Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs

Michael F. Dorman, Philipos C. Loizou and Dawne Rainey

ARTICLES YOU MAY BE INTERESTED IN

Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants
The Journal of the Acoustical Society of America **110**, 1150 (2001); https://doi.org/10.1121/1.1381538

On the number of channels needed to understand speech
The Journal of the Acoustical Society of America **106**, 2097 (1999); https://doi.org/10.1121/1.427954

A cochlear frequency-position function for several species—29 years later
The Journal of the Acoustical Society of America **87**, 2592 (1990); https://doi.org/10.1121/1.399052

Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility
The Journal of the Acoustical Society of America **142**, EL537 (2017); https://doi.org/10.1121/1.5016044

Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience
The Journal of the Acoustical Society of America **122**, 2376 (2007); https://doi.org/10.1121/1.2773993

The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels
The Journal of the Acoustical Society of America **104**, 3583 (1998); https://doi.org/10.1121/1.423940

CALL FOR PAPERS

JASA
THE JOURNAL OF THE
ACOUSTICAL SOCIETY OF AMERICA

Special Issue: Fish Bioacoustics: Hearing and Sound Communication

Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs

Michael F. Dorman[a)]

Department of Speech and Hearing Science, Arizona State University, Tempe, Arizona 85287-0102 and University of Utah Health Sciences Center, Salt Lake City, Utah 84132

Philipos C. Loizou[b)]

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

Dawne Rainey

Arizona State University, Tempe, Arizona 85287-0102

Vowels, consonants, and sentences were processed through software emulations of cochlear-implant signal processors with 2–9 output channels. The signals were then presented, as either the sum of sine waves at the center of the channels or as the sum of noise bands the width of the channels, to normal-hearing listeners for identification. The results indicate, as previous investigations have suggested, that high levels of speech understanding can be obtained using signal processors with a small number of channels. The number of channels needed for high levels of performance varied with the nature of the test material. For the most difficult material—vowels produced by men, women, and girls—no statistically significant differences in performance were observed when the number of channels was increased beyond 8. For the least difficult material—sentences—no statistically significant differences in performance were observed when the number of channels was increased beyond 5. The nature of the output signal, noise bands or sine waves, made only a small difference in performance. The mechanism mediating the high levels of speech recognition achieved with only few channels of stimulation may be the same one that mediates the recognition of signals produced by speakers with a high fundamental frequency, i.e., the levels of adjacent channels are used to determine the frequency of the input signal. The results of an experiment in which frequency information was altered but temporal information was not altered indicates that vowel recognition is based on information in the frequency domain even when the number of channels of stimulation is small. © *1997 Acoustical Society of America.* [S0001-4966(97)04010-1]

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

INTRODUCTION

Shannon *et al.* (1995) have reported nearly perfect scores on tests of speech recognition (vowels in /hVd/ context, consonants in /vCv/ context and sentences) when temporal information in the signals is preserved and spectral information is reduced to three or four bands of noise. This outcome is surprising from the point of view that vowel and consonant identity is specified by the location of formant frequencies and that small differences in formant frequencies can lead to changes in phonetic identity. For example, the formant frequencies of vowels can differ by as little as 100–200 Hz (Peterson and Barney, 1952) and 300–400 Hz differences in the onset frequency of the second-formant transition are sufficient to signal consonant place-of-articulation (/b d g/) in synthetic, two-formant syllables (Cooper *et al.*, 1952). These small, but critical, differences in formant frequencies would fall within a band of a 3- or 4-channel processor and, thus, would not be available to a listener for use in phonetic identification.

The Shannon *et al.* (1995) outcome is less surprising from the view that for each segmental phone there are multiple cues to identity. In addition to cues in the time domain, which are known to provide some information about vowel identity (House, 1961) and considerable information about consonant voicing (e.g., Liberman *et al.*, 1958) and consonant manner (e.g., Liberman *et al.*, 1956), there are cues in the frequency domain which are relatively wide band, e.g., the burst spectra of stop consonants and the noise spectra of fricative consonants. These cues may be represented adequately with only a few channels of stimulation. Other factors may also enhance intelligibility when speech is presented via a few channels. A small number of items in the test set (e.g., 8 vowels and 16 consonants) would contribute to high levels of performance. In tests of sentence intelligibility, ''top down'' processing, or the use of multiple levels of linguistic knowledge, would aid subjects in word identification even if phonetic information was reduced, because of very poor frequency resolution, to only broad phonetic categories (see, for example, Zue, 1985).

Our interest in Shannon *et al.* (1995) stems from our experience with cochlear implant patients who use 4- and 6-channel signal processors. When individual channels are stimulated the patients report, most generally, that the signals

---

[a)]Electronic mail: mdorman@imap2.asu.edu
[b)]Electronic mail: loizou@ualr.edu

sound like ''beep tones'' and not like bands of noise. This observation led us to wonder about the intelligibility of speech, for normal-hearing listeners, when the speech is processed through a signal processor which outputs a sine wave at the center of each frequency band instead of a band of noise. Previous research with speech coding systems which resynthesized speech as the summation of the sine wave outputs of fixed channels suggest an asymptote in speech recognition with six to ten channels of stimulation (e.g., Hill et al., 1968). This is approximately twice as many channels as Shannon et al. (1995) found to be necessary. The difference in the number of channels is striking and led us to investigate how the Shannon et al. (1995) processing scheme and a similar processing scheme using sine waves would compare when tested with the same materials. Thus, in experiment 1 we compared the intelligibility of vowels, consonants, and sentences when transmitted by a processor which output bands of noise, as in Shannon et al. (1995), and when transmitted by a processor which output sine waves at the center frequencies of the filters.

