PERCEPTION OF SYNTHETIC VOWELS AND STOP CONSONANTS BY COCHLEAR IMPLANT USERS

P.J. Blamey and G.M. Clark

Department of Otolaryngology
University of Melbourne

ABSTRACT - Three multiple-channel cochlear implant users were tested with speech sounds that were synthesized using electrical parameters representing the fundamental frequency of the voice, and the frequencies and amplitudes of the first and second formants. Using vowels of equal duration and loudness, it was shown that most of the vowel recognition could be attributed to the formant coding. Unvoiced stops with varying burst frequencies, voiced stops with varying second formant loci, bilabial stops with varying voice onset times, and bilabial consonants with varying formant transition durations were also synthesized. For each consonant set, the responses showed similar patterns to those observed with normally-hearing listeners for analagous acoustic stimuli. Interactions between amplitude and frequency cues were observed.

INTRODUCTION

In this study speech sounds were synthesized directly as electrical stimulus patterns to test some of the assumptions that underlie the speech coding used with the Cochlear Pty Ltd multichannel cochlear implant (Clark *et al*, 1984). The speech processor explicitly estimates selected acoustic parameters and codes them in terms of electrical stimulus parameters. Most other schemes attempt to present the whole waveform or the whole spectrum to the patient (Millar *et al*, 1984). Summerfield (1985) has published a very clear discussion of the speech processing alternatives that may be appropriate in different situations. Advocates of parameter extraction schemes claim that these parameters can be presented to the patient more effectively than they are with the whole waveform or whole spectrum schemes. This argument is plausible because of the electrical and neural interactions between simultaneously stimulated channels, and the reduced psychophysical abilities of patients compared with normally-hearing listeners. Parametric coding schemes may be devised to avoid these difficulties (Tong *et al*, 1983a, 1983b). The advocates of wholistic processing claim to present more cues to the patient. So far, neither of these claims has received much more than theoretical justification, although very successful users of both types of device have been reported.

The present study aims to discover whether the cues presented by the speech processor are used effectively and in the same way as the corresponding parameters are used by normally-hearing listeners. Although it is conceivable that postlinguistically deafenned cochlear implant patients may develop new perceptual mechanisms for understanding speech, it seems preferable to use coding schemes that are as natural as possible to avoid the need for extensive relearning of speech patterns and to take advantage of specialized structures that may exist in the human brain for speech processing. It is therefore of interest to compare the patients' perception of these stimuli with normally-hearing listeners' perception of the corresponding stimuli. It is also hoped that experiments of this type will suggest improved coding schemes that achieve a closer match.

METHOD

A multiple-electrode receiver-stimulator (Clark *et al*, 1984) was used for these experiments. The receiver-stimulator was implanted in the mastoid bone behind the patient's ear and the array of 22 platinum electrodes inserted in the scala tympani through the round window. Control data and power were transmitted through the skin to the device by a radio frequency signal from an external coil. The receiver-stimulator controlled charge-balanced biphasic electrical pulses applied to pairs of electrodes in the cochlea. Each phase of the biphasic pulse lasted 200μs and the maximum current used was approximately 1.5 mA. The electrodes were spaced evenly along a length of 17 mm. The most apical electrode was inserted up to 24 mm into the cochlea.

Three patients took part in this study. All three were profoundly deaf adults who scored zero preoperatively on an open-set word recognition test using an appropriately fitted hearing aid. The patients were selected for this study on the basis that they all had at least 12 months experience with the implant, they all had above average speech perception scores without lipreading, and they were available for experiments at regular times each week.

The structure and function of the wearable real-time speech processor used with the cochlear implant was described in detail by Blamey *et al* (1987). The processor estimated five parameters of the speech signal: the fundamental frequency (F0), the amplitude (A1) and frequency (F1) of the first formant and the amplitude (A2) and frequency (F2) of the second formant. These parameters were selected on the basis of their usefulness to normally-hearing listeners. In the present study, the F0, F1, F2, A1, and A2 parameters were specified by hand for each stimulus. In each F0 period, electrical pulses were applied to two electrode pairs in quick succession. The positions of the electrode pairs in the cochlea were chosen on the basis of the corresponding formant frequencies. The amplitude parameters were coded using a relationship that mapped a 30 dB range of amplitude onto the electrical current range from threshold to maximum comfortable level, separately for each electrode pair. The wearable speech processor set A2 equal to A1 whenever A1 exceeded A2. The reason for this was to avoid the possibility that F2 would be masked by F1. This rule was also applied to the synthetic stimuli described below. Usually, A2 is larger than A1 only for unvoiced sounds so that the independent presentation of A1 and A2 in this case adds a potential cue to voicing.

