. Sentences were presented once and the listeners were instructed to repeat as many words as they could. The listeners were trained on sample conditions to familiarize them with the testing environment; training continued until their performance stabilized, typically within two to three sessions, for a total of 8 to 10 hours. No feedback was provided in any of the test conditions.

Speech recognition performance on all three measures increased with the number of noise bands (Fig. 2). Changing the cutoff frequency of the envelope filter had a significant effect across all tests [$F(3,21) < 0.01$]. Paired t tests revealed no significant difference between the results with the 50-, 160-, and 500-Hz low-pass envelope filters, so these results were pooled for presentation in Fig. 2. A significant reduction in performance ($P < 0.01$) was observed with the 16-Hz envelope filter for consonants and sentences, but not for vowels. Thus, even under conditions of reduced spectral cues, slowly varying temporal information (<50 Hz) can yield relatively high speech recognition performance. This result is consistent with the observation of poor speech discrimination in children who have central processing disorders that disrupt temporal processing in the 20- to 50-ms range (*10*).

The specific reception of three speech features—voicing, manner, and place of articulation—was evaluated by information transmission analysis (*11*) on the consonant confusion matrix (Fig. 3). Information received on voicing and manner increased from one to two bands, to >90%, with no further improvement as the number of bands increased to three or four. Thus, binary information on the spectral distribution of energy, when combined with temporal cues, is sufficient to convey almost all information on voicing and manner. Voicing and manner have similar patterns of results as a function of the number of spectral bands, and both cues show maximum performance with only two spectral bands; these findings reinforce the hypothesis (*4*) that both categories of information, although labeled according to vocal produc-

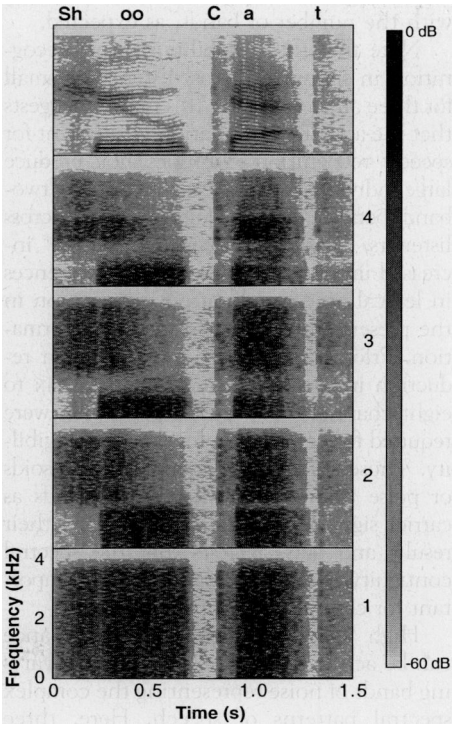

Fig. 1. Examples of spectral reduction for the speech token "shoo cat." The original narrowband spectrogram (top) shows energy as a function of frequency and time. Successively lower panels show spectrograms of the processed tokens with four, three, two, and one bands, respectively. Filter cutoff frequencies were 1500 Hz for the two-band processor, 800 and 1500 Hz for the three-band processor, and 800, 1500, and 2500 Hz for the four-band processor. All processors were low-pass–filtered at 4 kHz.

House Ear Institute, 2100 West Third Street, Los Angeles, CA 90057, USA.

*To whom correspondence should be addressed.

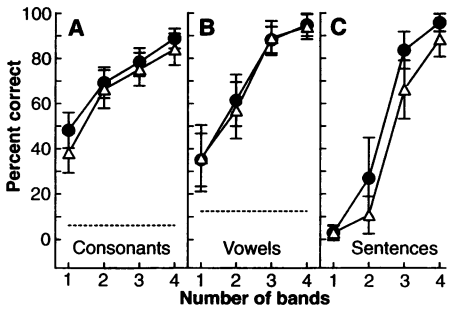

Fig. 2. Recognition scores for consonants (**A**), vowels (**B**), and sentences (**C**) for eight normal-hearing listeners are shown as a function of the number of noise bands. Chance performance scores in (A) and (B) are indicated by dashed lines. Results from envelope filters with frequencies of 50, 160, and 500 Hz were not significantly different and are pooled (●). Results from the 16-Hz envelope filter (△) were significantly lower than results from other envelope filter frequencies.

