Temporal cues for consonant recognition: Training, talker generalization, and use in evaluation of cochlear implants

Dianne J. Van Tasell, Donna G. Greenfield, Joelle J. Logemann, and David A. Nelson
Department of Communication Disorders, University of Minnesota, Minneapolis, Minnesota 55455

(Received 24 September 1991; revised 11 March 1992; accepted 5 May 1992)

Limited consonant phonemic information can be conveyed by the temporal characteristics of speech. In the two experiments reported here, the effects of practice and of multiple talkers on identification of temporal consonant information were evaluated. Naturally produced /aCa/ disyllables were used to create "temporal-only" stimuli having instantaneous amplitudes identical to the natural speech stimuli, but flat spectra. Practice improved normal-hearing subjects' identification of temporal-only stimuli from a single talker over that reported earlier for a different group of unpracticed subjects [J. Acoust. Soc. Am. **82**, 1152–1161 (1987)]. When the number of talkers was increased to six, however, performance was poorer than that observed for one talker, demonstrating that subjects had been able to learn the individual stimulus items derived from the speech of the single talker. Even after practice, subjects varied greatly in their abilities to extract temporal information related to consonant voicing and manner. Identification of consonant place was uniformly poor in the multiple-talker situation, indicating that for these stimuli consonant place is cued via spectral information. Comparison of consonant identification by users of multi-channel cochear implants showed that the implant users' identification of temporal consonant information was largely within the range predicted from the normal data. In the instances where the implant users were performing especially well, they were identifying consonant place information at levels well beyond those predicted by the normal-subject data. Comparison of implant-user performance with the temporal-only data reported here can help determine whether the speech information available to the implant user consists of entirely temporal cues, or is augmented by spectral cues.

PACS numbers: 43.71.Es, 43.66.Ts, 43.71.Ky

INTRODUCTION

A classification system for the temporal structure of speech proposed by Rosen (1989) includes three categories of speech cues, based on the periodicity of the dominant temporal structure in the category. In his scheme "envelope" cues exist from 2–50 Hz, and include acoustical aspects of phonetic segments such as duration and rise–fall time that can convey consonant manner and voicing information, vowel duration information, and prosodic information concerning syllabification and stress. "Periodicity" cues from 50 to 500 Hz signal the presence (and frequency) of vocal fold vibration, and therefore can convey information about consonant voicing and manner, as well as prosodic information about intonation and stress. The periodicity of temporal "fine-structure" information is much higher: from 0.6–10 kHz. These fine-structure aspects of the speech waveform carry information related to the spectral distribution of energy in vowels and consonants, and are commonly referred to as "spectral" cues for consonant place and vowel quality.

The actual availability to the central nervous system of temporal speech information depends, of course, on the temporal-resolving power of the auditory system. Normal-hearing subjects cannot detect amplitude modulation at modulation rates above about 1000 Hz (Viemeister, 1979); if we take this frequency to be the upper limit of temporal resolution for the speech cues described above, then it must be concluded that Rosen's "envelope" and "periodicity" cues are available to normal-hearing subjects, but that the bulk of temporal "fine-structure" cues are not.

In our earlier work (Van Tasell *et al.*, 1987) we removed the spectral information from /aCa/ disyllables by modulating pink noise with the envelope of the speech.[1] Subjects' identification of these "temporal-only" stimuli reflected the temporal-resolving limits discussed above. That is, performance increased from envelope bandwidth of 20 to 200 Hz, but did not improve further when envelope bandwidth was increased to 2000 Hz. At envelope bandwidths of 200–2000 Hz, subjects classified consonants into groups homogeneous with regard to voicing, and almost so with regard to consonant manner. They did not, however, seem to be able to distinguish consonants on the basis of place of articulation.

The objective of the work begun in the Van Tasell *et al.* study was to simulate for normal-hearing subjects the information content (not, it is important to note, the actual sound) of the speech signal received by users of a single-channel cochlear implant. A subsequent study by Rosen *et al.* (1989) confirmed that users of the single-channel House/3M cochlear implant device categorized consonant stimuli in the same way that the Van Tasell *et al.* subjects had: they seemed to be able to separate stimuli into perceptual categories related to consonant manner and voicing, but were unable to extract consonant place information from the speech signal. The similarities between the normal-subject temporal-only data of Van Tasell *et al.* (1987) and the

 0001-4966/92/091247-11$00.80

House/3M data of Rosen *et al.* (1989) encouraged us to pursue our simulation approach, but there remained many unanswered questions.

The central question arose from the large variance reported among implant users in their abilities to understand speech (even among users of the same implant device). Could the use of temporal cues account for the whole range of observed performance in users of single-channel cochlear implants? Before that question could be addressed, another had to be answered. Was the categorization performance of the normal subjects of Van Tasell *et al.* (1987) limited to the single talker used in that study, or would their temporal categories generalize to performance with multiple talkers? Our single-channel simulation data had been obtained from subjects who were listening to stimuli derived from the speech of only one talker, and who had received no feedback regarding the correctness of their responses. The results therefore did not reflect either the best performance an individual might achieve, or the performance that could be expected from an individual listening to the speech of multiple talkers.

The experiments reported here were undertaken in order to define the range of consonant recognition performance that might be expected from subjects listening to speech that contained only temporal information. The effects of practice with one talker, and the effects of additional talkers, were assessed. The results were compared with consonant-recognition data obtained by us and with data reported in the literature for users of various cochlear implant devices. These comparisons, when made appropriately, can be useful in determining whether the speech information conveyed to a multi-channel implant user is functionally single- or multi-channel, and in elucidating the type of speech information being transmitted to the user.

I. EXPERIMENT I: EFFECTS OF PRACTICE

In this experiment, stimulus conditions similar to those of Van Tasell *et al.* (1987) were reproduced. The temporal-only stimuli were derived using a different, all-digital, technique that resulted in more accurate representation of the speech temporal characteristics. The bandwidths of the stimulus sets were adjusted to provide: (1) only temporal information below 20 Hz, as in the Van Tasell *et al.* study, (2) temporal information below 150 Hz, which included the fundamental frequency of the talker but virtually omitted first-formant information, and (3) as much temporal detail as possible (limited only by the 4200-Hz anti-aliasing low-pass filter).

"Practice" in these experiments consisted only of providing correct-answer feedback; no other specific training techniques were employed. Rather, subjects were allowed to use the feedback in whatever way they found most useful.

A. Methods

1. Subjects

Twelve normal-hearing subjects participated in the experiment; each had pure-tone air conduction thresholds 15 dB HL (ANSI, 1989) or better at octave frequencies from 125–4000 Hz.

