
A Comparison of the Speech Understanding Provided by Acoustic Models of Fixed-Channel and Channel-Picking Signal Processors for Cochlear Implants

Michael F. Dorman

Arizona State University
Tempe
and
University of Utah
Health Sciences Center
Salt Lake City

Philipos C. Loizou

University of Texas at Dallas

Anthony J. Spahr

Erin Maloff

Arizona State University
Tempe

Vowels, consonants, and sentences were processed by two cochlear-implant signal-processing strategies—a fixed-channel strategy and a channel-picking strategy—and the resulting signals were presented to listeners with normal hearing for identification. At issue was the number of channels of stimulation needed in each strategy to achieve an equivalent level of speech recognition in quiet and in noise. In quiet, 8 fixed channels allowed a performance maximum for the most difficult stimulus material. A similar level of performance was reached with a 6-of-20 channel-picking strategy. In noise, 10 fixed channels allowed a performance maximum for the most difficult stimulus material. A similar level of performance was reached with a 9-of-20 strategy. Both strategies are capable of providing a very high level of speech recognition. Choosing between the two strategies may, ultimately, depend on issues that are independent of speech recognition—such as ease of device programming.

KEY WORDS: cochlear implants, acoustic models, signal processing, adults

Modern cochlear implants provide two strategies for coding speech (e.g., McDermott, McKay, & Vandali, 1992; Wilson et al., 1991; for reviews, see Loizou, 1998, and Wilson, 2000). Fixed-channel strategies use a relatively small number of input channels (6–12), and the number of output channels equals the number of input channels. Channel-picking strategies commonly use a relatively larger number of input channels (e.g., 20) and a smaller number of output channels (6–9). At issue in the present study is how many channels need to be implemented in fixed-channel strategies to provide the same level of speech understanding, in quiet and in noise, as provided by channel-picking strategies. We used acoustic stimulation and normal-hearing listeners for our experiments because in studies with cochlear implant patients, the number of functional channels of stimulation is, in most instances, far fewer than the number of electrodes. Patients commonly perform as well with 4–8 electrodes activated as with 20 electrodes activated (Fishman, Shannon, & Slattery, 1997; Wilson, 1997). Thus, the speech perception results from cochlear implant patients do not reflect the amount of information represented in a strategy's output signals.

The “fixed-channel” strategy has antecedents in the vocoders of the 1940s and 1950s. Vocoders divide speech into a number of bands, lowpass filter the bands, and then output the signals using buzz/hiss excitation. Early experiments suggested that 7–10 bands were sufficient to transmit speech of high quality (Halsey & Swaffield, 1948; Hill, McRae, & McClellan, 1968). More recent experiments using either sine wave outputs or noise band outputs (and thus not using the normal mix of buzz and hiss excitation) indicate that, in quiet, the number of channels needed for a high level of speech intelligibility varies between 4 and 10 channels, depending on the nature of the stimulus material and the age of the listener. For adults, 4 channels are sufficient for 90% recognition of “easy” sentences (Dorman, Loizou, & Rainey, 1997; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995), 8 channels are needed for “difficult” sentences (Loizou, Dorman, & Tu, 1999), and 10 channels are needed for recognition of single words from high-density lexical neighborhoods (Dorman, Loizou, Kemp, & Kirk, 2000). In noise, as the signal-to-noise ratio (SNR) becomes poorer, the number of channels needed for maximum performance increases (Dorman, Loizou, Fitzke, & Tu, 1998; Fu, Shannon, & Wang, 1998). Children generally need more channels than adults to reach the same level of performance (Dorman, Loizou, Kemp, et al., 2000; Eisenberg, Shannon, Martinez, Wygonski, & Boothroyd, 2000).

