

Speech Recognition in Analog Multichannel Cochlear Prostheses: Initial Experiments in Controlling Classifications

MARK W. WHITE, MARLEEN T. OCHS, MICHAEL M. MERZENICH, AND EARL D. SCHUBERT

Abstract—Computer-synthesized vowels were used to examine methods for controlling and measuring the perceptions elicited during electrical stimulation of the human cochlea. In the first experiment, we measured the importance of the second formant (F_2) in the identification of vowels, matched for duration, in a single subject with a multichannel cochlear implant. The subject never confused vowels having a “low” frequency F_2 with those having a “high” frequency F_2 . In the second experiment, identification functions were generated for a series of vowels varying only in F_2 . When the pattern of F_2 stimulation at the basilar membrane was manipulated, vowel identification functions were altered. For the categorization of vowels, the data indicate that the relative cochlear position of F_2 stimulation was more important than fine-grain temporal waveform cues. The data are supportive of cochlear implant coding strategies that make use of cochlear place information. In the later experiments, we manipulated filter passbands and channel gains to explore their effect on these classifications. These preliminary studies indicate that it is possible to “fine-tune” such classifications.

INTRODUCTION

OUR MOTIVATION for conducting this study was to develop methods for controlling and “fine-tuning” the percepts elicited by “analog” multichannel electrical stimulation of the cochlea in humans. Secondly, we needed to develop perceptual measures that would be useful in validating and optimizing such control strategies. This paper describes our initial progress in measuring and controlling the percepts elicited with such stimuli. Our study was further motivated by the efforts of Eddington [3] to determine what features of the stimulus could be used by analog, multichannel prosthesis recipients in speech recognition tasks. We wanted to estimate the relative importance of these features.

When designing speech processing and neural stimulation strategies for multichannel cochlear implants, it is

helpful to know which aspects of the speech signal and which aspects of the signals at the implanted electrodes are responsible for the subject’s speech identification ability. This information is difficult to extract by analyzing phonetic contrasts in a consonant or vowel confusion matrix since there are often several acoustic characteristics underlying the perception of each phoneme [5], [6], [11], [12]. Computer-generated speech has an advantage over natural speech in this regard. One can examine a particular acoustic aspect of the phoneme while controlling others. Computer-generated speech has been used with a single-channel cochlear implant subject [13], and more recently with multichannel devices [4], [1]. In all cases, the speech stimuli have been manipulated in an attempt to show how these subjects use cochlear stimulation patterns to decode speech. It is hoped that a more thorough understanding of these issues will enable improvements in speech processing strategies.

I. EXPERIMENT 1

The set of experiments described in this paper was designed to examine appropriate methods for processing the second formant frequency (F_2) using an “analog” processor driving an intracochlear electrode array. In experiment 1 isolated vowels were examined. These stimuli had many features of naturally produced stimuli but were modified slightly to remove some possible identification cues. The modifications were: 1) Vowel duration was fixed at 250 ms for each of the vowels. 2) The fourth formant frequency was fixed at 3300 Hz and the fifth formant frequency was fixed at 3750 Hz. 3) The first formant frequency (F_1) was not varied over the duration of the vowel. 4) F_1 ’s were manipulated slightly to reduce acoustic differences between vowels (see Table I). In spite of these controls, the vowels differed in several acoustic features. In experiment 2 only F_2 differed among stimuli. In experiment 2 we measured vowel classifications for F_2 ’s evenly distributed across the entire F_2 range. Also, in experiments 2 and 3 we explored the effect on vowel classification of certain manipulations of the processor.

A. Methods

1) *Subject*: Subject ET was 68 years old at the time of implantation and testing. He had a gradual onset of hear-

Manuscript received May 19, 1989; revised November 28, 1989. This work was supported by NIH Grant NS-11804 and N01-NS-9-2401 and the Office of Naval Research Grant N00014-89-J-1461. The majority of this work was conducted in the Coleman Laboratory, University of California, San Francisco.

M. W. White is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27695.

M. T. Ochs is with the Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN 37232.

M. M. Merzenich is with the Department of Otolaryngology, University of California at San Francisco, San Francisco, CA 94143.

E. D. Shubert is with the Department of Hearing and Speech Science, Stanford University, Stanford, CA 94305.

IEEE Log Number 9037730.

