

# Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution

Philipos C. Loizou<sup>a)</sup>

*Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688*

Michael Dorman

*Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287*

Oguz Poroy

*Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688*

Tony Spahr

*Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287*

(Received 8 April 2000; accepted for publication 28 August 2000)

The importance of intensity resolution in terms of the number of intensity steps needed for speech recognition was assessed for normal-hearing and cochlear implant listeners. In experiment 1, the channel amplitudes extracted from a six-channel continuous interleaved sampling (CIS) processor were quantized into 2, 4, 8, 16, or 32 steps. Consonant recognition was assessed for five cochlear implant listeners, using the Med-El/CIS-link device, as a function of the number of steps in the electrical dynamic range. Results showed that eight steps within the dynamic range are sufficient for reaching asymptotic performance in consonant recognition. These results suggest that amplitude resolution is not a major factor in determining consonant identification. In experiment 2, the relationship between spectral resolution (number of channels) and intensity resolution (number of steps) in normal-hearing listeners was investigated. Speech was filtered through 4–20 frequency bands, synthesized as a linear combination of sine waves with amplitudes extracted from the envelopes of the bandpassed waveforms, and then quantized into 2–32 levels to produce stimuli with varying degrees of intensity resolution. Results showed that the number of steps needed to achieve asymptotic performance was a function of the number of channels and the speech material used. For vowels, asymptotic performance was obtained with four steps, while for consonants, eight steps were needed for most channel conditions, consistent with our findings in experiment 1. For sentences processed through 4 channels, 16 steps were needed to reach asymptotic performance, while for sentences processed through 16 channels, 4 steps were needed. The results with normal-hearing listeners on sentence recognition point to an inverse relationship between spectral resolution and intensity resolution. When spectral resolution is poor (i.e., a small number of channels is available) a relatively fine intensity resolution is needed to achieve high levels of understanding. Conversely, when the intensity resolution is poor, a high degree of spectral resolution is needed to achieve asymptotic performance. The results of this study, taken together with previous findings on the effect of reduced dynamic range, suggest that the performance of cochlear implant subjects is primarily limited by the small number (four to six) of channels received, and not by the small number of intensity steps or reduced dynamic range. © 2000 Acoustical Society of America. [S0001-4966(00)04911-0]

PACS numbers: 43.71.Ky, 43.71.Es [CWT]

## I. INTRODUCTION

Normal-hearing listeners can detect amplitude changes over a 100-dB acoustic dynamic range with a high resolution. The total number of discriminable intensity steps for normal acoustic hearing, as calculated from Weber fractions reported by Schroder *et al.* (1994), was found to be over 80 (Nelson *et al.*, 1996). The large dynamic range coupled with fine intensity resolution and fine spectral resolution allows normal-hearing listeners to maintain high speech intelligibility in background noise and in extreme signal conditions lacking traditional spectral cues.

In contrast, cochlear implant listeners have a small dynamic range, typically 5–30 dB, and receive only a small number (four to six) of channels of spectral information through their device, despite the large number of electrodes stimulated (Fishman *et al.*, 1997; Wilson, 1997). The intensity difference limens (DLs) in electrical stimulation are comparable, and sometimes better, than the DLs in acoustic stimulation, depending on the intensity level. Implant listeners can detect 1–2-dB changes in intensity near threshold and 0.25–0.5-dB changes at higher intensity levels (e.g., Shannon, 1983; Nelson *et al.*, 1996; Pfungst *et al.*, 1983; Hochmair-Desoyer *et al.*, 1981; Dillier *et al.*, 1983). Although some implant listeners (e.g., Hochmair-Desoyer *et al.*, 1981) can detect smaller intensity changes than

<sup>a)</sup>Electronic mail: loizou@utdallas.edu

normal-hearing listeners, the difference is not enough to compensate for the large difference in dynamic range. The number of discriminable steps within the dynamic range is considerably smaller than the number of steps in acoustic hearing, and also varies considerably among subjects. Nelson *et al.* (1996) found that in some subjects the 30-dB dynamic range of speech was coded by fewer than 8 steps while in other subjects it was coded by less than 45 steps.<sup>1</sup> The three factors cited above, spectral resolution (number of channels), dynamic range, and intensity resolution (number of discriminable intensity steps across the dynamic range), are all factors that can potentially affect the performance of cochlear implant subjects. The number of channels available clearly affects the transmission of spectral cues (Dorman *et al.*, 1997; Loizou *et al.*, 1999; Shannon *et al.*, 1995), while the dynamic range and the number of discriminable steps may affect the coding of temporal-envelope cues. However, the extent to which these factors, separately and/or by interactions, affect speech understanding is not yet well understood.

We have addressed the effect of reduced dynamic range on speech understanding in a previous study (Loizou *et al.*, 2000). In that study, we simulated the effect of reduced dynamic range using normal-hearing listeners as subjects to avoid the confounding factors (e.g., surviving ganglion cells, electrode insertion depth, etc.) associated with cochlear implants. Signals were processed in a manner similar to a six-channel continuous interleaved sampling (CIS) processor (Wilson *et al.*, 1991) and output as the sum of sine waves with frequencies centered in the middle of each analysis band. The amplitudes of the sine waves were compressed to fit within a 6-, 12-, 18-, and 24-dB dynamic range. Our results showed a significant effect of compression for all test materials, although the effect of the compression was different for the three test materials (vowels, consonants, and sentences). Vowel recognition was affected the most by the compression, while consonant recognition was affected the least by the compression. Sentence recognition was moderately affected. Similar findings were reported by Zeng and Galvin (1999) with four implant listeners using the Nucleus-22 SPEAK speech processor. Two conditions of reduced dynamic range were examined. In the first condition, they reduced the electrode dynamic range to 25% of the full range by artificially raising the threshold levels to 75% of the dynamic range. In the second condition, they created a binary representation of the acoustic amplitudes by further decreasing the comfortable levels to 76% of the dynamic range. Their results showed that vowel recognition was significantly affected (although marginally) by the dynamic range reduction, whereas consonant recognition was not significantly affected.

The above studies demonstrated a significant effect, albeit mild, of dynamic range reduction on speech recognition. It is possible, however, that the number of discriminable steps available in the patient's dynamic range may be a more important factor for speech perception than the dynamic range itself. That is, a patient with a large dynamic range might only have a few discriminable steps, and, similarly, a patient with a small dynamic range might have a large number of steps. The variability in the number of steps among

the eight implant subjects in the Nelson *et al.* (1996) study led us to wonder whether the performance of the poorly performing subjects was limited by a small number of steps. To gain a purchase on the answer to this question, we created, for presentation to implant patients, VCV syllables that had been quantized into 2–512 intensity steps. In this manner we experimentally controlled the intensity resolution available to the patients. At issue in experiment 1 was how many quantized steps are necessary for patients to achieve an asymptotic level of consonant recognition. We chose consonant recognition as our measure since amplitude envelopes are significant cues to consonant identity.

