On the number of channels needed to understand speech

Philipos C. Loizou[a)]
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

Zhemin Tu
Department of Applied Science, University of Arkansas at Little Rock, Little Rock, Arkansas 72204-1099

(Received 5 December 1998; revised 7 April 1999; accepted 21 May 1999)

Recent studies have shown that high levels of speech understanding could be achieved when the speech spectrum was divided into four channels and then reconstructed as a sum of four noise bands or sine waves with frequencies equal to the center frequencies of the channels. In these studies speech understanding was assessed using sentences produced by a single male talker. The aim of experiment 1 was to assess the number of channels necessary for a high level of speech understanding when sentences were produced by multiple talkers. In experiment 1, sentences produced by 135 different talkers were processed through n ($2 \leqslant n \leqslant 16$) number of channels, synthesized as a sum of n sine waves with frequencies equal to the center frequencies of the filters, and presented to normal-hearing listeners for identification. A minimum of five channels was needed to achieve a high level (90%) of speech understanding. Asymptotic performance was achieved with eight channels, at least for the speech material used in this study. The outcome of experiment 1 demonstrated that the number of channels needed to reach asymptotic performance varies as a function of the recognition task and/or need for listeners to attend to fine phonetic detail. In experiment 2, sentences were processed through 6 and 16 channels and quantized into a small number of steps. The purpose of this experiment was to investigate whether listeners use across-channel differences in amplitude to code frequency information, particularly when speech is processed through a small number of channels. For sentences processed through six channels there was a significant reduction in speech understanding when the spectral amplitudes were quantized into a small number (<8) of steps. High levels (92%) of speech understanding were maintained for sentences processed through 16 channels and quantized into only 2 steps. The findings of experiment 2 suggest an inverse relationship between the importance of spectral amplitude resolution (number of steps) and spectral resolution (number of channels). [S0001-4966(99)01810-X]

PACS numbers: 43.72.Ar, 43.71.Es [JMH]

INTRODUCTION

Dudley (1939) provided one of the earliest demonstrations that speech understanding does not require a highly detailed spectral representation of the speech signal. After bandpass filtering the speech signal into ten spectral bands, Dudley (1939) estimated the envelopes of the bandpassed waveforms using rectification and low-pass filtering (20-Hz cutoff). Speech was synthesized by filtering an excitation signal (either buzz or hiss) through the same bandpass filters, and amplitude modulating the outputs of the filters by the envelopes of the bandpassed waveforms. The resulting speech was highly intelligible. Dudley (1939) concluded that much of the information in the speech spectrum is redundant. The channel vocoder approach, pioneered by Dudley, was later exploited for efficient transmission of speech over telephone channels (see review by Schroeder, 1966; Flanagan, 1972).

In the 1950s, researchers at Haskins Laboratories used a 50-component sine wave synthesizer to investigate the minimal cues necessary for the recognition of speech. Investigators showed that speech could be recognized with a high degree of accuracy when sine waves specifying only the first two or three formants of the signal were presented (e.g., Delattre *et al.*, 1952). In these experiments as few as four or six sine wave components (out of 50) were sufficient to create intelligible speech, if the sine wave components specified harmonics at or near the formant frequencies of the signal. Remez *et al.* (1981), elaborating on earlier work on syllable recognition by Cutting (1974) and Bailey *et al.* (1976), carried the minimal cues approach to one extreme by replacing the rich harmonic structure of speech with only three sine waves at the formant frequencies of the consonants and vowels in the words of sentences. Most listeners were able to identify the words with high accuracy.

The aforementioned studies, and many others (e.g., Hill *et al.*, 1968), provide overwhelming evidence that speech recognition does not require the fine spectral detail present in naturally produced utterances. This fortunate circumstance has proved essential in restoring speech understanding to deaf individuals fitted with cochlear implants, because it is not currently possible to provide fine spectral detail to im-

a)Electronic mail: loizou@utdallas.edu

plant patients. However, in the context of signal processing for cochlear implants, it is still unclear as to how little or how much spectral detail is necessary to allow speech understanding at a high level.