TABLE I
FORMANT FREQUENCIES OF VOWEL STIMULI USED IN EXPERIMENT 1

Vowel	F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	F5 (Hz)
<i>i</i>	320	2020-2070	2960-2980	3300	3750
<i>l</i>	450	1800-1600	2570-2600	3300	3750
<i>u</i>	320	1250-900	2500	3300	3750
<i>o</i>	450	1100-900	2500	3300	3750
<i>a</i>	700	1220	2600	3300	3750

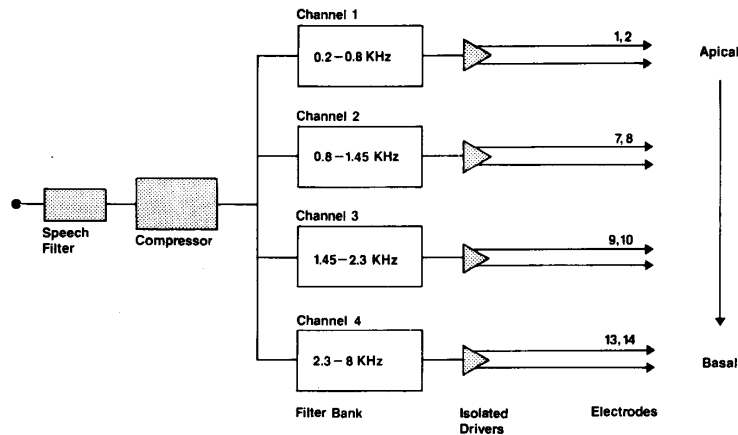


Fig. 1. Block diagram of the four channel speech processor and electrode hookup. Filter outputs were connected to the electrode array in a tonotopic fashion, i.e., low frequency energy was directed to the most apical pair of electrodes, and energy in the highest passband was directed to the most basal pair of electrodes.

ing loss due to otosclerosis until he became profoundly hearing impaired, 15 years before being implanted. Of the few patients that we have observed, this patient was one of the "better performers" on standard speech recognition tests. Testing was accomplished as part of the cochlear implant program of the University of California, San Francisco.

2) *Stimuli*: Five isolated vowels (*a*, *o*, *u*, *i*, *l*) were generated using a cascade synthesis routine [7] and a Data General Eclipse s/130 computer. Stimuli were output on a 12-b D/A converter at a rate of 10 kHz, and low-pass filtered at approximately 4.5 kHz. All vowels had five formants. To create natural sounding vowels, the second and third formant frequencies varied, in some cases, over the duration of the vowel. Formant frequencies and ranges are listed in Table I. The fundamental frequency began at 120 Hz and fell to 105 Hz over the duration of each vowel. To remove duration as a potential cue for vowel identification, all stimuli were 250 ms in length.

3) *Instrumentation*: Tape recorded speech stimuli were presented in a sound field and introduced to the speech processor via an environmental microphone. A block diagram of the speech processor is presented in Fig. 1. This four-band configuration is very similar to the processor studied by Eddington[3]. Incoming speech stimuli were

initially filtered by the "speech filter." This filter preemphasized the speech signal using a single-pole highpass filter (-3 dB at approximately 600 Hz). In addition, spectral components below 200 Hz were strongly attenuated using a three-pole Butterworth highpass filter with a -3 dB gain at 200 Hz. The resultant signal was amplitude compressed using a Gain Brain II, manufactured by Valley People Inc., Nashville, TN. The compression ratio was set to 5:1 with an attack time of approximately 0.2 ms and a release time of 5 ms. The compressed signal was passed through four bandpass filters, the output of each going to a single bipolar electrode pair via an optically-isolated, controlled-current source (i.e., the "isolated drivers" in Fig. 1). The gain of each isolated driver was patient-adjustable so that the loudness of each channel could be equated with the other channels (see the "Procedure" section). Each bandpass filter was constructed by cascading a three-pole Butterworth low-pass filter with a three-pole Butterworth highpass filter. These filters exhibited asymptotic "rolloffs" of 18 dB/octave on each side of the passband. Fig. 2 contains a graph illustrating how the gain of each filter varies with frequency.

The scala tympani electrode array consisted of eight bipolar pairs of electrodes located along the apical 14 mm

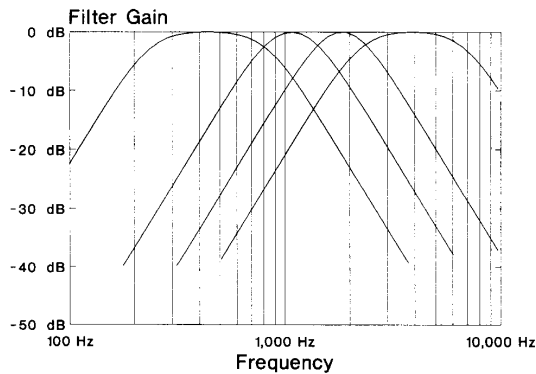


Fig. 2. Frequency response of the four bandpass filters in the speech processor. Vertical dimension represents the filter's gain in decibels.

of a 24 mm coiled Silastic insert. The apical-most bipolar electrode was inserted approximately 21–24 mm into the scala. Each electrode contact was mushroom-shaped to increase its surface area. The eight bipolar electrode pairs were spaced at 2 mm intervals. The intercontact spacing between bipolar contacts was approximately 700 μm , center-to-center. The bipolar electrode pairs were oriented approximately radial, and slightly diagonal to the axis of the cochlea. The electrode array and the implantation procedure are described in more detail by Loeb *et al.* [8] and Merzenich *et al.* [9].