2. Stimuli

The unprocessed stimuli were the same as those used by Van Tasell *et al.* (1987). They consisted of 19 /ɑCɑ/ disyllables (C = /p,t,k,b,d,g,f,θ,s,ʃ,v,ð,z,ʒ,m,n,r,l,j/) spoken by a male talker, digitized with 12-bit resolution at a sampling rate of 10 kHz. Three different sets of processed stimuli were created using the technique described by Schroeder (1968), in which each sample of the digitized signal is multiplied by either $+1$ or -1, with equal probability. The result is a noise with a flat spectrum (therefore containing no spectral information), but instantaneous amplitude identical to that of the signal. We will, as suggested by Rosen (1989), refer to the processed stimuli as "signal-correlated noise" (SCN). Three sets of SCN, each with differing temporal detail in the waveform, were created.

a. Unfiltered SCN. The SCN was derived from the digitized speech stimuli, as described above.

b. 150-Hz SCN. The 150-Hz envelope of the speech signal was derived by full-wave rectifying it and then digitally low-pass filtering it (third-order elliptical filter) at a cutoff frequency of 150 Hz. Each envelope sample was then multiplied by either $+1$ or -1 to produce the SCN.

c. 20-Hz SCN. This was obtained as described for the 150-Hz SCN, except that a digital filter with a cutoff frequency of 20 Hz was used to derive the speech waveform envelope.

The effects of the signal processing can be seen in Fig. 1, which shows the waveform of the unprocessed /ɑkɑ/, as well as those of the three processed versions derived from it. The periodic temporal structure related to the vocal fundamental frequency is clearly visible in the unprocessed, the unfiltered SCN and the 150-Hz SCN waveforms. It is absent in the 20-Hz SCN waveform, which preserves only slow periodicity related to word and syllable structure. No matter what the envelope bandwidth, the envelope of the speech waveform is fully "filled" with noise, effectively eliminating the interaction of the speech envelope modulator that can occur with the amplitude fluctuations characteristic of analog (e.g., Gaussian or pink) noise carriers.

The level of each stimulus set was expressed as the level of a steady-state random noise with rms amplitude equal to the average rms amplitude of the stimuli in the set. The nominal stimulus presentation level was 75 dB SPL, as measured at the TDH-49 earphone in an NBS 9A coupler. To prevent the small differences in rms amplitude among members of each stimulus set from providing loudness cues, the actual level of the stimulus was varied randomly from trial to trial in 1-dB steps, within a 6-dB range (nominal level ± 3 dB).

3. Procedures

Stimuli were output by a laboratory computer via a 12-bit D/A converter at 10 kHz, low-pass filtered at 4200 Hz (96 dB/octave), amplified, attenuated, and delivered to a transformer and then to a TDH-49 earphone in an MX-41/AR cushion. The subject was seated in a sound-attenuating booth in front of a video screen on which were displayed the 19 /ɑCɑ/ alternative responses, along with a sample word containing the target consonant sound for each. The subject

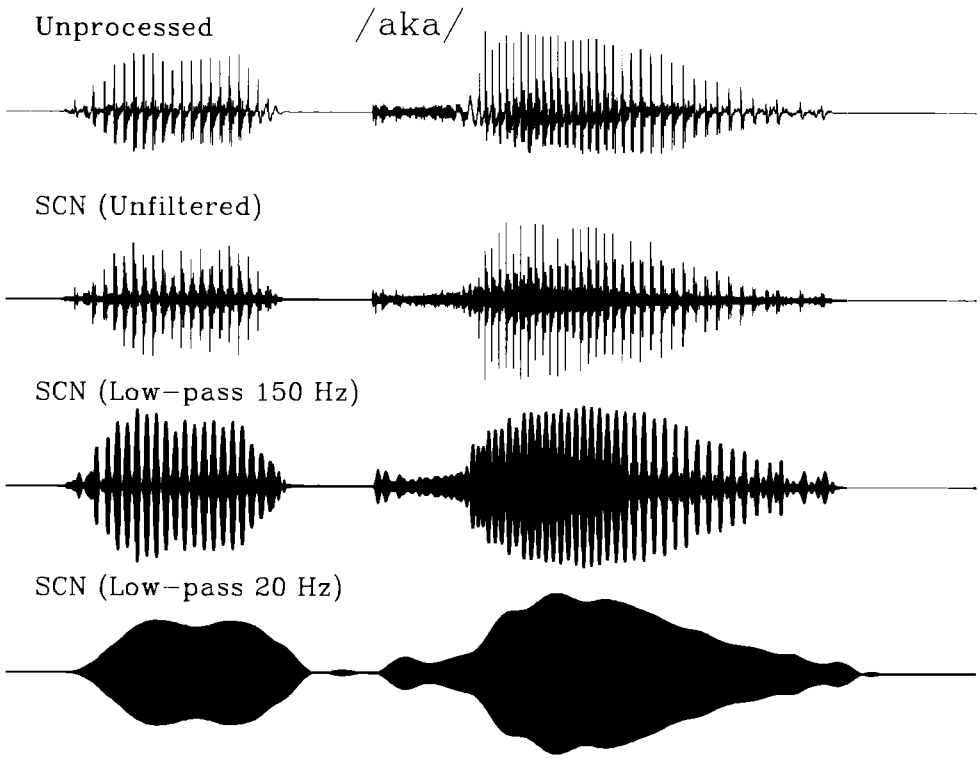

FIG. 1. Original (unprocessed) waveform of /ɑkɑ/, spoken by the male talker of experiment I, and the three signal-correlated-noise (SCN) stimuli derived from it. Total stimulus duration = 1152 ms.

used a computer mouse to highlight and select his/her response on each trial. Correct-answer feedback was provided after each trial by a line of asterisks that appeared next to the correct response on the video screen.

Each block of trials consisted of five randomly-ordered presentations of each of the 19 stimuli in a set, for a total of 95 trails per block. Each trial began with a 500-ms warning displayed on the video screen, followed by stimulus presentation, and a prompt to the subject requesting a response. After the subject responded and correct-answer feedback was provided, a new trial began. Responses were stored in the form of 19-by-19 confusion matrices. Stimulus presentation, timing, subject video display, and response collection were computer-controlled.

a. Order of stimulus conditions. All subjects completed the unprocessed condition first, in order to familiarize them with the task and the stimulus set. The order of the three SCN conditions was counterbalanced across the 12 subjects. Before each hour-long listening session subjects were familiarized with the stimuli by allowing them to request (via computer mouse) and listen to each stimulus item as many times as desired in a 5-min familiarization period.

b. Practice. During the first session the subject listened to two blocks of the unprocessed stimuli. In each subsequent session subjects listened to one of the three SCN conditions; four blocks were presented per session. Subjects continued testing with a specific stimulus set until performance had stabilized or until the subject had been tested with the set for four sessions (a total of 16 blocks). Stable performance was demonstrated by meeting one of two criteria: (1) mean envelope-feature correct identification (according to the envelope-feature categories of Van Tasell *et al.*, 1987) of 90% or greater, and a standard deviation of 3% or less across the four blocks of the first session; or (2) two successive sessions in which mean envelope-feature correct scores had a combined standard deviation across both sessions (eight blocks) of 4% or less and a standard deviation of 3% or less across the four blocks of the last session.