The “channel-picking” strategy implemented as the SPEAK strategy on Cochlear Corporation devices (McDermott et al., 1992) and as the “n-of-m” strategy on other devices (Wilson, 2000; Wilson et al., 1988), has antecedents in the channel-picking vocoders of the 1950s (Peterson & Cooper, 1957) and Haskins Laboratories’ Pattern Playback speech synthesizer (Cooper, Liberman, & Borst, 1950). The principle underlying the use of this strategy is that speech can be well understood when only the peaks in the short-term spectrum are transmitted. In the case of the Pattern Playback, only 4–6 of 50 sine-wave harmonics needed to be transmitted to achieve highly intelligible speech—as long as the “picked” harmonics defined the first two or three formants in the speech signal. One issue in the present experiment is how many channels need to be picked out of 20 channels (the potential number of electrodes/channels in a commonly used implant) to achieve a high level of speech understanding in quiet and in noise. Previous studies using acoustic simulations have shown performance maxima for sentences presented in noise when picking 4 out of 6 channels, when picking 6 out of 12 channels, and when picking 12 out of 16 channels (Loizou, Dorman, Tu, & Fitzke, 2000).

In the experiments described here, vowels, consonants, and sentences were processed by channel-picking strategies and by fixed-channel strategies and were

presented in quiet and in noise to normal-hearing listeners for identification. As noted above, we and others have previously conducted experiments of this type. This experiment differed from previous experiments in that (a) the same participants were used to test the intelligibility of speech processed by the two strategies, (b) the same test materials were used to test the intelligibility of speech processed by the two strategies, and (c) the signal-to-noise ratio was varied for the different test materials so that the level of performance in noise was similar for all test materials.

Method

Participants

The participants were 10 undergraduate and graduate students at Arizona State University. All reported that they had normal hearing and passed a pure-tone screening at 25 dB HL at frequencies of 1, 2, and 4 kHz.

Stimuli

The stimuli for the vowel tests were tokens of “heed, hid, head, had, hod, heard, who’d, hud, hay’ed, hoed, hood,” taken from Hillenbrand, Getty, Clark, and Wheeler (1994). Each word was spoken by three men and three women. The stimuli were randomized into a list for each test condition.

The stimuli for the tests of consonant identification were 16 male-voice consonants in the /aCa/ context taken from the Iowa laser video disk (Tyler, Preece, & Tye-Murray, 1986). There were three repetitions of each stimulus. The stimuli were randomized into a list for each test condition.

The stimuli for the tests of sentence intelligibility were taken from the Hearing in Noise Test (HINT) lists of Nilsson, Soli, and Sullivan (1994). Ten sentences were presented in each test condition and were scored for number of words correct. Different sentences were used in each test condition. In addition, participants were screened to ensure that they had not heard material from the HINT lists previously.

Two conditions were generated for each set of test material. In one condition, the stimuli were presented in quiet. In the other condition, the stimuli were presented in noise sufficient to allow a performance asymptote between 70 and 80% correct. Pilot experiments (using 20-channel processors and SNRs of –4, –2, 0, and 2 dB for vowels; 0, 2, 4, and 6 dB for consonants; and –2, 0, 2, and 4 dB for sentences) indicated that this level was –2 dB for vowels, +4 dB for consonants, and 0 dB for sentences. A different masking noise was created for each type of test material. The noises were created by filtering white noise through a 60th-order finite

impulse response (FIR) filter designed to match the averaged spectrum of the stimulus material. The noise level was scaled so that the difference between signal power and noise power was -2 dB for vowels, $+4$ dB for consonants, and 0 dB for sentences.