Numbering of electrodes began at the apical-most part of the array and progressed basally, such that the apical-most bipolar pair was labeled “(1–2)” and the basal-most bipolar pair was labeled “(15–16).” An odd-numbered electrode represents an electrode contact placed more towards the modiolus (medial) than the even-numbered (lateral) contacts.

Preliminary testing revealed that electrode pairs 1–2, 7–8, 9–10, and 13–14 (located apex to base, respectively), connected to the output of channels 1, 2, 3, and 4, respectively, provided a good level of word recognition. This arrangement was used throughout these experiments. Only four electrode channels were used because: 1) preliminary psychophysical measurements indicated that simultaneous stimulation of more than four channels could cause substantial electric field interactions among the channels, and 2) only four processing channels were available for patient testing.

It is likely that the electrically-evoked excitation pattern generated by this electrode array was located basal to where excitation would have occurred in a normal cochlea when driven with speech.

4) *Procedure:* The subject set the level of the signal at each electrode pair using a loudness scaling procedure. He was instructed that on a scale of 0 to 10, a sound given a loudness rating of 0 would denote an inaudible sound, whereas a rating of 10 would denote an uncomfortably loud sound. The subject indicated that those stimuli that he scaled at 5–5.5 were “at a most comfortable listening level.” While listening to a tape-recorded male narrative,

the subject was instructed to increase the level at each electrode independently until he reached a specified loudness. This target loudness for individual channels was chosen such that: when all channels were then used simultaneously, the sensation “summed” to a loudness of 5.0–5.5 on the subject’s loudness scale.

Vowel identification ability was examined in a closed-set fashion. Stimuli were presented in isolation and the subject was instructed to select the word whose vowel was most like what he heard. The subject was given five words to select from: bought, boat, boot, beet, and bit. The subject was permitted to indicate when the stimulus sounded like none of these choices; however, no such responses were observed. Each vowel was presented five times. The five choices were displayed on a computer terminal. Each of the choices was displayed with a unique single digit number. The subject selected a word by pressing the corresponding number on the terminal keyboard.

After the vowel identification portion of the experiment, the output level of each bandpass filter was measured. The rms amplitude of each bandpass filter’s output was computed from the time waveform, for each of the synthesized vowels using interactive laboratory system (ILS) routines. Because the measurements were taken prior to the isolated drivers, these measurements were unaffected by the patient’s adjustment of the gain of the isolated drivers.

B. Results

Fig. 3 presents a confusion matrix for the five vowels examined in this experiment. The results are tabulated with the rows representing the stimulus presented and the columns representing the response given. The numbers in each cell denote how often a particular response was given to each stimulus. The data in Fig. 3 reveal that although *i* was consistently mistaken for *I*, these two vowels were never confused with the low formant vowels. At the bottom of Fig. 3 the responses are collapsed to demonstrate the perfect identification of high versus low second formant vowels.

Fig. 4 displays the output levels of the four bandpass filters for the five synthetic vowels. (See the immediately following paragraphs for a detailed analysis.) In brief, the output levels of channels 2 and 3 carried enough information to determine whether vowels were “high F_2 vowels” (*i*, *I*) or “low F_2 vowels” (*a*, *o*, *u*). If we represent channel 2 and 3’s output levels as L_2 and L_3 , respectively, then those vowels with significantly larger L_3/L_2 ratios were “high F_2 vowels.” F_2 could be coarsely estimated using a monotonically increasing function of L_3/L_2 .

C. Discussion

We expected the first formants of each of the vowels to pass through channel 1’s filter without loss. As expected, the output level at channel 1 was essentially the same for all vowels.

As is indicated in Table I, *a*, *o*, and *u* had relatively low frequency second formants. That is, the second for-

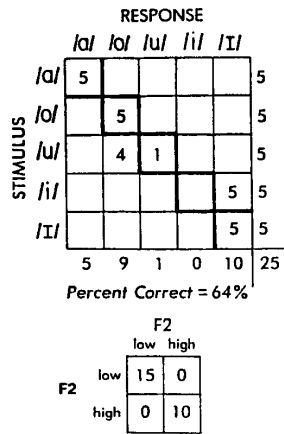


Fig. 3. Confusion matrix for five repetitions on the vowel identification task. The vowels *a*, *o*, and *u* had second formant frequencies that were low enough to pass through channel 2 of the processor. The vowels *i* and *I* had "high" frequency second formants which passed through channel 3 and stimulated a more basal pair of electrodes. The data are collapsed to reveal identification of high versus low second formant vowels at the bottom of the figure.

