FORMANT FREQUENCY DISCRIMINATION AND RECOGNITION IN SUBJECTS IMPLANTED WITH INTRACOCHLEAR STIMULATING ELECTRODES [a]

M. W. White

Coleman Laboratory
Department of Otolaryngology
University of California, San Francisco
San Francisco, California 94143

Introduction

Our group in the Coleman Laboratory has studied a subject implanted with 16 intracochlear scala tympani stimulating electrodes. A single-channel analog processor was tested. The processor consisted of a microphone, preamplifier, an instantaneous log compressor, and a linear filter, in that order. The linear filter reduced the low-frequency energy below 200 Hz and pre-emphasized the higher frequencies at about 6–12 dB/octave. Each of the two poles of the single-channel processor was attached to eight of the 16 wires of the electrode array. One pole was connected to the more medial eight electrodes and one pole was connected to the more lateral eight electrode contacts. The subject (L.Y.) was evaluated with standard speech intelligibility tests, basic psychophysical tests, and tests designed to determine what features of the speech signal were used by the subject to understand speech. The subject exhibited surprisingly good performance in understanding speech without the aid of lipreading. Without lipreading, the subject could recognize about 50% of the key words in CID (Central Institute for the Deaf) everyday sentences.[1] In a four-choice vowel-recognition task, the subject could consistently identify 50–60% of the monosyllabic words that were spoken ($n = 60$).

Experiments, reported here, were conducted to estimate which acoustic features were being utilized by this subject with a cochlear implant.

Synthetic-Vowel Paired Comparisons

Pairs of steady-state synthesized vowels were presented to the subject through an Electro-Voice E-V FIVE-C loudspeaker. The subject (L.Y.) was then asked to scale the difference between the two sounds on a scale of one ("identical") to seven ("very different"). The steady-state vowels were 300 msec in duration, with a fundamental frequency of 125 Hz. A digitally implemented, cascade speech synthesizer was implemented using infinite impulse response digital filters (Fig. 1). The formant frequencies were held constant throughout the 300-msec duration of the synthesized vowel. A 400-msec delay separated the first vowel from the second vowel in the pair.

[a] This work was supported by Grant NS–11804 and Contract NO 1–NS–7–2367 from the National Institutes of Health, and by Hearing Research, Inc., San Francisco, California.

0077–8923/83/0405–0348 $1.75/0

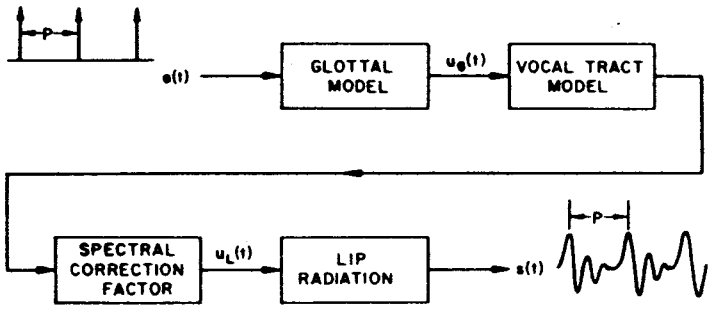

FIGURE 1. Block diagram of a cascade speech synthesizer. (From Markel and Gray.[6] Reprinted by permission.)

With synthetic speech, the experimenter can hold any or all speech variables constant. In this study, one variable in the vowel pair was changed and all the other variables were held constant. This is a considerable advantage in evaluating which speech features are discriminable by a subject. With natural speech, many parameters may co-vary and make interpretation more difficult. For example, vowel duration and the vowel's formant frequencies may co-vary.

Three pairs of synthetic vowels were generated in which the second formant frequency was significantly different in each pairing of the vowels, but the first formant frequency was held constant (TABLE 1). Two pairs of synthetic vowels were generated in which the first formant frequency was significantly different in each pairing of the vowels, but the second formant frequency was held constant within each pair (TABLE 2).

The five vowel pairs are listed in TABLE 3 with their respective formant frequencies. Each synthesized vowel was essentially identical in duration, fundamental frequency, third, fourth, and fifth formant frequencies (2500, 3400, and 3800 Hz, respectively), and the first through fifth formant bandwidths (100, 100, 170, 250, and 300 Hz, respectively).