In the work cited above, high levels of speech understanding were obtained if signals were filtered into a reasonably large number of frequency bands and/or a small number of sine waves were output at or near the formant frequencies. Such a strategy is implemented in one of the two current signal processing strategies used for cochlear implants (McDermott *et al.*, 1992; Loizou, 1998). The other signal processing strategy used for cochlear implants divides the speech spectrum into a small number of bands, 4 to 12 depending on the device, and, instead of picking high amplitude channels, transmits the energy in all of the bands. This strategy is the focus of this article. At issue is how many channels of stimulation are necessary to achieve a high level of speech understanding in quiet.

Shannon *et al.* (1995) showed that high levels of speech understanding (e.g., 90% correct for sentences) could be achieved using as few as four spectral bands. In Shannon *et al.* (1995) envelopes of the speech signal were extracted from a small number (1–4) of frequency bands, and used to modulate noise of the same bandwidth. The noise-modulated bands preserved the temporal cues within each band but eliminated the spectral details within each band. Dorman *et al.* (1997) synthesized speech as a sum of a small number of sine waves rather than noise bands. As in Shannon *et al.* (1995), sentence recognition using four channels was found to be 90% correct.

In Shannon *et al.* (1995) and Dorman *et al.* (1997) speech understanding was assessed using sentences produced by a single male speaker. It is very likely that the use of a single speaker overestimates the speech perception abilities of listeners in real-world situations because the use of a single speaker eliminates the need for listeners to accommodate to variability in the acoustic signal (e.g., Mullenix *et al.*, 1989; Sommers *et al.*, 1997). Variability in the acoustic signal arises from differences in the size and shape of vocal tracts, differences in phonetic realization (e.g., pronunciation), and differences in speaking rate. The aim of experiment 1 was to determine the number of channels of stimulation necessary to allow a high level of sentence understanding when speech was produced by 135 talkers, half of whom were female.

The aim of experiment 2 was to assess the intelligibility of speech processed through 6 and 16 channels and quantized into a small number of steps. The purpose of this experiment was to assess the importance of amplitude resolution for speech understanding when signals are processed into a relatively small, and a relatively large, number of channels. Our hypothesis was that a relatively high degree of amplitude resolution is a necessary condition for speech understanding when signals are processed into a small number of channels because, with a small number of channels, listeners must use differences in signal levels across channels to infer the location of formant frequencies (Dorman *et al.*, 1997; Loizou *et al.*, 1998). In contrast, when speech is processed into a large number of channels, a high level of spectral amplitude resolution is not necessary because the location of frequencies in the input spectrum are well specified by the channels which contain energy. The outcome of experiment 2 is of interest because a recent experiment by Nelson *et al.* (1996) with cochlear implant subjects showed that the total number of discriminable intensity steps varied from a low of 6 to a high of 45. If a high degree of amplitude resolution is necessary for frequency analysis when speech is processed into a small number of channels, then it is possible that speech perception in some cochlear implant subjects is constrained by a limited ability to resolve differences in signal level across channels.

I. EXPERIMENT 1

A. Method

1. Subjects

Nine graduate students from the Applied Science Department, UALR, 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. Sentence material

The multi-talker TIMIT database (Garofolo *et al.*, 1993) was used for testing. The TIMIT database contains speech from 630 speakers, representing 8 major dialect divisions of American English, each speaking 10 phonetically rich sentences. Some of the sentences were designed to provide a good coverage of pairs of phones with extra occurrences of difficult phonetic contexts and some of the sentences were designed to maximize the variety of allophonic contexts (Lamel *et al.*, 1986).

A total of 135 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 135 sentences were divided into 9 lists (1 list per channel condition), 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. Each subject listened to a total of 135 sentences ($=15$ sentences/condition$\times 9$ channel conditions).

3. Signal processing

Signals were first processed through a pre-emphasis filter (2000-Hz cutoff), with a 3-dB/octave rolloff, and then bandpassed into n frequency bands ($n=2,3,4,5,6,8,10,12,16$) using sixth-order Butterworth filters. Logarithmic filter spacing was used for $n<8$ and mel spacing[1] was used for $n\geqslant 8$. Logarithmic and semi-logarithmic (mel) filter spacing was used because: (1) the filter bandwidths can be computed systematically; and (2) it is the type of filter spacing used in current cochlear implant devices (e.g., Zierhofer *et al.*, 1994; Loizou, 1998).