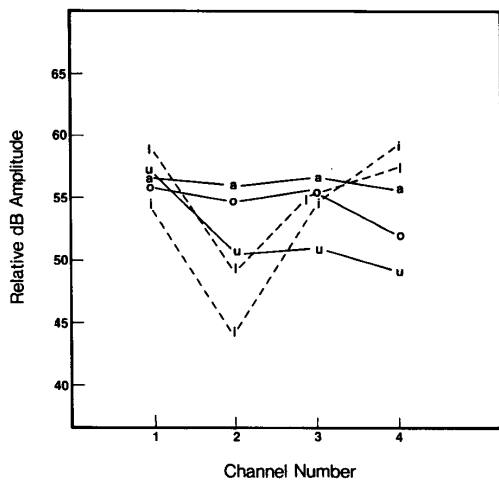


Fig. 4. Analysis of the relative amplitudes at the output of each channel of the processor for each vowel. Vowels connected by solid lines had "low" second formant frequencies (less than 1450 Hz), whereas vowel connected by dashed lines had "high" second formant frequencies (greater than 1450 Hz).

ment was in the frequency region of filter 2. The second formant of these vowels was expected to activate channel 2 of the processor. As expected the output level of channel 2 was higher for these vowels than for the high *F2* vowels, *i* and *I*. Initially we had expected that channel 3 would have relatively low output levels for the *a*, *o*, and *u* vowels (i.e., when compared to the output levels for the high *F2* vowels) since their *F2*'s are attenuated 7-12 dB by the channel 3 filter. However, the third formant (*F3*) also contributes substantially to the output of channel 3 since the channel 3 filter only attenuates *F3* by a few decibels. As a consequence, the output level of channel 3 can

be relatively high for both high *F2* and low *F2* vowels. The data in Fig. 4 are consistent with these expectations.

The third, fourth, and fifth formants of all vowels activate channel 4. Also, for the high *F2* vowels, *F2* contributes significantly to the output of channel 4. Again the data are consistent with this prediction since the channel 4 output levels are higher for the high *F2* vowels.

Although only of secondary importance, another factor affected the channel output levels. With the cascade synthesizer, the magnitudes of the formants decrease as the distance between the formants increase. This is a direct consequence of the cascading of bandpass filters: each bandpass filter attenuates the passband of the other filter more as the center frequencies of the two filters diverge. This is the most likely explanation for the relatively low channel output levels for *u* compared to those for *a* and *o*.

In summary, an examination of the channel output levels reveals that there is a relatively simple relationship between the channel output levels and the second formant frequency. For the high *F2* vowels the output levels of channels 3 and 4 are markedly higher than the output level of channel 2; whereas for the low *F2* vowels the output levels of channels 2, 3, and 4 are all about the same. Because vowels with high *F2* produced output level patterns different than those produced by low *F2* vowels, the output level patterns could be a cue for the recognition of vowels. Consistent with this observation, the confusion matrix data for this subject revealed that there was no confusion between high and low *F2* vowels (see confusion matrix at bottom of Fig. 3). However, other features of the electrical stimulus could have been used by the subject to classify the vowels in the manner observed. For example, fine-grain temporal waveform information could have been used by the subject. One of the purposes of experiment 2 was to determine the relative importance of "temporal" versus "place" cues.

Fig. 3 indicates that the frequency of the first formant may have played a minor role in the subject's ability to identify the vowels (e.g., The subject could correctly identify *a* versus *o*, *u*). However, there is not enough data to even conjecture on the roles of "temporal" or "place" cues in this identification.

II. EXPERIMENT 2

In experiment 1 the vowels differed in several acoustic features. In experiment 2, we synthesized vowels which only differed in *F2*. In this experiment, we measured vowel classifications for vowels with *F2*'s evenly distributed across the entire *F2* range. Also, we explored the effect on vowel classification of certain manipulations of the processor. One set of manipulations (i.e., channel reversal) was designed to determine the relative importance of "fine-grain temporal information" versus "place of excitation information." Other manipulations (i.e., changing filter passbands, changing channel gains) were used to explore methods for "fine-tuning" a subject's classifications.

A. Methods

1) *Stimuli*: Ten steady-state vowels were generated and presented to ET (see the previous experiment for synthesis details). All stimuli had the following formant frequencies: $F_1 = 320$ Hz, $F_3 = 2400$ Hz, $F_4 = 3300$ Hz, $F_5 = 3750$ Hz. The second formant frequency was set to one of the following: 1150, 1250, 1350, . . . , 2050 Hz. In all cases, formant frequency remained constant throughout the duration of the stimulus. All stimuli in the series were 250 ms in duration.

2) *Instrumentation*: As in experiment 1, the computer-generated stimuli were recorded on analog magnetic tape. Tape-recorded stimuli were presented in a sound field and were processed and delivered to the patient's electrode array in the fashion described in experiment 1 and summarized by Fig. 1. The mapping between bandpass filters and electrodes, as illustrated in Fig. 1, will be referred to as "tonotopic mapping."

In the second portion of the experiment, the outputs of channels 2 and 3 were reversed in terms of the area of the basilar membrane they stimulated (see Fig. 5). Low frequency second formant energy from channel 2 (below 1450 Hz) was directed to the more basal pair of electrodes devoted to second formant energy (electrode pair 9-10). Channel 3 information (i.e., second formant energy above 1450 Hz and third formant energy) was directed to the more apical position (electrode pair 7-8). This mapping between the bandpass filters and electrodes will be referred to as "channels 2 and 3 reversed."

