Speech recognition with altered spectral distribution of envelope cues

Robert V. Shannon,[a] Fan-Gang Zeng, and John Wygonski
House Ear Institute, 2100 West Third Street, Los Angeles, California 90057

(Received 12 September 1997; revised 1 May 1998; accepted 30 June 1998)

Recognition of consonants, vowels, and sentences was measured in conditions of reduced spectral resolution and distorted spectral distribution of temporal envelope cues. Speech materials were processed through four bandpass filters (analysis bands), half-wave rectified, and low-pass filtered to extract the temporal envelope from each band. The envelope from each speech band modulated a band-limited noise (carrier bands). Analysis and carrier bands were manipulated independently to alter the spectral distribution of envelope cues. Experiment I demonstrated that the location of the cutoff frequencies defining the bands was not a critical parameter for speech recognition, as long as the analysis and carrier bands were matched in frequency extent. Experiment II demonstrated a dramatic decrease in performance when the analysis and carrier bands did not match in frequency extent, which resulted in a warping of the spectral distribution of envelope cues. Experiment III demonstrated a large decrease in performance when the carrier bands were shifted in frequency, mimicking the basal position of electrodes in a cochlear implant. And experiment IV showed a relatively minor effect of the overlap in the noise carrier bands, simulating the overlap in neural populations responding to adjacent electrodes in a cochlear implant. Overall, these results show that, for four bands, the frequency alignment of the analysis bands and carrier bands is critical for good performance, while the exact frequency divisions and overlap in carrier bands are not as critical.
 [S0001-4966(98)02210-3]

PACS numbers: 43.71.Ky, 43.71.Es, 43.66.Ts [WS]

INTRODUCTION

Speech pattern recognition has been evaluated under a wide variety of conditions that alter and reduce spectral and temporal information. Speech has proven to be a robust signal that is resistant to many forms of distortion and information reduction. For example, speech can be recognized even when the spectral information is reduced to three sinusoids that track the formant transitions over time (Remez *et al.*, 1981). Even removing all spectral cues in speech results in a surprisingly high level of speech phoneme discrimination and recognition (Schroeder, 1968; van Tasell *et al.*, 1987, 1992; Rosen, 1992; Drullman, 1995; Turner *et al.*, 1995; Shannon *et al.*, 1995), and provides significant assistance for lip reading (Erber, 1972; Grant *et al.*, 1985, 1991).

Cochlear implants present an interesting case for understanding speech pattern recognition. In cochlear implants a relatively small number of electrodes activate tonotopic patches of neurons with a portion of the speech signal. However, the physiological response to electrical stimulation is quite different from the normal acoustic response. Among other differences, all neurons activated by an electrode are driven in a highly deterministic fashion (van den Honert and Stypulkowski, 1984, 1987; Hartmann *et al.*, 1984). Even with these abnormal properties in the physiological response, patients with as few as four electrodes have demonstrated high levels of speech recognition—higher levels than most researchers have predicted might be possible with such severe quantization of spectral information. This observation indicated how little was understood about recognition of speech under conditions of minimal or distorted spectral information.

In the design of prosthetic devices we face a dilemma which has both theoretical and practical implications: how do we maximize the speech information transmitted, given the limitations of the prosthetic device interface to the nervous system? Hill *et al.* (1968) partitioned the problem into two parts: analysis issues and presentation issues. The analysis problem is concerned with how to parcel speech information in a way that preserves the maximum amount of information, given the limitation of a small number of channels. The presentation problem is then to define the optimal mapping of the channels of speech information into perceptual/neural channels via the prosthesis.

A. The analysis problem

The analysis problem depends primarily on the distribution of critical speech pattern information in frequency and time. What temporal and spectral patterns of information are most critical for speech recognition? If we are limited in either the temporal or spectral domains, what are the most important speech cues remaining within those limitations? For example, if we could determine that a listener had only two potential receiving channels for speech information, what is the best way to divide speech information into two channels to preserve the maximum number of distinctions across different talkers and listening conditions? One indication comes from the original research on the articulation index (AI) which attempted to define an importance function for each spectral segment of speech (Fletcher and Steinberg,

[a]Electronic mail: Shannon@hei.org

1929; French and Steinberg, 1947). Fletcher and colleagues used high-pass and low-pass filtered speech to find that 1500 Hz was the frequency around which low-frequency and high-frequency contributions were equal for speech recognition. Articulation theory further defined the importance of each spectral region's contributions to the recognition of speech.

Using a different approach, Shannon *et al.* (1995) systematically reduced spectral information in speech to one, two, three, or four bands of modulated noise. The surprising result was that speech was highly recognizable with only three or four bands of noise, each modulated by the envelope from that same spectral band in speech. However, even with this result it is not clear that the four frequency bands chosen by Shannon *et al.* were the optimum for speech information content. One of the purposes of the present study is to study the importance of spectral parameters for speech pattern recognition.

The amount of temporal information required for speech recognition has recently been evaluated in conditions of reduced spectral information (Van Tasell *et al.*, 1987; Turner *et al.*, 1995; Shannon *et al.*, 1995; Dorman *et al.*, 1997a). In these studies envelope information was low-pass filtered at successively lower frequencies and used to modulate bands of noise or a sinusoid as a carrier. No change was observed in vowel, consonant, or sentence recognition as long as the low-pass envelope filter cutoff was 50 Hz or higher. Removing envelope fluctuations above 50 Hz had no effect on recognition, while removing envelope fluctuations between 20 and 50 Hz resulted in reduced phoneme discriminability and reduced speech recognition. Thus speech recognition is possible with only four bands of noise, even when each band is modulated with low-frequency envelope information below 50 Hz.

B. The presentation problem

In sensory prosthetic devices the goal is to convey information through an impaired sensory modality or through an alternate sensory modality. Thus we must understand the limitations and capabilities of the impaired or substitute sense to fully utilize its information carrying capacity. In the case of cochlear implants, acoustic information is being presented to the auditory system, but electrical stimulation activates the auditory system in a manner that produces highly unnatural patterns of neural activity. Some of the differences between acoustic and electrical stimulation are critical limitations for conveying speech information, while other differences may be of only secondary or no importance. To achieve the best match between the speech features extracted from the acoustic signal and the information transmitted to the electrically stimulated nerve, we must understand which patterns of neural activity are critical for speech recognition and which are secondary.