TABLE I. The center frequencies (Hz) of the filters.

No. of Channels	Channel 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	792	3392														
3	545	1438	3793													
4	460	953	1971	4078												
5	418	748	1339	2396	4287											
6	393	639	1037	1685	2736	4444										
8	394	692	1064	1528	2109	2834	3740	4871								
10	322	546	814	1137	1524	1988	2545	3213	4014	4976						
12	274	453	662	905	1190	1521	1908	2359	2885	3499	4215	5050				
16	216	343	486	647	828	1031	1260	1518	1808	2134	2501	2914	3378	3901	4489	5150

The center frequencies and the 3-dB bandwidths of the filters are given in Tables I and II, respectively. The envelope of the signal was extracted by full-wave rectification, and low-pass filtering (second-order Butterworth) with a 400-Hz cutoff frequency. Sinusoids were generated with amplitudes equal to the root-mean-square (rms) energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The phases of the sinusoids were estimated from the FFT of the speech segment[2] (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.

4. Procedure

The experiment was performed on a PC equipped with a Creative Labs SoundBlaster 16 soundcard. The subjects listened to the sentences via closed ear-cushion headphones at a comfortable level set by the subject. A graphical interface was used that allowed the subjects to type the words they heard. After listening to each sentence, subjects were asked to type in as many words as they could understand.

Before each channel condition, subjects were given a practice session with examples of ten sentences processed through the same number of channels in that condition. None of the sentences used in the practice was used in the test. A sequential test order, starting with sentences processed through a large number of channels ($n=16$) and continuing to sentences processed through a small number of channels ($n=2$), was employed. We chose this sequential test design to give the subjects time to adapt to listening to the altered speech signals. There is no doubt a ''warm-up'' effect when listening to sine wave speech of any kind.

B. Results and discussion

The subject's responses were scored as percentage of words correct. The results are shown in Fig. 1. A repeated measures analysis of variance indicated a main effect $[F(8,64)=261.94, p<0.0001]$ for number of channels. *Post hoc* tests according to Scheffe showed no statistically significant differences in scores when the number of channels was increased beyond eight. There was a significant difference ($p=0.001$) between the scores obtained with six and eight channels. There was no significant difference between the scores obtained with five and six channels. Speech recognition performance with four channels was 63% correct. This score was significantly lower than the score (90%) reported by Shannon *et al.* (1995) and the score (90%) reported by Dorman *et al.* (1997) using sentences from the H.I.N.T. database produced by a single male talker. This outcome, as well as others, demonstrates that the number of channels necessary to reach asymptotic performance varies as a function of the task and/or need for a listener to attend to acoustic/phonetic detail.

In our study, the task was recognition of speech produced by multiple speakers. Four channels did not seem to be sufficient for achieving high level of sentence understanding. To see why consider, in Fig. 2(a), the channel spectrum of the vowel [ɛ] (''head''), spoken by a male talker, and

TABLE II. The 3-dB bandwidths (Hz) of the filters.

No. of Channels	Channel 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	984	4215														
3	491	1295	3414													
4	321	664	1373	2842												
5	237	423	758	1356	2426											
6	187	304	493	801	1301	2113										
8	265	331	431	516	645	805	1006	1257								
10	204	244	293	352	422	506	607	729	874	1049						
12	165	193	225	262	306	357	416	486	567	661	771	900				
16	120	135	151	170	192	216	242	273	307	345	389	437	492	553	622	700

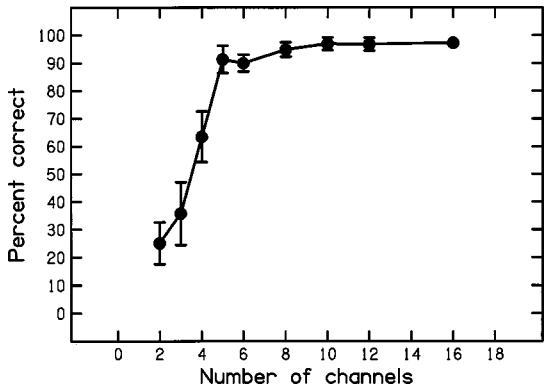

FIG. 1. Sentence understanding (percent correct) as a function of number of channels. Error bars indicate ±1 standard deviation.