The data used for the analyses that will be reported here are only those from each subject's final session with each stimulus set (20 observations per item per subject).

B. Results

The mean percent correct consonant recognition across all subjects for the unprocessed stimuli was 97.8%, with standard deviation of 1.8%. The stimuli were, therefore, highly intelligible in their unprocessed state.

1. Practice

One of the 12 subjects did not meet the performance criterion for any of the stimulus sets after four sessions with each. All other subjects met criterion for all stimulus sets in 1–4 sessions. From the data in Table I it can be seen that subjects took longer, on the average, to meet the criterion for the 20-Hz SCN stimuli than for the other two processed stimulus sets. There was also a stimulus order effect: subjects needed an average of 3.25 sessions to reach criterion on the first processed stimulus condition, but a decreasing number of sessions on the second and third sets.

TABLE I. Mean number of sessions (four blocks per session) to training criterion in experiment I.

	Stimulus condition		
	Unfiltered	LP 150	LP 20
Mean	2.41	1.91	3.50
Standard deviation	1.08	1.16	1.00
	Stimulus order		
	First	Second	Third
Mean	3.25	2.83	1.75
Standard deviation	0.87	1.27	1.14

2. Information transfer analyses

Information transfer calculations were performed on the confusion data of each individual subject. The average results for three separate transmitted-information quantities are shown in Table II.

a. Relative transmitted information for stimulus. This quantity was calculated as described by Miller and Nicely (1955). Transmitted stimulus information is a measure of the information transmitted from stimulus to response, in bits per stimulus. Relative transmitted information (RTI) relates the amount of transmitted information to the information contained in the stimulus set (in this case of 19 stimuli occurring equally often, the available stimulus information is 4.248 bits). It can be thought of as the proportion of available stimulus information that was transmitted to the subject. The quantity reported here is RTI multiplied by 100, thus expressed as a percentage.

A repeated-measures analysis of variance (ANOVA), with SCN condition as the within-subjects variable, confirmed that average RTI for stimulus differed significantly across stimulus conditions; post-hoc testing showed that the mean of the 20-Hz SCN data differed significantly ($p < 0.05$) from those of the other two conditions. In all conditions, the subjects' performance was substantially better than the 22%–35% reported by Van Tasell *et al.* (1987) for similar signals derived from the same speech stimuli. Because the SCN stimuli were derived differently from the speech-envelope noise signals used in that earlier study, however, it is not possible to conclude that the differences resulted entirely from practice.

b. Relative transmitted information for envelope. In Van Tasell *et al.* (1987) each of the three dimensions of the multidimensional scaling solution was treated as a separate envelope feature. Those three features uniquely defined four "envemes," or sets of consonants, within which confusions were frequent, but among which confusions were infrequent. In the analyses reported here, we have defined "envelope" as a single feature having four categories coincident with the "envemes" of Van Tasell *et al.* (see Table III for feature category membership of the stimuli). A subject's ability to classify /aCa/ stimuli correctly into the four "envelope" categories (as reflected in RTI for the envelope feature) is taken as a measure of his/her ability to extract and use temporal properties of the speech waveform that carry consonant information.

The data in Table II show that subjects were remarkably good at classifying the stimuli according to the four envelope categories. Repeated-measures ANOVA and subsequent post-hoc testing showed that the 20-Hz SCN mean was significantly lower than the other two.

c. Relative transmitted information for place. When phoneme confusion data are analyzed according to more than one feature, it is important that the effects of overlap among features (i.e., similarities in feature category membership of the stimulus items) be factored out. Sequential information analysis (SINFA; Wang, 1976) is a method for performing feature information transfer analyses while holding constant the effect of previously evaluated features; it was used to calculate RTI for place while holding constant the effects of the envelope feature. Table III shows the classification of stimuli into the three broad place categories of Pickett (1980). It should be noted that the envelope and place features used here are already relatively independent: that is, there is little duplication of category membership. That being the case, the envelope feature effects on RTI for place were minimal (i.e., the conditional RTI for place was never very different from unconditional RTI for place). This would not be the result, however, for a feature such as voicing, which is closely related to envelope.

The data in Table II show that conditional RTI for place was not different across stimulus conditions, and was considerably lower than RTI for the envelope feature. Nevertheless, there was some place information being transmitted to these subjects.

TABLE II. Average SINFA data for the 12 subjects of experiment 1 in each stimulus condition. Standard deviations in parentheses. The F ratios are results of repeated-measures ANOVAs. RTI = relative transmitted information.

	Stimulus condition			
	Unfiltered	LP 150	LP 20	$F(2,22)$
RTI stimulus	57.98 (7.71)	55.05 (6.01)	51.28 (5.54)	4.97[a]
RTI envelope	79.09 (11.64)	73.18 (12.88)	60.61 (12.08)	14.16[b]
RTI place (conditional)	19.71 (13.52)	19.72 (10.73)	16.53 (5.98)	0.75

[a] $p < 0.05$.
[b] $p < 0.01$.

TABLE III. Categorization of stimuli under envelope and place features for SINFA analyses.

	Feature	
Category #	Envelope	Place
1	/p,t,k/	/p,b,m,f,v/
2	/b,d,g,v,ð,z,ʒ/	/t,d,n,ð,θ,z,s,l,j/
3	/f,θ,s,ʃ/	/g,k,r,ʒ,ʃ/
4	/m,n,r,l,j/	

C. Discussion

Stable performance on a stimulus set was usually achieved within four hours of practice. Even though they received no specific training to do so, subjects classified the stimuli, with a high level of accuracy, into the four envelope categories defined by Van Tasell *et al.* About 20% of the place information available in the stimuli was also transmitted to subjects.

As reported by Van Tasell *et al.* (1987), performance on the 20-Hz envelope condition generally was poorer than that observed at wider envelope bandwidths. A likely explanation for this is that the voice fundamental frequency information (available in the unfiltered and 150-Hz SCN condition) was used by subjects to help sort stimuli into the correct envelope categories. The envelope-category breakdown in Table III shows that the members of each envelope category are homogeneous with regard to voicing. Therefore, the fundamental frequency periodicity cue to voicing would also serve as a useful partial envelope-categorization cue. It would not augment correct categorization of stimuli into place categories, however; consistent with this interpretation, subjects performed uniformly across the SCN conditions on the place feature.