To better understand the role of intensity resolution on speech recognition one also needs to take into account the number of channels available. There might be a trade-off relationship between spectral resolution (number of channels) and intensity resolution (number of quantization steps). This hypothesis is based on our view (Dorman *et al.*, 1997) that when speech is processed through a small number of channels, the relative differences in across-channel amplitudes must be used to code frequency information. In this view, if intensity resolution were to be distorted, then speech recognition ought to decline. On the other hand, when speech is processed through a large number of channels, fine intensity resolution might not be needed, since the frequency information can be coded by the channels which have energy. These questions are investigated in experiment 2 with normal-hearing listeners, where we assess speech intelligibility as a function of number of channels and as a function of number of intensity steps. Normal-hearing listeners were used because the channels and steps manipulations cannot be independently controlled with implant listeners due to the many confounding factors associated with electrical stimulation. The results of experiment 2 could also be used to benchmark the performance of cochlear implant listeners, i.e., to assess whether cochlear implant listeners extract the same amount of information as normal-hearing listeners under similar conditions of reduced intensity resolution. To produce speech with varying degrees of intensity resolution, we synthesized speech as a linear combination of sine waves and quantized the amplitudes of the sinewaves to 2–32 levels. The intelligibility of vowels, consonants, and sentences was assessed as a function of spectral resolution and as a function of intensity resolution.

## II. EXPERIMENT 1: ELECTRIC HEARING

### A. Method

#### 1. Subjects

The subjects were five postlingually deafened adults who had used a CIS processor for periods ranging from 3 to 4 years. Each of the patients had used a four-channel, compressed-analog signal processor (Ineraid) for at least 4 years before being switched to a CIS processor. The patients ranged in age from 40 to 68 years and they were all native speakers of American English. Biographical data for each patient are presented in Table I. Four patients were fitted with a six-channel CIS processor and one patient (S2) was fitted with a five-channel CIS processor.

TABLE I. Biographical data of the five cochlear-implant users who participated in this study.

Subject	Gender	Age (years) at detection of hearing loss	Age at which hearing aid gave no benefit	Age fit with Ineraid	Age at testing	Etiology of hearing loss	Score on H.I.N.T sentences in quiet	Score on NU-6 words in quiet
S1	M	20	46	63	68	unknown	88	46
S2	F	10	46	47	55	unknown	44	20
S3	M	5	43	48	58	unknown	92	43
S4	F	7	31	33	40	unknown/ hereditary	100	80
S5	F	23	48	51	57	unknown	100	71

**2. Speech material**

The consonant test consisted of 20 /vCv/ consonants in three vowel environments, /i a u/, produced by a single female speaker, and was taken from the consonant database recorded at the House Ear Institute (Shannon *et al.*, 1999). The 20 consonants were /b p d t g k f v s z ʃ ð t ʃ d ʒ m n r l j w/.

**3. Experimental setup**

This experiment was performed on our laboratory cochlear implant processor, which was based on the design of the Geneva/RTI/MEEI wearable processor (Francois *et al.*, 1994). Several modifications were made to the Geneva design, the most important of which was the addition of five current sources. The block diagram of the laboratory processor is shown in Fig. 1. The input analog circuit consists of an audio multiplexer that selects the source of the input signal to the processor, several fixed-gain amplifiers, one variable-gain amplifier (adjusted externally by a sensitivity knob), an antialiasing filter, and a 16-bit A/D converter. The sampling rate of the A/D converter is controlled by the DSP chip, and for this study it was fixed at 22 kHz. The cutoff frequency of the antialiasing filter was set at 6.7 kHz. Once the signal is digitized, it is transmitted to the Motorola DSP56002 chip, where it is processed through the CIS strategy (see description in the following section). The CIS outputs are finally fed through a SSI port to the current sources built around a digital-to-analog converter. Biphasic pulses are generated, with amplitudes equal to the CIS envelope outputs, and sent

to the electrodes for stimulation. The pulse width as well as the stimulation rate was controlled through software. More information about the hardware of the laboratory processor can be found in Poroy and Loizou (2000).

**4. Signal processing and amplitude quantization**

Signals were first processed through a high-pass preemphasis filter (1200-Hz cutoff), with a 3-dB/octave roll-off, and then bandpassed into six frequency bands using sixth-order Butterworth filters. The center frequencies of the six bandpass filters were 461, 756, 1237, 2025, 3316, and 5428 Hz. The envelopes of the filtered signals were extracted by full-wave rectification and low-pass filtering (second-order Butterworth) with a 400-Hz cutoff frequency. The six envelope amplitudes  $A_i$  ( $i=1,2,\dots,6$ ) were mapped to electrical amplitudes  $E_i$  using a power-law transformation:

$$E_i = cA_i^p + d,$$

where  $c$  and  $d$  are constants chosen so that the electrical amplitudes fall within the range of threshold and most-comfortable level, and  $p$  is the power exponent. The power exponent  $p$  was set equal to  $-0.0001$  to obtain a compression function similar to the logarithmic function found in the Med-El/CIS link device.<sup>2</sup> The electrical amplitudes were then systematically quantized into  $Q$  steps ( $Q=2,4,8,16,32,512$ ) to create six different conditions with varying degrees of intensity resolution. The 512-step condition corresponded to the number of steps used in our default CIS implementation, and was labeled the “unquantized”

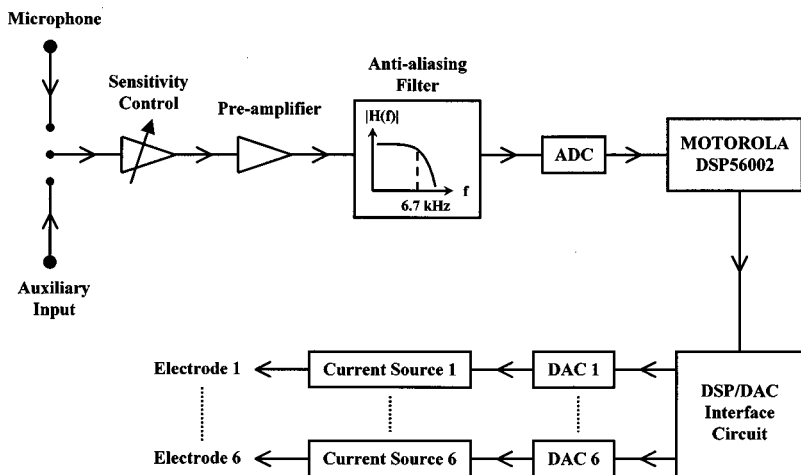


FIG. 1. Block diagram of the laboratory cochlear implant processor used in this study.