Shannon et al. (1995) interpreted their results as indicating that speech recognition, in the context of greatly reduced spectral information, can be achieved with primarily temporal cues. In experiment 2 we tested this hypothesis with the vowels used in experiment 1.

Consider the information available to a listener about vowel identity when signals are resynthesized as the summation of, for example, four bands of noise. The relative amplitudes of the noise bands indicates the approximate locations of the formants. Thus, the relative amplitudes across channels provide information, albeit very crude information, in the frequency domain. Temporal information is limited to a binary distinction along a continuum of vowel length, i.e., long versus short vowels. Now consider an experiment in which signal length is left unchanged, so short vowels are ''short'' and long vowels are ''long'', but in which the amplitudes of the four noise bands are inverted. That is, the output of channel 1 is directed to channel 4, the output of channel 4 is directed to channel 1, the output of channel 3 is directed to channel 2, and the output of channel 2 is directed to channel 3. If temporal information is central to vowel recognition, then, in the situation just described, vowel recognition ought to be little changed since the temporal information was left unchanged. If, however, vowel recognition is based primarily on information in the frequency domain, then recognition ought to be very poor. In experiment 2 the outcome of such an experiment is reported.

I. EXPERIMENT 1

A. Method

1. Subjects

The subjects were eight young adults (all female, age range 22–31 years, mean=25 years) and one 63-year-old female.[1] All of the subjects passed a hearing screening at 25 dB HL for frequencies of 0.5, 1, 2, and 4 kHz.

TABLE I. Channel center frequencies for signal processors with 2–9 channels.

No. of channels	Channel								
	1	2	3	4	5	6	7	8	9
2	792	3392							
3	545	1438	3793						
4	460	953	1971	4078					
5	418	748	1339	2396	4287				
6	393	639	1037	1685	2736	4443			
7	377	572	866	1312	1988	3013	4565		
8	366	526	757	1089	1566	2252	3241	4662	
9	357	493	682	942	1301	1798	2484	3431	4740

2. Speech materials

Three tests of vowel recognition were used. One was the vowel test employed by Shannon et al. (1995)—the Iowa vowel test which used 8 vowels from a single male speaker (Tyler et al., 1989). The second test was composed of 13 synthetic vowels in /bVt/ format (see Dorman et al., 1989). These stimuli were used because the vowels were of equal duration and had identical pitch contour. Thus, temporal cues were not a factor in identification. The third test was composed of 11 vowels in the words ''heed, hid, hayed, head, had, hod, hud, hood, hoed, who'd, heard,'' each produced by three men, three women, and three girls. The stimuli were drawn from a set used by Hillenbrand et al. (1995).

The consonant test was the Iowa constant test—16 consonants in /aCa/ environment spoken by a single male speaker (Tyler et al., 1986). This was the consonant test used by Shannon et al. (1995). The sentence material was from the H.I.N.T. test presented without competing noise (Nilsson et al., 1994). Examples of sentences in this test are, ''They met some friends at dinner,'' ''Yesterday he lost his hat,'' ''She spoke to her eldest son,'' and ''She is washing her new silk dress.''

All of the test materials were stored on computer disk and were output via custom software routines using MAT-LAB software and a 16-bit D/A converter.

3. Signal processing

The noise-band processor was implemented in the following manner. The signal was first processed through a pre-emphasis filter (low-pass below 1200 Hz, -6 dB per octave) and then bandpassed into N logarithmic frequency bands (where N varied from 2 to 9) using sixth-order Butterworth filters. The filter center frequencies and bandwidths, at 3 dB down from the passband level, are shown in Tables I and II. The envelope of the signal was extracted by half-wave rectification and low-pass filtering (second-order Butterworth) with a 160-Hz cutoff frequency. The envelope of each frequency band was used to modulate white noise, which was bandlimited with the same Butterworth bandpass filters. The noise-modulated envelopes of each band were finally combined, low-pass filtered (using sixth-order elliptic filters with 50-dB attenuation) at 5 kHz, and presented to the listeners at a comfortable level through headphones (Sennheiser HMD 410).

TABLE II. Channel bandwidths for signal processors with 2–9 channels.

No. of channels	Channel								
	1	2	3	4	5	6	7	8	9
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			
7	154	234	355	538	814	1234	1870		
8	131	189	272	391	563	810	1165	1676	
9	114	158	218	302	417	576	796	1099	1519

Our implementation of a noise-band processor differed from that of Shannon *et al.* (1995) in several ways. In Shannon *et al.* (1995) the filters overlapped at −15 dB instead of at −3 dB as in the present experiment. In Shannon *et al.* (1995) filter 1 and filter 2 overlapped at 800 Hz, filters 2 and 3 overlapped at 1500 Hz, and filters 3 and 4 overlapped at 2500 Hz.

The sine-wave processor was implemented as follows. The signal was first processed through a preemphasis filter (low-pass below 1200 Hz, −6 dB per octave) and then bandpassed into N logarithmic frequency bands (where N varied from 2 to 9) using sixth-order Butterworth filters (see Tables I and II). The envelope of the signal was extracted by full-wave rectification, and low-pass filtering (second-order Butterworth) with a 400-Hz cutoff frequency. [We used a 400-Hz cutoff frequency to conform to the cutoff frequency used in the Med El Corporation's cochlear-implant signal processor. Shannon *et al.* (1995) found no difference in performance for lowpass filters set at 160 Hz and above.] Sinusoids were generated with amplitudes equal to the root-mean-square (rms) energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed and presented to the listeners at a comfortable level.

4. Procedures

For the Iowa consonant, Iowa vowel, and synthetic vowel test sequences, five tokens of each stimulus were created. The stimuli were grouped into five blocks with each stimulus appearing once in a block. Stimulus order within each block was randomized. In the multitalker vowel test sequence, each stimulus appeared once. The stimuli were completely randomized. Ten sentences from the H.I.N.T. sentence lists were presented in each channel condition. Each ten-sentence list contained approximately 52 words. Different sentence lists were used for each condition. The sentence tests were conducted in open set format.