To aid interpretation of the results of this experiment, we combined the data from the two conditions of Van Tasell *et al.* (1987) shown in their report to be statistically equivalent (200- and 2000-Hz envelope bandwidths) and re-analyzed them with the feature set described above. The results are shown in Fig. 2, along with the combined data from the 150-Hz and unfiltered SCN conditions of the present experiment (the third data set depicted in Fig. 2 will be discussed in connection with experiment II). It can be seen that the net effects of practice and the digital processing SCN technique were to improve the average RTI for stimulus and envelope, without reducing the variance. This was an unexpected finding; we had anticipated that practice would make the subjects perform more uniformly than in the earlier study. On the other hand, both RTI for place and its associated variance increased (although the apparent increase in variance is due in part to the lower variance associated with the very low percentage scores in the data of Van Tasell *et al.*).

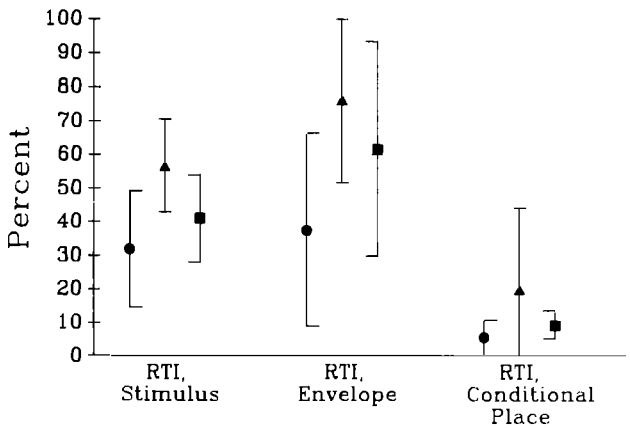

FIG. 2. Mean ± 2 standard deviations of SINFA quantites. Filled circles = one-talker, unpracticed-subject data of Van Tasell *et al.* (1987). Filled triangles = one-talker, practiced-subject data of experiment I. Filled squares = six-talker, practiced-subject data of experiment II.

II. EXPERIMENT II: EFFECTS OF MULTIPLE TALKERS

The purpose of this experiment was to evaluate the effects of multiple talkers on subjects' identification of the 19 /aCa/ temporal-only syllables. In this experiment we used only unfiltered SCN, results from which will be compared directly to the combined unfiltered and 150-Hz data of experiment I, which were not significantly different from one another. Rosen (1989) reported that the use of half-wave rectification in deriving the speech envelope improved his two subjects' performance on consonant identification over that observed for SCN derived from full-wave rectified speech, possibly because the lower "temporal density" of the half-wave rectified speech enhances auditory resolution of the temporal structure of the signal. In order to investigate this possibility, we tested subjects with SCN stimuli created using both half-wave and full-wave rectified speech.

A. Methods

1. Subjects

Twelve different normal-hearing subjects were recruited for this experiment. Their hearing characteristics were the same as those of the subjects in experiment I.

2. Stimuli

Two male talkers and three female talkers each recorded a set of the 19 /aCa/ stimuli described in experiment I. Their utterances were digitized with 12-bit resolution at a sampling rate of 10 kHz. Together with the stimuli from the original male talker of experiment I, this provided a total of 114 tokens (19 stimuli $\times$ 6 talkers).

From each talker's 19 utterances, two sets of SCN stimuli were created. For one set, the speech waveform was full-wave rectified and the SCN created as described in experiment I. For the other, the SCN was created from the half-wave rectified speech waveform. Samples from the vowel portions of the two waveform types are shown in Fig. 3, along with the corresponding sample from the unprocessed waveform. In both SCN waveforms, the periodicity of the voice fundamental frequency is plainly discernible, but it seems to be more evident, at least visually, in the half-wave rectified speech waveform.

3. Procedures

Procedures were the same as in experiment I, except that a block of trials in experiment II consisted of one observation each for all 114 stimuli (compared to 5 repetitions of 19 stimuli = 95 observations in experiment I). Six of the subjects listened to the full-wave rectified stimuli first; the other six listened to the half-wave stimuli first. Performance criteria were the same as in experiment I.

B. Results

The mean percent correct consonant recognition for the unprocessed stimuli was 91.0% (s.d. = 3.2%). Although performance dropped some 7% compared to the one-talker

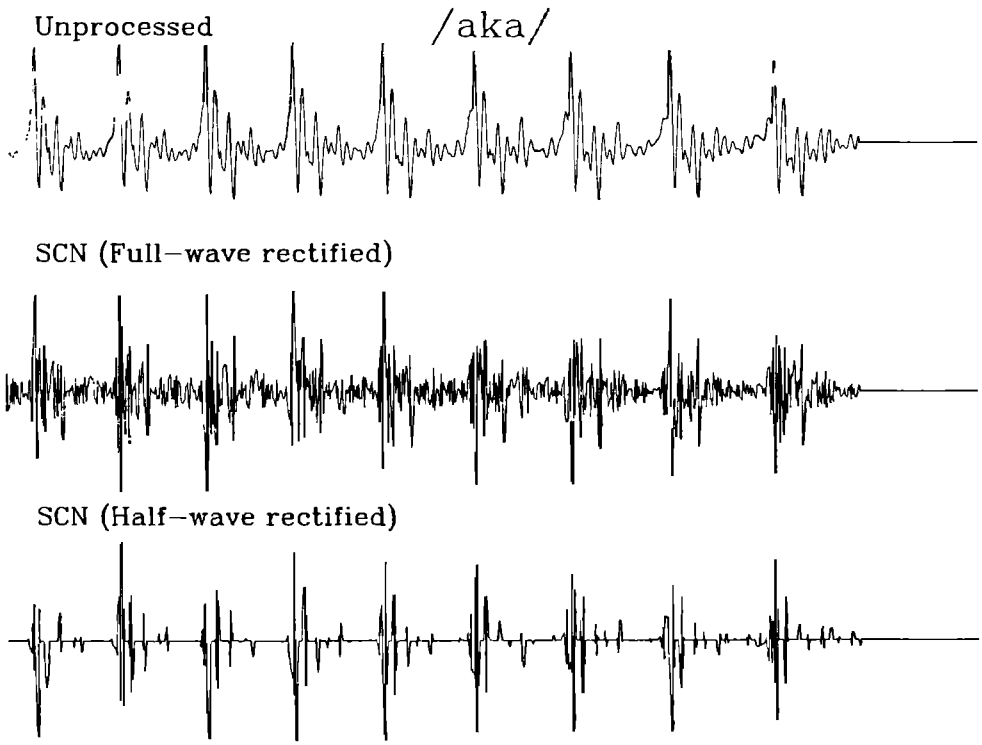

FIG. 3. The 90-ms segments from vowel portions /ɑkɑ/ waveforms (male talker) from experiment II.

stimuli of experiment I, the stimuli were still very intelligible.