Signal Processing

Using MATLAB-based software, the signals were processed into seven fixed-channel conditions (4, 6, 8, 10, 12, 16, and 20 channels) and three channel-picking conditions (3-of-20, 6-of-20, and 9-of-20 channels). Twenty input channels were used in the channel-picking conditions because that is the number of channels nominally implemented in the most widely used implant application. For the fixed-channel conditions, the signals (with bandwidth 0.25–6 kHz) were first processed through a preemphasis filter (high-pass filter with a 1200-Hz cutoff frequency and 6 dB/octave rolloff) and then bandpassed into N frequency bands (where N varied from 4 to 20) using sixth-order Butterworth filters. The spacing of the frequency bands was logarithmic for processors with 6 or fewer channels. For processors with 8 or more channels, semilogarithmic or mel spacing of channels was used. For these processors, the filter bandwidths were computed according to the equation $1100 \log (f/800 + 1)$, where f = frequency in hertz. Mel spacing of low-frequency channels was used for processors with 8 or more channels because logarithmic spacing resulted in channels with very small bandwidths and mel spacing resulted in bandwidths that better approximated critical bandwidths. The envelope of the signal was extracted by full-wave rectification and low-pass filtering (second-order Butterworth) with a 400-Hz cutoff frequency. Sinusoids were generated with amplitudes equal to the root-mean-square (rms) energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids were summed and presented to the listeners at a comfortable level.

To create stimuli for the channel-picking conditions, signals were processed in the manner of a channel-picking speech processor—one similar to the SPEAK strategy used in the Nucleus cochlear implant (McDermott et al., 1992). Signals were processed into 20 channels using linear spacing to 1 kHz and logarithmic spacing thereafter, and every 4 ms the 3, 6, or 9 maximum channel amplitudes were identified and output as sine waves at the center frequency of the analysis band.

Procedure

All participants were tested first with signals in quiet and then in noise. The order of vowel, consonant,

and sentence tests was quasi-randomized across participants. For all sets of stimulus material, the fixed-channel conditions were presented in descending order of number of channels, that is, 20 channels first and 4 channels last. This order was chosen to maximize the listeners' familiarity with the stimulus material. The 9-of-20 conditions were inserted into the test sequences between the 16 and 20 fixed-channel conditions, the 6-of-20 conditions were inserted between the 10 and 12 fixed-channel conditions, and the 3-of-20 conditions were inserted between the 4 and 6 fixed-channel conditions.

Results

The results are shown in Figures 1, 2, and 3 for vowels, consonants, and sentences, respectively. The left panel of each figure shows performance as a function of number of fixed channels for signals in quiet and in noise. The right panel of each figure shows performance as a function of the number of channels picked for output from 20 analysis channels. In these figures, the data points for the 20-of-20 conditions are taken from the 20 fixed-channel conditions. When $n = m$ in a channel-picking algorithm, there is no difference between the channel-picking algorithm and a fixed-channel algorithm with m channels.

For the majority of conditions, percent-correct scores increased with an increase in the number of channels of stimulation. The number of channels necessary to reach a performance maximum varied as a function of the type of stimulus material and the SNR. Repeated-measures analyses of variance indicated that in quiet the number of channels had a significant main effect for vowels [$F(9, 81) = 97.9, p < .000001$], consonants [$F(9, 81) = 23.29, p < .000001$], and sentences [$F(9, 81) = 15.7, p < .000001$]. According to post hoc tests (Tukey-Kramer with alpha at .01), a performance maximum for vowels was reached with 8 fixed channels and with a 3-of-20 processor. A performance maximum for consonants was reached with 6 fixed channels and with a 3-of-20 processor. A performance maximum for sentences was reached with 6 channels and with a 6-of-20 processor.

In noise, the number of channels had a significant main effect for vowels [$F(9, 81) = 97.8, p < .000001$], consonants [$F(9, 81) = 24.2, p < .000001$], and sentences [$F(9, 81) = xxx, p < .000001$]. A performance maximum for vowels was reached with 10 fixed channels and with a 6-of-20 processor. A performance maximum for consonants was reached with 6 fixed channels and with a 9-of-20 processor. A performance maximum for sentences was reached with 10 fixed channels and with a 9-of-20 processor.

Figure 1. Recognition of multitalker vowels. Left panel: Recognition as a function of the number of channels of stimulation. Right panel: Recognition as a function of the number of output (or “n”) channels out of 20 input channels. Performance in quiet is shown by the filled squares. Performance in noise at -2 dB SNR is shown by the open circles. Error bars indicate ± 1 standard deviation.