The tests were run in the order vowels (Iowa, synthetic, multitalker), consonants, and sentences. This test order was used because it was the order in which the stimulus materials became available for testing. After the experiment described here was conducted, we tested, in another experiment, other subjects with the consonant, multitalker vowel, and sentence material. In this experiment the order of tests was randomized. The mean scores were within 7 percentage points of the

mean scores in this report. For this reason we believe that the test order did not have a significant effect on performance.

For each type of material subjects were given practice before the test sequence began. Practice consisted of two repetitions of the test items with visual indication of item identity followed by a randomized sequence of stimuli with feedback of correct answers. Practice was given before each processor by channel-number condition. The test sessions were run in the order 9-channel processor to 2-channel processor, sequentially. For each type of material half of the patients were tested first with the noise-band processor and half were tested first with the sine-wave processor.

We chose a sequential test order starting with the largest number of channels, rather than a randomized test order, for two reasons. First, a completely randomized design would have required a heroic number of subjects. Second, we wanted to give the subjects time to adapt to the novel stimulation. There is, undoubtedly, a ''warm up''effect for listening to altered speech signals of any kind. This effect was offset, to some degree, by our use of a sequential, rather than randomized, test order and by our familiarization procedure before each test condition. It is likely the case that *absolutely naive* subjects would not perform as well as the subjects in our experiments.

Responses were collected with custom software using a computer display of response alternatives (except for the sentence material) and a mouse as a response key. The subjects were allowed to use a ''repeat'' key during the consonant and vowel tests as many times as they wished. For the tests of word intelligibility in sentences the subjects were presented a sentence once, and were instructed to repeat as many of the words as they could. Each word in the sentence was scored.

B. Results

The results for vowel, consonant, and sentence identification are shown in Fig. 1, panels A–E. For the Iowa vowels a repeated measures analysis of variance indicated a main effect for channels ($F[7,56]=74.2$, $p<0.0001$) but no main effect for processors ($F[1,8]=0.004$, $p=0.85$). *Post hoc* tests according to Scheffe (alpha=0.05) indicated no statistically significant differences in performance when the number of channels was increased beyond 6.[2] For the synthetic vowels a repeated measures analysis of variance indicated a main effect for channels ($F[7,56]=208.8$, $p<0.0001$) but no main effect for processors ($F[1,8]=3.05$, $p=0.12$). *Post hoc* tests according to Scheffe indicated no statistically significant differences in performance when the number of channels was increased beyond 8. For the multitalker vowels a repeated measures analysis of variance indicated a main effect for channels ($F[7,56]=330.3$, $p<0.0001$), a main effect for processors ($F[1,8]=8.98$, $p=0.02$), and a processor by channels interaction ($F[7,56]=4.88$, $p=0.0002$). The sine-wave processor produced slightly higher mean scores with 3–9 channels of stimulation. The largest difference between processors occurred with 7, 8, and 9 channels of stimulation. *Post hoc* tests according to Scheffe indicated no statistically significant differences in performance when the number of channels was increased beyond 8.
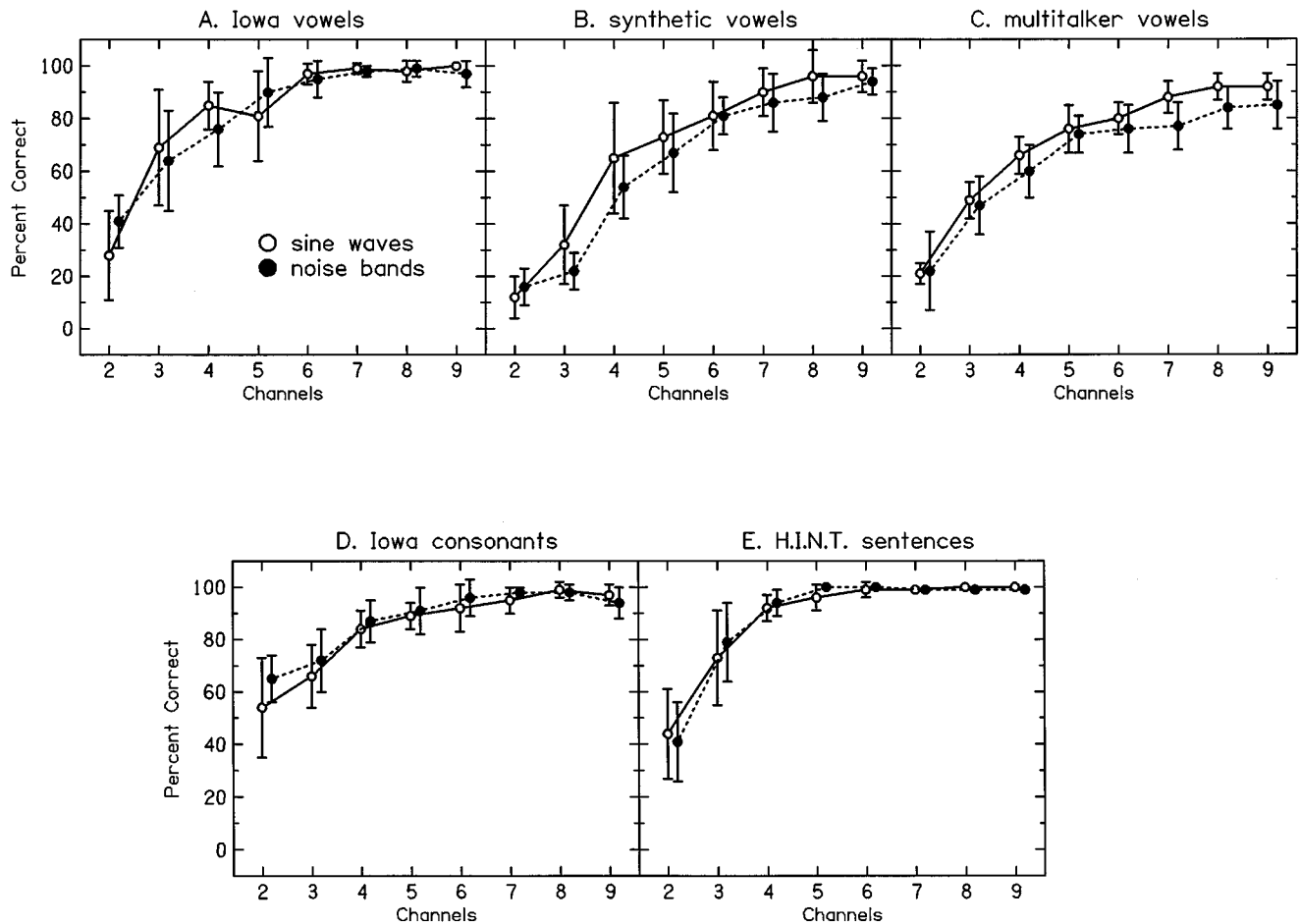
FIG. 1. Percent correct as a function of the number of channels of stimulation for vowels, consonants, and sentences. The parameter is processor type: sine-wave output (open circles) or noise-band output (filled circles). Error bars indicate ±1 standard deviation.