1. Practice

One subject failed to reach criterion for either stimulus set; another two subjects did not reach criterion with the full-wave stimuli. Table IV shows the mean number of sessions subjects took to reach the performance criterion. There was no difference across subjects in the number of sessions to criterion between the full-wave and half-wave rectified stimuli. As in experiment I, there was an order effect, with the mean number of sessions for the stimulus set presented second being lower than for the stimulus set presented first.

2. Information transfer analyses

As in experiment I, each subject's confusion matrix was submitted separately to SINFA analysis. The results, averaged across subjects, are shown in Table V. There was no significant difference, for any of the three SINFA measures, between the full- and half-wave rectified stimuli. For all further analyses, therefore, the data were pooled across stimulus sets.

Examination of the individual matrices of many of the subjects made it clear that they were using envelope categories slightly different from those defined by Van Tasell *et al.* (1987). Specifically, they were confusing the voiced fricatives /v,ð,z,ʒ/ with the sonorants instead of the voiced stops; that is, the voiced fricatives had moved from envelope category 2 (see Table III) into envelope category 4. We analyzed the pooled data according to the original four envelope categories, and again with the modified envelope categories in which the voiced fricatives were moved into the category

TABLE IV. Mean number of sessions (four blocks per session) to training criterion in experiment 2.

	Stimulus condition	
	Full wave	Half wave
Mean	3.08	3.08
Standard deviation	0.86	0.86
	Stimulus order	
	First	Second
Mean	3.42	2.75
Standard deviation	0.76	0.83

TABLE V. Average SINFA data [original envelope categories] for the 12 subjects of experiment 2 in each stimulus condition. Standard deviations in parentheses. The t statistics are results of t tests for correlated samples. RTI = relative transmitted information.

	Stimulus condition		
	Full	Half	$t(11)$
RTI stimulus	40.13 (7.68)	41.95 (5.22)	1.20
RTI envelope	52.45 (16.73)	55.27 (13.57)	0.97
RTI place (conditional)	9.21 (2.15)	9.94 (2.54)	0.94

with the sonorants. The results are shown in Table VI.

Mean RTI for envelope was higher for the modified envelope-category analysis than for the original envelope categories; the same was true of the individual data of every subject. Furthermore, the same effects were not observed in the data of experiment I. When those data were re-analyzed according to the modified envelope categories, RTI for envelope actually declined.

C. Discussion

For these subjects and these stimuli, half-wave rectification in the signal generation process provided no advantage over full-wave, contrary to the observation of Rosen (1989). Given the large variance among subjects in our data, it is likely that Rosen's outcome was an artifact of the small number of subjects he tested. Still, it seems reasonable to expect that half-wave rectification would be beneficial with speech materials containing information (prosodic, for example) conveyed specifically by fundamental frequency.

The SINFA data calculated using the modified envelope categories are shown in Fig. 2, along with the data from experiment I, and the re-analyzed data from Van Tasell *et al.* (1987) for comparison. The RTI for stimulus data show that expanding the number of talkers reduced subjects' identification performance; the mean RTI for stimulus was 41.3%, significantly lower than the same quantity measured for the combined unfiltered and 150-Hz SCN data in experiment I [$t(46) = 5.10$, $p < 0.01$]. Practice was not sufficient to overcome this talker generalization effect. Similarly, the RTI for envelope data show that subjects on the whole did not identify envelope characteristics of stimuli as well as subjects listening to one talker with feedback provided, but they did perform better than subjects listening to the same talker without correct-answer feedback in the Van Tasell *et al.* (1989) study. The important aspect of the RTI for envelope data is the large variance, even for the practiced subjects. With feedback provided, subjects still varied greatly in their abilities to extract the consonant information contained in the temporal characteristics of speech.

The conditional RTI for place results are particularly striking. The low mean scores and the small variance in the Van Tasell *et al.* study, and in experiment II reported here, indicate that subjects could extract virtually no consonant place information. The 9% RTI for place corresponds to the 9% RTI for place that Rosen *et al.* (1989) measured with normal-hearing subjects listening to /aCa/ stimuli consisting only of fundamental frequency patterns. In experiment I, however, there were subjects whose conditional RTI for place was as high as 56%. It seems clear that in experiment I some subjects had learned to take advantage of idiosyncratic characteristics of the 19 stimuli in the set in order to classify the stimuli according to what appears to be consonant place in the confusion analysis. When stimuli from six talkers were used, the individual tokens could not be learned, and the subjects' "true" consonant place identification abilities were revealed.

The learning of individual stimuli was likely responsible also for the change in envelope categories from experiment I to II; there was something about the voiced fricatives produced by the single talker of experiment I that allowed them to be differentiated from the sonorants, whereas this did not occur with multiple talkers.

TABLE VI. Average SINFA data [full- and half-wave data pooled] for the 12 subjects of experiment II, analyzed according to the original and the modified envelope categories. Standard deviations in parentheses. RTI = relative transmitted information.

	Envelope categorization	
	Original	Modified
RTI stimulus	41.04 (6.49)	41.04 (6.49)
RTI envelope	53.86 (14.97)	61.50 (15.94)
RTI place (conditional)	9.58 (2.33)	9.07 (1.92)

III. APPLICATION TO EVALUATION OF COCHLEAR IMPLANTS

Assuming that the data of experiments I and II define the range of performance that can be expected from subjects receiving only single-channel temporal speech information (e.g., users of single-channel, or functionally single-channel, cochlear implants), then comparison of our data with the performance of implant patients for similar VCV materials should provide some indication of the type of consonant information they are receiving. We tested four users of the Nucleus-22 cochlear implant with our stimulus set, and selected two other sets of consonant confusion data from the literature for analysis with the feature categories described in experiment II.

These anayses are not intended in any way as comparisons of the implant devices used by the different groups of subjects. Rather, they are intended to: (1) demonstrate how the SCN data obtained in experiments I and II might be used to support inferences about what consonant information is being conveyed to implant users, and (2) raise some important questions about what constitute appropriate comparisons between data obtained from normal-hearing subjects and implant users.