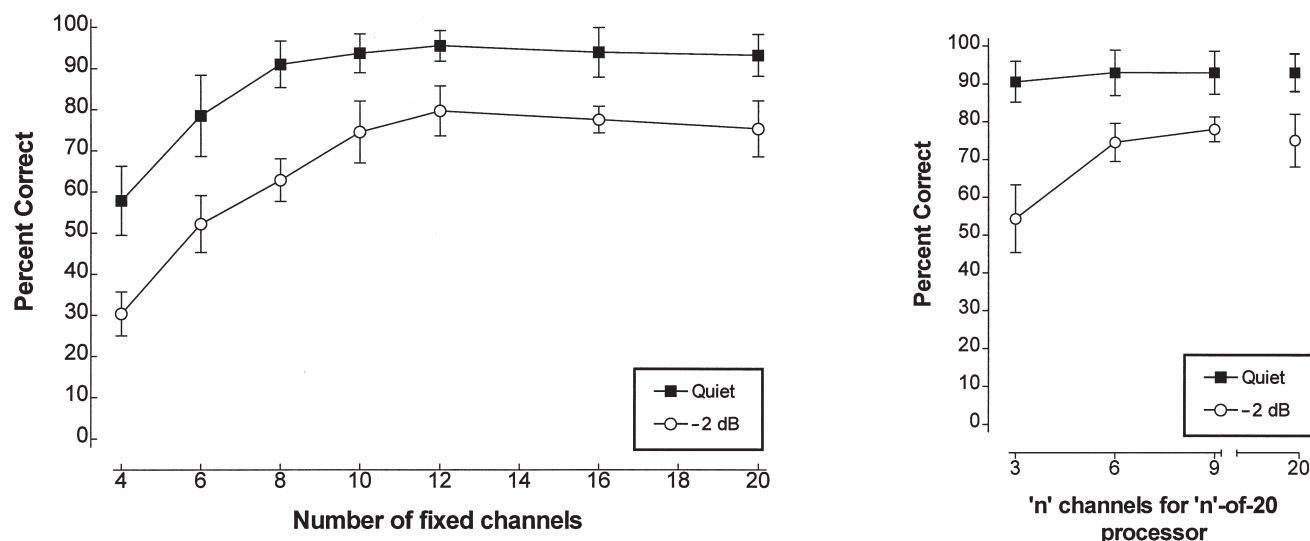
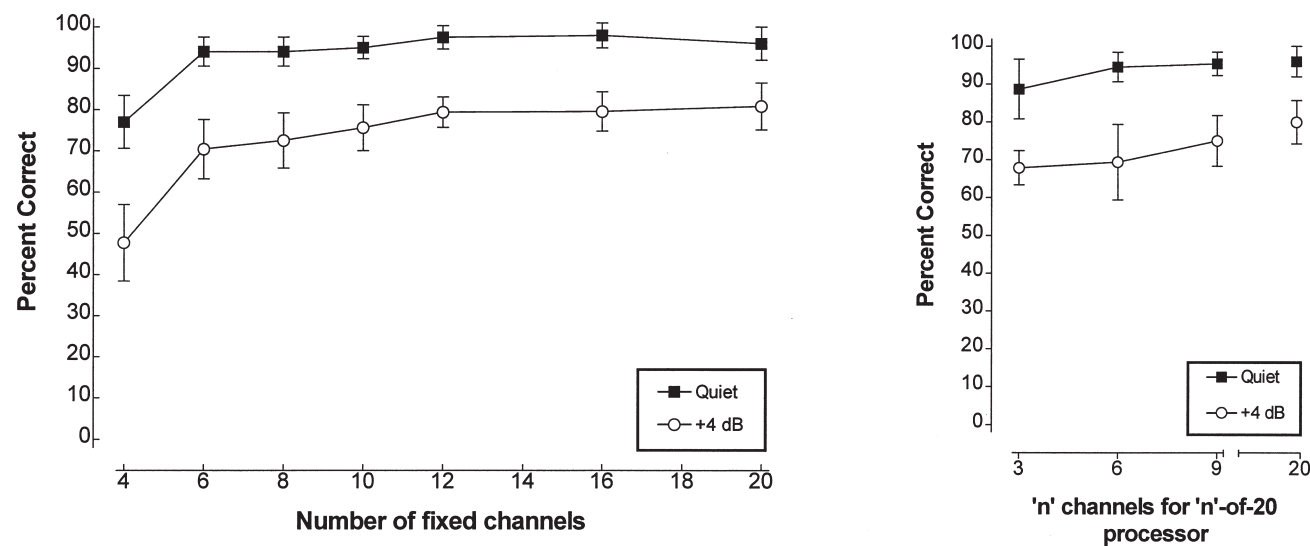