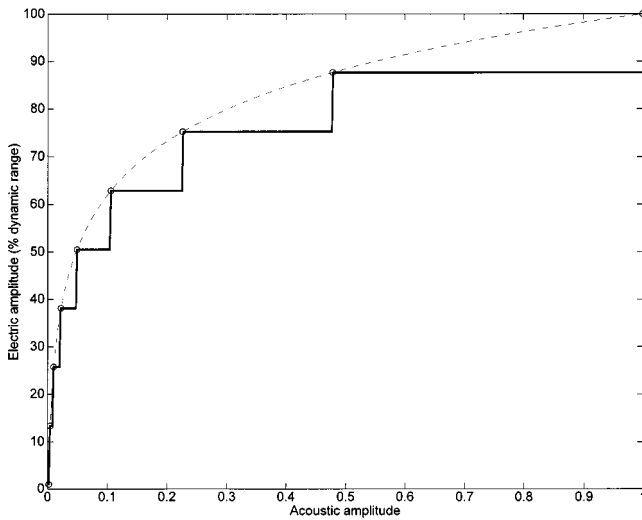


FIG. 2. Example of a logarithmic compression function quantized uniformly into eight steps. The solid curve shows the eight quantized levels and the dashed curve shows the original, unquantized, logarithmic function.

condition. [Note that there are some differences between the quantized steps of this study, and the psychophysical discriminable steps measured by Nelson *et al.* (1996) (see Discussion in Sec. II B)]. As an example, Fig. 2 illustrates the quantization of the output dynamic range into eight uniform steps. The quantization step sizes,  $\Delta_i$ , were estimated from the threshold (THR) and most-comfortable (MCL) levels of each electrode as follows:

$$\Delta_i = \frac{MCL_i - THR_i}{Q - 1}, \quad i = 1, 2, \dots, 6,$$

where  $THR_i$  and  $MCL_i$  are the threshold and most-comfortable levels of the  $i$ th electrode, respectively. Note that since each electrode had different values for  $THR_i$  and  $MCL_i$ , the step sizes  $\Delta_i$  were different for each electrode. The quantized envelope amplitudes were finally used to modulate biphasic pulses of duration 40  $\mu$ s/phase at a stimulation rate of 2100 pulses/s. The electrodes were stimulated in the same order as in the subject's daily processors. For all but one subject, the electrodes were stimulated in "staggered" order.

### 5. Procedure

The test was divided into six sessions, one for each step condition. The six conditions were counterbalanced among subjects to avoid any order effects. There were nine repetitions of each consonant, presented in blocks of three repetitions each. The consonants were completely randomized. All test sessions were preceded by one practice session in which the identity of the consonants was indicated to the listeners.

The stimuli were presented to the subjects through a direct electrical connection using our laboratory processor at a comfortable listening level. To collect responses, a graphical interface was used that allowed the subjects to identify the words they heard by clicking on the corresponding button on the graphical interface.

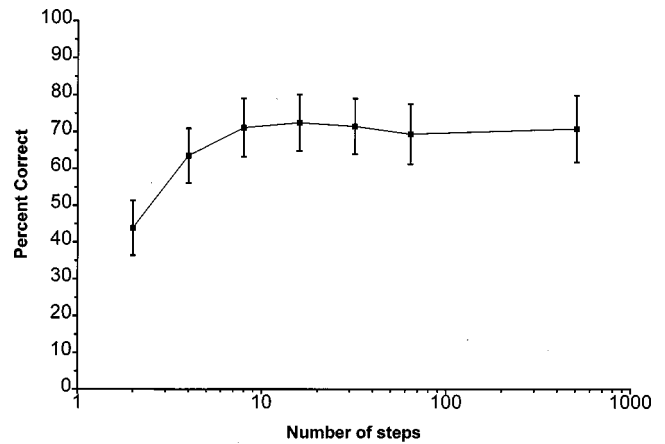


FIG. 3. Mean performance of cochlear-implant listeners on consonant recognition as a function of the number of quantized amplitude steps in the dynamic range. Error bars indicate  $\pm$  standard errors of the mean.

### B. Results and discussion

The results, scored in terms of percent correct, are shown in Fig. 3. Repeated measures analysis of variance showed a significant main effect [ $F(6,24) = 12.14$ ,  $p < 0.0005$ ] for the number of steps. *Post-hoc* analysis (Fisher's LSD—least significant difference) indicated that the asymptote in performance was obtained with eight steps, i.e., consonant recognition did not improve when eight or more steps were used.

The individual subjects' performances are shown in Fig. 4. The subjects' performances on consonant recognition varied considerably, from a low of 45% correct (subject S2) to a high of 90% correct (subject S4). For most subjects, performance improved substantially as the number of steps increased from two to eight, with only a small change in performance when more than eight steps were used. The fact that eight steps were sufficient for achieving asymptotic performance on the difficult test of consonant recognition demonstrates that a high degree of amplitude resolution is not necessary for consonant recognition. If eight steps are sufficient, then it is unlikely that intensity resolution limits many cochlear implant patients' performances on consonant recog-

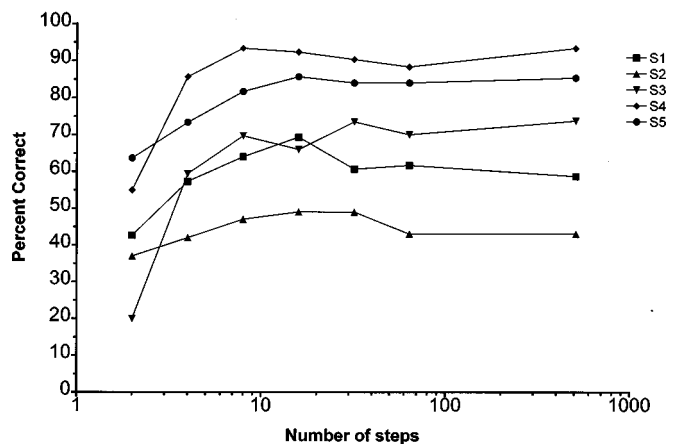


FIG. 4. Individual cochlear implant subjects' performance on consonant recognition as a function of the number of steps in the dynamic range.

