Comparative Studies of Speech Processing Strategies for Cochlear Implants

BS Wilson[1,2], CC Finley[1], JC Farmer, Jr.[2], DT Lawson[1], BA Weber[2], RD Wolford[2], PD Kenan[2], MW White[3], MM Merzenich[4] and RA Schindler[4]

[1] Neuroscience Program Office, Research Triangle Institute, Research Triangle Park, NC 27709.

[2] Division of Otolaryngology and Center for Speech and Hearing Disorders, Department of Surgery, Duke University Medical Center, Durham, NC 27710.

[3] Electrical and Computer Engineering Department, North Carolina State University, Raleigh, NC 27695.

[4] Department of Otolaryngology and Coleman and Epstein Laboratories, University of California at San Francisco, San Francisco, CA 94143.

Presented in part at the National Meeting of the Triological Society, April 28, 1987, Denver, CO.

Reprint or editing correspondence should be sent to:

Blake S. Wilson
Neuroscience Program Office
Research Triangle Institute
Research Triangle Park, NC 27709.

Telephone: (919) 541-6974

ABSTRACT

In studies of two patients implanted with the UCSF electrode array and fitted with percutaneous cables, our teams at UCSF and Duke have compared a wide variety of speech-processing strategies for multichannel auditory prostheses. Each strategy was evaluated using tests of vowel and consonant confusions, with and without lipreading. Included were the compressed-analog-outputs approach of the present UCSF/Storz prosthesis and a group of interleaved-pulses (IP) strategies in which the amplitudes of non-simultaneous pulses code the spectral variations of speech. For these two patients, each with psychophysical manifestations of poor nerve survival, scores were significantly higher with the IP processors than with any alternative strategy tested. We believe this superior performance results from (1) the substantial "release" from channel interactions provided by non-simultaneous stimuli, and (2) a fast enough rotation among the channels to ensure adequate temporal and spectral resolution. Such IP processors offer substantial improvement in the otherwise dismal performance of patients with poor nerve survival.

INTRODUCTION

In late 1983 we began a collaborative project among Research Triangle Institute (RTI), the University of California at San Francisco (UCSF) and Duke University Medical Center (DUMC) to develop speech processors for multichannel auditory prostheses. From the outset an important aim was to compare alternative processing strategies in tests with individual implant patients. In this way we could provide hitherto unrealized controls for differences among patients in (1) the patterns of neural survival at the periphery, (2) the integrity of the central auditory system, and (3) cognitive skill and language acquisition. In addition, comparisons of processing strategies with individual patients would allow us to use a single type of electrode array and to conduct tests in a uniform and consistent manner across strategies.

The initial period of our project was devoted to construction of advanced tools for comparisons of many different strategies in tests with individual patients. Primary among these tools is a software package for specification and simulation of a large variety of speech processors for auditory prostheses.[1] This software runs on Eclipse S-130 (at UCSF) and S-140 (at DUMC) computers in our cochlear implant laboratories. The outputs of the computer simulations are presented to the patient via a specially-designed hardware interface[2] that (1) supports a high bandwidth of information transmission to as many as eight stimulation channels, and (2) isolates the patient electrically from the computer equipment. Stimuli are delivered to the implanted electrode array either through the four-channel transcutaneous transmission system of the UCSF/Storz prosthesis[3] or through a percutaneous cable.[4] Use of the percutaneous cable allows access to all 16 electrode contacts in the array (which are usually configured as eight bipolar pairs) and control of the current or voltage waveforms of the stimuli. In contrast, alternating pairs of bipolar electrodes are assigned to the four channels of the transcutaneous system and the current and voltage waveforms of the stimuli depend complexly on the nonlinear impedances of the electrodes. We therefore prefer the cable for studies directed at measurements of psychophysical performance or at comparisons of alternative processing strategies.

In this report we will summarize findings from two cable patients in our collaborative study. Among the many processing strategies tested for both patients, large differences in performance were found between the compressed-analog-outputs (CAO) processor of the present UCSF/Storz prosthesis and a type of interleaved-pulses (IP) processor in which the amplitudes of non-simultaneous pulses code the short-time spectra of speech. To emphasize the importance of the processing strategy on the outcome for individual patients, we will restrict ourselves here to brief descriptions of tests related to these two types of processor. Detailed descriptions of the present tests, along with the results obtained with other processing strategies, are presented elsewhere.[5-7]

METHODS

Patients

Both patients in this study were selected and implanted according to procedures established by the UCSF team.[3,8,9] Patient LP was implanted at UCSF in May, 1985. His history included bouts of severe but currently inactive otitis media bilaterally, and simple mastoidectomies had been performed on both sides in 1941. LP experienced gradual loss of hearing in both ears thereafter. He was profoundly deaf by age 64, nine years prior to his implant operation.

As indicated in detail elsewhere,[5] the psychophysical performance of LP along almost every measured dimension was worse than any previous patient in the UCSF experimental series. Among the findings of the psychophysical studies were the following:

1. Thresholds for stimuli delivered to pairs of bipolar electrodes were much higher than thresholds for the same stimuli delivered to monopolar electrodes, for all channels;
2. Dynamic ranges from threshold to maximum comfortable loudness were extremely narrow compared to all other patients in the UCSF series (e.g., for 0.3 msec/phase biphasic pulses dynamic ranges were 4 dB or lower for five of the six channels used for speech studies);
3. Channel interactions, as measured by a loudness summation paradigm,[10] were severe for the middle channels of bipolar stimulation and somewhat less severe for the basal-most and apical-most electrode pairs;
4. LP was able to distinguish percepts elicited by stimulation with different bipolar pairs in the electrode array, if the stimuli were delivered one at a time;

5. Excitation of the middle channels strongly inhibited percepts elicited by excitation of the apical and basal channels; and
6. Thresholds and loudness levels were labile, changing both within and between testing sessions.

Not surprisingly, LP's case has been informally described as "one tough nut to crack" and "off the map." With the exception of the noise/voice test of the Minimal Auditory Capabilities Battery,[11] none of his scores on speech tests with the present UCSF/Storz processor was above chance; indeed, heroic efforts were required just to map the processor outputs into LP's useable dynamic range. Taken together, the results outlined above are consistent with a picture of very poor survival of peripheral neural elements along the middle portion of the electrode array and at least some survival in the apical and basal segments.

Unfortunately, LP's case was further complicated by a recurring, low-grade mastoiditis that placed his ear and implant at risk. Because LP was obtaining little benefit from his UCSF/Storz processor, and because applications of alternative processing strategies were only beginning to demonstrate benefit, a medical decision was made to explant the device and thereby minimize the risk of inner-ear infection. The practical consequence of this decision for the present study was that we had an extremely limited amount of time to work with LP. In all, we worked with him for 13 two-hour sessions.