processed through four channels. Four channels are sufficient to code the frequency of $F1$ and $F2$. The $F1$ of [ε] is coded by a high-amplitude in channel one, and a low-amplitude in channel two. The $F2$ of [ε] is coded by a high amplitude in channel three, and a low amplitude in channels two and four. Now, consider the four-channel spectrum [Fig. 2(b)] of the vowel [ε] produced by a female talker. In this case, four channels are not sufficient for coding $F2$ information, since channel three is no longer a peak in the spectrum. Figure 2(c) and 2(d) shows the channel spectra of the same vowels processed through five channels. The $F2$ information is coded adequately for both male and female vowels. The $F2$ is coded by a high amplitude in channel four, and a low amplitude in channels three and five [see Fig. 2(c)]. Most generally, four-channel processors use two channels (channels three and four) for coding $F2$ and the other high-frequency information needed for consonant recognition, while five-channel processors use three channels (channels three, four, and five). Overall, our results suggest that a minimum of three channels is needed to code $F2$ and/or high-frequency information for multi-talker speech recognition.

It is possible that four channels might yield higher levels of speech understanding if the filter spacing were optimized. Shannon *et al.* (1998) showed that there was a significant difference in sentence recognition scores as a function of three filter spacings (linear, logarithmic, and intermediate) of a four-channel processor. This filter optimization, however, can only be tailored for a particular speaker, e.g., a particular female or male, and is therefore not practical for real-world situations where multiple talkers must be accommodated.

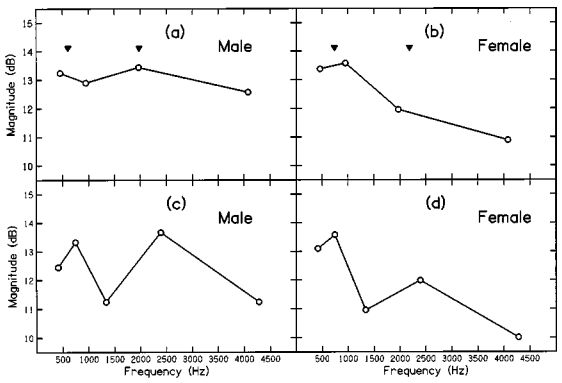

FIG. 2. The channel spectra of the vowel [ε] (''head'') produced by a male and a female talker. The spectra in (a) and (b) were generated using a four-channel processor, and the spectra in (c) and (d) were generated using a five-channel processor. The filled triangles indicate the formant frequencies of the vowels.

Although five channels achieved high levels (>90%) of intelligibility, asymptotic performance was not achieved until eight channels were used. Increasing the number of channels beyond eight did not improve speech intelligibility, but did improve the subjective quality of speech. The finding that eight channels are needed to reach asymptotic performance is consistent with the study by Dorman *et al.* (1997) who showed that eight channels were needed to reach asymptote for multi-talker vowel recognition.

Training (i.e., practice) is a factor that needs to be taken into account when interpreting the above results, since the normal-hearing listeners were not accustomed to listening to speech containing limited spectral/temporal information. The order of the test conditions was purposely confounded with the amount of experience in listening to the altered speech signals because it was felt that giving listeners additional practice before encountering signals with the least spectral information would maximize performance in the most difficult listening situations.

II. EXPERIMENT 2

Experiment 1 showed that a high level (90%) of intelligibility can be achieved using processors with five or more channels of stimulation. This finding is surprising given that the processors did not track or follow formant frequencies, like the pattern playback or the Remez *et al.* sine wave synthesizer. In the Remez *et al.* synthesizer, for instance, three sine waves trace out three formant frequencies in each update cycle. In contrast, the processors used in experiment 1 generated sine waves in each cycle (4 ms) at fixed frequencies (Table I). The only parameter that varied from cycle to cycle was the amplitudes of the sine waves. The frequencies of the sine waves coincided with the formant frequencies of speech only by chance and only rarely. This circumstance raises the question, ''How is information coded in the frequency domain with processors that do not track formant frequencies?'' As pointed out by Dorman *et al.* (1997), the relative differences in across-channel amplitudes must be used to code frequency information. On this view, if amplitude resolution were to be distorted, then speech recognition ought to decline. This hypothesis was tested in experiment 2 where the channel amplitudes of a six-channel processor were quantized to a finite number (2, 4, 8, 16) of steps. At issue was how many discriminable steps are needed to maintain high levels of speech intelligibility when speech is processed through a small number of channels. The answer to that question is of interest because it could provide some insight into whether the speech perception abilities of some cochlear implant patients are limited by electrode dynamic range or the number of discriminable intensity steps within the dynamic range (Nelson *et al.*, 1998).