VOWELS

Four sets of stimuli were used in the experiments, each representing the words "hid, head, had, hud, hod, hood" as spoken by an Australian male audiologist. The first set was recorded using the speech processor and a computer program that collected the estimated values of F0, F1, A1, F2, and A2 for each word. These values were then used to generate electrical stimuli corresponding to the six words with the F0F1F2 coding scheme. The other three stimulus sets were synthetic versions of the same six words. The speech parameter values were specified by hand and used to generate electrical stimuli with the F0F2, F0F1F2, and F0F1F2F3 coding schemes. The stimuli were tested in blocks containing ten of each word in a random order. Before each block, the six stimuli were presented once in order. The patient responded after each trial by pressing one of six buttons labelled with the words. No feedback was given to the patient. The patients received one block of each of the stimulus sets per session in weekly sessions.

The average results were 42% for F0F2, 54% for F0F1F2, 56% for F0F1F2F3, and 77% for the recorded vowels. Clearly, the place coding of formant frequencies in terms of electrode positions gave useful information for the identification of vowels. It is also clear that the recorded stimuli included more information than the synthesized stimuli. The possible extra cues were loudness differences, duration differences, F0 differences, and dynamic aspects of the stimuli such as the rate and extent of formant transitions. In real speech, these cues will be correlated to some extent with the F1 and F2 frequencies. Previous studies (Blamey *et al*, 1987) have documented the probable use of duration to distinguish the "long" vowels in the words "heed, heard, hard, hoard, who'd" from the "short" vowels in the words "hid, head, had, hud, hod, hood", but these studies were not able to separate the effects of the explicit formant frequency coding from the other possible cues within the set of short vowels. Because only a single recorded utterance of each word was used in the present study, the effect of the extra cues may have been exaggerated. If many utterances had been used, the natural variations in loudness, duration and F0 may have diluted the usefulness of these cues.

The addition of F1 information improved vowel recognition compared to the case when only F2 was presented. The addition of F3 did not produce a significant improvement in the recognition of the synthetic words. However, the scores for F0F1F2F3 were greater than those for F0F1F2 in every case so that it is unlikely that the stimulation of the third electrode had a detrimental effect on the recognition of the F1 and F2 information. F3 is known to contribute to the discrimination of some pairs of consonants such as /r, l/ as well as to the naturalness of vowels. For these reasons, it may be advantageous to include F3 in future coding schemes even if it does not produce an immediate improvement in vowel recognition. It should also be noted that the patients had no experience with the F0F1F2F3 coding scheme prior to this study.

CONSONANTS

Each of the experiments described below used a set of stimuli in which the onset characteristics of three steady state vowels were varied in a systematic way. For example, in the first stimulus set, a burst was inserted before each vowel, with the burst frequency chosen from a set of eight values. Thus there were 24 different stimuli (8 bursts X 3 vowels). The three vowels used in every stimulus set were /i, a, ɔ/, with the F1 and F2 values chosen to model the vowels of an Australian male audiologist who had worked extensively with each patient. The fundamental frequency, F0, and hence the electrical pulse rate, was fixed at 120 Hz for each stimulus. The duration of the steady state portion of the vowel was 300 ms. Each set of stimuli was modelled on acoustic stimuli used by researchers from the Haskins Laboratories (Cooper *et al* 1952; Liberman *et al* 1956). The different stimulus sets were presented in randomized blocks containing two of each stimulus. The patient responded by pressing a button labelled with the consonant that was closest to the stimulus heard. The possible responses were limited to a set of two or three, for example /p, t, k/ in the case of the burst stimuli described above. The patient was told that the stimuli would include examples of each consonant followed by each of the three vowels. No prior training with the stimuli was given and no feedback was given during the trial sessions. Each patient received at least five blocks of each stimulus set in weekly test sessions.