Psychophysical experiments with cochlear implants have demonstrated that temporal processing in implant patients is relatively normal (Shannon, 1983, 1986, 1989, 1990, 1992). Intensity processing is abnormal in implant listeners, but is still capable of restoring normal loudness growth with the proper loudness mapping function in the speech processor. However, implant listeners only have 20–40 discriminable intensity levels (Zeng and Shannon, 1992, 1994; Nelson *et al.*, 1997), which may be adequate for speech. In contrast, spectral resolution in an implant depends on many factors, including the number of electrodes, the proximity of the electrode to the remaining neurons, and the degree of neuronal survival in the individual patient. Because several of these factors are difficult to evaluate in implanted patients, we have developed an acoustic model of implant speech processing that allows us to parametrically measure the effect of tonotopic distortions in the speech pattern in normal-hearing listeners.

C. Rationale of present study

To simulate the limited spectral cues available to cochlear implant listeners Shannon *et al.* (1995) developed an acoustic simulation of a cochlear implant that systematically reduced spectral information. In this simulation speech was divided into several contiguous frequency bands. The envelope of each band was extracted by full-wave rectification and low-pass filtering at 160 Hz. This envelope was then used to modulate a band of noise. In the original experiments Shannon *et al.* (1995) changed the number of bands and the cutoff frequency of the envelope filter. High levels of speech recognition were possible with only four modulated bands of noise, modulated at relatively low rates. These results were obtained in normal-hearing listeners in whom the global pattern of speech information, while spectrally reduced, was at least presented to the appropriate tonotopic region of the cochlea via the modulated noise bands.

Recent studies (Dorman and Loizou, 1998; Dorman *et al.*, 1997a; Fishman *et al.*, 1997) validated the noise-band simulation of cochlear implants by comparing the simulation results to performance of cochlear implant patients with the same number of channels. The performance of normal-hearing listeners with the noise-band simulation sets an upper bound on the performance of implant listeners with the same number of channels. The results from the best performing implant listeners were similar to those from the normal-hearing listeners for the same number of channels. Poor performing implant listeners may be using fewer effective spectral channels, or there may be other processing deficits underlying their poor performance.

But how would speech recognition be affected by tonotopic alterations in the minimal pattern of information? When the central auditory system is already working with a minimal spectral representation of speech (as in implants), how would recognition be affected by changes in the spectral distribution of information along the nerve array? In cochlear implant listeners the neural population stimulated by a given electrode may be larger or smaller than for the adjacent electrode. If uniformly spaced electrodes stimulate nonuniformly spaced segments of nerve, this would result in a warping or stretching of the spectral distribution of envelope cues. In the present experiments we systematically alter the spectral location and extent of four bands of speech envelope information. Since four was the minimum number of bands that provided a high level of speech recognition, any detrimental manipulation of the pattern of envelope information should result in a reduction in speech performance. These manipu-

lations may not only provide insights into the difficulties faced by cochlear implant patients, but may also demonstrate the sensitivity of normal speech pattern recognition to alterations in the spectral distribution of information.

I. METHOD

A. Subjects

Subjects were eight native speakers of American English, normal-hearing adults, ranging in age from 22 to 50 years old. Their hearing was tested in a standard audiometric fashion to ensure that their thresholds were better than 20 dB HL. Two of the authors (RVS and JW) were two of the eight subjects.

B. Equipment and signal processing

All signals were digitized at a 10- or 20-kHz sample rate and passed through a pre-emphasis filter (−6 dB/octave below 100 Hz). Signals were then split into frequency bands (third to fourth order Elliptical IIR filters). For the standard condition (STD) the output from each filter was 15 dB down from the level in the passband at the frequency where adjacent filters overlapped. For conditions with increasing overlap between adjacent filters the same nominal cutoff frequencies were used as in the STD condition, but the slopes of the filters were reduced to the specified value. Thus conditions with shallower slopes had more energy at the crossover point between adjacent filters. The envelope was extracted by full-wave rectification and low-pass filtering (−6 dB/octave Elliptical IIR filter with cutoff frequency of 160 Hz, selected to be well above the 50-Hz value where previous research had shown no deterioration in temporal cues). The envelope derived from each band was then used to modulate a white noise. The modulated noise was frequency limited by filtering with a bandpass filter (third to fourth order Elliptical IIR filters). This last bandpass filtering reduced the modulation depth to some degree because it removed the modulation sidebands. The resulting modulated noise bands were then combined, low-pass filtered at 4 kHz (8 kHz for experiment III) (Kronhite 3343: 24 dB/oct), amplified (Crown D75) and presented to the listener through headphones (TDH-49). Overall levels were calibrated for each combination of parameters to produce an average A-weighted output level of 75 dB for continuous speech.

C. Test materials

Consonant and vowel stimuli were taken from the sound track of the Iowa audiovisual speech perception laser videodisc (Tyler *et al.*, 1989). A single 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 /b d g p t k l m n f s ʃ v z j θ/ presented in an /a/-consonant-/a/ context. Vowel confusion matrices were compiled from nine presentations of each of the eight vowels in an /h/-vowel-/d/ context (heed: /hi̱d/, hawed: /hɔd/, head: /hɛd/, who'd: /hu̱d/, hid: /hɪd/, hood: /hʊd/, hud: /hʌd/, had: /hæd/, heard: /hɜ˞d/, hoed: /ho̱d/, hod: /hʊd/, hayed: /he̱d/). Recognition of words in sentences was measured using the sound track from the City University of New York laser videodisc everyday sentences (Boothroyd *et al.*, 1985). Data were collected for 24 sentences, representing 100 key words, from each subject. The sentences were of easy-to-moderate difficulty, presented with no context, and no sentences were repeated to an individual listener. No feedback was provided other than the overall score for each condition.

D. Data analysis

In each experiment the experimental conditions were tested for significant differences from a comparison condition using a standard paired *t*-test. In addition, the results of experiments II were tested against a single-channel condition to evaluate the loss of spectral information. Results of all tests of significance are presented in Tables I–IV.