For the Iowa consonants a repeated measures analysis of variance indicated a main effect for channels ($F[7,56] = 130.9$, $p < 0.0001$) but no main effect for processors ($F[1,8] = 3.26$, $p = 0.11$). *Post hoc* tests according to Scheffe indicated no statistically significant differences in performance when the number of channels was increased beyond 6.

The results of feature analyses for the consonants are shown in Fig. 2. For the feature "place of articulation" a repeated measures analysis of variance indicated a main effect for channels ($F[7,56] = 87.21$, $p < 0.0001$) and a main effect for processors (noise band = 76% correct and sine wave = 70% correct; $F[1,8] = 8.37$, $p = 0.02$). *Post hoc* tests according to Scheffe indicated no statistically significant differences in performance when the number of channels was increased beyond 6. The reduction in mean score with 9 channels relative to 8 channels for the noise-band processor was the result of a single subject's low score in the 9-channel condition. This, most likely, reflects a warm-up effect as the 9-channel condition was always run first. For the feature "voicing" a repeated measures analysis of variance indicated a main effect for channels ($F[7,56] = 6.89$, $p < 0.0001$) but no main effect for processors ($F[1,8] = 0.01$, $p = 0.91$). *Post hoc* tests according to Scheffe indicated no statistically significant differences in performance

when the number of channels was increased beyond 3. Analysis of the data for the feature "manner" was complicated by a decrease in performance with 3 channels of stimulation for the noise-band processor. With the exception of this point, scores were 90% correct or better for 2–9 channels of stimulation for both processors. The dip in performance at three channels for the noise-band processor was due to errors on the nasal consonants which were identified as the semivowel /l/. This outcome appears to be an oddity of the cutoff frequencies for the three-channel condition because nasal manner was well identified with both fewer channels and a greater number of channels.

For the H.I.N.T. sentences a repeated measures analysis of variance indicated a main effect for channels ($F[7,56] = 80.1$, $p < 0.0001$) but no main effect for processors ($F[1,8] = 0.16$, $p = 0.91$). *Post hoc* tests according to Scheffe indicate no statistically significant differences in performance when the number of channels was increased beyond 5.

C. Discussion

The first issue to be considered is whether the Shannon *et al.* (1995) data are replicable. Our results indicate that they are. Our 4-channel, noise-band processor allowed intelligi-
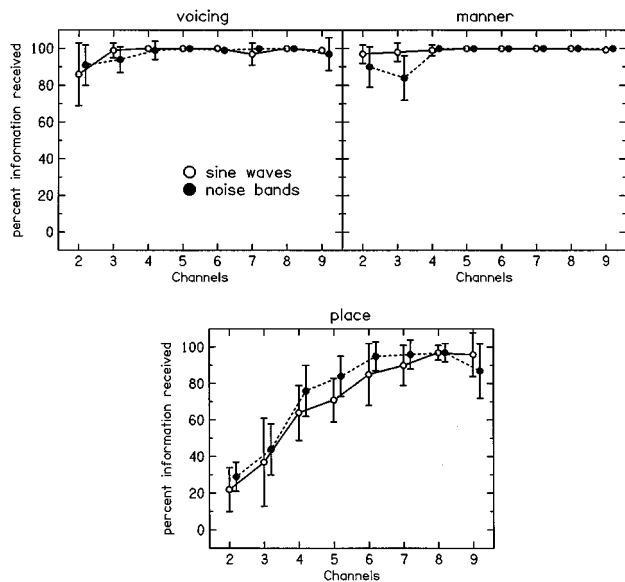
FIG. 2. Percent information received for the features ''voicing,'' ''manner,'' and ''place'' as a function of the number of channels of stimulation. The stimulus material was 16 consonants in ''aCa'' environment. Error bars indicate ±1 standard deviation.

bility scores of 87% for consonants, with 76% correct for place, and 94% correct for sentences. Shannon *et al.* (1995) report approximately 90% for consonants, with 65% for place, and 95% correct for sentences. However, our 4-channel, noise-band processor allowed only 76% correct for the Iowa vowels while Shannon *et al.* (1995) report a mean of approximately 95%. Two factors may have contributed to the difference in outcome. One factor is practice. Shannon *et al.* (1995) allowed his subjects 8–10 hours of practice before testing began. Our subjects had less practice. A second factor is the configuration of the filters. The filters in the present experiment had different corner frequencies from those in Shannon *et al.* (1995) and were chosen so that similar logarithmic spacing could be employed with any number of channels. In addition, our filters were broadband, while Shannon's filters were narrowband with small overlap between adjacent filters. Given the different results of the present experiment and Shannon *et al.* (1995) for the Iowa vowels, filter spacing deserves study in the design of signal processors with a small number of channels.