When presented at the same RMS amplitude, each of the synthesized vowels was different in loudness to the subject. In order to determine whether the subject could discriminate between vowel pairs on the basis of spectral as opposed to gross amplitude information, an adaptive procedure was used to equalize the loudness of the two vowels within each vowel pairing.

TABLE 1

PERCEIVED DIFFERENCES BETWEEN TWO SYNTHESIZED STEADY-STATE VOWELS

	Phoneme:	
	/u/	/i/
Sounds like vowel in:	boot	beet
First formant frequency:	250 Hz	250 Hz
Second formant frequency:	900 Hz	2250 Hz
Subject's comments:	"oo's"	"oo's"
Subject's rating of the difference between the two sounds (scale 1–7, where 1 = no difference and 7 = very different)	1.25 (mean)	

TABLE 2

PERCEIVED DIFFERENCES BETWEEN TWO STEADY-STATE VOWELS

	Phoneme:	
	/i/	/æ/
Sounds like vowel in:	beet	bat
First formant frequency	250 Hz	1000 Hz
Second formant frequency	2250 Hz	2250 Hz
Subject's comments:	"oo's"	"ah's"
Subject's rating of the difference between the two sounds (scale 1–7, where 1 = no difference and 7 = very different)	7 (mean)	

One of the two vowels in a vowel pairing was held constant in level with what the subject considered a comfortable listening level. A 2I-2AFC adaptive procedure was used to equate the loudness of the other vowel in the vowel pair with that of the original vowel. In the adaptive procedure, the order of the two vowels was randomized in order to reduce temporal effects (for example, the first of two identical stimuli may sound louder). The subject was asked to press the button corresponding to the interval of the loudest sound. The adaptive procedure automatically adjusted the level in such a manner as to reduce the difference in loudness between the two vowels. In this manner, the difference in loudness between the two sounds was reduced until the subject was unable to differentiate between the two sounds on the basis of loudness.

After the vowels' loudnesses were equalized, pairs of vowels were presented to the subject. The subject was asked to "rate" or scale how different the two sounds were on a scale of one to seven, where "one" meant that the two stimuli sounded identical and "seven" meant that the two sounds were very different.

TABLE 3

FIVE SYNTHETIC VOWEL PAIRINGS USED IN THE SYNTHETIC VOWEL EXPERIMENT

	F1 (Hz) (1–7)	F2 (Hz)	Phonetic Symbol	Difference (Scaled)
Vowel pairing #1	700	1750	(ae)	2.0
	700	1100	(aa)	
Vowel pairing #2	550	800	(ao)	1.3
	550	1900	(eh)	
Vowel pairing #3	250	2250	(iy)	1.25
	250	900	(uw)	
Vowel pairing #4	250	2250	(iy)	7.0
	1000	2250	(ae)	
Vowel pairing #5	300	750	(uw)	6.0
	550	750	(ao)	

NOTE: The mean of the subject's response for each vowel pair is listed in the last column.

The five vowel pairings were presented in a pseudo-random order. Each one of the five vowel pairs was presented at least ten times during a test set. Two test sets were devoted to each of three different levels of high-frequency pre-emphasis of the second formant frequency region. The second formant frequency region was "emphasized" or amplified relative to the first formant frequency region in an effort to improve the subject's second formant discriminative ability.

FIGURE 2 and TABLE 3 present the averages of the subject's scaling of the differences between the synthetic vowels. If only the first formant frequencies (F1) of the two synthesized vowels were different, the subject rated the difference between the two sounds between 5.5 and 7 on the difference scale. Such vowel pairings sounded quite different to the subject. However, when only the second formant frequencies (F2) were different, the subject rated the difference between the two sounds between 1 and 2.5 on the difference scale. Many times such vowel pairings would sound "identical" or "almost identical."

When the subject was asked to describe the sounds that he heard, he would consistently describe the steady-state vowels with low first formant frequencies as /u/ (that is, an "oo" sound, as in boot), regardless of the second formant frequency (TABLE 4). Subject L.Y. would consistently describe the steady-state vowels with high first formant frequencies as /ɔ/ (that is an "ah" sound, as in law). The second-formant frequency did not appear to have any effect on L.Y.'s description of these steady-state vowels.