Unvoiced stops with varying burst frequency

This set of stimuli was modelled on those used by Cooper *et al* (1952). There were 8 burst frequencies combined with three different two-formant vowels. The burst frequencies used were 290, 530, 770, 1010, 1410, 1870, 2500, and 3500 Hz. These values were chosen to stimulate electrode pairs that were evenly spaced in the cochlea. The burst lasted 40 ms and was separated from the vowel by 20 ms of silence. The F0, A1, and A2 values were the same for all stimuli in this set. No electrode was activated to represent F1 during the burst. Figure 1 shows the response patterns for patient 36 with each vowel. The tic marks on the horizontal axis of each diagram correspond to the eight stimuli with different burst frequencies. The proportions of /p/, /t/, and /k/ responses for each stimulus are shown. The burst frequency clearly affected the consonant chosen by the patient. The vowel that followed the burst also had a strong influence on the perceived consonant. The results are quite similar to those reported for normally-hearing listeners by Cooper who described the responses thus "... high frequency bursts were heard as t for all vowels. Bursts at lower frequencies were heard as k when they were on a level with, or slightly above, the second formant of the vowel; otherwise they were heard as p." The graphs for patient 36, in particular, follow closely the patterns observed by Cooper: the /i/ vowel showed no strong region of /k/ responses; the /a/ vowel had strong /t/, /k/, and /p/ responses at high, mid and low frequencies respectively; and the /ɔ/ vowel showed a double-peaked /p/ response with a strong /k/ response close to the F2 frequency of the vowel. The response patterns for patients 7 and 16 showed similar trends.

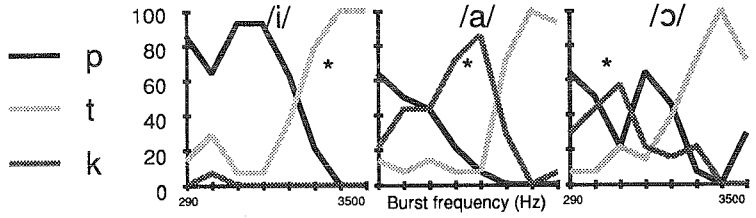

Figure 1. Responses of patient 36 to unvoiced stops. * indicates the vowel F2 frequencies.

Voiced stops with varying F2 locus

These stimuli were modelled on those used by Cooper *et al* (1952). The F2 loci were the same as the burst frequencies used in the previous experiment and were chosen for the same reasons. A1 and A2 were switched on when the formants were half-way between the loci and the vowel formant frequencies. Of the 9 response patterns (3 patients X 3 vowels), only one (the /b/ vowel for patient 16)

was similar to the response patterns observed by Cooper. There was a peak in the /g/ response distribution for flat and slightly falling transitions, ie F2 locus close to F2 for vowel; steeply falling transitions were mainly labelled as /d/; and rising transitions were predominantly /b/. For the /i/ and /a/ vowels, patient 16 was aware of the differing F2 loci but did not interpret the information in the manner described by Cooper. Patient 7 labelled nearly every stimulus as /b/. Patient 36 associated each consonant with a different vowel and ignored the F2 transition to respond with /bi/, /ga/, or /d /. To increase the differences between them, the stimuli were altered by starting A1 and A2 earlier and more gradually. This had two effects. Firstly, more of the F2 transition was audible, and secondly, the onset of the sound was not so abrupt. The responses for every patient showed that the F2 transitions were audible. Patient 16's responses (shown in Figure 2) were closest to those reported by Cooper. For the /i/ vowel, rising transitions were classified predominantly as /b/. Flatter transitions were classified mostly as /d/. Patient 16 classified flat and falling transitions as /g/ and patient 7 showed a similar tendency. Cooper reported a similar response pattern for the /a/ vowel, and this was observed for patients 16 and 7. Patient 36 showed quite a different response pattern for /a/. For the American /u/, /o/, and /ɔ/ vowels, Cooper reported mainly /b/ responses for rising transitions, /g/ for relatively flat transitions, and /d/ for falling transitions. All patients showed a tendency towards this pattern. There was an interaction between F2 locus, amplitude envelope, and formant frequencies of the following vowel, since the responses to the two sets of stimuli depended on all these parameters.

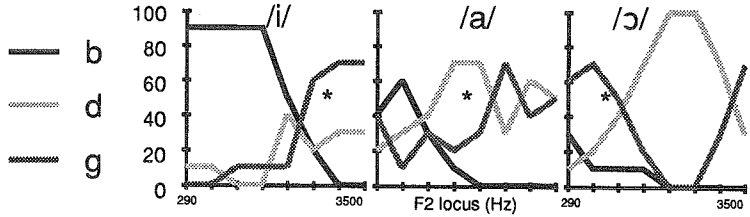

Figure 2. Responses of patient 16 to voiced stops. * indicates the vowel F2 frequencies.