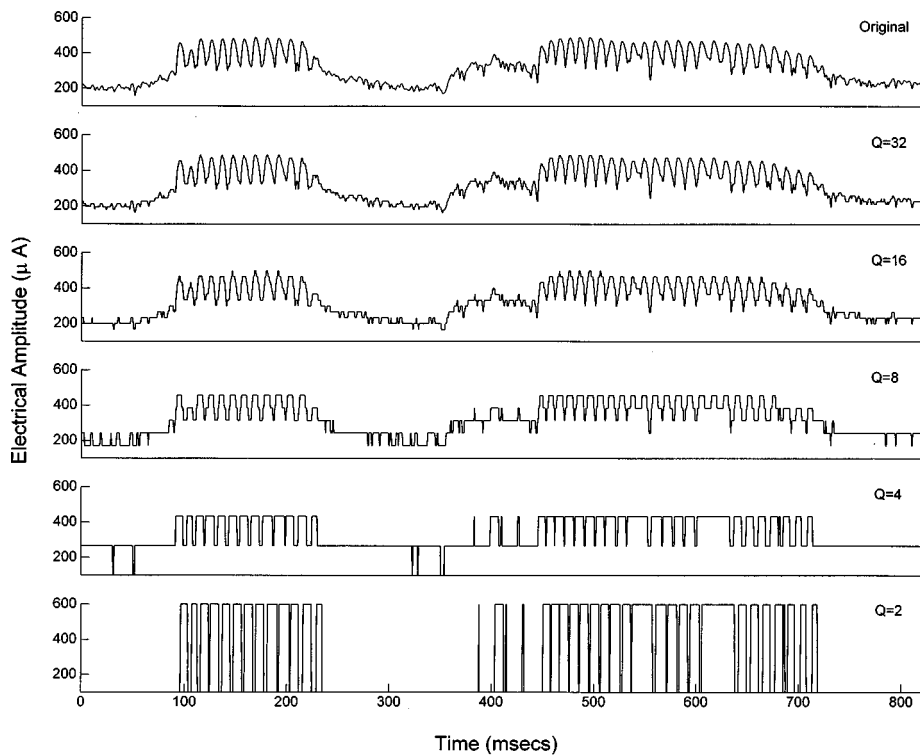


FIG. 5. Example of the waveform of the syllable /a p a/ quantized into 2, 4, 8, 16, and 32 steps. These waveforms were estimated by bandpass filtering speech into six frequency bands, computing the channel amplitudes in each band through envelope detection (400-Hz low-pass filter), and then quantizing the channel amplitudes into the indicated number of steps ( $Q = 2-32$ ). Only the (full-wave) rectified waveforms computed for channel 2 (center frequency=756 Hz) are shown in this figure.

nition since, judging from Nelson *et al.* (1996) and Zeng and Shannon (1999), most patients have at least eight discriminable steps in their dynamic range.

Wilson and colleagues (unpublished) have conducted an experiment similar in design to ours for a single Ineraid subject using a four-channel CIS processor. The subject was tested with a high rate (6900 pulses/s) and a low-rate (500 pulses/s) CIS processor. For the 24-consonant test (Tyler *et al.*, 1987), 8 steps were needed to attain asymptotic performance with the high-rate processor (consistent with our findings), and 16 steps were needed for the low-rate processor. Wilson's results suggest the possibility that the number of steps needed for consonant identification might be dependent on the stimulation rate.

Why is it that 8 steps, and not 16 or 32 steps, are sufficient for accurate consonant recognition? The answer can be found by examining closely the quantized waveforms in Fig. 5. As one might expect, quantization affects the coding of temporal-envelope cues. These cues become progressively less clear as the number of steps is reduced from 32 to 2 (Fig. 5). The envelopes obtained using two and four steps provide very little, and possibly not perceptually meaningful, information about manner of articulation. That is, the patients will most likely perceive the presence or absence of the signal, but would not be able to tell whether the signal is a consonant or a vowel, let alone identify the consonant or vowel. To visual inspection, manner cues improve significantly when the number of steps is increased to eight. With eight steps, the voiced segment (the vowel /a/, in our case) of the word /apa/ can be easily discriminated from the unvoiced segment (the closure, burst and aspiration of /p/, in our case). This information about manner of articulation contained in each channel coupled with the across-channel envelope information allowed the implant listeners to identify the con-

sonants accurately with only eight steps in the dynamic range.

The outcome that eight steps were enough for consonant recognition demonstrates that fine temporal-envelope cues are not needed for accurate consonant recognition when there are enough spectral cues. As shown in Fig. 5, the waveform that was quantized into eight steps lacks the temporal fine structure present in the original signal. Yet, the 8-step consonants were identified with roughly the same accuracy as the 512-step consonants. Drullman (1995) has also demonstrated that fine temporal cues are not needed when full spectral cues are available. He processed speech into 24  $\frac{1}{4}$ -octave channels, computed the envelopes of each channel, and modulated the envelopes with noise. The processed waveforms had the original speech envelopes but the temporal structure of noise. High speech intelligibility was obtained even when the envelopes were quantized into two levels. In his study, however, the normal-hearing listeners had more spectral information (24 channels) compared to the cochlear-implant listeners (at most 6 channels) of this study. Two steps were not enough, in our study, to maintain high consonant recognition. We suspect that a greater number (eight) of steps was needed to compensate for the lower spectral resolution. This hypothesis is investigated in experiment 2.

Contrary to the Zeng and Galvin (1999) study, we found that performance drops significantly when the number of steps is reduced to two (binary representation). We suspect that the difference in outcome is due to two factors. First, in the Zeng and Galvin study, the two output levels used in the binary condition were both above threshold, whereas in our study, only one of the two levels was above threshold. Second, the speech-processing strategies, SPEAK and CIS, that were used in the two studies were different. The SPEAK strategy uses a large number of electrodes (22) and extracts

mostly spectral peak information, which is well preserved even for the binary representation of electrical amplitudes (see Fig. 3 in Zeng and Galvin, 1999). On the other hand, the CIS strategy of this study uses fewer number (six) of electrodes and extracts primarily temporal-envelope information from the acoustic signal. The results of the two studies taken together suggest that the number of intensity steps needed for speech recognition might be dependent on the number input channels supported by the cochlear implant processor. This hypothesis is investigated in experiment 2.

Experiment 1 demonstrated that only a handful of steps are needed by cochlear implant listeners for consonant identification under quiet conditions. However, under more realistic conditions with background noise, a greater number of steps might be needed. Zeng and Galvin (1999) noted that although the dynamic range reduction degraded phoneme recognition in quiet marginally, it had a significant effect in noise. Similarly, Wilson and colleagues (unpublished) noted that more steps were required to reach asymptotic performance in noise. More studies are needed to investigate the effect of number of steps on speech understanding under noisy conditions.

It should be noted that the amplitude quantization steps ( $\Delta i$ ) of this study do not represent the discriminable steps measured by Nelson *et al.* (1996) using psychophysical methods. For some subjects in the Nelson *et al.* study, the amplitude DLs were not constant across the dynamic range but decreased as a function of the dynamic range, i.e., the step sizes were smaller in the upper dynamic range and relatively larger in the mid-to-low dynamic range. In our study, the step sizes ( $\Delta i$ ) were uniform across the dynamic range, much like the amplitude DLs in two of the subjects in the Nelson *et al.* study. It is important to note, however, that the dynamic range was defined differently in the two studies. In our case, the dynamic range was defined from threshold to most-comfortable level (MCL), as commonly implemented in commercial speech processors, while in the Nelson *et al.* study it was defined from threshold to maximum acceptable loudness (MAL) level. If we exclude the upper dynamic range, from MCL to MAL, where the step sizes are smaller, we are left with the mid-to-low dynamic range where the step sizes are more uniform. On this view, the uniform quantization steps used in this study are a reasonable approximation to the uniform perceptual steps found in the low- to mid-dynamic range of implant users.

### III. EXPERIMENT 2: ACOUSTIC HEARING

In experiment 1, we found that eight intensity steps within the dynamic range were enough for asymptotic consonant recognition by implant listeners who were using a six-channel CIS processor. In this experiment, we investigate whether this outcome holds when speech is processed through a larger (or smaller) number of channels. We hypothesize that there is a trade-off between spectral resolution (number of spectral channels) available and the intensity resolution (number of steps) needed. This hypothesis was motivated by our previous work on the need for accurate intensity resolution for processors using a small number of channels (Dorman *et al.*, 1997; Loizou *et al.*, 1998).