3) *Procedure*: Individual channels were balanced for loudness using the 0-10 scale described previously (see experiment 1 for details). When channels were activated simultaneously, all vowel stimuli were in the 4.0-5.0 range of loudness, which the subject considered comfortably loud.

While listening to a playback of the analog magnetic tape, normal hearing listeners perceived these stimuli as a continuum of vowels from *u* to *i* as second formant frequency increased from 1150 to 2050 Hz. Although the stimuli sounded somewhat artificial, they were clearly identifiable to a normal hearing listener. The subject was not told which vowels were involved and was allowed to supply his own phonetic labels to the end points of the continuum. The subject identified the vowel with a second formant of 1150 Hz as *o* and the stimulus at the other end of the continuum (2050 Hz) as *i*. These phonetic labels will be used to refer to the stimuli throughout this discussion.

The ten vowels were randomized into six lists of 30 items each. Two lists were presented as practice trials. Then, 18 trials of each vowel were presented randomly and the subject was asked to categorize each stimulus as either *o* or *i*. Prior to vowel identification testing, three of the lists were presented, and the subject was asked to give a loudness rating for each stimulus using the 0-10 scale. The average loudness of each of the ten vowels was calculated from the nine loudness ratings per vowel. Also,

using the nine loudness ratings per vowel, the standard deviation of the loudness ratings for each of the ten vowels was calculated. The averages were very similar when compared to the standard deviations for the individual vowels. Furthermore, the loudness of the vowels was not correlated with the second formant frequency of the vowels. As a consequence, it is very unlikely that the subject could accurately categorize vowels along a second formant continuum based on loudness.

B. Results

Fig. 6 summarizes the subject's vowel classifications for the two processor configurations illustrated in Figs. 1 and 5. In Fig. 6 the second formant frequency of each stimulus is represented along the abscissa and the percentage of stimuli identified as *i* is represented along the ordinate. The error bars in Figs. 6-8 represent one standard deviation around the mean. The standard deviation was calculated assuming a binomial distribution. The curve in Fig. 6 that is labeled "tonotopic mapping" refers to the situation where all four channels were presented to the electrode array in a tonotopic arrangement. Fig. 6 reveals that stimuli were seldom identified as *i* unless the second formant was greater than 1750 Hz. When the second formant was 1450 Hz or below, stimuli were consistently identified as *o*. This outcome is consistent with the findings of experiment 1, in that vowels identified as *i* had second formant frequencies of 1800 to 2020 Hz, and vowels identified as *a*, *u*, and *o* had second formant frequencies of 1250 Hz and below.

Fig. 6 also shows data collected when the output of channels 2 and 3 were reversed. When the electrode connections to channels 2 and 3 were reversed, a reversal in vowel identification was observed. Stimuli having second formants below 1350 Hz were consistently identified as *i* by the subject, whereas these stimuli were previously identified as *o*. During this experiment the electrode reversal procedure was not explained to the subject, and he did not report any change in the quality of the stimuli. The pattern of responses suggest, however, that the change in place of maximal stimulation of the basilar membrane resulted in a change in the identity of the vowel perceived.

For the vowel with the lowest F_2 (1150 Hz), the output levels of the bandpass filters were nearly identical to those for the vowel *u* in experiment 1 (see Fig. 4). This was not surprising since the two stimuli are nearly identical. For the vowel at the other extreme of the F_2 range (2050 Hz), the filter output levels were nearly identical to those for the vowel *i* in experiment 1. Again, this was to be expected since these two stimuli are nearly identical. As expected, the output levels for the other vowels (i.e., those vowels with intermediate values of F_2) were intermediate between the output levels of the two vowels representing the two extremes of F_2 . As in experiment 1, the bandpass filter output levels carried enough information to discriminate between vowels with different F_2 's. For

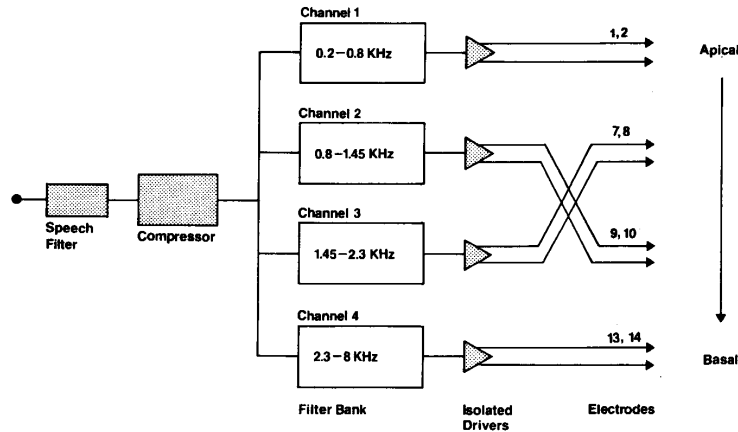


Fig. 5. Block diagram of the four-channel speech processor with the electrode connections to Channels 2 and 3 reversed.