The second patient, MH, was implanted at Duke in February, 1986. The etiology was otosclerosis, which produced profound bilateral deafness by age 40. MH was 51 at the time of her implant operation. When the cochlea was entered during surgery, it was discovered that the basal-most 4-5 mm of scala tympani was obliterated with otosclerotic bone. This bone had to be drilled out for insertion of the electrode array. Therefore, the two basal-most pairs of electrodes were probably more distant from the target neural tissue than in other patients implanted with the UCSF/Storz electrode array. Once drilled, the bone did not further impede the insertion of the electrode array. The array was inserted to a depth of approximately 25 mm, and clinically-indicated paranasal sinus X rays at a later date demonstrated that the implant followed the spiral course of the scala tympani.

An intensive series of tests was begun with MH in early March, 1986. A battery of psychophysical tests was conducted first, to assess the status of her implanted ear. These tests were a superset of those conducted with LP, and have been described elsewhere.[6] In brief, the results from the tests indicated generally poor survival of peripheral neural elements. Thresholds for bipolar stimulation varied significantly from pair to pair within the electrode array, and thresholds for bipolar stimulation were much higher than those for monopolar stimulation. In addition, interactions among most channels were severe, with some degree of isolation found for only one-third of the possible channel combinations. Finally, the dynamic ranges for pulsatile and sinusoidal stimuli were generally narrow (e.g., around 10 dB or less for 0.3 msec/phase biphasic pulses), although not as narrow as the dynamic ranges found for LP. Overall, MH presented a somewhat more favorable picture of psychophysical performance than the picture for LP. Like LP, MH had severe channel interactions and large differences in thresholds for bipolar and monopolar stimulation. Unlike LP, though, her thresholds and loudness levels were stable, and her dynamic ranges were only somewhat narrower than those found for typical patients.[12]

Processors

A block diagram of CAO processors of the type used in the UCSF/Storz prosthesis is shown in Fig. 1. Speech inputs are first high-pass filtered to flatten the speech spectrum and diminish the otherwise overwhelming influence of low-frequency components in speech (primarily the components of the fundamental frequency and the first formant frequency). The filtered signal is then compressed to map it onto the narrow dynamic range of electrically-evoked hearing. The frequency ranges of the band-pass and high-pass filters following the compressor encompass the first and higher formants of speech. In particular, the frequency ranges are selected so that the first formant is represented in channel 1 and the third and higher formants are represented in channel 4. The second formant is divided between channels 2 and 3, and this division is designed to emphasize discrimination of this critical component of speech.[4,9,13] The high-pass filter in the chain for channel 1 provides a first-order equalization of loudnesses for frequency components below 300 Hz. Specifically, it compensates for the large differences in the thresholds for low-frequency (e.g., around 100 Hz, where thresholds are low) and high-frequency (e.g., around 300 Hz, where thresholds are relatively high) stimuli. The adjustable clippers in the chains for channels 3 and 4 limit peak intensities of waveforms in these channels to levels below those that elicit "squeaky" or otherwise noxious percepts. Finally, adjustable gain controls are provided for each channel so that speech features signalled in that channel's band can be made clearly audible. For example, in the fitting of the prosthesis the gain of channel 4 is increased until "s" sounds are heard.

As indicated in Fig. 2, IP processors have a design that is quite different from the design of CAO processors. In the IP processor an automatic gain control (AGC) continuously adjusts the level of speech input so that a steady average level is presented to subsequent stages of the processor. Typical attack and release times for the AGC are 8 and 200 msec, providing a "slow AGC" action. The level-adjusted signal is then high-pass filtered as in the CAO processor to reduce the amplitudes of speech components below 1200 Hz. The output of the high-pass filter is fed to a bank of bandpass filters whose center frequencies span the combined range of the first and second formants of speech, along a logarithmic scale. The root-mean-square (RMS) energy in each band is sensed by a full-wave rectifier and low-pass filter connected in series to each bandpass filter output. Next, a "post processor" is programmed to scan the RMS outputs on a periodic basis. The output of a filter bank channel is coded for stimulation of its assigned electrode(s) only if the RMS energy is above a preset "noise threshold." The amplitudes of the pulses delivered to the selected channel(s) are derived with a logarithmic mapping law of the form:

$$\text{pulse amplitude} = A \times \log(\text{RMS level}) + k,$$

where the parameters "A" and "k" are determined for each channel according to the threshold and most-comfortable loudness level for that channel. Finally, the voicing detector senses the fundamental frequency of voiced speech sounds and whether a given speech input is voiced (periodic) or unvoiced (aperiodic). The output of the voicing detector can optionally be used by the post processor to control the timing of "round-robin" update cycles, as described below.

Variations of IP processors are produced with different choices of parameters for the post processor. These parameters include (1) the number of channels stimulated on each stimulus cycle; (2) the duration of stimulus pulses for each channel; (3) the interval between pulses on sequentially stimulated channels; (4) the order in which channels are to be stimulated; (5) the mapping law for each channel, as described above; (6) the waveforms of stimulus pulses; and (7) whether stimulus sequences are to cycle continuously or are to be timed according to information provided by the voicing detector. Parameters 1 through 4 define the basic sequence of stimulation across channels, which we term as one "round-robin" cycle. Round-robin cycles are repeated as rapidly as possible if voicing information is not to be explicitly coded. Alternatively, inputs from the voicing detector can be used to time the beginning of each round-robin cycle. If voicing information is to be explicitly coded, round-robin cycles are timed to start in synchrony with the fundamental frequency (F0) during voiced speech sounds and at either randomly-spaced or maximum-rate intervals during unvoiced speech sounds. Explicit coding of voicing information might be expected to improve a patient's perception of prosodic features associated with F0 contours and to help the patient make voice/unvoice distinctions for consonants (e.g., improve the ability to distinguish an "s" from a "z" or a "t" from a "d"). Also, an explicit representation of voicing information might be expected to improve the "naturalness" of speech percepts and the ability to make man/woman/child distinctions.

To illustrate the fundamental differences in CAO and IP processors, Figs. 3 and 4 show typical waveforms for each. In each panel of each figure the top trace is the input to the processor and the remaining traces are channel outputs. The input is the word "BOUGHT." The initial consonant occurs at about 180 msec and the vowel follows immediately thereafter. An expanded display of waveforms well into the vowel is shown in the lower-left panel of each figure. Next, the "t burst" of the final consonant begins slightly before 640 msec, and an expanded display of waveforms beginning at 640 msec is shown in the lower-right panel of each figure. The lower panels thus exemplify differences in waveforms for voiced and unvoiced intervals.