The second issue to be considered is whether the sine-wave processor and the noise-band processor produced different results. Different results were expected given the different stimulation along the cochlear partition produced by the two processors and given the results of Shannon *et al.* (1995), on the one hand, and Hill *et al.* (1968), on the other. However, differences in mean performance between the two processors were generally small and nonsignificant. Visual inspection of Fig. 1 indicates that the noise-band processor consistently allowed slightly lower vowel recognition scores than those allowed by the sine-wave processor. For one type of material—multitalker vowels—the scores were significantly lower. On the other hand, the noise-band processor allowed slightly, but significantly, higher scores for the feature place of articulation. The small disadvantage accruing to

the noise-band processor for vowels may have been due to the loss of frequency specific information within the noise-bands (see the discussion below). The same mechanism could mediate the *better* performance of the noise-band processor for consonant place of articulation. That is to say, energy distributed over the width of a band is more appropriate for many place cues than energy concentrated in a very narrow frequency region.

The third issue to be considered is how many channels are necessary to approach optimum performance with fixed-channel signal processors. The number of channels varies with test material. For material such as sentences, in which multiple levels of linguistic or ''top-down'' knowledge can be used, 5 channels allowed essentially 100% accuracy. For other material, such as multitalker vowels, or synthetic vowels, which needed fine-grained acoustic or ''bottom-up'' analyses, asymptotic performance was reached with 8 channels. It is possible, of course, that performance would reach asymptote at a larger number of channels if subjects were tested with more difficult materials, e.g., words with unreleased final stops.

If the present data with normal-hearing listeners can be extrapolated to cochlear implant patients, then it would be useful if signal processors for cochlear implants had 8 channels. More channels would not add greatly to intelligibility (at least in quiet) and fewer channels would detract from fine-grained acoustic analysis, although sentence understanding would not suffer with 5 channels, or even 4 channels, of stimulation.

The fourth issue to be considered is how information in the frequency domain is coded by processors which do not track formant frequencies. It is possible that the auditory system views the channels of fixed-channel processors in the same manner as harmonics of a high-pitched glottal source and derives an estimate of formant frequency from the amplitudes of adjacent channels in the same manner as formant frequencies are estimated from the amplitudes of adjacent harmonics of the glottal source. In the instance of normal speech signals with a high fundamental frequency, the location of formant peaks cannot be completely dependent on the location of the highest amplitude harmonics in the spectrum because the harmonics are too far apart. The relatively small difference limens for formant frequencies (e.g., 12 Hz at 550 Hz) indicates that the relative amplitudes of harmonics around a formant peak are used in the estimate of formant frequency (e.g., Sommers and Kewley-Port, 1996). Consider now the channel outputs in Fig. 3 for the 8-channel, sine-wave processor which allowed asymptotic performance for the multitalker vowel set. (For these plots the channel outputs were normalized into a 15-dB range. This allowed a common y axis for all of the plots in this figure and in Fig. 4). The formant frequencies of the vowels are indicated by the solid triangles. The formant frequencies are well represented by the relative amplitudes of the channels to the either side of the channels with the highest amplitude. For example, the low $F1$ of /i/ is coded by a high-amplitude channel 1 and a low-amplitude channel 2. The higher frequency $F1$ of /ɪ/ is coded by a high-amplitude channel 1 and a similar amplitude for channel 2. For /ɛ/ $F1$ is higher still and now channel 2 is
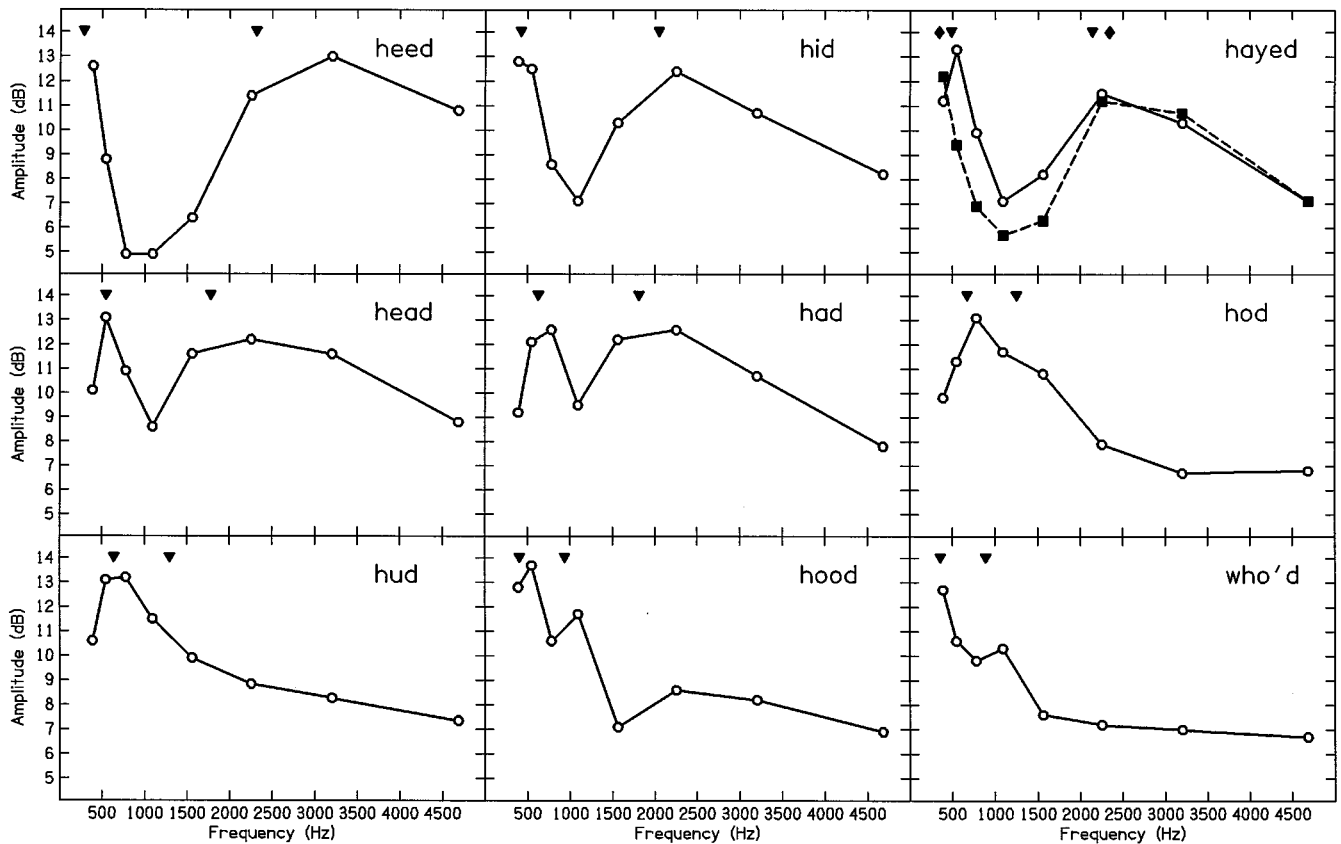
FIG. 3. Signal amplitude as a function of channel number for nine vowels produced by male speakers. The filled triangles and diamonds indicate the formant frequencies of the signal derived from 22nd-order LPC analyses. For "hayed" two sets of channel outputs are displayed. One was taken during the on-glide (open circles) and one (filled squares) was taken during the off-glide of the diphthong.