To produce speech with varying degrees of spectral resolution, speech was filtered through 4–20 frequency bands, and synthesized as a linear combination of sine waves with amplitudes extracted from the envelopes of the band-passed waveforms, and frequencies equal to the center frequencies of the bandpass filters. To produce speech with varying degrees of intensity resolution, we quantized the amplitudes of the sinewaves to 2–32 levels. The intelligibility of vowels, consonants, and sentences was assessed as a function of spectral resolution and as a function of intensity resolution in normal-hearing listeners.

## A. Method

### 1. Subjects

Nine graduate students from Arizona State University served as subjects. All of the subjects were native speakers of American English and had normal hearing. The subjects were paid for their participation.

### 2. Speech material

The test material included sentences, consonants in /aCa/ context and vowels in /hVd/ context. The consonant test was a subset of the Iowa consonant test (Tyler *et al.*, 1987) and consisted of 16 consonants in /aCa/ environment spoken by a single male speaker. Five repetitions of each consonant were used in a blocked and randomized test sequence. The vowel material consisted of the vowels in the words “heed, hid, hayed, head, had, hod, hud, hood, hoed, who’d, heard.” Each word was produced by three women and three girls. The stimuli were drawn from a set used by Hillenbrand *et al.* (1995). All the stimuli were presented in a completely randomized test sequence.

The sentence material was from the TIMIT database (Garofolo *et al.*, 1993). A different set of 15 sentences was used for each condition. A total of 540 sentences were randomly selected from the TIMIT database from the DR3 (north midland) dialect region. The sentences were produced by an equal number of female and male speakers—one sentence per speaker. The 540 sentences were divided into 36 lists (for 6 channel conditions  $\times$  6 quantization conditions), with 15 sentences in each list. Fifteen sentences were used for the first channel condition, 15 different sentences were used for the second channel condition, etc. There were eight sentences spoken by eight different male speakers and seven sentences spoken by seven different female speakers within each list. Each sentence contained, on the average, 7 words, and the 15 sentences in each list contained, on the average, a total of 100 words.

### 3. Amplitude quantization

Critical to the quantization process is the determination of the quantization step sizes, which are, in turn, dependent on the amplitude dynamic range. The dynamic range itself is a function of the frequency band (e.g., Boothroyd *et al.*, 1994). We therefore determined the quantization step sizes for each band separately.

We filtered the speech material (vowels, consonants, and sentences) through  $n$  (4–16) bandpass filters, and estimated

the envelopes of the bandpassed waveforms in each frequency band. We then determined the envelope-amplitude dynamic range (i.e., the difference between the maximum and minimum amplitudes) of each channel by computing the envelope histograms of the speech material. The maximum envelope amplitude in each channel, denoted as  $X_{\max}^i$  where  $i$  is the channel number, was chosen to include 99% of all amplitude counts in that channel. The minimum envelope amplitude ( $X_{\min}^i$ ) was set 0.5 dB above the rms value of the noise floor. The  $X_{\max}^i$  and  $X_{\min}^i$  values were then used to estimate the quantization step size,  $\Delta_i$ , of each channel as follows:

$$\Delta_i = \frac{X_{\max}^i - X_{\min}^i}{Q - 1}, \quad i = 1, 2, \dots, n,$$

where  $Q$  is the number of quantization levels or steps, and  $n$  is the number of channels. Note that each channel had a different value for  $X_{\max}^i$  and  $X_{\min}^i$  since the envelope dynamic range of each channel was different. Consequently, the step sizes  $\Delta_i$  were different in each channel. The  $X_{\max}^i$  and  $X_{\min}^i$  values are analogous to the most-comfortable (MCL) and threshold values in cochlear implants.

#### 4. Speech synthesis

After estimating the quantization step sizes for each frequency band, we processed the speech material as follows. Signals were first processed through a preemphasis filter (1200-Hz cutoff), with a 3-dB/oct roll-off, and then bandpassed into  $n$  frequency bands ( $n = 4, 6, 8, 12, 16$ ) using sixth-order Butterworth filters. Logarithmic filter spacing was used for  $n < 8$  and mel spacing was used for  $n \geq 8$ . The center frequencies and the 3-dB bandwidths of the filters can be found in Loizou *et al.* (1999). The envelopes of the signal were extracted by full-wave rectification, and low-pass filtering (second-order Butterworth) with a 400-Hz cutoff frequency. The envelope amplitudes were estimated by computing the root mean-square (rms) energy of the envelopes every 4 ms. The envelope amplitudes were then uniformly quantized to  $Q$  discrete levels ( $Q = 2, 4, 8, 16, 32$ ). Sine waves were generated with amplitudes equal to the quantized envelope amplitudes and with frequencies equal to the center frequencies of the bandpass filters. The phases of the sinusoids were estimated from the fast Fourier transform (FFT) of the speech segment (McAulay and Quatieri, 1986). The sinusoids of each band were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment. We found the normalization of the synthesized speech segments to be absolutely necessary to maintain high levels of speech intelligibility. Without this normalization the synthesized consonant segments could have the same rms levels as the vowel segments, thereby affecting the consonant-to-vowel (CV) ratios. Pilot experiments showed that this normalization can have a dramatic effect on performance for the two- and four-step conditions. Therefore, to avoid having improper vowel-to-consonant ratios as a factor that might potentially confound

our results, we decided to normalize the synthesized speech segments in order to preserve the original (natural) consonant-to-vowel ratios.

We also processed the speech material through a simulation of the SPEAK strategy used in the Nucleus-22 device (McDermott *et al.*, 1992; Loizou, 1998). We did that so that we could compare our results with those reported by Zeng and Galvin (1999). The signal was processed into 20 channels the same way as before, except that only the six (out of 20) highest channel amplitudes were used for synthesis in each 4-ms segment. We adopted the filter spacing used in the SPEAK strategy (Seligman and McDermott, 1995). Note that this signal processing strategy is sometimes referred to as the 6-of-20 strategy, since only 6 out of the 20 amplitudes are used for stimulation in each cycle.

In addition to the quantized speech material, we also processed speech as described above but without quantizing the envelope amplitudes. We used this condition for comparative reasons and refer to it as the ‘‘unquantized’’ condition.

#### 5. Procedure

The experiment was performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. The subjects listened to the speech material via closed ear-cushion headphones at a comfortable level set by the subject. A graphical interface was used that allowed the subjects to select the vowel or consonant they heard using a mouse. For the sentence material, subjects were asked to type in as many words as they could understand.

Before each condition, subjects were given a practice session with examples of speech (vowels, consonants, or sentences) processed through the same number of channels and the same number of steps in that condition. None of the sentences used in the practice was used in the test. A sequential test order, starting with speech material processed through a large number of channels ( $n = 16$ ) and continuing to speech material processed through a small number of channels ( $n = 4$ ), was employed. We chose this sequential test design to give the subjects time to adapt to listening to the altered speech signals. The test order for the different number of steps in each channel condition was counterbalanced between subjects.