II. EXPERIMENT I: LOCATION OF BAND DIVISIONS

The first experiment investigated the importance of the spacing of the cutoff frequencies that define the four bands. Three conditions were tested: linear spacing of the four bands (LIN), tonotopic spacing (LOG), and one intermediate spacing (STD). In the linear (LIN) condition the crossover frequencies (−15 dB down) of the four filters were nominally 1000, 2000, and 3000 Hz. The filter crossover frequencies in the standard (STD) condition were 800, 1500, and 2500 Hz which were selected to be the same as the filters used by Shannon *et al.* (1995). In the tonotopic (LOG) condition the frequency spectrum from 80 to 4000 Hz was divided into four equal distances on the basilar membrane, corresponding to crossover frequencies of 332, 845, and 1886 Hz (assuming a 35-mm cochlear length and using the formula of Greenwood, 1990). Fourth-order IIR elliptical filters were then designed to implement these crossover frequencies. In this experiment the same filters were used for the analysis and for filtering the noise carrier bands. The STD data are taken from the results of Shannon *et al.* (1995), collected earlier with the same eight subjects with the same test materials. The LIN and LOG conditions were run in randomized order for each listener.

Experiment I: Results and discussion

Figure 1 presents the results of experiment I for sentences, vowels, and consonants, and Table I presents the tests of significance between the experimental conditions and the STD condition. In the STD condition sound-only performance was 90%–95% correct on all three tests. Performance on sentences and vowels was significantly lower in the LIN–LIN condition and performance on all measures was significantly lower in the LOG–LOG condition. Although these scores are lower than the STD condition, they still represent high levels of speech recognition. Consonant recognition was slightly (but significantly) better in the LIN–LIN condition than in LOG–LOG, while vowel recognition was slightly better in the LOG–LOG condition.

These results indicate that the exact crossover frequencies in a four band representation are relatively minor factors in speech recognition. We speculate that the STD condition

TABLE I. Paired *t*-tests of significance for the data from experiment I.

	LIN–LIN vs STD–STD	LOG–LOG vs STD–STD	LIN–LIN vs LOG–LOG
Sentences	7.26[a]	4.76[a]	−0.21
Vowels	8.77[a]	3.33[b]	−3.41[b]
Consonants	0.85	4.11[a]	2.41[b]

[a]Significant at <0.01 level (df=7).
[b]Significant at <0.05 level (df=7).

had slightly higher scores because there was a band transition at 1500 Hz, a frequency known to divide high from low second formants and a frequency division point commonly used in automatic speech recognition systems to separate voiced from unvoiced stimuli. This is probably due to the fact that the central pattern recognition system has adapted to the variety of talkers and listening conditions. Information provided by second formant varies across talkers, room acoustics, and competing noise conditions. A specific, fixed set of frequency divisions might be best for a given talker, but a specific set of filter divisions may not be optimal across a range of different talkers or listening conditions. For example, the best dividing point may be different for male and female talkers due to differences in their formant frequencies. In a normal ear with many effective processing channels this is probably not a problem, but in processors with a reduced number of channels 1500 Hz probably represents a compromise frequency which provides the best dividing point for distinguishing coarse speech features (Zue, 1985).

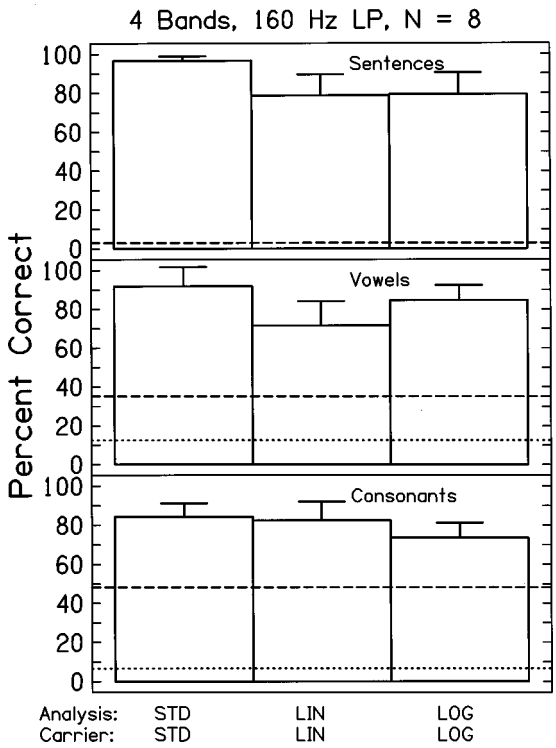

FIG. 1. Results of experiment I: Location of band divisions. Percent correct recognition of sentences (top), vowels (middle), and medial consonants (lower) for three frequency divisions. All conditions used matched analysis bands and noise carrier bands, 4-band processors and a 160-Hz, low-pass envelope filter. In all figures error bars indicate one standard deviation. Dotted line indicates chance level of performance and dashed line indicates the level of performance with no spectral cues (single-channel results from Shannon *et al.*, 1995).

III. EXPERIMENT II: WARPING THE SPECTRAL DISTRIBUTION OF ENVELOPE CUES

In cochlear implants the extent of auditory nerve excited by each electrode depends on many factors, including the electrode bipole orientation, the proximity of the electrode to surviving neurons, and the location and uniformity of surviving neurons. For example, some pathologies may produce poor spiral ganglion survival in the basal portion of the cochlear and better survival in the apical region. Other pathologies may produce relatively uniform nerve survival along the entire cochlea. Most cochlear implant speech processors divide the acoustic spectrum into equal tonotopically spaced bands and the information from each of these bands is routed to an electrode in the scala tympani. However, if the electrodes activate nonuniform regions of nerve due to a combination of the above factors, this strategy creates a mismatched condition in which the speech envelope information extracted from uniformly wide tonotopic bands are being presented to sectors of nerve of varying widths. In the example above, the basal electrode may stimulate a broad sector of nerve in a region with sparse neuronal survival, while the apical electrode may stimulate a narrow sector of nerve in a region with good survival. Since these two electrodes are presenting envelope information from comparable bands, the result is a distortion of the spatial (tonotopic) distribution of envelope information: The envelope information from high-frequency regions will activate a disproportionately wider section of the cochlea than the envelope from low-frequency regions.

To simulate such a mismatch a different frequency extent was used for the analysis and carrier bands. In this experiment the speech envelope information was extracted from linearly spaced bands and used to modulate tonotopically spaced noise bands (LIN-LOG), and vice versa (LOG-LIN). In these conditions the same envelope information was extracted as in experiment I, but the tonotopic extent of the cochlea that received that information was different from the analysis band. Thus the tonotopic representation of speech envelope cues was ''warped.''

Figure 2 presents a comparison of spectrograms for the STD–STD, LIN-LOG, and LOG-LIN conditions for the utterance ''shoo cat.'' Note how the spectral distribution of energy over time is altered for the conditions in which the analysis and carrier filters are mismatched. For a speech sound that contains a sweep in spectral energy crossing from one band to another, the transition in the processed version occurs at the right time, but in the wrong tonotopic location.