It is reasonable to expect that the number of steps used to code amplitude information within a channel will be less important when speech is processed through a large number of channels than when processed through a small number of

channels. This is because in the case of a large number of channels, signal frequency will be indicated by the channel or channels with significant energy. To test this hypothesis we processed speech through 16 channels, and quantized the channel amplitudes into 2–16 steps. At issue was whether the same number of steps are needed to maintain high levels of speech intelligibility for speech processed through a large number (16) of channels and through a small number (6) of channels.

A. Method

1. Subjects

The same subjects as in experiment 1 were used.

2. Sentence material

One hundred and fifty new sentences from the TIMIT database, produced by an equal number of female and male speakers, were randomly selected. Seventy-five sentences were used for the 6-channel processor and 75 sentences for the 16-channel processor. The 75 sentences used in each experiment were divided into five lists with 15 sentences in each list—one list was used for each of the four quantized conditions ($Q=2,4,8,16$ levels), and one list was used for the unquantized condition. The subjects listened to a total of 150 sentences, 75 sentences processed through 6 channels, and 75 sentences processed through 16 channels.

3. Quantization and signal processing

The envelope dynamic range of speech processed through a finite number of channels differs from channel to channel. For that reason, different quantization step sizes are needed for each channel. We first determined the amplitude dynamic range of each channel by computing envelope histograms of 100 TIMIT sentences. [The TIMIT sentences were scaled so that all sentences had the same peak amplitude.] The maximum envelope amplitude in each channel, denoted as $X^i_{\max}$ where i is the channel number, was chosen to include 99% of all amplitude counts in that channel. The minimum envelope amplitude ($X^i_{\min}$) was set 0.5 dB above the rms value of the noise floor. The $X^i_{\max}$ and $X^i_{\min}$ values were then used to estimate the quantization step size, Δi, of each channel as follows:

$$\Delta_i = \frac{X^i_{\max} - X^i_{\min}}{Q-1} \quad i=1,2,...,N,$$

where Q is the number of quantization levels or steps, and N is the number of channels (6 or 16 in our case). Note that each channel had a different value for $X^i_{\max}$ and $X^i_{\min}$ since the envelope dynamic range of each channel was different. Consequently, the step sizes Δi were different in each channel.

The quantized version of the six-channel sine wave processor was implemented as follows. Six envelope amplitudes were computed as before by pre-emphasizing the signal, bandpass filtering the signal into six logarithmic frequency bands (Table I), full-wave rectifying the bandpassed waveforms, and low-pass filtering (400 Hz) the rectified waveforms. The envelope amplitudes were then uniformly quantized to Q discrete levels ($Q=2,4,8,16$). Sine waves were generated with amplitudes equal to the quantized envelope amplitudes, and frequencies equal to the center frequencies of the bandpass filters. The phases of the sinusoids were estimated from the 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.

The quantized version of the 16-channel sine wave processor was implemented as follows. Sixteen envelope amplitudes were computed as before by pre-emphasizing the signal, bandpass filtering the signal into 16 frequency bands (Table I), full-wave rectifying the bandpassed waveforms, and low-pass filtering (400 Hz) the rectified waveforms. Of the 16 envelopes computed, the six envelopes with the largest amplitude were selected in each 4-ms cycle.[3] The six selected envelope amplitudes were then uniformly quantized to Q discrete levels ($Q=2,4,8,16$). Sine waves were generated with amplitudes equal to the quantized envelope amplitudes, and frequencies equal to the center frequencies of the selected bandpass filters. The phases of the sinusoids were estimated from the FFT of the speech segment. The sinusoids of the six selected bands were finally summed and the level of the synthesized speech segment was adjusted to have the same rms value as the original speech segment.