Waveforms for the CAO processor are presented in Fig. 3. In the voiced interval the relatively-large outputs of channels 1 and 2 reflect the low-frequency formant content of the vowel and in the unvoiced interval the relatively large outputs of channels 2 through 4 reflect the high-frequency noise content of the "t." In addition, the clear periodicity in the waveforms of channels 1 and 2 reflects the fundamental frequency of the vowel during the voiced interval, and the lack of periodicity in the outputs of channels 2 through 4 reflects the noise-like quality of the "t" during the unvoiced interval. These represented features are likely to be perceived to varying degrees by different implant patients. A principal concern is that simultaneous stimulation of the channels can exacerbate interactions between channels, particularly for patients who require high stimulation levels. Also, summation of stimuli between channels depends on the phase relationships of the waveforms. Because these relationships are not controlled in a CAO processor, the representation of the speech spectrum usually will be further distorted by continuously-changing patterns of channel interactions. Therefore, one might expect that CAO processors would work best for patients with low thresholds and good isolation between channels.

The problem of channel interactions is addressed in the IP processor through the use of non-simultaneous stimuli. This eliminates direct summation of the stimuli across channels. Further, secondary interactions produced by temporal integration at neural membranes[14] may be reduced by increasing the interval between pulses delivered to sequential channels.

Typical waveforms for an IP processor are shown in Fig. 4. A striking difference between the stimuli for this processor and those for the CAO processor is the relative sparseness of stimulation resulting from the use of non-simultaneous stimuli. In the particular variation of IP processors presented in Fig. 4, the greatest 4 of 6 channels are updated on every round-robin cycle and voicing information is explicitly coded. During voiced speech sounds the round-robin cycles are timed to begin in synchrony with the detected fundamental frequency, while during unvoiced speech segments the cycles are initiated at randomly-spaced intervals. The periodicity of cycle updates can be seen for a voiced speech sound in the lower-left panel of Fig. 4 and the randomly-spaced cycle updates can be seen for an unvoiced speech sound in the lower-right panel. As mentioned before, the amplitudes of the pulses reflect the RMS energy levels in each channel's frequency band. Thus the timing of round-robin updates codes F0 for voiced speech sounds and also indicates whether a given speech sound is voiced or unvoiced. The upper spectrum of speech above F0 is coded by the amplitudes of stimulus pulses and by the selection of channels. Many other variations of IP processors are available through manipulations of the parameters for the post processor.[6,7]

Procedure

The performance of each processing strategy was measured with confusion-matrix tests. The confusion matrix for vowels included the tokens "BOAT," "BEET," "BOUGHT," "BIT," and "BOOT," and the confusion matrix for consonants included the nonsense tokens "ATA," "ADA," "AKA," "ASA," "AZA," "ANA," "ALA," and "ATHA." All processing strategies were implemented with computer simulations as previously described. The presentation of each processed token was accompanied by a display of response options on a computer console used by the patient. When the patient responded, his or her response was used to update a matrix display on the investigator's computer console (not seen by the patient), and the next token was drawn from a randomized list. Five presentations of each processed token were included in the vowel test and three presentations of each processed token were included in the consonant test. At the end of a test we usually would give the patient the overall correct score and an indication of the principal confusions made during the test. No feedback was given during the test itself.

Four tests were given for each processing strategy evaluated with patient MH: vowel recognition with lipreading; vowel recognition without lipreading; consonant recognition with lipreading; and consonant recognition without lipreading. Lipreading information was provided by miming tokens in synchrony with stimulus presentations. The same investigator (CCF) presented lipreading information for all tests. Finally, presentations of processed tokens usually were repeated at regular intervals until the patient responded. Although there is evidence that repetition of test tokens can increase scores (particularly for tests using open set material, such as tests of spondee recognition),[8,9] we did not find statistically-significant differences in the scores of several tests of consonant recognition for single- and multiple-trial conditions. Retests at various intervals under the same conditions validated the use of these brief confusion matrices to identify processor strengths and weaknesses.

Because time was extremely limited with patient LP, most formal tests with him were restricted to vowel recognition without lipreading. In addition to a direct measure of vowel recognition, these tests provided good indications of whether the percepts elicited by a given processing strategy sounded like speech and whether loudnesses could be balanced across tokens. These latter determinations were particularly important for LP's case inasmuch as his percepts with the CAO processor were not described as speechlike and his dynamic range for loudness mapping was both narrow and labile. Our first task with LP was to demonstrate that use of any processing strategy would put him into the "speech mode" of auditory perception. We then could evaluate in greater detail that strategy and closely-related alternatives. As described in the RESULTS section, the first task was accomplished for LP, but the second task was only partially completed before our time with him expired.

The speed with which totally new processor designs could be simulated in software allowed the evaluation of any one design to influence the choice of strategies for the next testing session. The combination of great flexibility in the range of possible designs and the very short time required for a diagnostic-prescriptive cycle in processor optimization comprise an extremely powerful tool not only for this research but also for clinical fitting of highly customized processors.

RESULTS

Patient LP

As indicated above, LP presented a tremendous challenge to our team. His psychophysical performance was the worst of any patient in the UCSF series. In addition, percepts elicited with several variations of 4- and 6-channel CAO processors were not described by LP as speech-like in character. Instead, processed speech tokens sounded like "bumps" with little or no variation within the bumps. The percepts were further described as "mushy," "drawn out" or "on all the time." The general picture that emerged from these anecdotal remarks was one of a poor representation of temporal events, possibly produced by LP's severe channel interactions. In no case was a speech token spontaneously identified as the word delivered to the speech processor; the tokens included most of those from the vowel and consonant confusion tests.

The suggestion that channel interactions might have been largely responsible for these disappointing results led us to evaluate processors in which non-simultaneous stimuli were used. Two IP processors were tested. The design and evaluation of other processors using non-simultaneous stimuli are described elsewhere.[5] In the first of the two IP processors, the greatest two of six channels were selected for stimulation in each round-robin cycle (see METHODS for further description of processors). Balanced biphasic pulses were interleaved so that the onset of a pulse on one channel would never follow the offset of a pulse on another channel within an interval of less than 1.0 msec. Because short-term temporal integration fell off rapidly at and beyond 1.0 msec for LP, we thought this interleaving of stimuli would eliminate channel interactions produced by simultaneous current summation and greatly reduce interactions produced by temporal integration of non-simultaneous stimuli at neural membranes.[14] The duration of stimulus pulses was 0.3 msec/phase, so that the maximum rate of stimulation on any single channel was 313 Hz. Finally, stimuli were presented one after another in this processor and therefore voicing and voice/unvoice boundaries were not explicitly coded.