Experiment II: Results and discussion

Figure 3 presents the results for the tonotopic warping conditions on sentence, vowel, and consonant materials. Note that a dramatic and significant reduction in performance was evident in all conditions, compared to the STD–STD, LIN–LIN, or LOG–LOG conditions (Table II). The consonant scores for the mismatched conditions were only 50%–55% correct, which were not significantly different (Table II) from the consonant scores with the same subjects

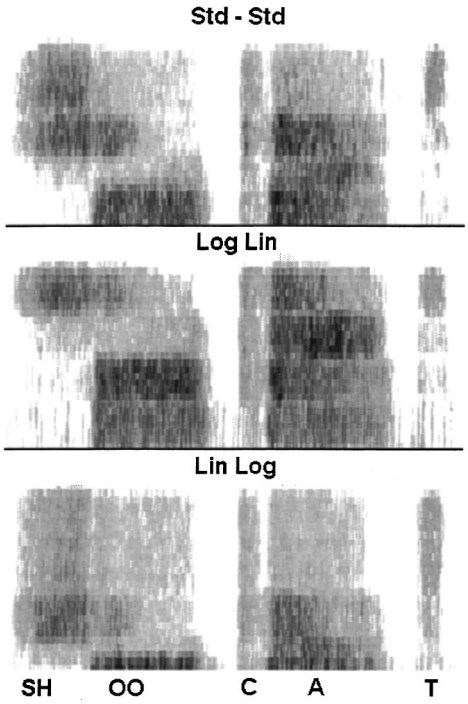

FIG. 2. Example spectrograms of the utterance ''SHOO CAT'' from conditions of experiment II: Tonotopic warping of envelope cues. Upper panel shows spectrogram from STD–STD condition: a four band processor with no tonotopic warping. In the middle panel envelope cues were extracted from logarithmically spaced filters and used to modulate noise bands that were spaced linearly (LOG-LIN) and in the lower panel the reverse (LIN-LOG).

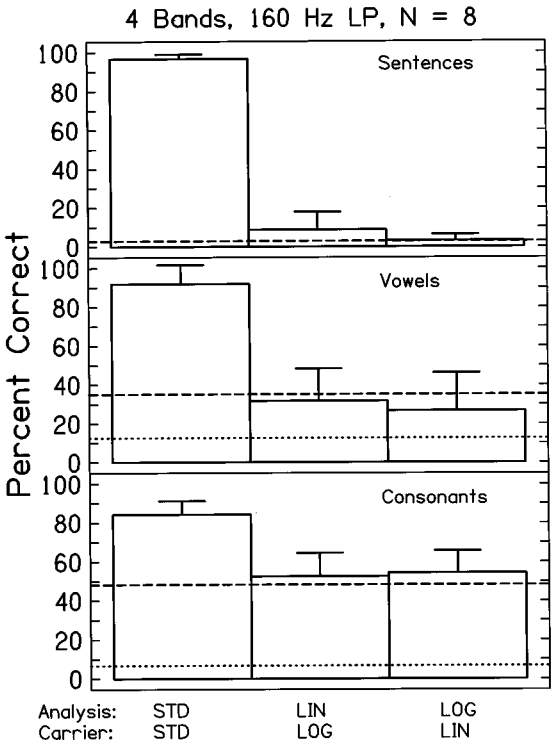

FIG. 3. Results of experiment II: Tonotopic warping of envelope cues. Dashed and dotted lines are the same as in Fig. 1.

and the same materials for a single band processor with no spectral cues (level indicated in each panel of Figs. 1, 3, 4, and 5 by a horizontal dashed line). Consonant scores in the single band condition are higher than those observed by Van Tassel *et al.* (1987) because a single male talker was used and the subjects had considerable practice with this stimulus set, which probably resulted in some incidental learning (Van Tassel *et al.*, 1992). Vowel recognition was reduced to 25% correct. Like the consonant score, this level of performance was no better (and possibly a bit worse) than the score with no spectral cues (Table II). Sentence recognition scores in the mismatched conditions were less than 10% correct.

These results indicate that the mismatching of the spectral extent of envelope cues eliminates the ability of the listener to utilize the minimal spectral cues provided by these stimuli. If the envelope cues are in their proper tonotopic locations and extents, they can be integrated for speech recognition, even with only four bands. But when the tonotopic distribution of envelope information is warped, performance falls to the level achieved in the single band case, which provides no spectral cues at all. In this case the same spectral envelope information is present as in the four band LIN–LIN or LOG–LOG conditions and the information is still distributed tonotopically. However, the central pattern recognition system apparently cannot utilize the tonotopic distribution of envelope information when the pattern is distorted. From this result we infer that the relative tonotopic distribution of envelope information is a critical dimension for recognition of speech when access to the full range of speech cues is restricted.

We speculate that people may be able to learn to recognize speech even with these distortions, because the information is still present. Prior studies of the learning time for altered or distorted speech (Blesser, 1972) indicate that several days or even weeks may be necessary for learning to understand running speech, similar to the relearning time for altered vision (Kohler, 1964).

IV. EXPERIMENT III: FREQUENCY SHIFTING ENVELOPE CUES

Cochlear implant electrodes are inserted through the round window into the scala tympani for a length of 20–25 mm. The multiple stimulating electrodes in a fully inserted device are located at tonotopic locations corresponding to frequencies of 800 Hz and above. If the electrode carrier is not fully inserted, the tonotopic location of the most apical electrode corresponds to an even higher frequency. However, all present commercial implant speech processors divide the audio spectrum into multiple, contiguous frequency bands. A processed version of the information in each of these bands (usually the envelope) is presented to a single electrode pair.

TABLE II. Paired *t*-tests of significance for the data from experiment II.

	LIN-LOG vs STD–STD	LOG-LIN vs STD–STD	LIN-LOG vs Single channel	LOG-LIN vs Single channel
Sentences	32.19[a]	83.71[a]	2.24	0.50
Vowels	15.50[a]	10.49[a]	−1.70	−3.17[a]
Consonants	9.58[a]	9.80[a]	0.66	1.22

[a]Significant at <0.01 level ($df=7$).