the highest amplitude channel. Finally, the high $F1$ of /æ/ is coded by high amplitudes in channels 2 and 3. Note also how the downward shift in frequency of $F1$ over the course of /e/ is coded by the change in the relative amplitudes of the first two channels. The location of $F2$ is coded in the same fashion as $F1$. For the high $F2$ in /i/ channel 7 has the highest amplitude. $F2$ is lower in /ɪ/ and channel 6 has the highest amplitude. $F2$ is lower still in /ɛ/ and channels 5 and 6 have similar amplitudes. The very low $F2$ of /u/ is coded by high amplitude in channel 4.

Not all vowels were coded by two peaks, corresponding to two formants, in the 8-channel outputs. Only one peak is present for "hod" and "hud." It is not surprising that these vowels were well identified, in spite of the single peak, since experiments dating to Delattre *et al.* (1952) have shown that vowels with $F1$ and $F2$ close together in frequency space can be synthesized with a single formant.

As the number of channels is reduced, the definition of spectral peaks in the channel outputs is, of course, reduced. This is illustrated in Fig. 4 which shows the channel outputs for the vowels in "hid," "head," and "had" for processors with 8, 6, and 4 channels. As noted previously, 8 channels provide good resolution of the formant peaks, if we assume that the levels in adjacent channels are used in the estimation of formant frequency. For the 6-channel processor the relative amplitudes of the channels continue to provide a good estimate of formant frequencies and the mean score for mul-

titalker vowels remains high (80% correct). When the number of channels is reduced to 4, the output patterns look very different than those for 6 and 8 channels. However, adjacent channels, e.g., channels 1 and 2, continue to provide information about the relative frequencies of the input signal. Thus, the low $F1$ of /ɪ/ is coded by a large difference in signal level between channels 1 and 2. The higher $F1$ in /ɛ/ is coded by a smaller difference between channels 1 and 2. The still higher $F1$ of /æ/ is coded by a slightly higher signal level in channel 2 than channel 1. These differences, although not as visually salient as those for the 6- and 8-channel conditions, appear to be used by listeners since the mean score for the 4-channel condition was 66% correct. In this instance, e.g., for discriminating between /ɪ/ and /ɛ/, vowel length may be especially useful.