Similar results were obtained when the higher frequencies (that is, the second formant frequency region) of the synthesized vowels were amplified or "pre-emphasized" by an additional 6-dB/octave highpass filter. The third level of pre-emphasis involved the elimination of the first formant in the three vowel pairs in which only the second formant frequency was different. This was an attempt to reduce any "confounding" signals that might interfere with the second formant information. Unfortunately, the results were similar to those just reported. The subject could hear little, if any, difference between these vowels.

ERROR ANALYSIS OF A FOUR-CHOICE VOWEL TEST

The results of the synthetic vowel experiment prompted an analysis of the errors made in a natural-voice four-choice vowel test. In the test, one word was spoken and the subject was asked to choose which of the four words listed was the word spoken. For example, the subject might have to choose from the

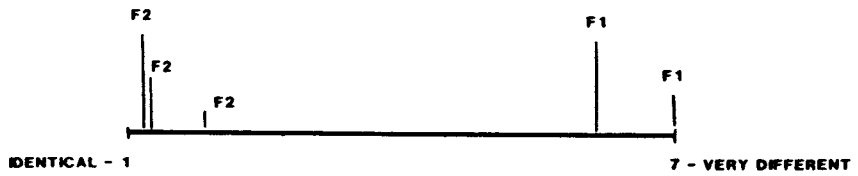

FIGURE 2. Averages of the subject's scalings of the differences between vowel pairs. The vertical lines labeled F2 represent scalings in which only the second formant frequencies are different in the two vowels, while the vertical lines labeled F1 represent scalings in which only the first formant frequencies are different in the two vowels.

TABLE 4

PERCEIVED DIFFERENCES BETWEEN TWO SYNTHESIZED STEADY-STATE VOWELS

	Phoneme:	
	/u/	/ɔ/
Sounds like vowel in:	boot	law
First formant frequency:	300 Hz	550 Hz
Second formant frequency:	750 Hz	750 Hz
Subject's comments:	"oo's"	"ah's"
Subject's rating of the difference between the two sounds (scale 1–7, where 1 = no difference and 7 = very different)	6 (mean)	

following list of four words—*not, note, net, night*—in which only one of the four words was spoken. The subject consistently obtained scores of 50–60% correct in these four-choice, 60-trial tests.[2] When the subject made an error, the similarity between the features of the selected word and that of the spoken (that is, "target") word may be an indicator of which features are used by the subject in the recognition task.

For example, if the chosen vowel's first formant frequency was consistently similar to the target vowel's first formant frequency, one might conclude that the first formant frequency was either a perceptually significant feature or that it was correlated with such a perceptually significant feature.

The similarity between the target vowel's first and second formant frequencies and the selected (or "chosen") vowel's first and second formant frequencies was measured. The similarity was ranked by comparing the target vowel's formant frequency with each of the three possible vowel formant frequencies.

TABLE 5 illustrates the method by which distance ranks were calculated. In the four-choice vowel test, one word of the four was spoken. The subject was to pick which of the four was spoken. When the subject made an error in his choice, the error analysis determined whether the subject had chosen a word with characteristics similar to the spoken word's features. The spoken word's first formant frequency (F1) was compared with the chosen word's F1.

TABLE 5

METHOD BY WHICH DISTANCE RANKS WERE CALCULATED

Description	Word	First Formant Frequency (F1)	F1-(F1 of Target Word)	Distance Rank
Word spoken (that is, target word):	bean	270 Hz		
Word selected by subject:	boon	300 Hz	30 Hz	1 (nearest)
Word not selected:	burn	490 Hz	220 Hz	2
Word not selected:	ban	660 Hz	390 Hz	3 (farthest)

NOTE: The selected word's first formant distance rank therefore is 1.

The difference between the spoken word's F1 and the chosen word's F1 was calculated. Also, the differences between the spoken word's F1 and the F1 of the other two words were calculated. The three differences were ranked relative to each other. The word with the F1 farthest from the spoken word's F1 was ranked "3". Each trial in which an error was made was assigned a number equal to the rank of the word chosen within that trial. The average of these numbers (that is, distance or similarity ranks) was calculated and compared with the probability that such an average could be obtained if the words were chosen by chance. The same ranking procedure was used to rank the similarities between the spoken word's F2 and the chosen word's F2.