#### B. Results

The results, scored in percent correct for vowels and consonants and in percent words correct for sentences, are shown in Fig. 6. A two-factor (channels and steps) repeated measures analysis of variance (ANOVA) on the sentence data showed a significant main effect of number of channels [ $F(5,45) = 371.6, p < 0.0005$ ], a significant effect of number of steps [ $F(5,45) = 586.9, p < 0.0005$ ], and a significant interaction between number of channels and number of steps [ $F(25,225) = 34.1, p < 0.0005$ ]. *Post hoc* analysis, according to Fisher’s LSD, showed that asymptotic performance for the 4-channel condition was reached with 16 steps, for the 6-, 8- and 12-channel conditions with 8 steps, and for the 16-channel and 6-of-20 channel conditions with 4 steps.

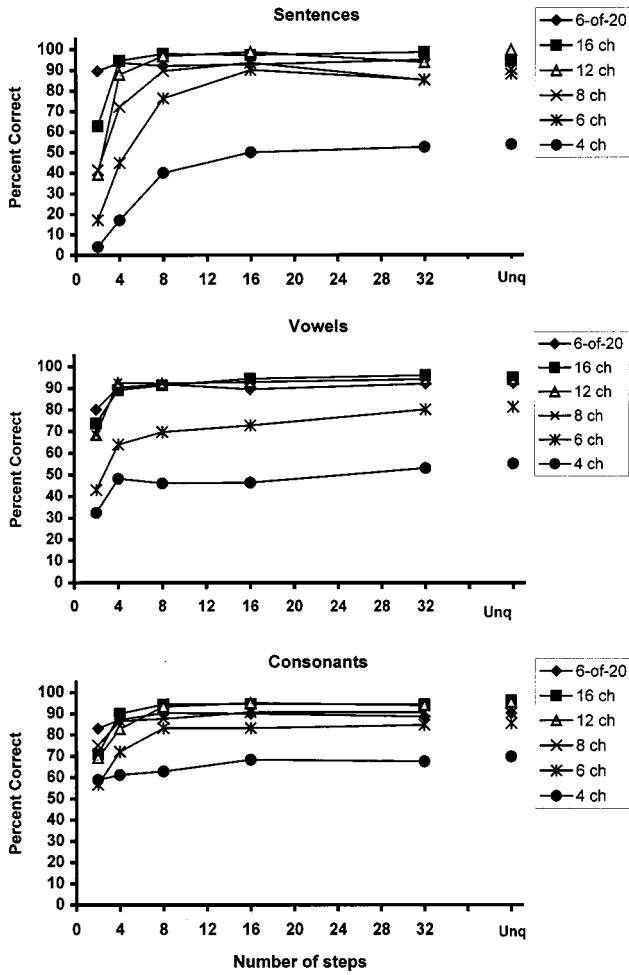


FIG. 6. Performance of normal-hearing listeners on sentence, consonant, and vowel recognition as a function of the number of channels and as a function of the number of steps. In the “Unq” condition, the speech materials were left unquantized.

A two-factor (channels and steps) repeated measures ANOVA on the vowel data showed a significant main effect of number of channels [ $F(5,45) = 100.3, p < 0.0005$ ], a significant effect of number of steps [ $F(5,45) = 82.1, p < 0.0005$ ], and a significant interaction between number of channels and number of steps [ $F(25,225) = 4.5, p < 0.0005$ ]. *Post hoc* analysis, according to Fisher’s LSD, showed that asymptotic performance for all channel conditions, except for the six-channel condition, was reached with four steps. For the six-channel condition, the vowel performance with 8 steps did not differ significantly ( $p > 0.05$ ) from the 4-, 16-, 32-step or the unquantized conditions.

Similar repeated measures ANOVA on the consonant data showed a significant main effect of number of channels [ $F(5,45) = 29.4, p < 0.0005$ ], a significant effect of number of steps [ $F(5,45) = 112.3, p < 0.0005$ ], and a significant interaction between number of channels and number of steps [ $F(25,225) = 8.3, p < 0.0005$ ]. *Post hoc* analysis, according to Fisher’s LSD, showed that asymptotic performance for the 6-, 12-, and 16-channel conditions was reached with eight steps. Consonant performance for the four-channel condition was low and uniform across the number of steps used. The score obtained with 2 steps did not differ significantly ( $p$

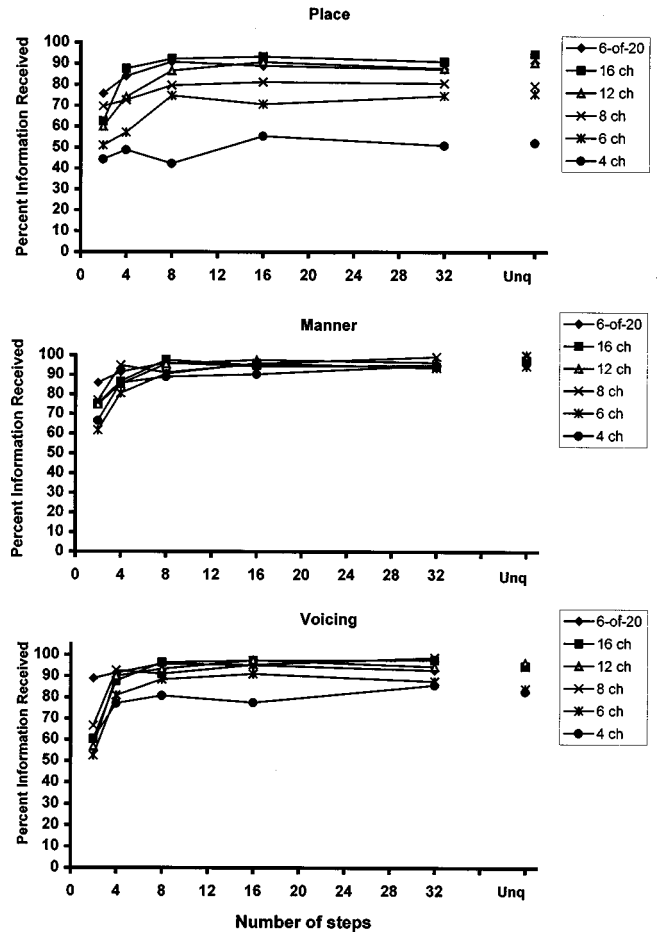


FIG. 7. Percent information received for the features “place,” “manner,” and “voicing” as a function of the number of channels and as a function of the number of steps.

$= 0.068$ ) with the 32-step and unquantized conditions. Asymptotic performance for the 8-channel and 6-of-20 channel conditions was reached with four steps.