II. EXPERIMENT 2

The aim of experiment 2 was to test the hypothesis that vowel recognition, in the condition of a small number of channels of stimulation, is based primarily on temporal cues. As noted in the Introduction, the temporal cue to vowel identity is vowel length. In the experiment which follows the identification of three sets of vowels (Iowa vowels, synthetic vowels, and multitalker vowels) was assessed in a condition of "normal" 4-channel stimulation and in a condition in which vowel length was left unchanged but the signal level
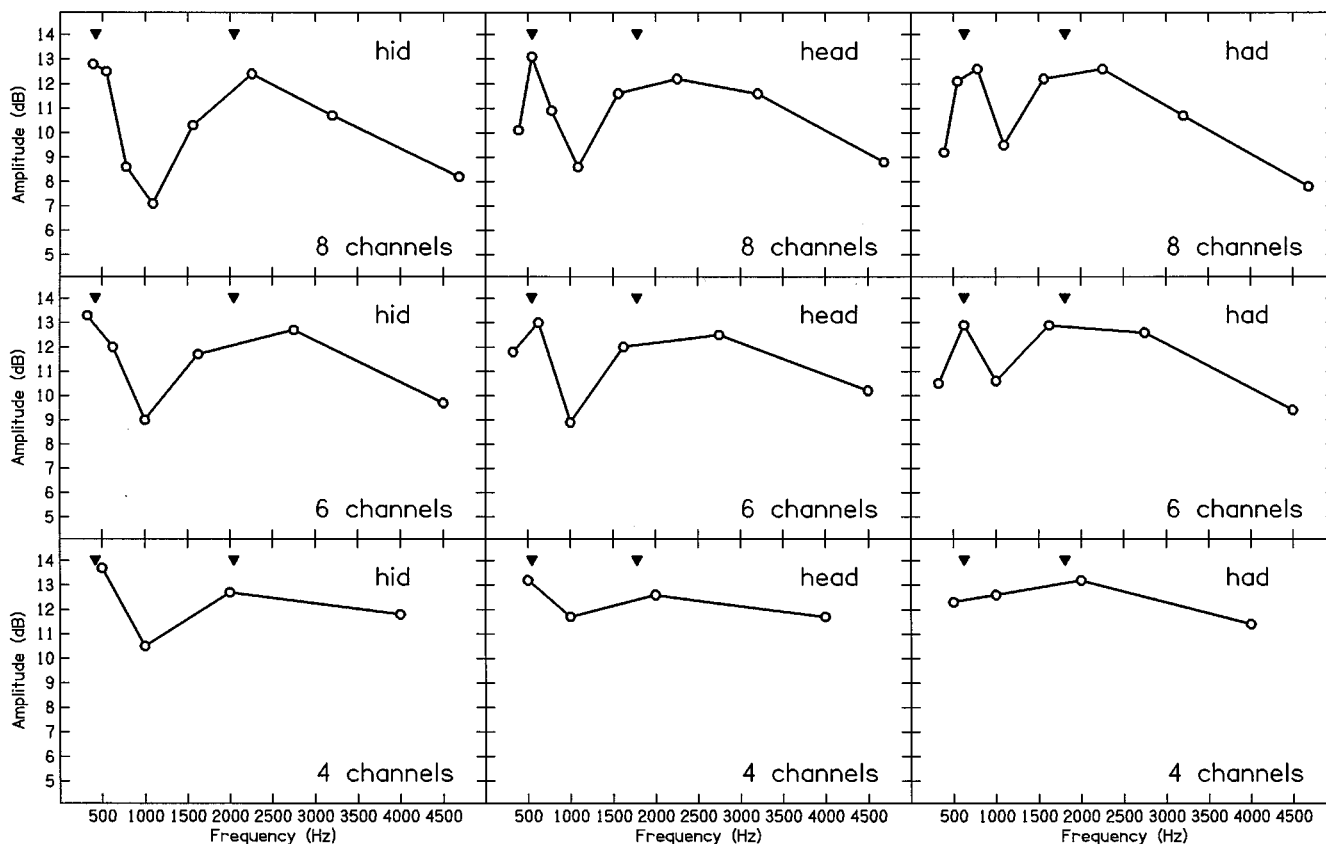
FIG. 4. Signal amplitude as a function of channel number for the vowels in ''hid,'' ''head,'' and ''had.'' In each column the output of processors with 8, 6, and 4 channels is shown. The stimuli were tokens of vowels produced by male speakers.

in each channel was altered. In the latter condition temporal information was normal but the information which specified frequency (the relative levels of the channels) was altered. If vowel recognition is determined primarily by temporal cues, then the altered stimuli should be well identified. If, however, vowel recognition is determined principally by information in the frequency domain, then the altered stimuli should not be well identified.

A. Method

1. Subjects

Eight subjects participated in the tests with the Iowa vowels and synthetic vowels. Six of the eight participated in the test with the multitalker vowels. The number of subjects varied solely as a function of availability for testing. All of the subjects had participated in experiment 1.

2. Speech materials

The three tests of vowel recognition used in experiment 1 were used in this experiment.

3. Signal processing

All stimuli were first processed through the 4-channel, noise-band processor described in experiment 1. To create stimuli with an altered representation of frequency the energy level of channel 1 was mapped onto channel 4, the energy level of channel 2 was mapped on channel 3, the

level of channel 3 was mapped onto channel 2, and the level of channel 4 was mapped onto channel 1. Vowel length was left unchanged.

4. Procedures

Following testing with the materials in experiment 1 the subjects were tested with the frequency altered stimuli. The patients were given the familiarization and practice with the altered 4-channel stimuli in the manner of the ''normal'' stimuli described in experiment 1. In the test sequence only the frequency-altered stimuli were presented. The order of testing for the Iowa vowels, the synthetic vowels and the multitalker vowels varied among the subjects in quasi-random fashion. Since all subjects had participated in experiment 1 the scores from that condition were used for the ''normal'' 4-channel scores.

Responses were collected in the same manner as in experiment 1.

B. Results and discussion

The averaged identification scores for the Iowa vowels, synthetic vowels, and multitalker vowels in the normal and frequency-altered stimulus conditions are shown in Fig. 5. For the Iowa vowels the mean score for the normal 4-channel stimuli was 76% correct. The mean score for the altered vowels was 35% correct. The two scores differed significantly [$t(7) = 6.8$, $p < 0.0002$]. For the synthetic vowels the mean score for the normal 4-channel stimuli was 54% cor-
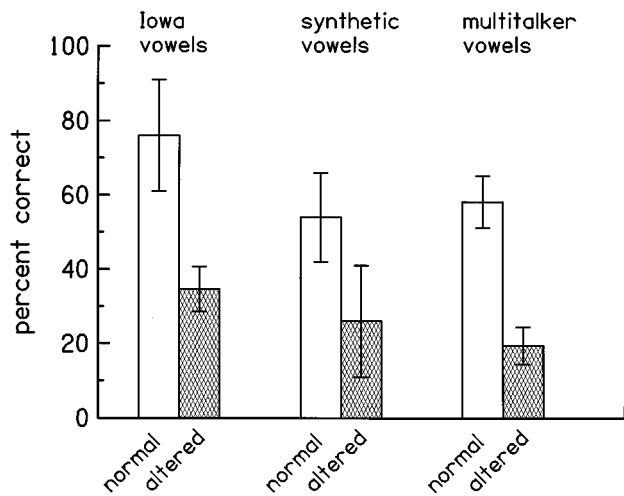
FIG. 5. Percent correct identification as a function of stimulus condition (normal or frequency altered) for three sets of 4-channel, noise-band vowels. The error bars indicate ±1 standard deviation.