Figure 2. Recognition of consonants. Left panel: Recognition as a function of the number of channels of stimulation. Right panel: Recognition as a function of the number of output (or “n”) channels out of 20 input channels. Performance in quiet is shown by the filled squares. Performance in noise at 4 dB SNR is shown by the open circles. Error bars indicate ± 1 standard deviation.

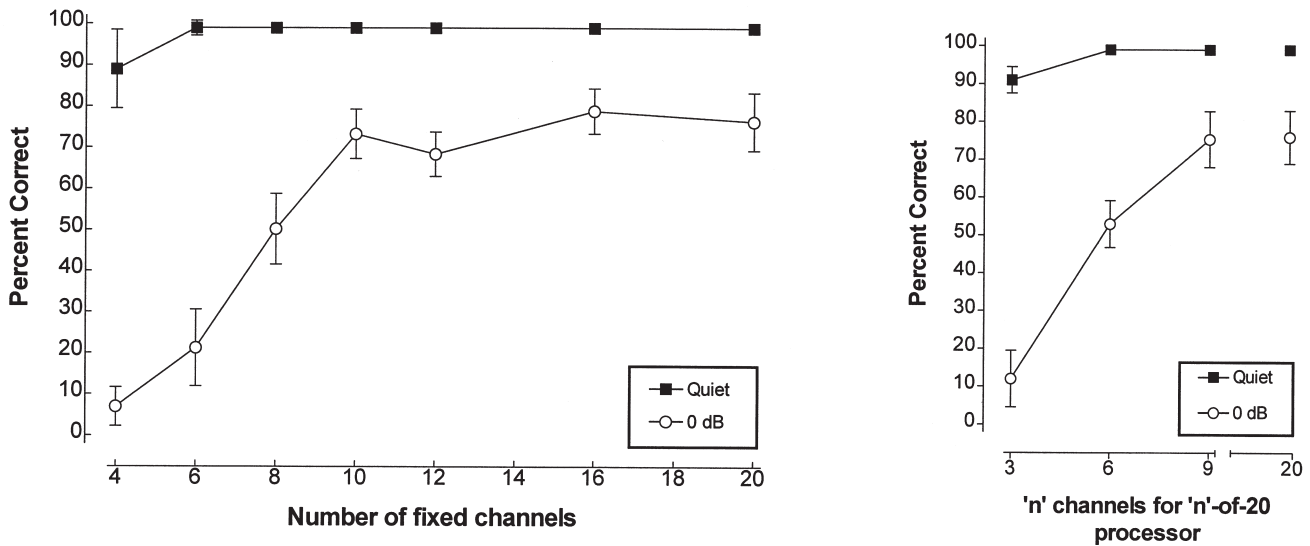


Discussion

The results of our experiments indicate that both fixed-channel strategies and channel-picking strategies can provide a very high level of speech understanding. For speech material presented in quiet, only a few channels—three in our experiments—need to be picked from a set of 20 channels to provide approximately 90% recognition of vowels, consonants, and words in sentences. A performance maximum was reached for all material

with no more than a 6-of-20 processor. This outcome suggests that in a quiet environment, channel-picking strategies with 20 input channels could be programmed with 6 output channels. However, when operating in a noisy environment, channel-picking strategies should be programmed with more output channels. This conclusion follows from the increase in speech recognition for consonants and sentences presented against a noisy background when the number of output channels was increased from 6 to 9. This outcome suggests that for adults,