We are pleased to report that the percepts elicited with the 6 channel IP processor were all in the "speech mode," that most of the tokens in the vowel confusion test were spontaneously recognized as the correct words, and that half of the six tokens we presented in the consonant confusion test were spontaneously recognized as the correct nonsense syllables. The improvement over the results obtained with the CAO processors was immediate and compelling. Moreover, the use of pulsatile stimuli in the IP processor produced (for the first time) a tolerable range of loudnesses across tokens. Although formal tests were not conducted, these loudnesses also appeared to have far greater stability than the loudnesses of percepts produced with the CAO processors. In all, it was clear to us and clear to the patient that speech information was making its way onto the nerve. A record of LP's initial reports in listening to the outputs of the 6-channel IP processor is presented in Table I. Of the 11 tokens presented after our first adjustment of processor outputs (to bring the outputs into an audible range), 7 were immediately and spontaneously recognized as the correct words or syllables. Unfortunately, time ran out in the session before we were able to conduct matrix tests of vowel and consonant confusions.

The second IP processor tested with LP was a reduced 4-channel version of the processor just described. Evaluation of the 4 channel processor was motivated by the need to identify a more-or-less "optimal" configuration for a processor that LP could use with the 4-channel UCSF/Storz transcutaneous transmission system. In particular, we needed information on the benefit LP might receive from his cochlear implant after the transcutaneous transmission system was installed. Because LP was then exhibiting signs of mastoiditis, we were concerned that the long-term management and risks associated with recurring mastoiditis might outweigh the potential benefit of the prosthesis.

Unfortunately, recognition of the vowel and consonant tokens with the 4-channel processor seemed to be far less salient than recognition with the 6-channel processor. In a formal test of vowel recognition with the 4-channel processor, LP correctly identified 56% of the randomly-presented tokens. Although this score was significantly above the chance level of 20% for this test, it was also well below the level of performance that might be expected for the 6-channel processor on the basis of the reports in Table I. Indeed, the speech percepts elicited with the 4-channel processor were described by LP as being "distorted" and "less distinct" compared with the percepts he remembered from the 6-channel processor. Finally, informal tests of consonant identification with the 4-channel processor indicated that LP would have great difficulty in distinguishing the tokens in a matrix test.

The generally-disappointing results obtained with the 4-channel processor supported a medical decision to remove LP's implanted device. Testing ended at this point and the explant operation was performed shortly thereafter.

Although we were greatly saddened by the fact that LP's device had to be removed, we regarded the overall findings with him as most encouraging. In particular, his case demonstrated the potential of IP processors for patients with very poor nerve survival. The switch from a 4- or 6-channel CAO processor to a 4- or 6-channel IP processor immediately placed LP in the "speech mode" of auditory perception. Moreover, his scores on tests of vowel recognition were significantly above chance with the 4-channel IP processor. Finally, LP's reports during the first application of a 6-channel IP processor suggested that performance might be substantially improved with a modest increase in the number of stimulation channels.

Patient MH

With the encouraging but preliminary results from LP's case in hand, we were of course anxious to conduct fully-controlled comparisons of processing strategies with another patient. Fortunately, our studies with LP at UCSF were closely followed by the start of our studies with MH at Duke. We have now worked with her for more than a year since her implant operation in February, 1986. During this period we have completed an extensive series of psychophysical studies and have evaluated a very wide range of processing strategies. In this report we will present the results from evaluations of CAO and IP processors. Results obtained with other processors are available elsewhere.[6,7]

The main results from the evaluations of CAO and IP processors with MH are summarized in Fig. 5. For each processor tested, at least 4 variations were evaluated to optimize processor parameters. The rationale and procedures for parametric manipulations have been presented in detail elsewhere.[6] The scores for each processor in Fig. 5 are those for the parametric set that produced the highest overall percent-correct score in the four tests of vowel and consonant recognition.

Before describing the results for each processor in Fig. 5, we note a few general features of the data. First, high scores are consistently found for the tests of vowel identification with lipreading. MH got 92% correct on a test we administered to measure her performance with lipreading alone, a score that is not significantly different from most of the scores shown in Fig. 5 for vowel identification with lipreading. Therefore, the scores for this test are not a sensitive indicator of processor performance.

Next, we note that scores on the tests of consonant identification with lipreading are a sensitive indicator of processor performance. In multiple tests of consonant identification with lipreading only, MH got an average score of 52% correct. With the exception of the CAO processor, all scores in Fig. 5 for consonant identification with lipreading are significantly above this level.

Third, test/retest reliability was good for MH. When we retested a processor that produced low scores on a previous occasion MH always would obtain low scores again, and when we retested a processor that had produced high scores on a previous occasion MH always would repeat her high scores. The standard deviation of overall percent-correct scores from seven repeated trials of the last (rightmost) processor shown in Fig. 5, for example, was slightly less than 3%.

Finally, it is noteworthy that MH's anecdotal remarks were stable across repeated tests of a single processor. When a "good" processor was retested MH would immediately identify it as such, usually in terms like "this is a good processor," "this processor sounds natural and like speech I remember," "this processor doesn't sound simulated," or "this processor is very clear." In contrast, a retest of a processor that produced low scores on a previous occasion would elicit comments like "this is a lousy processor," "this processor sounds like a man in a barrel," "the speaker sounds like he is talking through the telephone with a handkerchief or towel over the mouthpiece," or "this processor is not as clear as some you have tried." MH's anecdotal remarks were always consistent with her test scores on confusion-matrix material.

Perhaps the most striking feature of the data in Fig. 5 is the large difference in performance found across processing strategies. The results range from poor levels of performance to outstanding levels of performance.

The lowest scores in every category are found for the 4-channel CAO processor. The scores for the tests of processor performance with lipreading are about the same as the scores obtained for lipreading alone. The scores for the test conditions without lipreading, while significantly above chance, are clearly lower than these scores for all other processors. This picture of relatively poor performance with the CAO processor is consistent with the observations made in tests with LP. That is, both patients have psychophysical manifestations of poor nerve survival and both patients receive little benefit from the CAO processor. Presumably, the severe interactions between channels for simultaneous stimulation limit the performance of CAO processors in such patients.

The remaining results presented in Fig. 5 show the performance levels of various IP processors. These results allow direct comparisons of (1) 4-channel CAO and IP processors; (2) 4- and 6-channel IP processors; and (3) 4- and 6-channel IP processors with and without explicit coding of voicing information. The comparisons indicate that:

1. Performance is markedly improved when a 4-channel IP processor is used instead of a 4-channel CAO processor;
2. Scores are much higher in all categories except vowel identification with lipreading (where scores are about the same) when a 6-channel IP processor is used instead of a 4-channel IP processor; and
3. Explicit coding of voicing information improves the performance of IP processors, particularly in the categories of vowel identification without lipreading (4-channel processor), consonant identification without lipreading (6-channel processor) and consonant identification with lipreading (both processors).