rect. The mean score for the altered vowels was 26% correct. The two scores differed significantly [$t(7) = 4.75$, $p < 0.001$]. For the multitalker vowels the mean score for the normal 4-channel stimuli was 58% correct. The mean score for the altered vowels was 19% correct. The two scores differed significantly [$t(5) = 9.71$, $p < 0.0006$]. Overall, our result, i.e., the significant and large drop in performance when frequency information was altered, suggests that information in the frequency domain is the principal factor determining the identification of vowels when vowels are processed and presented through a small number of channels. We suspect that this conclusion extends also to the identification of consonants. It is reasonable to suppose that the amplitude envelope in a channel specifies *when* energy is *where* in frequency space and it is the frequency domain information, *no matter how sparse*, on which the recognition routines for speech operate. There is no reason to believe that the nature of the recognition process changes in a fundamental fashion when the number of channels of stimulation becomes small. From this point of view, interest in the results of Shannon *et al.* (1995) and Hill *et al.* (1968) stems from a consideration of how sparse the representation of frequency can be and still support high levels of speech recognition.

In the forgoing discussion we have used the term "temporal cue" in the fashion commonly used to describe the acoustic cues for vowel and consonant identification, i.e., a portion of an acoustic signal which, when varied in duration, alters phonetic identification. However, the term "temporal cue" could also be used in another sense. As noted above, when speech signals are reduced to a small number of bands, the amplitude envelope in a channel specifies *when* energy is *where* in frequency space. The within- and across-channel changes in amplitudes over time specify changes in formant frequencies and, thus, are "temporal cues" for speech recognition.

III. CONCLUSIONS

The results of the present investigation indicate, as previous investigations have suggested, that high levels of speech understanding can be obtained using signal processors with a small number of channels. The nature of the output signal, noise bands or sine waves, makes only a small difference in performance. The number of channels needed for high levels of performance varies with the nature of the test material. For the most difficult material—vowels produced by men, women and girls—8 channels were necessary to approach asymptotic performance. For the least difficult material—sentences—5 channels were sufficient. We suggest that the mechanism mediating the high levels of speech recognition achieved with only few channels of stimulation is the same one that mediates the recognition of signals produced by speakers with a high fundamental frequency, i.e., signal levels in adjacent channels are used to estimate the frequency of the input signal. Finally, our results suggest that vowel recognition is based principally on information in the frequency domain even when the number of channels of stimulation is small.

[1]The 63-year-old subject performed as well as the younger subjects.

[2]A reviewer has pointed out that a conservative *post hoc* test would require that two mean scores be very different to find a difference between means. This would work against finding differences between mean scores when the number of channels is fairly large because performance is near asymptote. This, in turn, would lead to a conclusion of fewer, rather than more, channels being needed to achieve high levels of speech identification. The Scheffe test was chosen because it controls the experimentwise, or overall, error rate. Another option would be to use a *post hoc* test that uses an error rate that is comparisonwise. To see if this option would make a difference in outcome, we reran the *post hoc* tests with Fisher's LSD—a comparisonwise test. While there were some differences in outcome, there were no differences which affected the issue of asymptote.

Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**). "Some experiments on the perception of synthetic speech," J. Acoust. Soc. Am. **24**, 597–606.

Delattre, P., 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 patterns," Word **8**, 195–210.

Dorman, M., Dankowski, K., McCandless, G., and Smith, L. (**1989**). "Identification of synthetic vowels by patients using the Symbion multichannel cochlear implant," Ear Hear. **10**, 40–43.

Hill, J., McRae, P., 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.

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

House, A. (**1961**). "On vowel duration in English," J. Acoust. Soc. Am. **33**, 1174–1178.

Liberman, A., Delattre, P., and Cooper, F. (**1958**). ''Some rules for the distinction between voiced and voiceless stops in initial position,'' Lang. Speech **1**, 153–167.

Liberman, A., Delattre, P., Gerstman, L., and Cooper, F. (**1956**). ''Tempo of frequency change as a cue for distinguishing classes of speech sounds,'' J. Exp. Psychol. **52**, 127–137.

Nilsson, M., Soli, S., and Sullivan, J. (**1994**). ''Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and noise,'' J. Acoust. Soc. Am. **95**, 1085–1099.

Peterson, G., and Barney, H. (**1954**). ''Control methods used in a study of vowels,'' J. Acoust. Soc. Am. **24**, 175–184.

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

Sommers, M., and Kewley-Port, D. (**1996**). ''Modeling formant frequency discrimination of female vowels,'' J. Acoust. Soc. Am. **99**, 3770–3781.

Tyler, R., Preece, J., and Tye-Murray, N. (**1986**). ''The Iowa audiovisual speech perception laser videodisc,'' *Laser Videodisc and Laboratory Report*, Department of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City, IA.

Zue, V. (**1985**). ''The use of speech knowledge in automatic speech recognition,'' Proc. IEEE **73**, 1602–1615.