Two methods for ranking were used: In one case, the absolute value of the difference between the target vowel's formant frequency and that of the other three vowels' formant frequencies were used to determine the "distance rank" (as in TABLE 5). The second method was identical to the first, except that the logarithms of the vowels' formant frequencies were used instead of a linear scale, as in the first method. The second method, in effect, meant that the distance ranks were determined from the ratio of the target vowel's formant

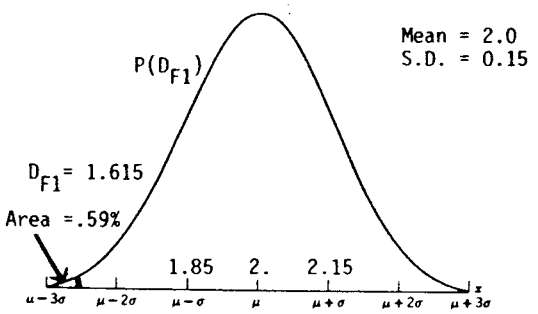

FIGURE 3. Probability of the average distance rank ($n = 26$), if F1 did not affect the subject's choice. DF 1 ($= 1.615$) was the actual average of the distance ranks for an average of 26 ranks. If F1 did not affect the subject's choice, a rank average of 1.615 or less would occur only in 0.59% of the averages.

frequency and that of each of the three possible vowel formant frequencies. Interestingly enough, only in one of the 26 trials (in which the subject made an error) were the distance ranks different, as determined by the two methods. Because there was so little difference between the two methods, the ranks obtained from the linear frequency scaling method were arbitrarily chosen for the analysis.

The mean of the first formant distance ranks was 1.615 for the 26 error trials that were examined. If the subject's errors were not at all related to the similarity of the target and selected vowels' first formant frequencies, one would expect an average distance rank that was more nearly equal to "2.0". In fact, an average distance rank of 1.615 could only very rarely occur (approximately only 0.6% of the time) if the subject was not basing his decisions on some feature that was correlated with the vowel's first formant distance ranks (FIG. 3).

FIGURE 4 is a graph of the probabilities of second formant ranks, if the subject was not basing his decisions on some feature that was correlated with the vowel's second formant ranks. On this graph, a vertical line was drawn to illustrate the second formant distance rank average (DF $2 = 2.153$) obtained for this sample of 26 trials. This average distance rank is only slightly greater than 1 standard deviation from the mean rank average calculated by assuming

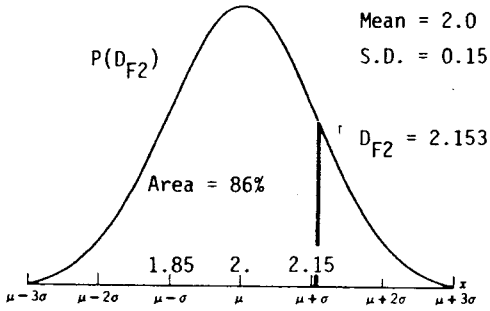

FIGURE 4. Probability of the average distance rank ($n = 26$), if F2 did not affect the subject's choice.

that F2 distance had no correlation with the subject's choice. The subject may not have used any second formant information. Or, if the subject did use second formant information, it is more likely that he used it in an inappropriate manner because he was more likely to choose a word containing a vowel with a second formant frequency farther from the spoken vowel's F2 than a word chosen by chance.

Another possibility exists. The second formant frequency distance ranks may be correlated (or negatively correlated) with the first formant frequency distance ranks. From the F1 error analysis described earlier, it seems likely that the subject's choices were correlated with the similarity between the first formant frequencies of the target and the selected vowel. If the F2 distance ranks are positively or negatively correlated with the F1 distance ranks, the second formant distance rank average could significantly deviate from what would be expected by chance—even if the similarity between the second formant frequencies of the target vowel and the selected vowel does not affect the subject's choice.

FIGURE 5 is a plot of the conditional probability of second formant frequency distance rank averages conditioned on a first formant distance rank average of 1.615. The second formant frequency's distance rank average of 2.153 is only about one-half a standard deviation from the "corrected" mean second formant distance rank average. This "corrected" or conditional probability distribution is a graph of the estimated probability of second formant distance rank averages if the subject did not use second formant frequency information. The "corrected" distribution function compensates for the *negative* correlation between the first and second formant distance ranks.