A. Patient testing

Four adult users (age 44–65) of the Nucleus-22 cochlear implant device were tested. Each subject used the Wearable Speech Processor with the $F0$–$F1$–$F2$ coding scheme (Blamey *et al.*, 1987). All subjects were regular participants in psychoacoustic experiments in the Hearing Research Laboratory at the University of Minnesota. They were selected only on the basis of availability, with no regard to their particular success as users of the device.

The stimuli were the *unprocessed*, one-talker stimuli of experiment I. Each subject listened to one practice block and

four test blocks of trials while seated in the sound booth with his/her speech processor in the settings normally used for everyday communication. The stimulus on each trial was presented twice in succession before a response was requested. Subjects did not receive correct-answer feedback. Otherwise, testing was performed as it was in experiment I. Each subject's data consisted of the summed matrices from the four test blocks, for a total of 20 observations per stimulus. Individual subject data were submitted to SINFA analysis, with the original envelope categories of experiment I.[2]

Data of the individual Nucleus patients are denoted by the Ns in Fig. 4. For comparison, the means, ± 2 standard deviations, for the combined 150-Hz and unfiltered SCN data of experiment I are shown on the left. With two exceptions, all the individual data points lie within the ranges of temporal-only performance expected on the basis of the SCN data. The two "outlier" data points (for RTI stimulus and envelope) were from the same patient: her limited ability to use envelope information reduced her ability to identify the items. The place data are especially revealing. Even though these subjects' place performance was relatively good, it was still not better than could be achieved by a trained subject with temporal-only versions of the same stimuli. The comparison supports the conclusion that the performance of these four patients, even though each wore a multi-channel cochlear implant, was not better than that theoretically obtainable with a single-channel implant for these speech materials.

B. Comparisons with published data

Two sets of confusion matrices were selected for comparison with the data of experiment II. In both instances group rather than individual data had been presented, so the *group* matrices were analyzed. It should be noted that this group analysis probably underestimates the individual performance achieved by most of the subjects. On the other hand, both sets of data were obtained using 16-item sets; performance with this smaller set may be inflated relative to what it might have been with our 19-item set. Nevertheless, the technique of information analysis does allow us to make limited direct comparisons of our data with these published implant data.

Wilson *et al.* (1990) tested seven users of the multichannel Ineraid cochlear implant *who had been selected for their unusually good speech recognition abilities.* Each subject was tested while wearing one of two speech processors: (1) a standard four-channel "compressed analog" (CA) processor, and (2) an experimental, five- or six- channel "continuous interleaved sampler" (CIS) processor (Wilson *et al.*, 1991). The CIS processor was designed to minimize interaction of information on separate electrode channels, thereby maximizing reception of channel-related (i.e., place) cues.

Subjects listened to a set of 16 recorded /aCa/ stimuli spoken by multiple talkers; no correct-answer feedback was given. Because multiple talkers had been used, group confusion matrices for the CA and the CIS processors were submitted to SINFA analysis using the modified-envelope categories of experiment II.[3] The results also are shown in Fig. 4, labeled "CA" and "CIS." For comparison with these data, the means ± 2 standard deviations of the multiple-talker data of experiment II (analyzed with the modified-envelope categories) are shown in the right brackets. Group performance with both processors was within the predicted range for envelope, but higher for stimulus. This is a result of the much-better-than-predicted RTI for place, which may be taken as an indication that: (1) both processors (particularly the CIS) were providing consonant place information beyond that which could be obtained from solely single-channel temporal speech information; and (2) these exceptionally successful Ineraid users were able to make use of that information.

Dorman *et al.* (1990) tested 10 users of the Ineraid cochlear implant (standard processor) with a set of 16 /aCa/ syllables similar to those used in experiment I. There were eight tokens of each stimulus, each uttered by the same male talker. Subjects did not receive correct-answer feedback. The authors provided separate confusion matrices for a set of three "better"-performing subjects (their Table II, p. 2076) and a set of seven "poorer"-performing subjects (their Table III, p. 2077). We submitted both sets of data to SINFA analysis, using the modified envelope categories of experiment II.[4] Results are indicated in Fig. 4 by "IB" for the better subjects, and "IP" for the poorer subjects.

The results highlight the importance of choosing the proper temporal-only comparison. Should the single-talker (but multiple-token) data of Dorman *et al.* be compared to the single-talker data of experiment I, or the multiple-talker data of experiment II? This decision is crucial to the interpretation of the RTI for place data: if the one-talker data are the correct comparison, then only the IB subjects' performance was better (and then only slightly better) than could be achieved with temporal-only stimuli. If the multiple-talk-

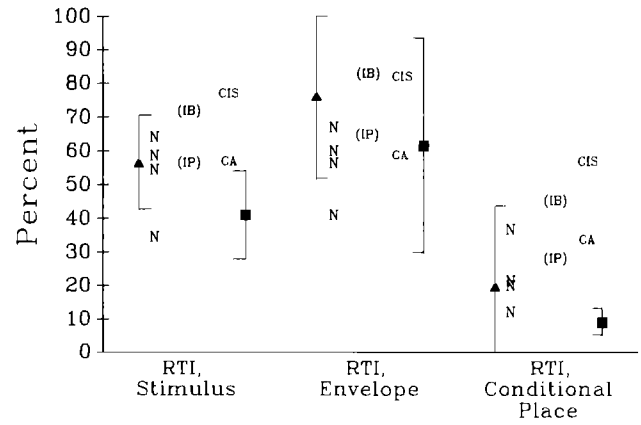

FIG. 4. Mean ± 2 standard deviations of SINFA quantities from trained subjects of experiment I (filled triangles) and experiment II (filled squares). N = data from individual Nucleus implant users tested with unprocessed experiment I stimuli. CIS and CA = group data of Wilson *et al.* (1970) from 7 Ineraid cochlear implant users tested with a continuous interleaved sampler (CIS) and a compressed-analog (CA) experimental speech processer. IP and IB = group data of Dorman *et al.* (1990) from 7 "poorer" (IP) and 3 "better" (IB) users of the Ineraid implant. Data from experiment I and Nucleus users analyzed with original envelope categories, all others with modified envelope categories.

er data are used, then Place performance clearly suggests that both groups of subjects were extracting spectral information from the signal.

IV. DISCUSSION

The results of experiments I and II support the following conclusions about the consonant information carried by certain temporal characteristics of speech:

A. Envelope bandwidth

As shown in previous studies, there is consonant information at envelope bandwidths of 150–200 Hz that is not present in stimuli with 20-Hz envelope bandwidth. Furthermore, increases in bandwidth beyond 150–200 Hz do not improve subjects' performance significantly, even with practice. The obvious implications are that: (1) voice fundamental frequency can be extracted from temporal-only representations of speech (providing it has not been removed during the signal processing) and used to support consonant phonemic identification; and (2) higher-frequency temporal information in the speech waveform cannot be used for consonant identification.