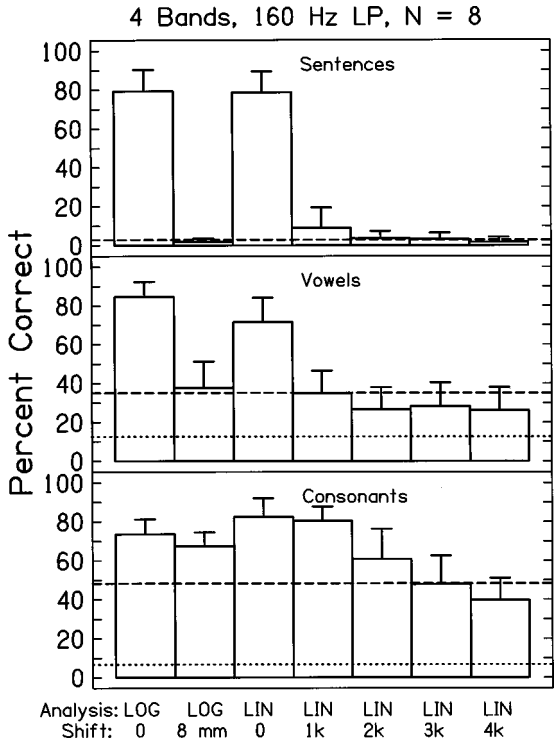

FIG. 4. Results of experiment III: Frequency shift. Dashed and dotted lines are the same as in Fig. 1.

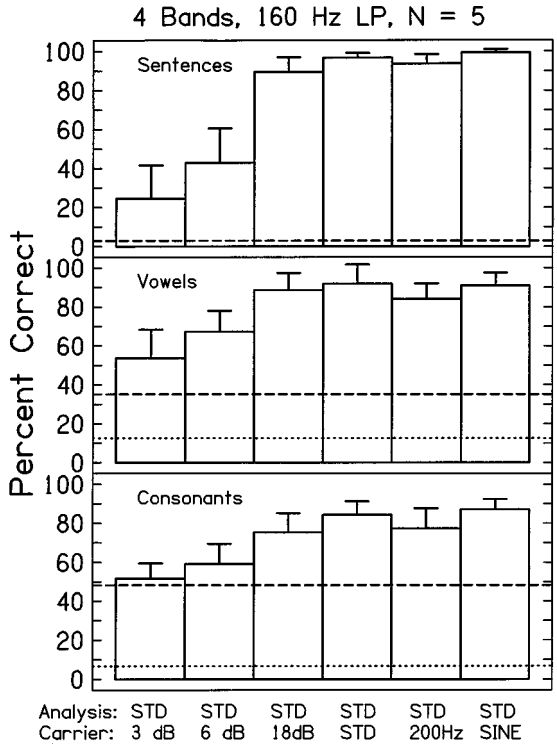

FIG. 5. Results of experiment IV: Band overlap. Dashed and dotted lines are the same as in Fig. 1.

The output of the lowest-frequency analysis band is presented to the most apical electrode pair. The output from the next lowest band is presented to the second most apical pair, and so on, until all bands are assigned to all usable electrode pairs. If we assume that each electrode is stimulating neurons near its cochlear location, this stimulation scheme will result in a basal shift of the speech pattern. If the neuron survival is uniform and the electrode proximity to those surviving neurons is uniform, the speech pattern will simply be shifted from its normal location. However, if either neuron survival or electrode location are nonuniform the speech pattern on the auditory nerve will be both shifted *and* warped.

To simulate these conditions we shifted the noise carrier bands relative to the analysis bands. In one condition the analysis bands were arranged tonotopically (LOG) and the noise carrier bands were shifted 8 mm basally in cochlear location using the Greenwood (1990) formula, which was a shift of nearly two octaves. This resulted in a condition similar to a cochlear implant that was not fully inserted. In this case the noise carrier bands were at the following frequency locations: band 1: 500–1156 Hz; band 2: 1156–2640 Hz; band 3: 2640–5052 Hz; and band 4: 5052–10 200 Hz. In another set of conditions linearly spaced analysis bands (LIN) were used and the output noise carrier bands were shifted up in frequency by 1, 2, 3, or 4 kHz from the analysis bands. This manipulation resulted in a basal shift of the speech pattern and a compression of the pattern in terms of tonotopic distance.

Experiment III: Results and discussion

Figure 4 presents the results from the frequency shift conditions for sentences, vowels and consonants. The results from experiment I (LIN–LIN and LOG–LOG) are reproduced here for comparison. For the LOG-8 mm shift condition the consonant score was not significantly reduced, but the vowel score was dramatically reduced and the sentence recognition score was reduced to essentially zero (statistical significance results are presented in Table III). Although all of the temporal cues in the four bands were preserved and the four bands maintained their relative tonotopic spacing and extent, speech recognition was completely disrupted.

In the LIN shift conditions there was a graded effect as a function of the amount of frequency shift. For a 1-kHz shift (LIN-1 k) the consonant score was similar to the unshifted condition, while the vowel score was significantly reduced. The sentence score was strongly reduced to only 10% correct from the unshifted score of 80% correct. As the frequency shift increased the consonant score decreased, while the vowel and sentence scores remained at low levels for all frequency shifts, probably indicating a floor effect.

We anticipated that the detrimental effect would be less in the LOG shift condition than for the LIN shift condition because the relative tonotopic pattern was preserved, and simply shifted in mm along the cochlea. If the central mechanisms for speech pattern recognition stored templates of *relative* speech envelope patterns in terms of mm along cochlea or along the nerve array (Miller, 1989), we would expect that the LOG shift condition would result in relatively little decrease in performance. This was clearly not the case, suggesting that the central representation is stored in terms of *absolute*, not relative cochlear place. The LIN shift conditions may have resulted in a smaller decrease in performance because the shift in terms of mm in the cochlea was smaller

TABLE III. Paired *t*-tests of significance for the conditions of experiment III.

	LIN 1 k vs LIN–LIN	LIN 2 k vs LIN–LIN	LIN 3 k vs LIN–LIN	LIN 4 k vs LIN–LIN	LOG 8 mm vs LOG–LOG
Sentences	22.24[a]	28.64[a]	27.58[a]	31.44[a]	21.27[a]
Vowels	9.77[a]	9.26[a]	11.15[a]	8.64[a]	11.43[a]
Consonants	0.74	4.33[a]	6.56[a]	9.43[a]	2.28

[a]Significant at <0.01 level ($df=7$).
[b]Significant at <0.05 level ($df=7$).

in the high-frequency regions than for the LOG shift condition.