DISCUSSION

Interpretation of Results

While the results of the vowel and consonant tests described in this paper are most encouraging, it should be clearly understood that these tests sample a rather limited set of attributes associated with speech perception. The confusion matrix tests were selected because they could be rapidly applied and because they provide valuable diagnostic information (in the patterns of confusions) for improving processor design. Moreover, the matrix tests emphasize measurement of perception at a peripheral level in the auditory system. To be specific, good performance on the vowel and consonant identifications indicates that these speech features are represented at the periphery by a given speech processor. Such a representation can support, but does not guarantee, good performance on more complex tasks such as open-set recognition of continuous discourse. A host of cognitive and linguistic skills may influence performance on open-set tasks. Although open-set recognition is the ultimate goal of research on auditory prostheses, a first and important step is to demonstrate representation of fundamental elements of speech at the periphery.

The primary finding of the present study is that such a representation can be provided for patients with poor nerve survival. This finding offers the realistic expectation that performance for these patients might be substantially improved with applications of the right type of processing strategy. The next steps to confirm and extend the generality of the present findings are to increase the range of speech perception studies and to test more patients. We are now comparing the performance of CAO and IP processors in an extensive series of tests with seven patients. One of these patients is MH and the remaining six are all implanted with the

4-channel UCSF/Storz transcutaneous transmission system. The tests include the vowel and consonant tests described in this paper; all subtests of the Minimal Auditory Capabilities (MAC) Battery;[11] the Diagnostic Discrimination Test of consonant confusions;[15] speech tracking[16,17] with and without the aid of the prosthesis; and the IOWA test of medial consonant identification with speechreading cues.[18] The results of these studies will be published in a future report.

Applications of IP Processors

In the absence of data comparing CAO and IP processors for patients with varying degrees of nerve survival, experience with LP and MH suggests that IP processors might be best for patients with poor nerve survival while CAO processors might well be best for patients with good nerve survival. This expectation is based on the observations that (1) IP processors provided better performance for the two patients of the present study, both of whom had psychophysical manifestations of poor nerve survival; (2) approximately half of the patients in the UCSF/Storz clinical series have truly excellent results with their CAO processors, possibly because these patients have good survival of peripheral neurons; (3) it has been demonstrated that in some patients continuous, "analog-type" waveforms can provide temporal and frequency information up through the range of first formant frequencies for speech;[19,20] and (4) at least some of this information is discarded when an IP processor is used. These comparisons between processing strategies are summarized in Table II. Briefly, the CAO processor may be superior for patients with good nerve survival because such patients might perceive substantial temporal and frequency information in analog waveforms and because the lower stimulus intensities required for these patients, along with survival of ganglion cells and/or dendrites over each active pair of electrodes, greatly minimizes channel interactions produced by simultaneous stimulation. On the other hand, the IP processor may be superior for patients with poor nerve survival because isolation between channels for such patients is tremendously improved with the use of non-simultaneous stimuli.

Finally, we note that the type of IP processor used is critically important to the outcome for the patients we have studied. Measurements of performance changes with parametric manipulations in IP processors have indicated that good performance appears to depend on the following, in approximate order of importance:[6,7]

1. Total nubmer of channels (large increases in performance are found when the number of channels is increased from 2 to 4 and from 4 to 6);
2. Number of channels updated per round-robin cycle (performance in tests of consonant identification declines precipitously if this number falls below 4);
3. Total duration of each round-robin cycle (performance gets better as duration is decreased, and is markedly better when the duration is less than 4-5 msec);
4. Time between pulses (performance improves as the time between pulses is increased, up to the point at which the total duration of the round-robin cycle begins to exceed 4-5 msec); and
5. Explicit coding of voicing information (performance is better with explicit coding of voicing information, and the percepts elicited with processors that use such coding are described as more natural and speechlike);

Among these factors, the second is perhaps the most surprising and significant in terms of processor design. In our parametric studies with

6-channel IP processors we found that good recognition of vowels could be maintained even if only two channels were updated on each round-robin cycle. However, scores on the tests of consonant recognition declined precipitously if fewer than four channels were updated on each cycle. The relative improvement in consonant recognition when the number of updated channels is increased may have resulted from an improved representation of the complex spectra and temporal dynamics of consonants. Vowels generally have steady or slowly-varying spectra that can be well-characterized by two or three formant frequencies. In contrast, consonant recognition depends on much more rapid variations and on broad noise bands that do not lend themselves to formant representation. Consonant recognition may involve a host of features including (1) voicing/frication; (2) amplitude envelope; (3) loci and shapes of broad spectral peaks; and (4) rapid formant transitions from a leading vowel into a following consonant or from a leading consonant into a following vowel. These features other than steady-state formants are probably best represented with rapid updates of information on all channels of a multichannel array.

CONCLUSIONS

The most general conclusion to be drawn from the results presented in this report is that manipulations in the processing strategy used in an auditory prosthesis can have huge effects on recognition of consonants and vowels. This basic finding demonstrates the importance of selection of an appropriate processing strategy for individual implant patients. MH, for example, attains outstanding levels of recognition with certain processing strategies, and poor-to-moderate levels of recognition with others. Because our studied population of patients is limited, we do not know at this time whether one processing strategy will emerge as superior for all patients. For patients LP and MH, processors that represented the RMS energies in five or six frequency bands with interleaved pulses provided much better performance than the other strategies we have evaluated. We note, though, that both these patients had psychophysical manifestations of poor (MH) or extremely-poor (LP) nerve survival. It may be that a completely different class of processors would work best for a more-fortunate patient with good nerve survival. For example, the excellent results from approximately half of the patients in the UCSF series strongly indicate that a compressed-analog-outputs strategy may be as good as or superior to an interleaved-pulses strategy for cases in which nerve survival is good.[3,8,9] As mentioned before, this hypothesis is being tested for a variety of patients. Pending those more detailed and general results, we conclude that:

1. Different processing strategies can produce widely-different outcomes for individual implant patients;
2. Interleaved-pulses processors are far superior to other processors for at least two patients with poor nerve survival;

3. Processors other than the interleaved-pulses processors may be superior for patients with good nerve survival; and
4. Therefore it is important not to have an "adopted religion" for a single strategy of speech processing for auditory prostheses.

ACKNOWLEDGEMENTS

We thank the patients who participated in this study for their dedicated effort and pioneering spirit. We also thank the many people who contributed to the study, but whose names are not listed among the authors. Notable and important contributions were made by F.T. Hambrecht, G.E. Loeb, R.V. Shannon, L. Vurek, J. Molinari and J.A. Moore. This work was supported by NIH contracts N01-NS-3-2356 and N01-NS-5-2396, from the Neural Prosthesis Program.

BIBLIOGRAPHY

1 Wilson, B.S. and Finley, C.C.: Speech Processors for Auditory Prostheses. Fourth Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-3-2356, 1984.

2 Wilson, B.S. and Finley, C.C.: Speech Processors for Auditory Prostheses. Second Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-3-2356, 1984.

3 Schindler, R.A., Kessler, D.K., Rebscher, S.J., *et al*.: The UCSF/Storz Multichannel Cochlear Implant: Patient Results. *Laryngoscope*, 96: 597-603, 1986.