4. Procedure

The experiment was run in two independent $1\frac{1}{2}$-h sessions. In the first session, the listeners were presented with a list of 75 sentences processed through the 6-channel processor, 60 quantized sentences (15 for each of the 4 conditions) and 15 unquantized sentences. In the second session, the listeners were presented with a list of 75 sentences processed through the 16-channel processor, 60 quantized sentences (15 for each of the 4 conditions) and 15 unquantized sentences. The quantized and the unquantized sentences, in both experiments, were completely randomized. A practice session preceded each test session, in which the listeners were presented with ten examples of sentences from each quantized condition. None of the sentences used in the practice session were used in the test session.

B. Results and discussion

The results for the 6- and 16-channel processors are shown in Fig. 3. A repeated measures analysis of variance on the data for the six-channel processor indicated a main effect [$F(4,32)=112.54$, $p<0.0001$] for the number of quantization steps. The mean scores were 41% correct for the 2-step condition, 52% correct for the 4-step condition, 80% correct for the 8-step condition, 83% correct for the 16-step condition, and 92% correct for the unquantized condition. *Post hoc* tests indicated that 4 steps allowed better performance than 2, 8 allowed better performance than 4 steps, 8 and 12 steps produced scores which did not differ, and the unquantized signal allowed better scores than the signal processed into 16 steps. Relatively high levels of intelligibility were achieved using 8 levels (mean score=80% correct) and 16

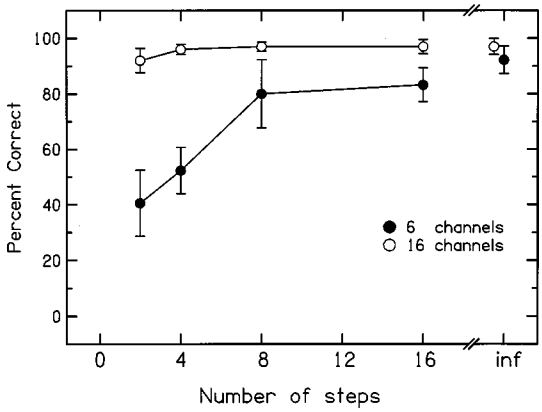

FIG. 3. Speech recognition with 6-channel (filled circles) and 16-channel (empty circles) processors as a function of the number of the steps used to quantize the spectral amplitudes. ''Inf'' refers to the condition in which the spectral amplitudes were not quantized. Error bars indicate ±1 standard deviations.

levels (mean score=83% correct). These results are similar to the results found with early-model ten-channel vocoders (David, 1956), e.g., 82% correct with six levels of intensity quantization.

A repeated measures analysis of variance on the data for the 16-channel processor indicated a main effect [$F(4,32) = 7.67, p < 0.0001$] for the number of quantization steps. *Post hoc* tests according to Scheffe showed that there was a significant difference ($p = 0.002$) between the scores obtained with two and four steps, 92% correct and 96% correct, respectively. There was no statistically significant difference between the scores obtained with four steps and greater number of steps. Thus two steps were sufficient for achieving a high level (92%) of performance. This outcome is consistent with the findings of Drullman *et al.* (1995) that reported nearly perfect intelligibility when speech was processed through 24 $\frac{1}{4}$-octave bands, and the amplitude envelopes were quantized into two levels. Our results and those of Drullman *et al.* (1995) suggest that poor amplitude resolution (defined in terms of the number of steps) does not have a large effect on intelligibility when speech is processed through a large number of channels.

In contrast, when speech was processed into a small number (six) of channels, performance was poor (<55% correct) when the number of levels was smaller than eight. This outcome can be accounted for by the view that that listeners must rely on relative amplitude differences across channels to infer frequency information when speech is processed into a small number of channels. If amplitude differences are distorted, then recognition accuracy will suffer. On this view, cochlear implant patients who are able to use only a few channels of stimulation, and who are able to discriminate only a small number of intensity differences on each channel (Nelson *et al.*, 1996), should find speech recognition relatively difficult.