FIGURE 5 was obtained by fitting a continuous gaussian curve to the estimated discrete probability distribution. The fit was very good and simplifies the illustration. The discrete distribution was obtained by using a Monte Carlo

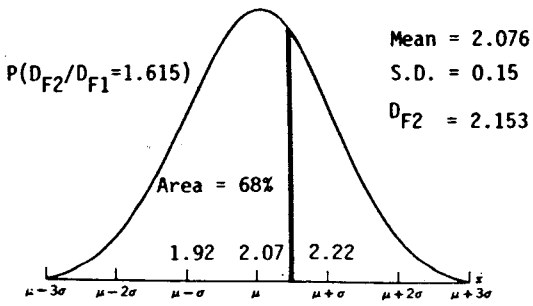

FIGURE 5. Probability of the average distance ranks of F2, if F2 did not affect the subject's choice of words, when DF1 was fixed at 1.615.

simulation technique. Ten thousand second formant distance rank averages were calculated. Each of the 10,000 second formant distance rank averages had a first formant distance rank average of 1.615. Otherwise, the vowel was randomly selected by the computer program in order to simulate a subject that was not using second formant information (but was using information that was correlated to first formant distance ranks). The 10,000 distance rank averages were used to generate a histogram that forms an estimate of the conditional probability distribution.

FIGURE 5 indicates that the subject's performance could be modeled as if he were not using second formant information, but only first formant information or information correlated with the first formants of the vowels.

SUMMARY OF RESULTS WITH SUBJECT L.Y.

It may be possible for some subjects with single-channel stimulation to use information that is carried in the frequency of the first formant. These experiments point strongly to the subject's ability to discriminate differences in steady-state vowels with significantly different first formant frequencies.

Can the subject use this discriminative ability in a natural speech-recognition task? The result just described was found to be consistent with the vowel identification task described earlier. The errors made by the subject in this four-choice vowel identification task were analyzed (error rate approximately 45%). When the subject made an error, he generally chose the word containing the vowel with a first formant frequency closest to the first formant frequency of the spoken word's vowel. The subject appeared to have no preference for vowels with a second formant frequency near the spoken vowel's second formant frequency.

One cochlear prosthesis group's [3] speech processor deletes any information pertaining to the first or the second formant frequency. Our data strongly imply that subject L.Y. was quite capable of using first formant frequency information.

In addition to first formant information, some voicing and nasality information may be obtained by some patients having implants.[2, 3] The Vienna group [4] has reported that some second formant information may be used by some of their patients.

MULTICHANEL SPEECH PROCESSORS

Recently we have tested a multichannel speech processor with another subject, C.B. The results have been encouraging. Both a single-channel speech processor (FIG. 6 and TABLE 6) and a three-channel speech processor (FIG. 7 and TABLE 7) were tested with this subject. The single-channel processor was similar to that used by subject L.Y.

Subject C.B. did not perform nearly as well as subject L.Y. did while using a single-channel processor. With her single-channel processor, C.B. performed in a manner similar to that of the majority of single-channel users who have been evaluated by Owens *et al.*[5]

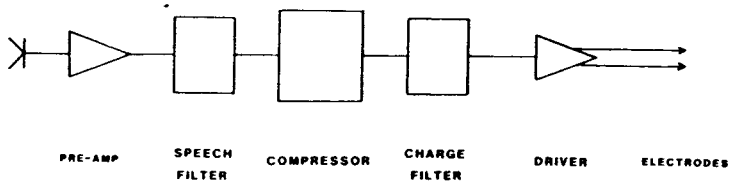

FIGURE 6. Block diagram of CB's single-channel speech-processor. TABLE 6 contains a list of the processor's specifications.

However, when the three-channel processor was utilized with subject C.B., her performance on a set of speech tests improved considerably (see Owens *et al.*[5]).

It is not safe to conclude that any given class (such as multichannel processors versus single-channel processors) of processors is optimal. It is very difficult even within individual subjects to ascertain which classes of speech-processors are optimal. It may be even more difficult to determine whether given classes of speech-processors are optimal for all, or groups, of implanted subjects. In subject C.B., a particular single-channel processor did not perform as well as a particular three-channel processor. However, additional adjustments of the single-channel processor variables may have improved her performance. Indeed, entirely different single-channel processors may have significantly improved her performance. The same can be said for the optimization of the three-channel processor.