4 Merzenich, M.M., Rebscher, S.J., Loeb, G.E., *et al*.: The UCSF Cochlear Implant Project. Cochlear Implants in Clinical Use. W.D. Keidel and P. Finkenzeller (Eds.). *Adv. Audiol*., 2: 119-144, 1984.

5 Wilson, B.S., Finley, C.C. and Lawson, D.T.: Speech Processors for Auditory Prostheses. Seventh Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-3-2356, 1985.

6 Wilson, B.S., Finley, C.C., and Lawson, D.T.: Speech Processors for Auditory Prostheses. Second Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-5-2396, 1986.

7 Wilson, B.S., Finley, C.C. and Lawson, D.T.: Speech Processors for Auditory Prostheses. Fourth Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-5-2396, 1986.

8 Schindler, R.A., Kessler, D.K., Rebscher, S.J., *et al*.: Surgical Considerations and Hearing Results with the UCSF/Storz Cochlear Implant. *Laryngoscope*, 97: 50-56, 1987.

9 Schindler, R.A., Kessler, D.K., Rebscher, S.J., et al.: The University of California, San Francisco/Storz Cochlear Implant Program. Otolaryngol. Clin. North Am., 19: 287-305, 1986.

10 White, M.W., Merzenich, M.M. and Gardi, J.N.: Multichannel Cochlear Implants: Channel Interactions and Processor Design. Arch. Otolaryngol., 110: 493-501, 1984.

11 Owens, E., Kessler, D.K., Raggio, M., et al.: Analysis and Revision of the Minimal Auditory Capabilities (MAC) Battery. Ear Hear., 6: 280-287, 1985.

12 Shannon, R.V.: Threshold and Loudness Functions for Pulsatile Stimulation of Cochlear Implants. Hearing Res., 18: 135-143, 1985.

13 Loeb, G.: The Functional Replacement of the Ear. Scientific American, 252: 104-111, 1985.

14 Wilson, B.S., Finley, C.C. and Lawson, D.T.: Speech Processors for Auditory Prostheses. Eighth Quarterly Progress Report. National Institutes of Health, Contract #N01-NS-3-2356, 1985.

15 Grether, C.B. and Kessler, D.K.: The Diagnostic Discrimination Test (DDT). Laboratory Report. University of California at San Francisco, Department of Otolaryngology, 1985.

16 De Filippo, C.L. and Scott, B.L.: A Method for Training and Evaluating the Reception of Ongoing Speech. J. Acoust. Soc. Am., 63: 1186-1192, 1978.

17 Owens, E. and Telleen, C.: Tracking as an Aural Rehabilitative Process. J. Acad. Rehabil. Audiol., 14: 259-273, 1981.

18 Tyler, R.S., Preece, J.P. and Lowder, M.W.: The Iowa Cochlear-Implant Test Battery. Laboratory Report. University of Iowa at Iowa City, Department of Otolaryngology-Head and Neck Surgery, 1983.

19 White, M.W.: Formant Frequency Discrimination and Recognition in Subjects Implanted with Intracochlear Stimulating Electrodes. Ann. N.Y. Acad. Sci., 405: 348-359, 1983.

20 Eddington, D.K.: Speech Recognition in Deaf Subjects with Multichannel Intracochlear Electrodes. Ann. N.Y. Acad. Sci., 405: 241-258, 1983.

TABLE I.

Initial Reports Made by LP in Listening to the Outputs of an Interleaved-Pulses Processor

Token*	Report
BOOT	near threshold ("not loud enough to make it out")
BOUGHT	spontaneous recognition ("a perfect BOUGHT")
BOAT	spontaneous recognition ("you're saying BOAT; the sound is nice and has a good loudness")
BIT	spontaneous recognition
BEET	spontaneous recognition ("the EE is high in pitch; BEET is very clear")
ADA	spontaneous recognition ("close to ATA, but is clearly ADA; that's a good ADA")
AKA	not recognized ("could be ADA or ATA")
ANA	spontaneous recognition ("sounds just like ANA; a beautiful ANA!")
ASA	not recognized ("can't tell")
ATA	spontaneous recognition
AZA	not recognized ("could be ASA or AZA; there's no way I could tell the difference between those two")

*Tokens ALA and ATHA were not presented in the initial tests with this first interleaved-pulses processor.

TABLE II.

Characteristics of Processors*

ANALOG	PULSATILE
continuous waveforms, presented simultaneously	non-simultaneous pulses
severe interactions between channels for patients with poor nerve survival	improved channel isolation, especially for patients with poor nerve survival
in some patients, continuous waveforms can provide good temporal and frequency information (F0, voice/unvoice boundaries, F1, possible F2)	limited transmission of temporal and frequency information (F0, voice/unvoice boundaries)

*Symbols used in this Table are F0 for the fundamental frequency of voiced-speech sounds, F1 for the first formant frequency of speech, and F2 for the second formant frequency of speech.

FIGURE LEGENDS

Fig. 1. Block diagram of 4-channel, compressed-analog-outputs (CAO) processors. See text for details.

Fig. 2. Block diagram of 6-channel, interleaved-pulses (IP) processors. See text for details.

Fig. 3. Waveforms of a compressed-analog-outputs (CAO) processor. The top trace in each panel is the input to the processor and the remaining traces are channel outputs. The input is the word "BOUGHT." An expanded display of waveforms during the initial portion of the vowel is shown in the lower-left panel and an expanded display of waveforms during the "T" is shown in the lower-right panel. Characteristics of the filters in the processor are the same as those indicated in Fig. 1. The adjustable gain controls are set at the same level to demonstrate the pattern of channel outputs before the outputs are mapped into audible ranges for individual patients. Finally, the compression ratio is set at 3.0, and the threshold for the onset of compression is approximately 3% of the full scale deflection of the input signal.

Fig. 4. Waveforms of an interleaved-pulses (IP) processor. The top trace in each panel is the input to the processor and the remaining traces are channel outputs. The input is the word "BOUGHT." An expanded display of waveforms during the initial portion of the vowel is shown in the lower-left panel and an expanded display of waveforms during the "T" is shown in the lower-right panel. Characteristics of the filters in the processor are the same as

those indicated in Fig. 2. The rolloff frequency for the smoothing filters in the RMS (root-mean-square) energy detectors is set at 25 Hz. In this particular processor the greatest 4 of 6 channels are updated on each round-robin cycle, and the cycles are timed to start in synchrony with the fundamental frequency during voiced speech sounds and at randomly-spaced intervals during unvoiced speech sounds. The initial phase of stimulus pulses is 0.5 msec in duration and the second phase is 3.0 msec in duration. The amplitude of the second phase of each pulse is chosen to make the net charge transferred by the pulse zero. Finally, the amplitudes of the pulses are set according to mapping parameters derived for patient MH; this is the processor used for the last processor condition indicated in Fig. 5.