III. GENERAL DISCUSSION

A. Number of channels

The results in experiment 1 showed that five channels are needed to achieve high levels of sentence understanding and eight channels are needed to reach asymptotic performance. The task at hand was recognition of TIMIT sentences produced by multiple speakers. It is very likely that the number of channels needed to reach asymptotic performance as well as the shape of the performance-channels function will depend on the speech material and whether listeners will be required to rely on phonetic detail. A different asymptote would be expected, for instance, if the task were nonsense-syllable recognition since the listeners will need to attend to fine acoustic/phonetic detail in order to understand what was being said. The results of experiment 1 do not support a general conclusion that eight channels are needed for all types of speech material, but rather for recognition of syntactically well-formed and meaningful sentences produced by multiple speakers.

Other factors that could affect the number of channels needed to achieve a high level of sentence intelligibility include speaking rate, speaking style (conversational versus clear) and background noise. Higher speaking rates are often associated with reduced sentence understanding, and speaking clearly is associated with improved sentence understanding in noise for normal-hearing listeners (Tolhurst, 1955) and improved sentence understanding in quiet for hearing impaired listeners (Picheny *et al.*, 1985). Both speaking style and speaking rate deserve further study in the context of the number of channels necessary for speech understanding. Speech understanding in noise has been studied by Dorman *et al.* (1998) and by Fu *et al.* (1998). More channels are needed in noise than in quiet to achieve high levels of speech understanding.

B. Number of steps and number of channels

The findings obtained in experiment 2 with the 6- and 16-channel-processors suggest an inverse relationship between the importance of spectral amplitude resolution and spectral resolution (defined in terms of the number of spectral channels available). Two levels of amplitude resolution were sufficient for nearly perfect intelligibility (92%) when speech was processed through 16 channels. However, eight or more levels were needed for high intelligibility when speech was processed through six channels. We have only investigated the effect of quantization on two extreme cases, i.e., a small number of channels and a large number of channels. Further studies are needed to complete our understanding of the effects of spectral amplitude resolution and spectral resolution on speech understanding.

IV. CONCLUSIONS

These studies have provided yet another demonstration that speech understanding does not require a detailed spectral representation of the speech signal. In experiment 1 we found that five channels of fixed frequency stimulation allowed 90% identification accuracy for sentences produced by multiple speakers. Asymptotic performance was achieved with eight channels. In experiment 2 we found that the number of levels used to code spectral amplitude information has a significant effect on speech understanding. If speech is processed into a large number of channels, two levels of ampli-

tude resolution are sufficient to achieve a high level of speech understanding. However, when speech is processed into a small number of channels, eight or more levels are necessary. Thus the number of channels of stimulation and the resolution of amplitude information within those channels trade off in determining the level of speech understanding allowed by signal processors which reduce the speech signal to a relatively small number of fixed-frequency channels.

ACKNOWLEDGMENTS

The authors would like to thank James Hillenbrand, Robert Shannon, and Steve Greenberg for providing valuable suggestions on earlier drafts of this paper. This research was supported by a Shannon award (R55 DC03421) from the National Institute of Deafness and other Communication Disorders, NIH.

[1]For $n \geqslant 8$, the filter bandwidths were computed according to the equation: $1100 \log(f/800+1)$, where f indicates the frequency in Hz. This is similar to the technical mel scale of Fant (1973), which is a variant of the critical band scale. As shown in Table II, the channel filter bandwidths, for $n \geqslant 8$, are approximately 1/4 of an octave wide, which is roughly the bandwidth of the critical band. Logarithmic spacing was used for $n < 8$ to conform with the spacing used in current cochlear implant devices (e.g., Zierhofer *et al.*, 1994).

[2]The phases of the sinusoids were computed from the FFT of the speech segment as follows. Let $\phi(k)$ be the phases of the FFT of a (4-ms) speech segment. The phases $\theta(j)$ of the N sinusoids, N being the number of channels, were set equal to the phases of the FFT spectrum evaluated at frequencies closest to the center frequencies of the bandpass filters, i.e.,

$$\theta(j) = \phi\left(\left\lceil \frac{f_j}{r} \right\rceil\right), \quad j = 1,2,...,N,$$

where f_j is the center frequency (Hz) of the jth bandpass filter (Table I), r is the FFT resolution (r = sampling frequency/FFT length) in Hz, and $\lceil \cdot \rceil$ denotes the nearest integer. Due to the limited FFT resolution, the above equation only provides a rough estimate of the underlying sinewave phases. This estimate seems to be sufficient in our case, however, since we are only concerned with speech intelligibility rather than speech quality (see McAulay and Quatieri, 1995, for a discussion on alternative sinewave phase representations).