The results for feature analysis (Miller and Nicely, 1955) for the consonants are shown in Fig. 7. A two-factor (channels and steps) repeated measures ANOVA performed on the three features (place, manner, and voicing) separately indicated a significant main effect of number of channels ( $p < 0.005$ ), a significant effect of number of steps ( $p < 0.005$ ) and a significant interaction ( $p < 0.005$ ) between number of channels and number of steps. For the feature “place of articulation,” *post hoc* analysis, according to Fisher’s LSD, showed that asymptotic performance for the 6-, 8-, and 12-channel conditions was reached with eight steps, while for the 16-channel and 6-of-20 conditions it was reached with four steps. Performance for the four-channel condition was flat across the different numbers of steps used, in line with the percent-correct results. For the feature “manner of articulation,” *post hoc* analysis (Fisher’s LSD) showed that asymptotic performance for the 6-, 12-, and 16-channel condition was reached with eight steps, while for the 4- and 8-channel conditions it was reached with four steps. Performance was flat for the 6-of-20 condition, with no statistically significant differences between the scores obtained with 2 and 32 steps. For the feature “voicing,” *post hoc*

analysis (Fisher's LSD) showed that asymptotic performance for the 6-, 8-, 12-, and 16-channel conditions was reached with only four steps. Performance was flat for the 4-channel and 6-of-20 conditions.

### C. Discussion

The statistical analysis indicated, consistent with our original hypothesis, a significant interaction between spectral resolution and intensity resolution for all test materials. The minimal number of steps needed to reach asymptotic performance was a function of the number of channels available and the speech material used. For the sentence test, the number of steps needed to obtain high levels of intelligibility was found to be inversely proportional to the number of channels available. Sixteen steps were needed to reach asymptotic performance for speech processed through four channels. In contrast, four steps were sufficient for speech processed through 16 or more channels. High sentence intelligibility scores (90%) were obtained even with two steps when speech was processed through the 6-of-20 strategy. The sentence results confirm our original hypothesis about the tradeoff relationship between spectral resolution and intensity resolution. When the spectral resolution is poor (i.e., a small number of channels is available) relatively fine intensity resolution is needed to achieve a high level of speech understanding. Conversely, when the spectral resolution is good (i.e., a large number of channels is available) fine intensity resolution is not needed to achieve high levels of understanding.

A different pattern of performance was obtained for vowels. Asymptotic performance was obtained with four steps in all channel conditions except for  $n=6$ . Vowel performance seemed to be more dependent on the number of channels available, which was not surprising since fine spectral resolution is needed for accurate vowel identification. High vowel identification scores (>90%) were obtained with eight or more channels, in agreement with our previous findings (Dorman *et al.*, 1997). The vowel performance obtained with eight or more channels and four steps was not significantly different than the vowel performance obtained in the unquantized condition.

Compared to vowels, a greater number of steps was needed to identify consonants. Eight steps were needed for most channel conditions to reach asymptotic performance. Eight steps were needed for consonants processed through six channels, consistent with our findings with cochlear implant listeners in experiment 1. The finding that more steps are needed to identify consonants than vowels is not surprising, since the quantization affects the coding of temporal-envelope cues, which are more important for consonant identification than vowel identification. Feature analysis also indicated that eight steps are needed for the features "place" and "manner" to reach asymptotic performance in most channel conditions, while only four steps are needed for the feature "voicing." Overall, eight steps are needed to reliably transmit information about place, manner, and voicing, consistent with the number of steps needed to reach asymptotic performance on consonant recognition.

The data from the Zeng and Galvin (1999) study are entirely consistent with the data from normal-hearing listeners (experiment 2) if we assume that the Nucleus-22 implant listeners have 20 functional channels. However, several experiments have shown that this is not the case. Fishman *et al.* (1997) and Wilson (1997) have demonstrated that the performance of N22 implant listeners reaches asymptote when only four to seven channels are activated. This also seems to be the case with the N22 implant listeners in the Zeng and Galvin study. For vowels, the N22 patients' level of performance at asymptote (20 electrodes activated and full dynamic range) was approximately 78% correct, and for consonants it was approximately 70% correct, which is equivalent to the performance of normal-hearing listeners using four to six channels of stimulation (Fig. 6). Furthermore, increasing the number of electrodes activated from 4 to 20 improved the consonant score by only 10% (50% versus 60% correct), whereas in our study (Fig. 6) the mean consonant score improved by approximately 40% (60% versus 100% correct) when the number of channels increased from 4 to 20. This suggests that, most likely, the N22 patients were not using 20 channels of information when all 20 electrodes were activated, but rather 4 to 6 channels. Thus there is converging evidence that (i) the N22 patients receive the equivalent of 4 to 6 channels of information, and (ii) the N22 patients are not performing at the level of a 6-of-20 processor. If we then assume that the N22 patients were using only four channels, then the pattern of performance of our vowel and consonant data (Fig. 6) falls in line with the data in the Zeng and Galvin study. With four channels, there is hardly any difference in vowel or consonant performance as we increase the number of steps from 2 to 32 (Fig. 6), much like the data in the Zeng and Galvin study on vowel and consonant recognition.

It is clear from our studies on dynamic range reduction and number of channels that degrading amplitude information and degrading frequency information have different effects on sentences, as compared to consonants and vowels. Degrading frequency information (i.e., reducing number of channels) has a large effect on consonants and vowels but a smaller effect on sentences (Dorman *et al.*, 1997; Shannon *et al.*, 1995). In contrast, degrading amplitude information has a large effect on sentences but a smaller effect on consonants and vowels. The latter outcome may be the result of the loss of amplitude-based word and syllabic segmentation cues in the time waveform (see Zue, 1985).

### IV. CONCLUSIONS

- (1) For five cochlear implant listeners using a six-channel CIS processor, eight quantized amplitude steps were sufficient for achieving asymptotic levels of consonant recognition in quiet. Eight steps were sufficient for both better performing and poorly performing patients. These results suggest that amplitude resolution is not a major factor in determining consonant identification.
- (2) The outcome that eight steps were enough for consonant recognition provides yet another demonstration (e.g., Drullman, 1995) that fine temporal-envelope cues are not needed for accurate consonant recognition when

there are enough spectral cues (a maximum of six channels, in our case, for the implant listeners).

- (3) The implication of these findings for the design of cochlear implant processors is that the compression function, which is used to map the acoustic input to electrical output, need not be defined with high resolution. Consequently, the size of the compression tables stored in processor memory could be greatly reduced.
- (4) For normal-hearing listeners listening to speech processed through 4–20 channels and channel amplitudes quantized to 2–32 steps, the number of steps needed to achieve asymptotic performance was a function of the number of channels and the speech material used.
- (5) For vowels, asymptotic performance was obtained with four steps and was independent of the number of channels. For consonants, eight steps were needed for most channel conditions to reach asymptotic performance. This is consistent with our findings with cochlear implant listeners in experiment 1.
- (6) For sentences 16 steps were needed to reach asymptotic performance with 4 channels, while 4 steps were needed for sentences processed through 16 or more channels. High sentence intelligibility scores (90%) were obtained even with two steps when speech was processed through a 6-of-20 strategy (SPEAK type). These results reaffirm our hypothesis that there is an inverse relationship between spectral resolution and intensity resolution. When spectral resolution is poor (i.e., a small number of channels is available) relatively fine intensity resolution is needed to achieve high levels of understanding. Conversely, when the intensity resolution is poor, a high degree of spectral resolution is needed to achieve asymptotic performance.
- (7) The results of this study, taken together with previous findings (Zeng and Galvin, 1999; Loizou *et al.*, 2000) on the effect of reduced dynamic range, suggest that the performance of cochlear implant subjects is primarily limited by the small number (four to six) of channels received, and not by the small number of steps or reduced dynamic range. Therefore, more research is needed to find ways to increase the number of channels transmitted by cochlear implants.