Fig. 5. Results of vowel and consonant confusion tests for patient MH. Diagonally-hatched bars indicate results obtained with lipreading and cross-hatched bars indicate results obtained without lipreading. The table at the bottom of the figure indicates the type of processor used (abbreviations are CAO for "compressed analog outputs" and IP for "interleaved pulses"); the number of stimulation channels; whether voicing information was explicitly coded for the IP processors; and the overall percent-correct scores from the four test conditions for each processor. The horizontal line in each panel shows the level of chance performance for that test.

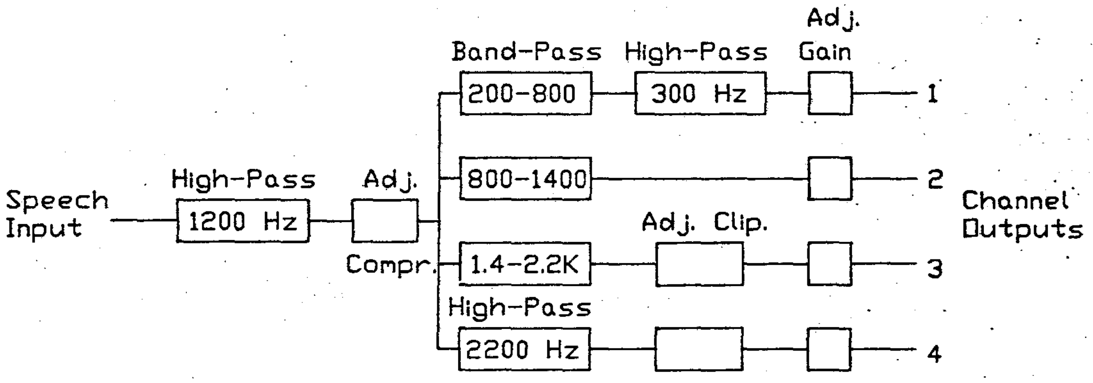
Speech Input
High-Pass
1200 Hz
Adj.
Compr.
Band-Pass
200-800
High-Pass
300 Hz
Adj. Gain
1
800-1400
2
1.4-2.2K
Adj. Clip.
3
High-Pass
2200 Hz
4
Channel Outputs

Fig. 1

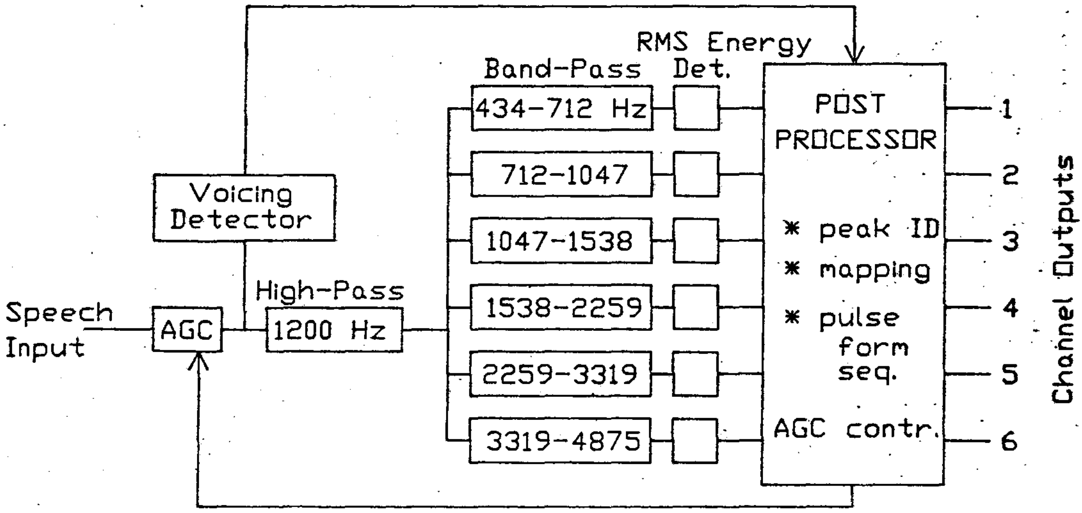
RMS Energy
Band-Pass
Det.
434-712 Hz
712-1047
1047-1538
1538-2259
2259-3319
3319-4875
POST
PROCESSOR
* peak ID
* mapping
* pulse
form
seq.
AGC contr.
Voicing
Detector
Speech
Input
AGC
High-Pass
1200 Hz
1
2
3
4
5
6
Channel Outputs

Fig. 2

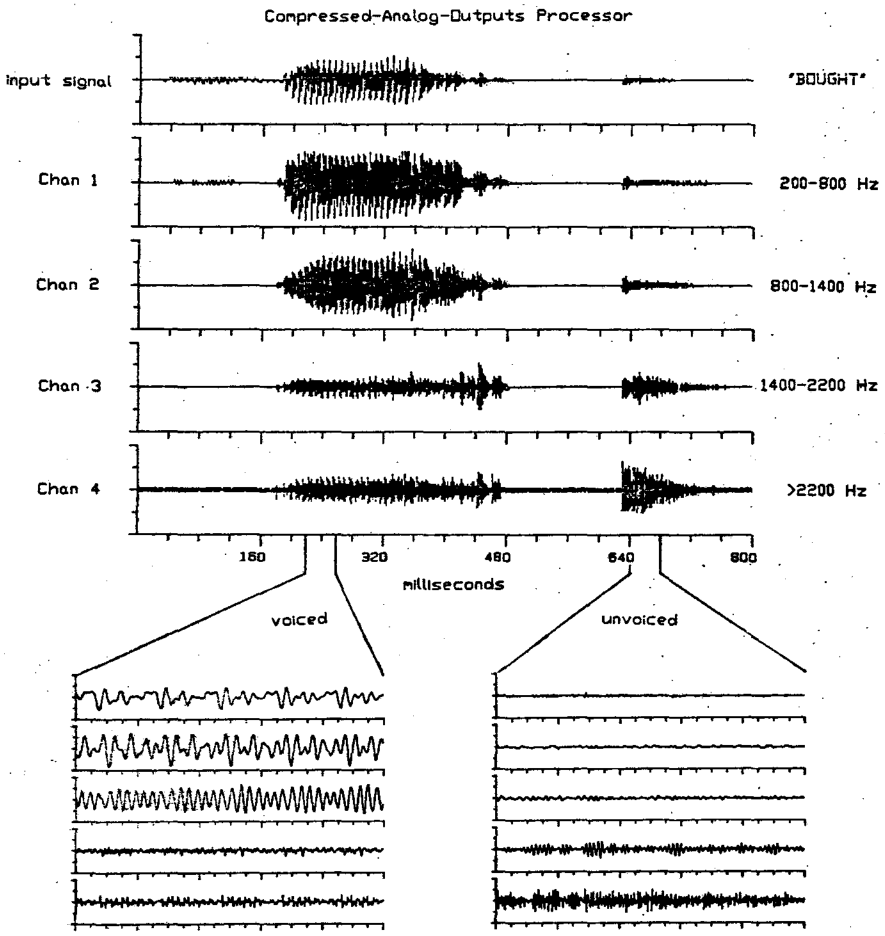
Compressed-Analog-Outputs Processor
Input signal
'BOUGHT'
Chan 1
200-800 Hz
Chan 2
800-1400 Hz
Chan 3
1400-2200 Hz
Chan 4
>2200 Hz
160
320
480
640
800
milliseconds
voiced
unvoiced