In our processor adjustment and comparison procedure, we have used a vowel test and an initial consonant test to guide us in the optimization procedure. The tape-recorded tests are two-choice monosyllable word tests. These tests are designed to give us information relevant to processor design and adjustment. These natural speech tests evaluate to what degree the subject is able to utilize certain distinctive features (and/or other features correlated with these particular distinctive features). After many processor adjustments and quick evaluations (using the two speech tests described earlier and subject quality judgments), we then evaluate the subject for a much wider range of

TABLE 6

SINGLE-CHANNEL PROCESSOR SPECIFICATIONS FOR SUBJECT C.B.

Speech filter:	Single-pole highpass filter; —3 dB at 170 Hz.
Compressor:	Two-stage, cascaded compressors.
First stage:	150-μsec attack time. 300-μsec release time.
Second stage:	Instantaneous log compressor.
Compression ratio:	5:1 over the 60-dB input range. 10:1 over the upper 40 dB input range. The compression ratio increases dramatically at the higher input levels. The compressor clips at the higher input levels.
Charge filter:	Two-pole highpass filter; —3 dB at 300 Hz and —12 dB at 140 Hz.

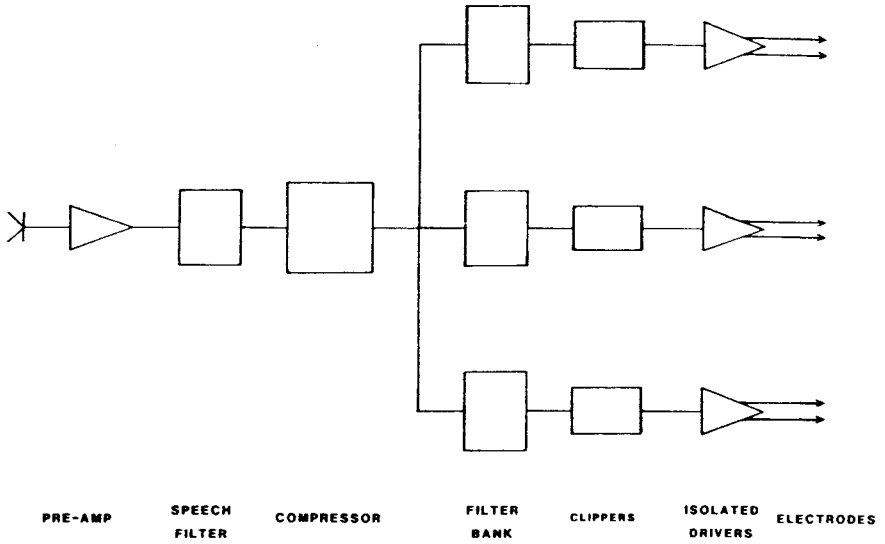

FIGURE 7. Block diagram of CB's three-channel speech-processor. TABLE 7 contains a list of the processor's specifications.

speech skills. The MAC battery [5] and several other speech tests have been developed for this purpose. This much larger and comprehensive test battery is our "benchmark" of performance for the implant subjects.

In the initial consonant test, subject C.B. was able to use voiced–unvoiced information (or information correlated with the voiced–unvoiced distinction) much more effectively with the three-channel processor than with the single-

TABLE 7

THREE-CHANNEL PROCESSOR SPECIFICATIONS FOR SUBJECT C.B.

Speech filter:	Single-pole highpass filter; —3 dB at 670 Hz.
Compressor:	Single stage, noninstantaneous compression. 0.5-msec attack time; 1.0 msec release time. 5–10 compression ratio over 60-dB input range. Compression greatly increases at the higher input levels.
Filter bank:	Three-pole highpass section; three-pole lowpass section. Both sections were designed for a filter function between a bessel and a butterworth characteristic.
	Low filter:[a] 200–800 Hz (—3dB intercepts). Mid filter: 850–1450 Hz. High filter: highpass filter; —3dB at 1450 Hz.
Limiters:	The bipolar clipping level is set approximately 0–20% above the maximal excursions of the speech signal.

[a] A "charge" filter was inserted immediately after the "low filter" to further reduce low-frequency components. The charge filter is a two-pole highpass filter (—3 dB at 300 Hz and —12 dB at 140 Hz).

channel processor. In the vowel test, C.B. was able to use first formant information (and/or information correlated with first formant frequency) more effectively with the three-channel processor than with the single-channel processor.