B. Rectification

Decreasing the temporal density of the SCN waveform by using half instead of full-wave rectification in its derivation did not enhance subjects' performance. A different outcome might be observed for speech materials containing information conveyable only by voice fundamental frequency.

C. Envelope information

The envelope categories defined by Van Tasell *et al.* (1987) using one talker and unpracticed subjects changed only minimally when six talkers and practiced subjects were used. We take this to indicate that there is something about the temporal structure of the waveforms of individual category members that causes listeners to perceive them as similar. (It is important to note that one would expect this behavior only from listeners with well-developed English phonological systems.)

From the envelope category membership it can be inferred that voicing (defined here as the presence or absence, and duration of, periodicity at the voice fundamental), makes an important contribution to envelope perception. Note that voicing is not the only cue that listeners can use, as demonstrated by the fairly good performance observed even at the 20-Hz envelope bandwidth; other complex durational and amplitude cues must also be important.

The findings of Rosen *et al.* (1989) attest to the validity of the envelope categories for description of consonant recognition by subjects with single-channel cochlear implants. They reported that a five-category voicing/manner feature with category members "voiced nasals, voiced fricatives, voiced plosives, voiceless plosives, and voiceless fricatives" accounted for 75% of the stimulus information transmitted to their subjects for a set of 12 /ɑCɑ/ stimuli; subjects were four users of the House/3M single-channel cochlear implant device.

D. Place information

When multiple talkers' utterances are included in the stimulus set, very little consonant place information can be extracted from the temporal characteristics of the /ɑCɑ/ stimulus set. This suggests that the bulk of the consonant place information in the speech waveform is spectral, and therefore requires some sort of amplitude-by-frequency representation.

The results reported in this study also provide the following information on the expected performance of subjects with temporal-only speech stimuli, and on the effects of various test parameters.

E. Subject variability

The subjects used in experiments I and II had normal hearing, and it can be assumed that their central auditory nervous systems were all receiving roughly the same information. Yet the variance among subjects in extraction and use of temporal speech information was very large, and did not diminish with practice. The only reasonable explanation is that cognitive factors underlie this variance: perhaps subjects differ substantially in their talents as extractors of information from minimal-cue stimuli. Whatever the exact nature of the cognitive differences, there is no reason not to expect them to occur, in equal magnitude, among recipients of cochlear implants. It seems likely, therefore, that the oft-reported variance among implant users in their abilities to understand speech are attributable as much to cognitive factors as they are to differences among devices or the physiological status of the electrically stimulated auditory system.

As Rosen *et al.* (1985) have noted, large variance among even trained normal subjects, such as that reported for the set of /ɑCɑ/ stimuli used here, makes the test of limited use for inferring the effectiveness of cochlear implant devices themselves. This is one reason that establishing the range of expected performance for a given type of speech information (in this case, single-channel temporal) is important: performance beyond the limits expected, in either direction, can provide useful insights into the speech information received by the listener.

F. Stimulus set

Comparison of experiment I with experiment II leads to the conclusion that the characteristics of individual members in a set of stimuli (even as many as 19) can be learned. This can lead, as shown in experiment II, to erroneous conclusions regarding the phonetic information available to listeners. Adding talkers (in this case, adding five more talkers) removes, or at least greatly reduces, the possibility that listeners can learn all the items in the set.

From the data reported here it is not possible to determine whether the "learnability" of the stimuli was a function of *talkers* or of *tokens*. If the stimulus characteristics that subjects learned to identify arose from the speech production characteristics peculiar to the talker, then increasing the number of tokens produced by a single talker might not reduce appreciably the learnability of the stimuli. If, on the other hand, subjects in experiment I were responding to

aspects of the individual items that would not be present in other tokens produced by the same talker, then increasing the number of tokens might be sufficient to insure that subjects' responses were generalizable across talkers. Until this is established empirically, data collected using one-talker stimulus sets should be interpreted with caution.

G. Practice

Subjects in experiment I performed much better than the unpracticed subjects of Van Tasell *et al.* (1987), with SCN derived from the same set of speech stimuli. To the extent that the differences can be ascribed to practice (rather than to differences in signal processing), practice can be said to have enhanced subjects' performance. In addition, the stimulus order effects observed in both experiments show that subjects could transfer their learning from one type of stimuli to another.

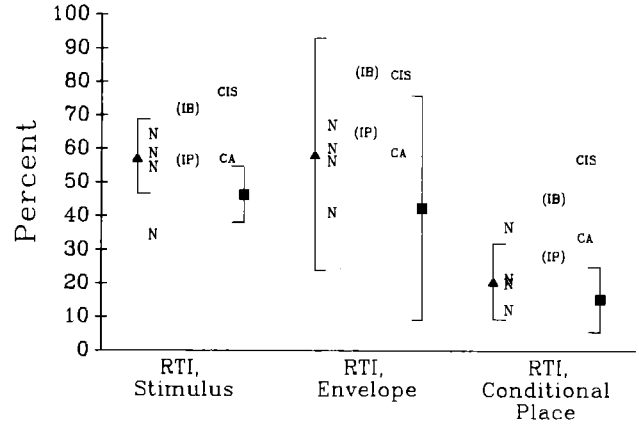

FIG. 5. Mean ± 2 standard deviations of SINFA quantities of minimally trained subjects from experiment I (filled triangles) and experiment II (filled squares). Cochlear implant user data indicated by letters, as in Fig. 4. Data from experiment I and Nucleus users analyzed with original envelope categories, all others with modified envelope categories.

H. Comparison with implant data

In order to prevent subjects from learning the speech materials, it is usually the case that correct-answer feedback is not provided during speech testing with cochlear implant users. For example, neither the subjects of Wilson *et al.* (1991) nor of Dorman *et al.* (1990) received feedback, and to facilitate comparisons with data from other implant users, we did not provide feedback to the Nucleus patients we tested. This practice does present some interesting problems for selection of the appropriate normal-subject data set with which to compare the implant data. Given the intensive pre- and post-implant testing undergone by most implant users, and their experience in using their implants for communication purposes, should they be considered to be "practiced" subjects? If they should, then the comparisons with the data of practiced normals in Fig. 4 are appropriate. But if the provision of correct-answer feedback has actually resulted in superior training for the normal-hearing subjects, then such a comparison places the implant users at a disadvantage.