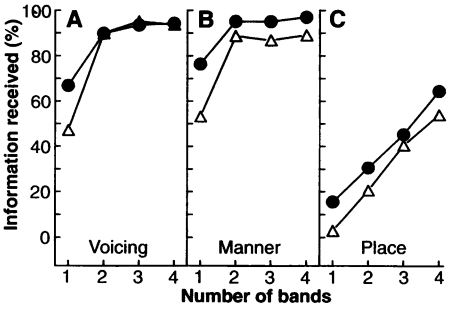

Fig. 3. Information transmission analysis for consonants. Consonant confusion matrices from eight normal-hearing listeners were summed and the aggregate matrix was analyzed in terms of the percentage of transmitted information received for medial consonant voicing (**A**), manner (**B**), and place of articulation (**C**) (*11*). Information received is shown as a function of the number of noise bands, with envelope filter frequency as a parameter (△, 16 Hz; ●, average of 50, 160, and 500 Hz).

tion, are perceptually primarily related to temporal envelope information. Recognition of consonantal place of articulation, which is primarily a spectral cue, increased with the number of bands, as expected.

Note that the variability of word recognition in sentences across listeners is small for three and four bands; this finding suggests that the acoustic information is sufficient for speech recognition but does not produce large individual differences. In contrast, two-band performance is quite variable across listeners, presumably because of the increased importance of individual differences in lexical access and pattern recognition in the presence of such poor spectral information. Prior experiments on information reduction in speech (*3*) indicated that six to eight channels of spectral information were required for comparable levels of intelligibility. Those studies used modulated sinusoids or pulse trains rather than noise bands as carrier signals; the difference between their results and ours suggests that the spectral contiguity of temporal information is important for central pattern recognition.

High speech recognition performance can be achieved with only three time-varying bands of noise representing the complex spectral patterns of speech. Here, three bands provided a severely degraded spectral representation of vowel and consonant formants and allowed only rudimentary spectral shape information to be transmitted. No formant structure was present, and formant frequency transitions either were lost completely (if they took place wholly within one of the present analysis bands) or were conveyed as a temporal change in the relative sound level of two adjacent noise bands. The harmonic structure of voiced speech was not present in the noise-band simulations. Despite this reduced spectral content, the temporal cues were sufficient to produce 90% correct identification of words. This result indicates that minimal spectral information is required for speech recognition as long as temporal cues are available in a few contiguous spectral regions.

Speech presents a difficult pattern recognition problem for the auditory system. The message content must be retrieved from speech in a wide variety of listening conditions, including different talkers, environments, and amounts of distortion. Our results suggest that speech pattern recognition is a robust process that can make use of both spectral and temporal cues. Because impaired or absent spectral resolution often is a consequence of hearing impairment, the finding that speech recognition can be achieved with primarily temporal cues suggests alternative signal-processing strategies for auditory prostheses.