Figure 3. Recognition of words in sentences. Left panel: Recognition as a function of the number of channels of stimulation. Right panel: Recognition as a function of the number of output (or “n”) channels out of 20 input channels. Performance in quiet is shown by the filled squares. Performance in noise at 0 dB SNR is shown by the open circles. Error bars indicate ± 1 standard deviation.



channel-picking strategies with 20 input channels should be programmed with at least 9 output channels to maximize patient performance in quiet and in noise.

The 90% correct level of speech understanding for stimulus material presented in quiet, met by a 3-of-20 processor for all stimulus material, was met by fixed-channel processors with 4, 6, and 8 channels for sentences, consonants, and vowels, respectively. In noise, 10 channels allowed a performance maximum for all materials. A similar outcome for cochlear implant patients has been reported by Zeng and Galvin (1999). These data suggest that for adults, fixed-channel processors should be programmed with a minimum of 10 input and output channels—a suggestion strikingly similar to that derived from work conducted nearly 50 years ago using channel vocoders.

Our conclusion about the number of channels needed to reach a performance maximum for both fixed-channel and channel-picking strategies is, of course, constrained by the signal-to-noise levels used in our experiments. If we had used a poorer signal-to-noise ratio, then the number of channels needed in each strategy would have changed. We chose the 0-dB signal-to-noise ratio for sentences because that noise level lowered performance from the “ceiling” but was not so low as to leave participants guessing what they were hearing.

In the foregoing discussion, we have emphasized the similarity in outcomes for channel-picking and fixed-channel processors when the number of output channels is relatively large. However, for a small number of

output channels, the outcomes differ significantly. For example, a 3-of-20 processor allows higher consonant and vowel scores than a processor with 4 fixed channels. Similarly, a 6-of-20 processor allows higher scores for vowels and for sentences in noise than a processor with 6 fixed channels. These outcomes illustrate the benefits of having a large number of analysis channels. If the number of functionally independent channels in cochlear implants were equal to the number of analysis channels, then implant patients would enjoy very high levels of speech understanding in quiet and in noise. Unfortunately, current cochlear implants only provide a relatively small number of independent channels (Fishman et al., 1997; Wilson, 1997; Zeng & Galvin, 1999). Increasing the number of functionally independent channels in cochlear implants should be a high priority for researchers.

The data discussed above indicate that, for adults, equivalent levels of speech understanding are provided by a 9-of-20 channel-picking strategy and a fixed-channel strategy with 10 channels. Our results do not necessarily predict results for children (see Dorman, Loizou, Kemp, et al., 2000, and Eisenberg et al., 2000) or for congenitally deaf children in contrast to late-deafened adults. For adults, there is no inherent advantage of one strategy over the other in terms of speech understanding—at issue is only the number of channels that are implemented for a given strategy. In this view, a choice between the two strategies may revolve around issues such as the ease of device programming for adults, children, and especially infants.

Acknowledgment

This work was supported by grants from the NIDCD to the first author (RO1 DC00654-9) and the second author (RO1 DC 03421-2). The data on sentence recognition described in this article were presented at the Ninth IEEE DSP workshop (Dorman, Loizou, Spahr, & Maloff, 2000).