Fig. 3

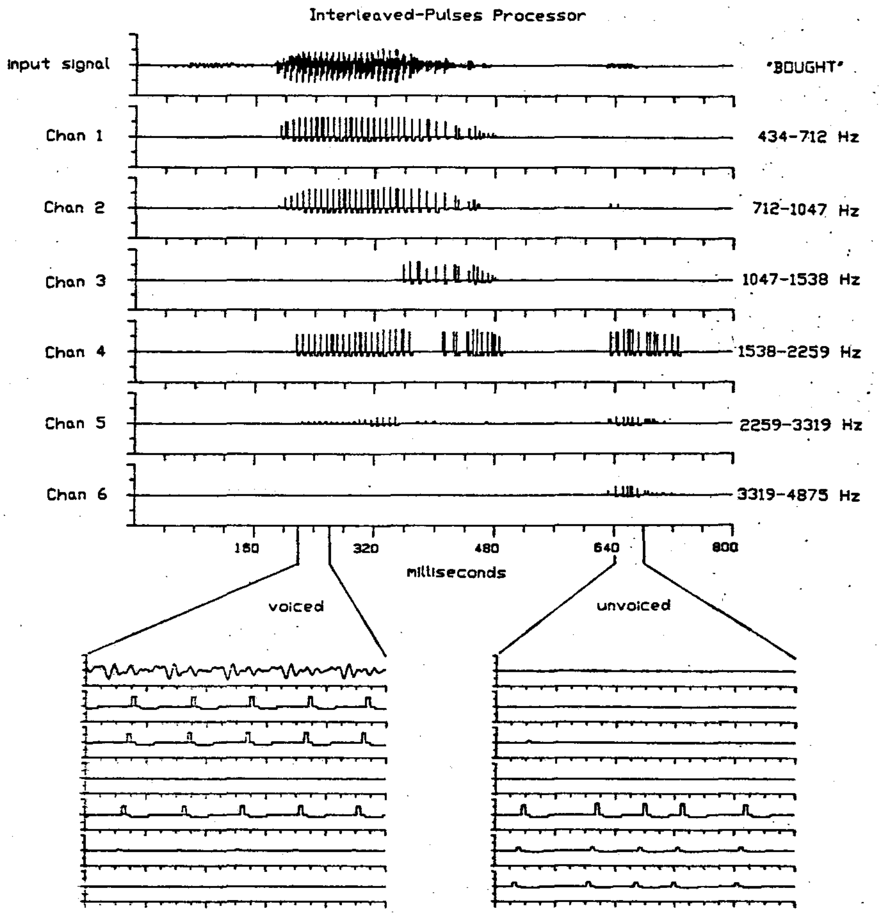
Interleaved-Pulses Processor
Input signal
"BOUGHT"
Chan 1
434-712 Hz
Chan 2
712-1047 Hz
Chan 3
1047-1538 Hz
Chan 4
1538-2259 Hz
Chan 5
2259-3319 Hz
Chan 6
3319-4875 Hz
160
320
480
640
800
milliseconds
voiced
unvoiced

Fig. 4

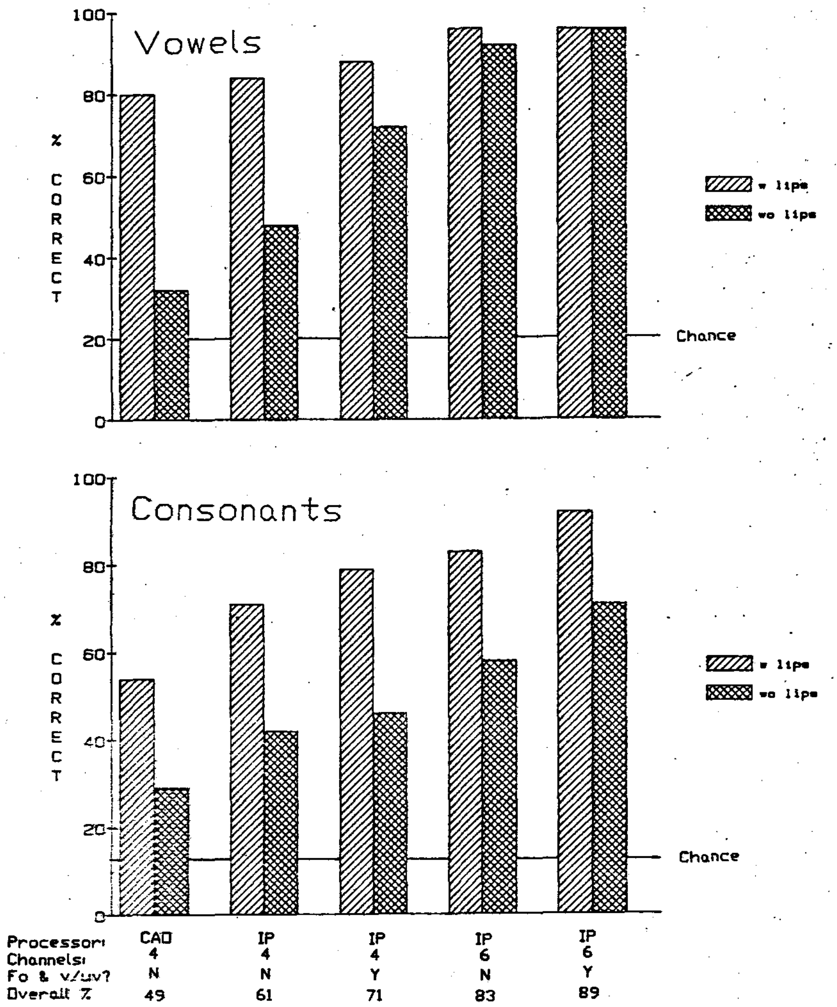
Vowels
% CORRECT
100
80
60
40
20
0
w lips
wo lips
Chance
Consonants
% CORRECT
100
80
60
40
20
0
w lips
wo lips
Chance
Processor: CAD IP IP IP IP
Channels: 4 4 4 6 6
Fo & v/uv? N N Y N Y
Overall % 49 61 71 83 89

Fig. 5