In order to evaluate this possibility, subsets of the data from experiments I and II were selected. The data from only the first two blocks of the first SCN stimulus type heard by each subject were extracted; in this way, the learning effects across stimulus type were eliminated, and the learning effects within stimulus type were confined to only those that occurred within the first two blocks of trials. SINFA quantities were calculated for each subject on these minimal-practice data as described earlier. The means, plus and minus two standard deviations, are shown in Fig. 5, along with the same implant-subject data that appear in Fig. 4. If the Fig. 5 data from minimally trained normal subjects had been used for comparison, then conclusions about performance of two of the Nucleus patients would have been different: one would have been performing better than expected on the place feature, and one would have performed within the expected range on envelope rather than below it. Also, the performance of the subjects of Wilson *et al.* (1991) with the CIS processor would have been slightly better than expected on envelope, rather than within the expected range. All other conclusions about implant patients' performance would have remained the same. Nevertheless, future comparisons between normal and implant subjects' performance could be made with more confidence if investigators would begin using correct-answer feedback with multiple-talker nonsense stimuli when testing implant patients. This would resolve the uncertainty about what level of training implant patients should be assumed to have received. Furthermore, the data of experiment II suggest that subjects will not be able to learn the individual stimulus items when enough talkers are used.

The inevitable training differences between our normal-hearing subjects and implant users whose data can currently be found in the literature are a source of variance that makes strictly valid comparisons between the two groups impossible. To maximize the validity of comparisons with existing implant data: (1) only the normal-subject data obtained with *multiple talkers* should be used; (2) only data from implant users obtained with a *similar set of stimuli* should be evaluated, and (3) the data should be analyzed with the *modified envelope and conditional place* categories described in experiment II. When these guidelines are followed, an implant user's use of spectral information may be indicated when the subject performs beyond the expected range on RTI for conditional place. The range of expected performance on envelope is so large that data from most implant patients would fall within it; relative placement within the range would indicate how much temporal information the subject was extracting. If a subject falls below the expected performance range, then this may be an indication that the device is not functioning properly, that the subject does not have the physiological substrate necessary for effective use of the implant, or that he/she need some specific aural rehabilitation aimed at extraction and use of speech temporal cues.

Finally, it is importnat to emphasize that the SCN stimuli used in these experiments are useful simulations only of single-channel cochlear implants, because the instantaneous envelope fluctuations of the SCN are the same at all spectral locations. Although they can be used to help infer whether an implant patient is or is not functioning like a single-channel listener, they will be of limited use in determing how

speech information is conveyed by a multi-channel implant. That application will require stimuli composed of temporal speech properties extracted separately from various spectral regions and presented simultaneously, as they have been in experiments reported by Breeuwer and Plomp (1984;1986) and Grant and his colleagues (Grant *et al.*, 1985; Grant *et al.*, 1991).

ACKNOWLEDGMENTS

We gratefully acknowledge the assistance of the staff of the University of Minnesota Hearing Research Laboratory in obtaining the data from the Nucleus implant users. This work was supported by NIDCD DC00110 and by the Bryng Bryngelson Fund of the Department of Communication Disorders, University of Minnesota.

[1]Here the term "envelope" takes on a meaning different from the one defined by Rosen. In our general definition, "envelope" refers to the time-intensity characteristics of speech, obtained by rectifying and low-pass filtering the signal. By varying the low-pass cut-off frequency the bandwidth of the envelope can be varied.

[2]The unmodified envelope categories were used because the stimuli heard by the Nucleus patients were the same 19 stimuli from which the SCN stimuli of experiment 1 had been derived, and for which use of the unmodified envelope categories had been shown to be appropriate.

[3]The group matrices of Wilson *et al.* (1991) for both the CA and CIS processors were also analyzed with the original envelope categories. RTI for envelope improved only slightly: 1.9% (CIS) and 4.7% (CA). Use of the modified envelope categories was therefore considered appropriate for these data.

[4]The group matrices from the better subjects and the poorer subjects of Dorman *et al.* (1990) were each analyzed using both original and modified envelope categories. RTI for the modified envelope categories was higher by 3.3% for the better group and by 7.4% for the poorer group; modified envelope categories were therefore used to compute the information transfer data shown in Figs. 4 and 5.

ANSI (**1989**). ANSI S3.6-1989, "Specifications for audiometers" (American National Standards Institute, New York).

Blamey, P. J., Dowell, R. C., Clark, G. M., and Seligman, P. M. (**1987**). "Acoustic parameters measured by formant-estimating speech processor for a multiple-channel cochlear implant," J. Acoust. Soc. Am. **82**, 38–47.

Breeuwer, M., and Plomp, R. (**1984**). "Speechreading supplemented with frequency-selective sound-pressure information," J. Acoust. Soc. Am. **76**, 686–691.

Breeuwer, M., and Plomp, R. (**1986**). "Speechreading supplemented with auditorily presented speech parameters," J. Acoust. Soc. Am. **79**, 481–499.

Dorman, M. F., Soli, S. F., Dankowski, K., Smith, L. M., McCandless, G., and Parkin, J. (**1990**). "Acoustic cues for consonant identification by patients who use the Ineraid cochlear implant," J. Acoust. Soc. Am. **88**, 2074–2079.

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

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

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

Pickett, J. M. (**1980**). *The Sounds of Speech Communication* (University Park, Baltimore), p. 104.

Rosen, S. (**1989**). "Temporal information in speech and its relevance for cochlear implants," in *Cocohlear Implant: Acquisitions and Controversies*, edited by B. Fraysee and N. Cochard (Toulouse Implant Conference Proceedings, Toulouse), pp. 3–26.

Rosen, S., Fourcin, A. J., Abberton, E., Walliker, J. R., Howard, D. M., Moore, B. C. J., Douek, E. E., and Frampton, S. (**1985**). "Assessing assesment," in *Cochlear Implants*, edited by R. A. Schindler and M. M. Merzenich (Raven, New York), pp. 479–498.

Rosen, S., Walliker, J., Brimacombe, J. A., and Edgerton, B. J. (**1989**). "Prosodic and segmental aspects of speech perception with the House/3M single-channel implant," J. Speech Hear. Res. **32**, 93–111.

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

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152–1161.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.

Wang, M. D. (**1976**). "SINFA: Multivariate uncertainty analysis for confusion matrices," Behav. Res. Methods Instrum. **8**, 471–472.

Wilson, B. S., Finley, C. C., Lawson, D. T., Wolford, R. D., Eddington, D. K., and Rabinowitz, W. M. (**1991**). "Better speech recognition with cochlear implants," Nature **352**, 236–238.

Wilson, B. S., Lawson, D. T., and Finley C. C. (**1990**). "Speech processors for auditory prostheses," Fourth Quarterly Progress Report, NIH Project N01-DC-9-2401. Neural Prostheses Program, National Institutes of Health.