REFERENCES AND NOTES

1. K. N. Stevens and S. E. Blumstein, in *Perspectives on the Study of Speech*, P. D. Eimas and J. L. Miller, Eds. (Erlbaum, Hillsdale, NJ, 1981), pp. 1–38.
2. J. C. R. Licklider and I. Pollack, *J. Acoust. Soc. Am.* **20**, 42 (1948); J. C. R. Licklider and G. A. Miller, in *Handbook of Experimental Psychology*, S. S. Stevens, Ed. (Wiley, New York, 1951), pp. 1040–1074.
3. R. E. Remez, P. E. Rubin, D. B. Pisoni, T. D. Carrell, *Science* **212**, 947 (1981); F. J. Hill, L. P. McRae, R. P. McClellan, *J. Acoust. Soc. Am.* **44**, 13 (1968); M. R. Schroeder, *Proc. IEEE* **54**, 720 (1966); J. Allen, *IEEE Trans. Speech Audio Proc.* **2**, 567 (1994).
4. S. Rosen, *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **336**, 367 (1992); D. J. Van Tasell, S. D. Soli, V. M. Kirby, G. P. Widin, *J. Acoust. Soc. Am.* **82**, 1152 (1987); D. J. Van Tasell, D. G. Greenfield, J. J. Logemann, D. A. Nelson, *ibid.* **92**, 1247 (1992); R. L. Freyman, G. P. Nerbonne, H. C. Cote, *J. Speech Hear. Res.* **34**, 415 (1991); S. M. Rosen, A. J. Fourcin, B. C. J. Moore, *Nature* **291**, 150 (1981).
5. M. Dorman, K. Dankowski, G. McCandless, L. Smith, *Ear Hear.* **10**, 288 (1989); M. F. Dorman, L. Smith, M. Smith, J. Parkin, *J. Acoust. Soc. Am.* **92**, 3428 (1992); R. V. Shannon, F.-G. Zeng, J. Wygonski, in *The Auditory Processing of Speech: From Sounds to Words*, M. E. H. Schouten, Ed. (Mouton-DeGruyter, New York, 1992), pp. 263–274; R. V. Shannon, *Hear. Res.* **11**, 157 (1983); *J. Acoust. Soc. Am.* **85**, 2587 (1989); *ibid.* **91**, 1974 (1992); D. K. Eddington, *ibid.* **68**, 885 (1980); B. S. Wilson *et al.*, *Nature* **352**, 236 (1991).
6. S. Rosen, J. Walliker, J. A. Brimacombe, B. J. Edgerton, *J. Speech Hear. Res.* **32**, 93 (1989); M. Dorman, M. Hannley, G. McCandless, L. Smith, *J. Acoust. Soc. Am.* **84**, 5011 (1988); M. F. Dorman *et al.*, *ibid.* **88**, 2074 (1990); B. S. Wilson, C. C. Finley, D. T. Lawson, in *Models of the Electrically Stimulated Cochlea*, J. M. Miller and F. A. Spelman, Eds. (Springer-Verlag, New York, 1990), pp. 339–375.
7. The signal was digitized at a 10-kHz sampling rate and passed through a preemphasis filter to whiten the spectrum (low-pass below 1200 Hz, −6 dB per octave). The signal was then split into frequency bands (third-order elliptical IIR filters). Adjacent filters overlapped at the point at which the output from each filter was 15 dB down from the level in the pass-band. The envelope was extracted by half-wave rectification and low-pass filtering (elliptical IIR filters with cutoff frequencies of 16, 50, 160, or 500 Hz, −6 dB per octave). The envelope derived from each band was then used to modulate a white noise. The modulated noise was frequency-limited by filtering with the same bandpass filters used in the original analysis band. This last bandpass filtering reduced the modulation depth to some degree because it removed the modulation sidebands. The resulting modulated noises from each band were combined, low-pass-filtered at 4 kHz, amplified (Crown D75), and presented to the listener through headphones (TDH-49). Overall levels were calibrated from each combination of parameters to produce an average output level of 75 dBA for continuous speech.
8. All listeners participated with full informed consent. The Institutional Review Board of the House Ear Institute and St. Vincent's Medical Center approved the study protocol and the informed consent form.
9. Consonant and vowel stimuli were taken from the sound track of the Iowa audiovisual speech perception laser videodisc (R. S. Tyler, J. P. Preece, M. W. Lowder, Department of Otolaryngology, University of Iowa, 1989). The male talker was used for both vowels and consonants. Three exemplars of each token were selected randomly. Consonant confusion matrices were compiled from 10 presentations of each of the 16 medial consonants (a/C/a, for example "aba") for each listener. Vowel confusion matrices were compiled from nine presentations of each of the eight vowels in a h/V/d (for example "hood") context. Sentences were taken from the sound track of the City University of New York laser videodisc of everyday sentences (A. Boothroyd, L. Hanin, T. Hnath, Speech and Hearing Sciences Research Center, City University of New York, New York, 1985). Data were collected for 24 sentences, representing 100 key words, from each listener. The sentences were of easy to moderate difficulty and no sentences were repeated to an individual listener.
10. P. Tallal *et al.*, Eds., *Temporal Information Processing in the Nervous System (Ann. N.Y. Acad. Sci.* **682**, 1993).
11. G. A. Miller and P. E. Nicely, *J. Acoust. Soc. Am.* **27**, 338 (1955).

12. We thank S. Rosen, D. Van Tasell, R. Diehl, S. Nittrour, B. Wilson, S. Soli, and M. Dorman for their help and comments. Supported by National Institute on Deafness and Other Communication Disorders (NIDCD) grant RO1 DCO1526.

21 April 1995; accepted 23 August 1995