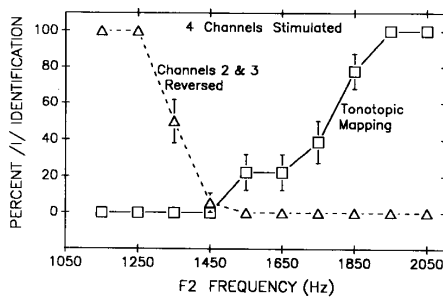


Fig. 6. Identification functions for a vowel continuum varying in second formant frequency. All four channels of the speech processor were used in this condition.

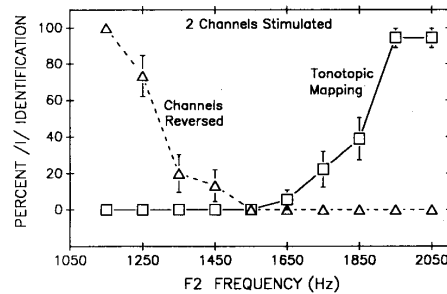


Fig. 7. Identification functions for a vowel continuum with stimulation only on channels 2 and 3.

example, if we denote channel 2 and 3's output levels as L_2 and L_3 respectively, then those vowels with larger L_3/L_2 ratios had higher frequency second formants.

To further simplify the stimulation pattern at the basilar membrane, identification of the vowel continuum was also done with channels 1 and 4 disconnected from the electrode array (these data are shown in Fig. 7). This arrangement ensured that *only* the two channels initially assigned to represent the second formant were stimulated. In other words, the two-channel configuration eliminated possible cues about the second formant's frequency that could be obtained from channel 4 and possibly channel 1. The subject reported that the stimuli sounded somewhat unnatural, but that they were still recognizable as speech. He elected to use the same phonetic labels for the end points of the continuum, although he noted that the *o* sounded somewhat like an *a*, and the *I* could also be identified as ϵ . Given the limited amount of information available to the subject, these "confusions" are not surprising. In the tonotopic mapping condition, stimuli having second formant frequencies of 1950 Hz and above were routinely identified as *I*, whereas second formants less

than 1650 Hz were identified as *o*. Reversal of the position of stimulation at the basilar membrane again resulted in a reversal of the vowel identification function.

In this two-channel mode the filter cutoffs for channels 2 and 3 were changed to explore methods for modifying or "fine-tuning" the vowel identifications of subjects. Whereas filter 2 had previously passed 800-1450 Hz, this was changed to pass a band of 800-1750 Hz. Filter 3 was changed from 1450-2300 Hz to 1750-2300 Hz. In this configuration, Fig. 8 reveals a corresponding shift in the category boundary. Presumably, the expansion of the passband of filter 2 increased the number of stimuli activating the more apical pair of electrodes, and resulted in an increase in the number of stimuli identified as *o*.

C. Discussion

In experiment 2, when the second formant of vowels was systematically varied across the F_2 continuum, this subject experienced a change in perceived vowel. Stimuli having second formants greater than 1750 Hz were almost always identified as *I* and those with lower frequency second formants identified as *o*. When the electrode stimulation pattern for the second formant was reversed so that

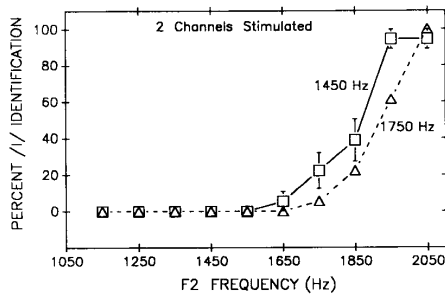


Fig. 8. Identification functions for a vowel continuum in the two-channel condition when the filter "crossover" frequency between channels 2 and 3 was changed from 1450 Hz to 1750 Hz.

tonotopicity was inverted, vowel identification functions reversed. In this case, stimuli that were previously identified as *I* were heard as *o*. Similar identification functions and reversals were observed when potential cues from other channels were removed. These data suggest the importance of relative cochlear position of stimulation for the identification of vowels. These data are supportive of cochlear implant coding strategies that make use of cochlear place information. Also, the results of the "fine-tuning" experiment (i.e., when the $F2$ and $F3$ filter cutoffs were changed) support this interpretation.

The channel reversal experiment is also useful for comparing the saliency of "the place of excitation cue" versus "the fine-grain temporal waveform cue." When channels 2 and 3 were reversed, the place cue was reversed, but fine-grain temporal patterns remained the same. It appears that "the place of excitation cue" was more important, because the subject's vowel identifications essentially reversed when the electrode connections were reversed.

Vowel stimuli with $F2$'s in the midrange of frequencies could evoke either identification from the subject. It may be that the rather gradual filter slope contributed some confusion in this frequency region. The tonotopic and reversed functions in Fig. 6 are not mirror images. It is possible that additional cues from channels 1 and 4 could have contributed to the asymmetry of the two identification functions. Consistent with this concept, when channels 1 and 4 are disconnected the tonotopic and reversed functions become considerably more symmetrical (see Fig. 7).