Recent clinical surveys (Bredberg and Lindstrom, 1996; Hartrampf *et al.*, 1996; Kumakawa *et al.*, 1997) have shown poorer performance in cochlear implant patients with short electrode insertion depths, even after a substantial period of experience. Dorman *et al.* (1997b) simulated the effect of electrode insertion depth in normal-hearing listeners with the noise-band technique. They found that the performance level of normal-hearing listeners was similar to that of cochlear implant listeners when the noise carrier bands were shifted in mm to simulate typical electrode insertion depths. This result implies that the implanted listeners did not improve their performance with practice, but were performing similarly to normal-hearing listeners with the same signal processing and tonotopic shift and no practice.

Rosen *et al.* (1997) recently measured performance as a function of training in normal-hearing listeners with a four noise-band processor and a 6-mm tonotopic shift. Initial performance was reduced to chance by the 6-mm shift, a result similar to the 8-mm shift in the present experiment III. After only 3 h of training, they found substantial improvement, but performance at that point was still only half of the level achieved with the STD–STD condition. The Rosen *et al.* result indicates that some relearning is possible for tonotopically shifted speech patterns, but the clinical studies suggest that complete accommodation may not be possible even after prolonged experience.

V. EXPERIMENT IV: SPECTRAL SMEARING

In a cochlear implant it is widely assumed that good speech recognition requires selective stimulation of distinct neural populations by each electrode. It is also assumed that neural overlap is inversely correlated with electrode discrimination, i.e., that electrodes that stimulate different neural populations are easily discriminable while electrodes that stimulate overlapping neural populations are poorly discriminated. In acoustic listening, the analogous assumption is that broadened auditory filters observed in some hearing impaired listeners are the primary cause of reduced speech intelligibility in those listeners (Boothroyd *et al.*, 1996; ter Keurs *et al.*, 1992, 1993; Moore *et al.*, 1992).

Electrode interaction in cochlear implants or broadened auditory filters in impaired hearing was simulated by broadening the filter slopes on the noise carrier bands. The analysis bands were fixed at the same frequencies and filter slopes as in the STD condition. However, the slopes of the noise carrier bands were systematically varied to produce conditions of varying degrees of overlap. At one extreme the carriers were simply sinusoids near the center frequency of each band. In another condition (STD-200 Hz) the noise carrier bands were 200-Hz-wide bands of noise for each channel. In these two conditions there was no overlap in the output of the carrier bands and, in fact, the output stimulus had spectral regions in which no stimulus was present. To simulate increasing overlap between output channels the carrier noise bands were fixed at the nominal crossover frequencies of the STD configuration and the slopes of the filters broadened from -24 to -18 to -6 to -3 dB/octave. Since the filters were generally separated by an octave or less this last condition resulted in overlap such that the output level at the center of a band was only 3 dB above the level from the adjacent carrier bands.

Experiment IV: Results and discussion

Figure 5 presents the results of the carrier overlap condition for sentences, vowels and consonants, and Table IV presents the statistical comparisons. In general, the degree of overlap in the output carrier bands had relatively little effect on speech recognition. Broadening the carrier bands from sinusoids to bands of noise with -18 dB/octave slopes produced a significant effect on consonant and sentence recognition, but the absolute performance level was still quite high. For the -3 and -6 dB/octave slope conditions all results were significantly lower than the STD condition.

Consider the limiting conditions for such manipulations. We know that as the filter slopes are broadened to 0 dB/octave the output is reduced to a single-band processor and the performance would be reduced dramatically, to less than 5% correct on sentences. Yet at -3 dB/octave, which results in extensive overlap, listeners still were able to correctly identify words in sentence material at 25%. This is an encouraging result for cochlear implant speech processor design. It indicates that speech recognition is possible even with considerable overlap between electrodes. Apparently almost total overlap is necessary to reduce performance to the single-band level.

At the other end of the scale, 200-Hz-wide carrier bands and sinusoidal carriers, which result in a sparse spectral representation, still allow nearly perfect recognition of sentence material with only four modulated sinusoids, replicating a recent result by Dorman *et al.* (1997a). Earlier studies with sinusoidal vocoders required 8–10 bands to achieve similar levels of performance (Hill *et al.*, 1968), in some cases because they were using low-pass filters for envelope extraction of only 15–20 Hz. With this severe filtering, some of the more rapid temporal cues and voicing periodicity was lost, and more spectral channels had to be added to make up for

TABLE IV. Paired *t*-tests of differences from the STD–STD condition for the conditions of experiment IV.

	STD 3 dB	STD 6 dB	STD 18 dB	STD 200 Hz	STD Sine
Sentences	9.98[a]	9.20[a]	3.27[b]	2.09	−5.41[a]
Vowels	16.33[a]	9.77[a]	0.96	2.43[b]	0.29
Consonants	10.81[a]	10.35[a]	3.14[b]	2.29	−1.90

[a]Significant at <0.01 level ($df=7$).
[b]Significant at <0.05 level ($df=7$).

the loss of temporal information. In a normal cochlea, a sinusoid or narrow band of noise is represented in the neural response across a considerable tonotopic extent due to the normal cochlear processing (Rose *et al.*, 1971). So the ''sparse'' spectral representation of the sinusoidal and 200-Hz noise-band stimuli actually resulted in an internal representation on the auditory nerve that was more contiguous than the spectrum.

VI. GENERAL DISCUSSION

The present series of experiments builds on earlier work with signal correlated noise (Schroeder, 1968; Rosen, 1992; van Tassel *et al.*, 1987, 1992; Turner *et al.*, 1995; Shannon *et al.*, 1995). Those experiments removed all spectral information in speech and preserved only the broadband temporal envelope. Recognition of consonants was surprisingly high even in the complete absence of any spectral information. Turner *et al.* (1995) and Shannon *et al.* (1995) extended this technique to two bands. In their experiments speech was filtered into a high-pass band and a low-pass band, divided at 1500 Hz. The envelope from the high-pass band of speech was used to modulate a high-pass band of noise and similarly for the low-pass band. Recognition of vowels and consonants improved dramatically with two modulated noise bands compared to one. This result demonstrated that temporal envelope information, quantized into two spectral bands, was sufficient for overall high recognition performance on consonants and was sufficient for recognition of almost 100% of the voicing and manner information. In addition, Shannon *et al.* (1995) expanded the technique to three and four bands, which resulted in high levels of sound-only speech recognition. Information transmission analysis (Miller and Nicely, 1955) showed that the improvement from two to four bands was primarily due to increased information on place of articulation, as might be expected.