## ACKNOWLEDGMENTS

The authors would like to thank Fan-Gang Zeng and the anonymous reviewers for providing valuable suggestions. This research was supported by Grant Nos. R01 DC03421 and R01 DC00654 from the National Institute of Deafness and other Communication Disorders, NIH.

<sup>1</sup>The range in number of discriminable steps (8–45 steps) estimated in the Nelson *et al.* (1996) study is actually an overestimate of the true number of steps that an implant listener might have with a commercial speech processor. This is because Nelson *et al.* calculated the number of DLs over the dynamic range between threshold and maximum acceptable loudness (MAL) levels. In commercial speech processors, however, speech is typically mapped between threshold and most comfortable level (MCL), which is smaller than the MAL level. And, since the smallest DLs were found between MCL and MAL levels, the number of discriminable steps associated with dynamic range in speech processors may be significantly smaller than those estimated by Nelson *et al.* (1996).

<sup>2</sup>The power exponent  $p$  was set equal to  $-0.0001$  to match the logarithmic compression function, of the form  $y = A \log(1 + cx) + B$ , used in the Med-El device, where  $A$  and  $B$  are constants used for mapping the acoustic signal  $x$  between threshold and most comfortable level, and  $c$  is a constant that controls the shape of the compression function ( $c = 512$ , in our case). It should be noted that the value of the power exponent  $p$  corresponding to logarithmic mapping depends on the input range, and, in particular, on the minimum value of  $x$  ( $X_{\min}$ ). If  $X_{\min} = 0$ , then the value of  $p = 0.2$  will yield a logarithmic mapping similar to the one used in the Med-El device, whereas if  $X_{\min} = 1$ , then the value of  $p = -0.0001$  will yield a logarithmic mapping. In our implementation,  $X_{\min} = 1$ .

- Boothroyd, A., Erikson, F., and Medwetsky, L. (1994). "The hearing aid input: A phonemic approach to assessing the spectral distribution of speech," *Ear Hear.* **15**, 432–442.
- Dillier, N., Spillman, T., and Guntensperger, J. (1983). "Computerized testing of signal encoding strategies with round window implants, in *Cochlear Prostheses: An International Symposium*, edited by C. W. Parkins and S. W. Anderson, Ann. N.Y. Acad. Sci. **405**, 360–369.
- Dorman, M., Loizou, P., and 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," *J. Acoust. Soc. Am.* **102**, 2403–2411.
- Drullman, R. (1995). "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Am.* **97**, 585–592.
- Fishman, K., Shannon, R., and Slattery, W. (1997). "Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant processor," *J. Speech Hear. Res.* **40**(5), 1201–1215.
- Francois, J., Tinembart, J., Bessat, C., Leone, P., Rossman, F., and Pelizzone, M. (1994). "Implants cochleaires: Un processeur portable pour le developpement de l'algorithme CIS," *Actes de la conference DSP 94*, Paris.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1993). "DARPA TIMIT: Acoustic-phonetic continuous speech corpus," *NIST Technical Report* (distributed with the TIMIT CD-ROM).
- Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.* **97**, 3099–3111.
- Hochmair-Desoyer, I., Hochmair, E., and Fischer, R. (1981). "Four years of experience with cochlear prostheses," *Med. Prog. Technol.* **8**, 107–119.
- Loizou, P. (1998). "Mimicking the human ear: An overview of signal processing techniques for converting sound to electrical signals in cochlear implants," *IEEE Signal Process. Mag.* **15**(5), 101–130.
- Loizou, P., Dorman, M., and Fitzke, J. (2000). "The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants," *Ear Hear.* **21**(1), 25–31.
- Loizou, P., Dorman, M., and Powell, V. (1998). "The recognition of vowels produced by men, women, boys and girls by cochlear implant patients using a six-channel CIS processor," *J. Acoust. Soc. Am.* **103**, 1141–1149.
- Loizou, P., Dorman, M., and Tu, Z. (1999). "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.* **106**, 2097–2103.
- McAulay, R., and Quatieri, T. (1986). "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.* **ASSP-34**(4), 744–754.
- McDermott, H., McKay, C., and Vandali, A. (1992). "A new portable sound processor for the University of Melbourne/Nucleus Limited multi-electrode cochlear implant," *J. Acoust. Soc. Am.* **91**, 3367–3371.
- Miller, G., and Nicely, P. (1955). "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.* **27**, 338–352.
- Nelson, D., Schmitz, J., Donaldson, G., Viemeister, N., and Javel, E. (1996). "Intensity discrimination as a function of stimulus level with electric stimulation," *J. Acoust. Soc. Am.* **100**, 2393–2414.
- Pfingst, B., Burnett, P., and Sutton, D. (1983). "Intensity discrimination with cochlear implants," *J. Acoust. Soc. Am.* **73**, 1283–1292.
- Poroy, O., and Loizou, P. (2000). "Development of a speech processor for laboratory experiments with cochlear implant patients," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey.
- Schroder, A., Viemeister, N., and Nelson, D. (1994). "Intensity discrimination in normal-hearing and hearing-impaired listeners," *J. Acoust. Soc. Am.* **96**, 2683–2693.
- Seligman, P., and McDermott, H. (1995). "Architecture of the Spectra 22 speech processor," *Ann. Otol. Rhinol. Laryngol. Suppl.* **166**, 139–141.

- Shannon, R. (1983). "Multichannel electrical stimulation of the auditory nerve in man. I. Basic Psychophysics," *Hear. Res.* **11**, 157–189.
- Shannon, R., Jansvold, A., Padilla, M., Robert, M., and Wang, X. (1999). "Consonant recordings for speech testing," *J. Acoust. Soc. Am.* **106**, L71–L74.
- Shannon, R., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Tyler, R., Preece, J., and Lowder, M. (1987). The Iowa audiovisual speech perception laser videodisc, *Laser Videodisc and Laboratory Report*, Dept. of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City.
- Wilson, B. (1997). "The future of cochlear implants," *Br. J. Audiol.* **31**, 205–225.
- Wilson, B., Finley, C., Lawson, D., Wolford, R., Eddington, D., and Rabinowitz, W. (1991). "Better speech recognition with cochlear implants," *Nature (London)* **352**, 236–238.
- Zeng, F-G., and Galvin, J. (1999). "Amplitude mapping and phoneme recognition in cochlear implant listeners," *Ear Hear.* **20**, 60–74.
- Zeng, F-G., and Shannon, R. (1999). "Psychophysical laws revealed by electric hearing," *NeuroReport* **10**, 1931–1935.
- Zue, V. (1985). "The use of knowledge in automatic speech recognition," *Proc. IEEE* **73**, 1602–1615.