The data indicate that the subject classified the stimuli as *o* more often than *I*. Perhaps the relative gains of channels during experiment 2 "biased" the results in this manner (see experiment 3). Or perhaps the subject was somehow "biased" for *o* responses.

Although the elimination of channels 1 and 4 made the stimuli sound less speech-like to the subject, the experimental results support the notion that changes in second formant frequency can be conveyed by relatively small shifts in place of excitation (in the two-channel experiment, the two stimulated bipolar electrodes were only 2 mm apart).

III. EXPERIMENT 3

In experiment 3 we tried to manipulate second formant discrimination by changing the relative gains of channels 2 and 3. In this experiment, we used the two-channel processor described in the previous section. The idea behind the experiment was simple: If we increase the gain of channel 3 (with channel 2's gain unaltered) we would expect the subject to classify more of the vowels as high $F2$ vowels. This is one of the simplest methods for "fine-tuning" vowel discriminations.

A. Methods

1) *Stimuli and Instrumentation*: Stimuli and instrumentation were the same as those described for experiment 2 with the two-channel processor (i.e., channels 1 and 4 disconnected) connected in the normal "tonotopic" configuration.

2) *Procedure*: The procedure was similar to that described in experiment 2, but with the following addition: the subject's identifications were measured for two settings of channel 3's gain. The first setting for channel 3's gain was determined using the same procedure as described in experiments 1 and 2. This gain is referred to as the "normal gain" setting. The second setting was "slightly" higher (i.e., about 2 dB) than the first, and is referred to as the "high gain" setting. We verified that channel 3's output was increased appropriately by monitoring the channel's output with a battery-powered oscilloscope while adjusting the channel's gain. When both channels were stimulated, the loudness of the vowels for both settings was reported to be the same by the subject. In this experiment, loudness judgments were made after the subject listened to 5–10 vowel presentations at each of the two gain settings. This procedure was less accurate than that used in experiment 2 where the loudness of many randomly presented vowel tokens was scaled by the subject.

B. Results

Fig. 9 summarizes the results of experiment 3. Each bar represents the percentage of times that the subject classified the vowel stimuli as *I* and not *o*. The left bar represents the trials when channel 3's gain was "normal" and the right bar represents the trials when channel 3's gain was "high."

C. Discussion

Fig. 9 indicates that increasing channel 3's gain caused the subject to classify a much larger portion of the vowels as *I*. This, of course, is what one would predict on the basis of a "place coding" hypothesis.

Although unlikely, there is a possibility that small loudness differences for the two gain settings could have affected the subject's identifications. In hindsight, it would have been useful to do an additional experiment, in which the gain of channel 2 was increased instead of channel 3. By comparing the identifications for these two "symmet-

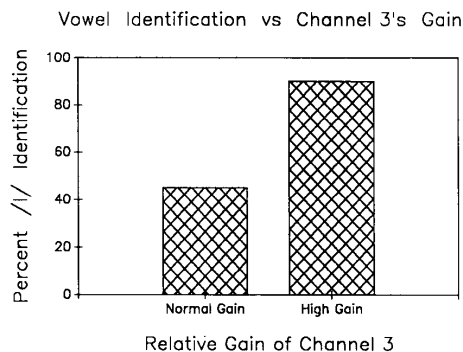


Fig. 9. Bar graph summarizing the results of experiment 3. Each bar represents the percentage of times that the subject classified the vowel stimuli as *I* and not *o*. The left bar represents the trials when channel 3's gain was "normal" and the right bar represents the trials when channel 3's gain was "high."

rical" stimulus conditions, it should be possible to determine the relative importance of the two potential cues, i.e., "the place of excitation cue" versus "the loudness cue."

It is interesting that about 45% of the vowels were classified as *I* for the "normal gain" setting in experiment 3. In contrast, only 27% of the vowels were classified as *I* in experiment 2, under supposedly "identical conditions." These differences could not be solely the result of statistical variation. This difference in identification results probably indicates that the procedure for adjustment of channel gains is not repeatable to the desired degree of accuracy. Thus, a procedure with better repeatability should be developed.

IV. CONCLUSION

Experiment 1 revealed that this subject was capable of identifying vowels differing primarily in their second formant frequency. Vowels having high frequency second formants were confused with each other, but never with vowels having low frequency second formants. In experiment 2, stimuli forming a continuum in terms of second formant frequency were utilized. The stimuli were identical except for second formant frequency. When the stimulation pattern was reversed at the basilar membrane, or the filter cutoffs were manipulated, or the relative gains of the channels were manipulated, the percept changed in the expected ways. These data suggest that relative changes in the place of stimulation at the basilar membrane can be important in the perception of vowels. This conclusion is consistent with another research group's experience with a pulsatile processor [2]. Experiment 2 has demonstrated the relative importance of "place" versus "fine-grain temporal" cues. With our analog processor both types of information were available and therefore we were able to compare their importance by "putting the two cues in conflict." We did this by reversing the electrode connections to channels 2 and 3. Although the place cue "dominated the identifications" in this experiment, it

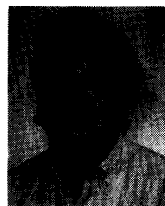
is nonetheless possible that fine-grain temporal information can be useful in vowel recognition. In fact, evidence from another study [13] indicates that fine-grain temporal information can be useful for recognizing vowels with significantly different first formant frequencies.