The present series of experiments investigated speech recognition by starting with the minimal spectral representation that resulted in good speech recognition (4 bands: STD–STD). This limited set of cues was then perturbed/distorted to see which dimensions were most critical for speech recognition. We might expect considerable robustness to small distortions in amplitude or spectral cues because these types of distortions occur commonly in everyday listening conditions. However, in listeners with limited capacity for processing acoustic signals, like the severely hearing impaired or cochlear implant listeners, the remaining cues they have available may be more delicate and more sensitive to distortion. The present experiments extend the results of previous experiments to demonstrate the robustness of speech pattern recognition to some distortions in the tonotopic pattern but not to others.

The data of experiments II and III on spectral warping and shifting demonstrate some conditions which have a more detrimental effect on vowels than on consonants. Drullman *et al.* (1994) found that conditions of temporal smearing had a larger detrimental effect on consonants than on vowels. As speech cues are reduced to a minimum it appears that reductions in temporal and spectral cues have differential effects on consonants and vowels. Shannon *et al.* (1995) demonstrated that consonantal voicing and manner cues are correctly perceived even when spectral cues are reduced to only two bands of modulated noise, indicating that those consonant contrasts require only minimal spectral information. However, vowel recognition and consonantal place of articulation required more spectral detail for high levels of recognition, indicating that vowel recognition depends more on spectral cues than on temporal cues. In the present experiments temporal cues were preserved across all conditions and spectral cues were distorted by shifting or warping. In several conditions consonant recognition was only slightly degraded while vowel recognition was reduced to single-channel levels. This result is consistent with those previous studies in that vowel recognition is more sensitive to spectral distortions than is consonant recognition. In conditions where consonant performance was relatively good and vowel performance was poor (Fig. 4, LOG-8 mm condition), sentence recognition was completely disrupted. This suggests either that both vowel and consonant phonemic recognition must achieve a certain minimal level before the linguistic retrieval can be successfully accomplished or that vowels are more important than consonants for sentence recognition.

The present experiments measured speech recognition performance without any practice in conditions of shifted and warped spectral distribution of envelope cues. Certainly, experience with the altered spectral distributions will improve performance, as demonstrated by Blesser (1972) and recently by Rosen *et al.* (1997). However, our intention was to take a ''snapshot'' of the effect of spectral alterations compared to existing normal acoustic pattern recognition. Whether or not these altered patterns can be fully learned and whether performance will ultimately return to the unaltered level is a separate and important question. From the present results it is not possible to predict which pattern of alteration from the present experiments might be learnable.

VII. SUMMARY AND CONCLUSIONS

Speech recognition is a complex and robust process that can be accomplished under conditions of severe distortion in

the original signal. Under ideal listening conditions the central pattern recognition and linguistic access mechanisms have a rich and redundant set of peripheral cues. As listening conditions deteriorate, or under conditions of peripheral pathology, the central pattern recognition must work with a reduced set of cues from the periphery. The present series of experiments tries to quantify the effect of corrupted and limited peripheral information on speech recognition.

Previous work (Shannon *et al.*, 1995) demonstrated that four bands of modulated noise were sufficient for high levels of speech recognition. The present experiments expand this previous result to define which parameters of such a reduced representation are most critical for speech recognition. Experiment I demonstrated that the exact cutoff frequencies which define the four bands were not critical for speech recognition. Experiment II demonstrated that warping the spectral distribution of envelope cues renders speech completely unintelligible. Experiment III demonstrated that a tonotopic shift of the envelope pattern resulted in poor intelligibility, even when the relative cochlear distribution of envelope cues was preserved. This result indicates that the central pattern recognition mechanisms are not robust to even linear translations of the pattern along the neural array, at least without training. Experiment IV demonstrated that the selectivity of the envelope carrier bands was not critical for speech recognition; performance deteriorated only when the bands were broadly overlapping, smearing the tonotopic specificity of envelope cues.

ACKNOWLEDGMENTS

We appreciate the help of Franco Portillo and Vivek Kamath in running subjects, and Alena Wilson and Steve Otto for their perseverance. This work was supported by Grant No. DC01526 from NIH/NIDCD.

Blesser, B. (**1972**). ''Speech perception under conditions of spectral transformation: I. Phonetic characteristics,'' J. Speech Hear. Res. **15**, 5–41.

Boothroyd, A., Hnath-Chisolm, T., and Hanin, L. (**1985**). ''A sentence test of speech perception: Reliability, set-equivalence, and short-term learning,'' City University of New York, Report No. RCI10.

Boothroyd, A., Mulhearn, B., Gong, J., and Ostroff, J. (**1996**). ''Effects of spectral smearing on phoneme and word recognition,'' J. Acoust. Soc. Am. **100**, 1807–1818.

Bredberg, G., and Lindstrom, B. (**1995**). ''Insertion length of electrode array and its relation to speech communication performance and nonauditory side effects in multichannel-implanted patients,'' Ann. Phys. (Leipzig) **104**, 256–258.

Dorman, M. F., and Loizou, P. C. (**1998**). ''Identification of consonants and vowels by cochlear-implant patients using a 6-channel continuous interleaved sample processor and by normal-hearing subjects using simulations processors with two to nine channels,'' Ear and Hearing **19** (2), 162–166.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997a**). ''Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs,'' J. Acoust. Soc. Am. **102**, 2403–2411.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997b**). ''Simulating the effect of cochlear-implant electrode insertion depth on speech understanding,'' J. Acoust. Soc. Am. **102**, 2993–2996.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). ''Effect of temporal envelope smearing on speech perception,'' J. Acoust. Soc. Am. **95**, 1053–1064.

Drullman, R. (**1995**). ''Temporal envelope and fine structure cues for speech intelligibility,'' J. Acoust. Soc. Am. **97**, 585–592.

Erber, N. (**1972**). ''Speech envelope cues as an acoustic aid to lipreading for profoundly deaf children,'' J. Acoust. Soc. Am. **51**, 1224–1227.

Fishman, K., Shannon, R. V., and Slattery, W. H. (**1997**). ''Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor,'' J. Speech Hear. Res. **40**, 1201–1215.

Fletcher, H., and Steinberg, J. C. (**1929**). ''Articulation testing methods,'' Bell Syst. Tech. J. **8**, 806–854.