References

- Cooper, F., Liberman, A., & Borst, J.** (1950). Preliminary studies of speech produced by a pattern playback. *Journal of the Acoustical Society of America*, 22, 678.
- Dorman, M.** (2000). Speech perception by adults. In S. Waltzman & N. Cohen (Eds.), *Cochlear implants*. New York: Thieme.
- Dorman, M., Loizou, P., Fitzke, J., & Tu, Z.** (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *Journal of the Acoustical Society of America*, 104, 3583–3585.
- Dorman, M., Loizou, P., Kemp, L., & Kirk, K.** (2000). Word recognition by children listening to speech processed into a small number of channels: Data from normal-hearing children and children with cochlear implants. *Ear and Hearing*, 21, 590–596.
- Dorman, M., Loizou, P., & Rainey, D.** (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, 102, 2403–2411.
- Dorman, M., Loizou, P., Spahr, T., & Maloff, E.** (2000). *Performance of spectral-maxima and fixed-channel algorithms for coding speech in quiet and in noise*. Ninth DSP Workshop, Hunt, TX, October 2000.
- Eisenberg, L., Shannon, R., Martinez, A., Wygonski, J., & Boothroyd, A.** (2000). Speech recognition with reduced spectral cues as a function of age. *Journal of the Acoustical Society of America*, 107, 2704–2710.
- Fishman, K., Shannon, R., & Slattery, W.** (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *Journal of Speech, Language, and Hearing Research*, 40, 1201–1215.
- Fu, Q.-J., Shannon, R., & Wang, X.** (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *Journal of the Acoustical Society of America*, 104, 3586–3596.
- Halsey, R., & Swafeld, J.** (1948). Analysis-synthesis telephony, with special reference to the vocoder. *Institute of Electrical Engineers (London)*, 95 Pt. III, 391–411.
- Hill, J., McRae, P., & McClellan, R.** (1968). Speech recognition as a function of channel capacity in a discrete set of channels. *Journal of the Acoustical Society of America*, 44, 13–18.
- Hillenbrand, J., Getty, L., Clark, M., & Wheeler, K.** (1994). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97, 3099–3111.
- Loizou, P.** (1998). Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants. *IEEE Signal Processing Magazine*, 15(5), 101–130.
- Loizou, P., Dorman, M., & Tu, Z.** (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, 106, 2097–2103.
- Loizou, P., Dorman, M., Tu, Z., & Fitzke, J.** (2000). The recognition of sentences in noise by normal-hearing listeners using simulations of SPEAK-type cochlear implant processors. *Annals of Otolaryngology and Laryngology*, 109(12, Suppl. 185), 67–68.
- McDermott, H., McKay, C., & Vandali, A.** (1992). A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant. *Journal of the Acoustical Society of America*, 91, 3367–3371.
- Nilsson, M., Soli, S., & Sullivan, J.** (1994). Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and Noise. *Journal of the Acoustical Society of America*, 95, 1085–1099.
- Peterson, G., & Cooper, F.** (1957). Peakpicker: A band-width compression device. *Journal of the Acoustical Society of America*, 29, 777(A).
- Shannon, R., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M.** (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303–304.
- Tyler, R., Preece, L., & Tye-Murray, N.** (1986). *The Iowa Phoneme and Sentence Tests*. Iowa City: The University of Iowa.
- Wilson, B.** (1997). The future of cochlear implants. *British Journal of Audiology*, 31, 205–225.
- Wilson, B.** (2000). Cochlear implant technology. In J. K. Niparko (Ed.), *Cochlear implants: Principles and practices* (pp. 109–118). Philadelphia: Lippincott, Williams & Wilkins.
- Wilson, B., Finley, C., Farmer, J., Lawson, D., Weber, B., Wolford, R., et al.** (1988). Comparative studies of speech processing strategies for cochlear implants. *Laryngoscope*, 98, 1069–1077.
- Wilson, B., Finley, C., Lawson, D., Wolford, R., Eddington, D., & Rabinowitz, W.** (1991). Better speech understanding with cochlear implants. *Nature*, 352, 236–238.
- Zeng, F.-G., & Galvin, J.** (1999). Amplitude mapping and phoneme recognition in cochlear implant listeners. *Ear and Hearing*, 20, 60–74.

Received July 25, 2001

Accepted February 8, 2002

DOI: 10.1044/1092-4388(2002)XXX

Contact author: M. F. Dorman, PhD, Department of Speech and Hearing Science, Arizona State University, Tempe, AZ 85287-0102. E-mail: mdorman@asu.edu