ACKNOWLEDGMENT

The authors wish to thank all members of the UCSF cochlear implant team for their efforts in developing a cochlear prosthesis for the deaf. We would also like to thank the Computer Systems Laboratory at North Carolina State University for their support.

REFERENCES

- [1] M. F. Dorman and G. McCandless, "Auditory/phonetic categories in a patient using a multichannel cochlear implant," *J. Acoust. Soc. Amer.*, 81 (Supplement), vol. 55, 1987.
- [2] R. C. Dowell, Y. C. Tong, P. J. Blamey, and G. M. Clark, "Psychophysics of multiple-channel stimulation," in *Cochlear Implants*, R. A. Schindler and M. M. Merzenich, Eds. New York: Raven, 1985, pp. 283-290.
- [3] D. K. Eddington, "Speech recognition in deaf subjects with multichannel intracochlear electrodes," *Ann. NY Acad. Sci.*, vol. 405, pp. 348-359, 1983.
- [4] L. J. Hood, M. A. Svirsky, and J. Cullen, "Discrimination of complex speech related signals with a multichannel electronic cochlear implant as measured by adaptive procedures," *Ann. Otol. Rhinol. Laryngol.*, Supplement 128, pp. 38-41, 1987.
- [5] A. S. House, "On vowel duration in English," *J. Acoust. Soc. Amer.*, vol. 33, pp. 1174-1178, 1961.
- [6] D. Kewley-Port, D. B. Pisoni, and M. Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, pp. 1779-1793, 1983.
- [7] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Amer.*, vol. 67, pp. 971-995, 1980.
- [8] G. E. Loeb, C. L. Byers, S. J. Rebscher, D. E. Casey, M. M. Fong, R. A. Schindler, R. F. Gray, and M. M. Merzenich, "The design and fabrication of an experimental cochlear prosthesis," *Med. Biol. Eng. Comput.*, vol. 21, pp. 241-254, 1983.
- [9] M. M. Merzenich, C. L. Byers, M. W. White, and M. C. Vivion, "Cochlear implant prostheses: Strategies and progress," *Ann. Biomed. Eng.*, vol. 8, pp. 361-368, 1980.
- [10] M. E. H. Schouten, "The case against a speech mode of perception," *Acta Psychologica*, vol. 44, pp. 71-98, 1980.
- [11] K. N. Stevens and D. H. Klatt, "Role of formant transitions in the voice-voiceless distinction for stops," *J. Acoust. Soc. Amer.*, vol. 55, pp. 653-659, 1974.
- [12] Q. Summerfield, "Speech-processing alternatives for electrical auditory stimulation," in *Cochlear Implants*, R. A. Schindler and M. M. Merzenich, Eds., New York: Raven, 1985, pp. 195-202.
- [13] M. W. White, "Formant frequency discrimination and recognition in subjects implanted with intracochlear stimulating electrodes," *Ann. NY Acad. Sci.*, vol. 405, pp. 348-359, 1983.



Mark W. White received the Ph.D. degree in electrical engineering and computer science from the University of California, Berkeley, in 1978.

He is currently an Associate Professor in Electrical and Computer Engineering at North Carolina State University, Raleigh. His research is concerned with cochlear implants, hearing aids, and learning algorithms for neural network signal processors.



Marleen T. Ochs received the Ph.D. degree in hearing and speech sciences from Vanderbilt University, Nashville, TN, in 1983. She had a post-doctoral fellowship in sensory physiology at the University of California, San Francisco.

She is currently an Assistant Professor in the Division of Hearing and Speech Sciences, School of Medicine at Vanderbilt University. Her research centers around speech perception in normally hearing and hearing-impaired subjects. She is particularly interested in changes in speech per-

ceptual strategy that accompany developing hearing loss in elderly subjects.



Michael M. Merzenich received his doctoral training in neurophysiology at The Johns Hopkins University, and in auditory neuroscience at the University of Wisconsin.

Since 1971, he has been a member of the faculty at the University of California, where he participates in Neuroscience, Speech and Hearing Science, and Biomedical Engineering doctoral training programs. His research has focused on basic features of organization of the central auditory nervous system; the development of electri-

cal stimulation cochlear prostheses and other aids for the profoundly deaf; and cortical representational plasticity underlying learned behaviors and the acquisition of skill. He is currently a Professor and the Vice Chairman for Research in the Department of Otolaryngology at the University of California at San Francisco.

Earl D. Schubert is Professor Emeritus of Hearing and Speech Science at Stanford University, Stanford, CA. He is affiliated with the Center for Computer Research in Music and Acoustics at Stanford University. His current interest is in auditory considerations in the perception of music, particularly their application to computer-generated music.