French, N. R., and Steinberg, J. C. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Grant, K. W., Ardell, L., Kuhl, P., and Sparks, D. (**1985**). ''The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speech reading in normal subjects,'' J. Acoust. Soc. Am. **77**, 671–677.

Grant, K. W., Braida, L. D., and Renn, R. J. (**1991**). ''Single band amplitude envelope cues as an aid to speechreading,'' Q. J. Exp. Psychol. **43A**, 621–645.

Greenwood, D. D. (**1990**). ''A cochlear frequency-position function for several species—29 years later,'' J. Acoust. Soc. Am. **87**, 2592–2605.

Hartmann, R., Topp, G., and Klinke, R. (**1984**). ''Discharge patterns of cat primary auditory fibers with electrical stimulation of the cochlea,'' Hearing Res. **13**, 47–62.

Hartrampf, R., Dahm, M. C., Battmer, R. D., Gnadeberg, D., Straus-Schier, A., Rost, U., and Lenarz, T. (**1995**). ''Insertion depth of the Nucleus electrode array and relative performance,'' Ann. Phys. (Leipzig) **104**, 277–280.

Hill, F. J., McRae, L. P., and McClellan, R. P. (**1968**). ''Speech recognition as a function of channel capacity in a discrete set of channels,'' J. Acoust. Soc. Am. **44**, 13–18.

Kohler, I. (**1964**). *The Formation and Transformation of the Perceptual World*, translated by H. Fiss (International Universities Press, Vienna).

Kumakawa, K., Tadeka, H., and Ujita, N. (**1997**). ''Determining the optimum insertion length of electrodes in the Cochlear 22-channel implant: Results of a clinical study,'' Adv. Otolaryngol. **52**, 129–134.

Miller, G. A., and Nicely, P. E. (**1955**). ''An analysis of perceptual confusions among some English consonants,'' J. Acoust. Soc. Am. **27**, 338–352.

Miller, J. D. (**1989**). ''Auditory-perceptual interpretation of the vowel,'' J. Acoust. Soc. Am. **85**, 2114–2134.

Moore, B. C. J., Glasberg, B. R., and Simpson, A. (**1992**). ''Evaluation of a method of simulating reduced frequency selectivity,'' J. Acoust. Soc. Am. **91**, 3402–3423.

Nelson, D. A., Schmitz, J. L., Donaldson, G. S., Viemeister, N. F., and Javel, E. (**1997**). ''Intensity discrimination as a function of stimulus level with electric stimulation,'' J. Acoust. Soc. Am. **100**, 2393–2414.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). ''Speech perception without traditional speech cues,'' Science **212**, 947–950.

Rose, J. E., Hind, J. E., Anderson, D. J., and Brugge, J. F. (**1971**). ''Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey,'' J. Neurophysiol. **34**, 685–699.

Rosen, S. (**1992**). ''Temporal information in speech: Acoustic, auditory and linguistic aspects,'' Philos. Trans. R. Soc. London, Ser. B **336**, 367–373.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1997**). ''Perceptual adaptation by normal listeners to upward shifts of spectral information in speech and its relevance for users of cochlear implants,'' in *Speech, Hearing, and Language: Work in Progress* (Phonetics and Linguistics, University College, London), Vol. 10.

Schroeder, M. R. (**1968**). ''Reference signal for signal quality studies,'' J. Acoust. Soc. Am. **44**, 1735–1736.

Shannon, R. V. (**1983**). ''Multichannel electrical stimulation of the auditory nerve in man: Basic psychophysics,'' Hearing Res. **11**, 157–189.

Shannon, R. V. (**1986**). ''Temporal processing in cochlear implants,'' in *Sensorineural Hearing Loss: Mechanisms, Diagnosis, and Treatment*, edited by M. J. Collins, T. Glattke, and L. Harker (University of Iowa Press, Iowa City), pp. 323–334.

Shannon, R. V. (**1989**). ''Detection of gaps in sinusoids and pulse trains by patients with cochlear implants,'' J. Acoust. Soc. Am. **85**, 2587–2592.

Shannon, R. V. (**1990**). ''Forward masking in patients with cochlear implants,'' J. Acoust. Soc. Am. **88**, 741–744.

Shannon, R. V. (**1992**). ''Temporal modulation transfer functions in patients with cochlear implants,'' J. Acoust. Soc. Am. **91**, 1974–1982.

Shannon, R. V., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (**1995**). ''Speech recognition with primarily temporal cues,'' Science **270**, 303–304.

ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). ''Effect of spectral

envelope smearing on speech reception,'' J. Acoust. Soc. Am. **91**, 2872–2880.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). ''Effect of spectral envelope smearing on speech reception,'' J. Acoust. Soc. Am. **93**, 1547–1552.
Turner, C. W., Souza, P. E., and Forget, L. N. (**1995**). ''Use of temporal envelope cues in speech recognition by normal and hearing-impaired listeners,'' J. Acoust. Soc. Am. **97**, 2568–2576.
Tyler, R. S., Preece, J. P. and Lowder, M. W. (**1987**). The Iowa audiovisual speech perception laser videodisc, Laser Videodisc and Laboratory Report, Dept. of Otolaryngology, University of Iowa, Iowa City, IA.
Van den Honert, C., and Stypulkowski, P. H. (**1984**). ''Physiological properties of the electrically stimulated auditory nerve. II. Single fiber recordings,'' Hearing Res. **14**, 225–243.
Van den Honert, C., and Stypulkowski, P. H. (**1987**). ''Single fiber mapping of spatial excitation patterns in the electrically stimulated auditory nerve,'' Hearing Res. **29**, 195–206.
Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). ''Speech waveform envelope cues for consonant recognition,'' J. Acoust. Soc. Am. **82**, 1152–1161.
Van Tasell, D. J., Greenfield, D. G., Logemann, J. J., and Nelson, D. A. (**1992**). ''Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants,'' J. Acoust. Soc. Am. **92**, 1247–1257.
Zeng, F. G., and Shannon, R. V. (**1992**). ''Loudness balance between electric and acoustic stimulation,'' Hearing Res. **60**, 231–235.
Zeng, F.-G., and Shannon, R. V. (**1994**). ''Loudness coding mechanisms inferred from electric stimulation of the human auditory system,'' Science **264**, 564–566.
Zue, V. W. (**1985**). ''The use of speech knowledge in automatic speech recognition,'' Proc. IEEE **73**, 1602–1615.