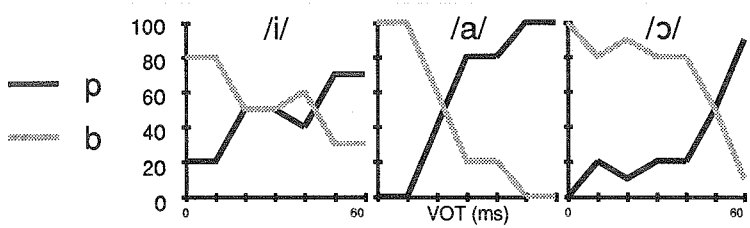

Figure 3. Responses of patient 7 to stops with varying voice onset times.

Bilabial stops with varying voice onset time

These stimuli used the onset time of the A1 parameter to model the voice onset time of natural speech. The F1 and F2 frequencies began at 200 and 500 Hz to represent transitions appropriate to a bilabial stop. The A1 parameter rose abruptly to equal A2 at times 0, 10, 20, 30, 40, 50, or 60 ms after the start of the stimulus. A majority of /b/ responses was expected for the short voice onset times and a majority of /p/ responses at long voice onset times. Only patient 16 showed this pattern, and only for the /i/ vowel. Most of the stimuli were classified as /b/, although patient 36 responded /p/ to nearly all stimuli with the vowel /i/. A second set of stimuli were synthesized with A2 parameters that covaried with the voice onset times. The onset value for A2 was reduced by 3 dB for every 10 ms of voice onset time. A2 increased linearly to 60 dB at the 50 ms point. A1 was equal to A2 at all times after the voice onset time. Thus the intended /b/ stimuli began more abruptly than the intended /p/ stimuli. Patient 7's responses are shown in Figure 3. The patterns for all patients were much closer to the

expected ones. The different responses to these two stimulus sets show interactions between amplitude envelope and voice onset time for implant patients.

Formant transitions with different durations

These stimuli were based on an experiment reported by Liberman *et al* (1956). The F1 and F2 frequencies began at 200 and 500 Hz to represent a bilabial place of articulation. The F1 and F2 frequencies rose linearly to their steady state values over the duration of the transition which was 28, 40, 56, 80, 112, 160, 224, or 320 ms. The transition was followed by a steady state vowel lasting 300 ms. The F0, A1, and A2 parameters were fixed. Liberman showed that normally-hearing listeners classified acoustic stimuli mainly as /b/ for transitions shorter than 50 ms, /w/ between 50 and 150 ms, and /u-/ for transitions longer than 150 ms. The longer transitions sounded like diphthongs beginning with the vowel colour of /u/. Patients 16 and 36 tended to classify the stimuli in the expected way, but the boundaries were much less clearly defined than those reported by Liberman and the durations at the boundaries appeared to be longer. Patient 7 responded with /b/ to nearly every stimulus. A second set of stimuli were synthesized with A1 and A2 starting from 45 dB and increasing steadily to 60 dB over the same duration as the formant transitions. The response patterns of patient 16 are shown in Figure 4. The category boundaries for these stimuli were steeper for all patients, and patient 7 gave some /w/ responses for the /i/ and /a/ vowels and some /u-/ responses for the /ɔ/ vowel. Thus the amplitude envelope had a strong modifying effect.

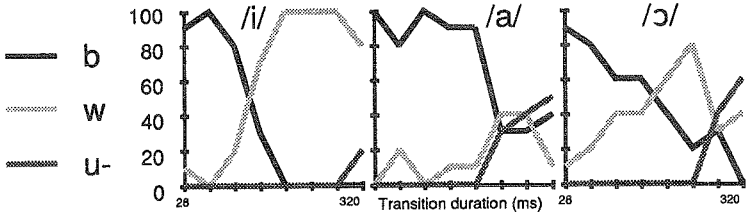

Figure 4. Responses of patient 16 to CVs with varying formant transition duration.