[3]This spectral-maximum implementation was chosen to mimic the signal processing used in the Nucleus 22 cochlear implant processor (McDermott *et al.*, 1992). In this processor, speech is processed through 16 channels, and the 6-channel amplitudes with the largest energy are selected in each cycle for electrical stimulation.

Bailey, P., Summerfield, Q., and Dorman, M. (**1977**). ''On the identification of sine-wave analogues of certain speech sounds,'' Haskins Laboratories Status Report on Speech Perception, **SR 51–52**, 1–26.

Cutting, J. (**1974**). ''Two left-hemisphere mechanisms in speech perception,'' Percept. Psychophys. **16**, 601–612.

David, E. (**1956**). ''Naturalness and distortion in speech-processing devices,'' J. Acoust. Soc. Am. **28**, 586–589.

Delattre, F., Liberman, A., Cooper, F., and Gerstman, L. (**1952**). ''An experimental study of the acoustic determinants of vowel color: Observations on one- and two-formant vowels synthesized from spectrographic displays,'' Word **8**, 195–210.

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.

Dudley, H. (**1939**). ''Remaking speech,'' J. Acoust. Soc. Am. **11**, 169–177.

Fant, G. (**1973**). *Speech Sounds And Features* (MIT Press, Boston).

Flanagan, J. (**1972**). *Speech Analysis, Synthesis And Perception* (Springer Verlag, New York).

Fu, Q-J., Shannon, R., and Wang, X. (**1998**). ''Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing,'' J. Acoust. Soc. Am. **104**, 3586–3596.

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).

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

Lamel, L., Kassel, R., and Seneff, S. (**1986**). ''Speech database development: Design and analysis of the acoustic-phonetic corpus,'' Proc. of the DARPA Speech Recognition Workshop, Report No. SAIC-86/1546.

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**, 101–130.

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.

McAulay, R., and Quatieri, T. (**1986**). ''Speech analysis/synthesis based on a sinusoidal representation,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-34**, 744–754.

McAulay, R., and Quatieri, T. (**1995**). ''Sinusoidal coding,'' in *Speech Coding and Synthesis*, edited by W. Kleijn and K. Paliwal (Elsevier Science, New York).

McDermott, H., McKay, C., and Vandali, A. (**1992**). ''A new portable sound processor for the University of Melbourne/Nucleus Limited multielectrode cochlear implant,'' J. Acoust. Soc. Am. **91**, 3367–3371.

Mullenix, J., Pisoni, D., and Martin, C. (**1989**). ''Some effects of talker variability on spoken word recognition,'' J. Acoust. Soc. Am. **85**, 365–378.

Nelson, D., Schmitz, J., Donaldson, G., Viemester, N., and Javel, E. (**1996**). ''Intensity discrimination as a function of stimulus level with electric stimulation,'' J. Acoust. Soc. Am. **100**, 2393–2414.

Picheny, M., Durlach, N., and Braida, L. (**1985**). ''Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech,'' J. Speech Hear. Res. **28**, 96–103.

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

Schroeder, M. (**1966**). ''Vocoders: Analysis and synthesis of speech,'' Proc. IEEE **54**, 720–734.

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

Shannon, R., Zeng, F.-G., and Wygonski, J. (**1998**). ''Speech recognition with altered spectral distribution of envelope cues,'' J. Acoust. Soc. Am. **104**, 2467–2476.

Sommers, M., Kirk, K., and Pisoni, D. (**1997**). ''Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format,'' Ear Hear. **18**, 89–99.

Tolhurst, G. (**1955**). ''The effect of intelligibility scores of specific instructions regarding talking,'' USAM Report No. NM 001 064 01 35 (Naval Air Station, Pensacola, FL).

Zierhofer, C., Peter, O., Bril, S., Pohl, P., Hochmair-Desoyer, I., and Hochmair, E. (**1994**). ''A multichannel cochlear implant system for high-rate pulsatile stimulation strategies,'' in *Advances in Cochlear Implants*, edited by I. Hochmair Desoyer and E. Hochmair (International Interscience Seminars, Vienna), pp. 204–207.