Initial measurements of an earlier version of the multichannel processor indicated that the subject was not using second formant information or information correlated with second formant frequency. Two changes were made to the multichannel processor to improve the subject's performance within this distinctive feature category. The upper 3-dB breakpoint of the "midrange filter" was adjusted from approximately 2000 Hz down to 1450 Hz and the lower 3-dB breakpoint of the "high filter" was adjusted from approximately 2000 Hz down to 1450 Hz. The purpose of this modification was to improve the separation of high-frequency second formants from low-frequency second formants. The second modification reduced the spectral distortion generated by the compressor that was located ahead of the filter bank. These modifications enhanced the capacity of the two upper channels to resolve differences in the second formant frequencies of speech segments. The improved resolution was measured by comparing the peak-to-peak output level of the two channels on a dual-trace oscilloscope for selected steady-state vowels. The vowels were chosen such that their second formant frequencies were considerably different. After these improvements, the subject C.B. performed significantly above chance on some of the tests used to measure her ability to use second formant information and/or information that was correlated with the second formant frequency. These tests do not give conclusive evidence that the subject was able to use the second formant frequency information in speech-recognition tasks. The subject may have used information that was correlated with second formant information. For example, the duration of vowel segments may have given a useful cue to the subject. However, the changes made in the processor that improved C.B.'s performance were primarily those that would affect only second formant resolution.

Summary

In a four-choice, vowel-identification task, the implanted subject (L.Y.) identified the spoken word in 50 to 60% of the trials ($n = 60$). Experiments were conducted to determine which acoustic features were being utilized by the implanted subject.

Pairs of steady-state synthesized vowels were presented to the subject. A difference-scaling procedure indicated that the subject could easily discriminate between synthetic vowels with different first formant frequencies. However, the subject had great difficulty in discriminating between synthetic vowels with different second formant frequencies when the first formant frequencies were identical.

This result was found to be consistent with an identification task using natural vowels in a single-word context. The errors made by the subject in a four-choice vowel-identification task were analyzed (error rate = 45%). When the subject made an error, he generally chose the word containing the vowel with a first formant frequency closest to the first formant frequency of the spoken vowel's first formant frequency. In contrast, when the subject made an error, he appeared to have no preference for vowels with a second formant frequency near the spoken vowel's second formant frequency. Consequently,

speech processors that do not transmit first formant frequency information may be less useful to implanted subjects.

With another subject (C.B.), a single-channel and a three-channel speech-processor were evaluated. The subject's performance with the three-channel processor was considerably better than that obtained with the single-channel processor in a series of speech tests.[5]

Acknowledgments

The entire staff of the Coleman Laboratory at the University of California, San Francisco, has contributed to this study. We wish to thank Pat Jones, Joe Molinari, Chuck Byers, Steve Rebscher, David Casey, Lindsay Vurek, Bob Shannon, Mike Merzenich, Robin Michelson, Bob Schindlar, Marcia Raggio, Dorcas Kessler, John Gardi, Earl Schubert, Elmer Owens, Chris Telleen, and Bill Garret for their valuable skills in helping us with our cochlear implant project.

References

1. Owens, E., D. Kessler & E. Schubert. 1982. Interim assessment of candidates for cochlear implants. Arch. Otolaryngol. **108:** 478–483.
2. Owens, E. & C. C. Telleen. 1981. Speech perception with hearing aids and cochlear implants. Arch. Otolaryngol. **107:** 160–163.
3. Fourcin, A. J., S. M. Rosen, B. C. J. Moore, E. E. Douek, G. P. Clark, H. Dodson & L. H. Bannister. 1979. External electrical stimulation of the cochlea: Clinical, psychophysical, speech-perceptual, and histological findings. Br. J. Audiol. **13:** 85–107.
4. Hochmair-Desoyer, I. J., E. S. Hochmair, R. E. Fischer & K. Burian. 1980. Cochlear prostheses in use: Recent speech comprehension results. Arch. Otorhinolaryngol. **229:** 81–98.
5. Owens, E., D. Kessler & M. Raggio. 1983. Results for some patients with cochlear implants on the minimal auditory capabilities (MAC) battery. This volume.
6. Markel, J. D. & A. H. Gray. 1976. Linear Prediction of Speech. Springer-Verlag. Berlin.