DISCUSSION

Each experiment described above indicated that the patients were capable of distinguishing between the the stimuli and producing response patterns that were similar to those reported for normally-hearing subjects. This occurred in spite of the unnatural nature of the electrical stimuli. Several factors contributed to the unnaturalness: Only two formants were presented rather than the complex spectrum of natural speech. The electrodes stimulated did not correspond exactly to the positions where the first and second formants would produce a maximum of intensity for normally-hearing listeners. The formant frequencies were quantized because of the discrete electrode positions available.

In some respects, the response patterns were easily disturbed. For example, the amplitude envelope parameters had a strong modifying influence on the responses to different formant transitions. There were also quite large differences between the patients, although they had all been chosen as above average performers on speech recognition tasks. If one were to rank the three patients according to how close their responses were to those of normally-hearing listeners, they would appear in different orders for the different stimulus sets. This suggests that the experiments required different perceptual skills that were possessed by the patients in different degrees. Another way of looking at this might be to suppose that real speech sounds contain multiple cues, even when presented via the cochlear implant, and that different patients pay attention to different cues. The effectiveness of the synthetic stimuli would depend on the variation of the cues expected by each patient. Factors that might be linked with the differences between patients include the number and distribution of nerve fibres surviving in the cochlea, the details of the frequency to electrode map, prior experience with hearing aids, lipreading, and the implant itself, and the willingness of the patient to interpret synthetic stimuli as speech sounds.

The results also suggest that there are interactions between the amplitude and frequency cues, and that the relations between them may differ for different patients. In particular, the abruptness of the onset of the stimuli was a stronger influence towards a /b/ response for patient 7 than for the other patients. The existence of these interactions may be important for the fine-tuning of cochlear implant speech processors because they imply that the coding of different parameters cannot be considered completely independently. For example, compression of the amplitude range may upset an interaction between amplitude envelope and formant transition duration. If speech parameters are encoded independently, the covariance of phonetic cues that allows the recognition of phonemes produced in different contexts by different speakers and with different stress may be changed so much that they become unrecognisable. In an extreme case, where a parameter that normally enters an interaction is not coded at all, the patient might be faced with a range of sounds that are no longer phonemically equivalent because the modifying influence of the uncoded parameter is absent. This problem may become a limiting factor affecting parametric speech processors used with cochlear implants if the set of parameters is incomplete. While this may be considered an argument in favour of wholistic processing, it must be noted that relevant information can be lost in the perceptual stages as well as in the speech processor. It remains to be shown that wholistic processors do provide the appropriate cues and that the interactions are similar to those for normally-hearing listeners.

CONCLUSIONS

The coding of formant frequencies and amplitudes by electrical parameters in cochlear implant speech processors can result in response patterns for speech recognition that are similar to those of normally-hearing listeners. Interactions between acoustic cues that are observed with normally-hearing listeners also have counterparts for cochlear implant patients. The interactions impose constraints on the independent coding of speech parameters.

REFERENCES

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

Clark, G.M., Tong, Y.C., Patrick, J.F., Seligman, P.M., Crosby, P.A., Kuzma, J.A. & Money, D.K. (1984) *A multi-channel hearing prosthesis for profound-to-total hearing loss*, J. Med. Eng. Technol. 8, 3-8.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M. & Gerstman, L.J. (1952) *Some experiments on the perception of synthetic speech sounds*, J. Acoust. Soc. Am. 24, 597-606.

Liberman, A.M., Delattre, P.C., Gerstman, L.J. & Cooper, F.S. (1956) *Tempo of frequency change as a cue for distunguishing classes of speech sounds*, J. Exp. Psychol. 52, 127-137.

Millar, J.B., Tong, Y.C. & Clark, G.M. (1984) *Speech processing for cochlear implant prostheses*, J. Sp. Hear. Res. 27, 280-296.

Summerfield, Q. (1985) *Speech-processing alternatives for electrical auditory stimulation*, in Schindler, R.A. & Merzenich, M.M., *Cochlear Implants*, p. 195-222, (Raven Press: New York).

Tong, Y.C., Blamey, P.J., Dowell, R.C. & Clark, G.M. (1983a) *Psychophysical studies evaluating the feasibility of a speech processing strategy for a multiple-channel cochlear implant*, J. Acoust. Soc. Am. 74, 73-80.

Tong, Y.C., Dowell, R.C., Blamey, P.J. & Clark, G.M. (1983b) *Two-component hearing sensations produced by two-electrode stimulation in the cochlea of a deaf patient*, Science 219, 993-994.