Perceptual Performance of Subjects with Cochlear Implants Using the Spectral Maxima Sound Processor (SMSP) and the Mini Speech Processor (MSP)

Colette M. McKay and Hugh J. McDermott

Abstract

A portable sound processor has been developed at the University of Melbourne for use with the 22-electrode cochlear implant manufactured by Cochlear Pty Limited. In this report, the performance of the Spectral Maxima Sound Processor (SMSP) is compared with the Mini Speech Processor (MSP) presently used clinically with this implant. Results of tests with 5 subjects demonstrate that the performance of the SMSP significantly exceeds that of the MSP for recognizing vowels and consonants in closed-set confusion studies and open-set monosyllabic word tests, although recognition of speaker identity and intonation patterns is similar for both processors. Information transmission analysis of phoneme features confirms that most features, such as vowel formant frequencies and consonant manner and place of articulation, are better conveyed with the SMSP. Possible reasons for its improved performance are discussed.

The Nucleus 22-electrode cochlear implant, manufactured by Cochlear Pty Limited, is now used extensively to assist adults and children with a profound to total hearing loss. The prosthesis includes a microphone, speech processor, implanted receiver-stimulator and electrode array. Sounds transduced by the microphone are modified by the speech processor, which extracts and enhances certain acoustic features. Digitally encoded signals are transmitted by the speech processor to the receiver-stimulator via a transcutaneous inductive link. The receiver-stimulator delivers pulses of electric current to selected intracochlear electrodes as specified by the decoded signals. The residual auditory neurons thus stimulated convey to the implantee hearing sensations that are intended to provide information particularly about speech.

Speech processor design for the 22-electrode implant has evolved over many years of research undertaken by the University of Melbourne and Cochlear Pty Limited. Currently the speech processor used clinically with the Nucleus implant is the Mini Speech Processor (MSP), usually programmed with the MULTIPEAK speech coding strategy, described below. In 1989 a novel speech processor, called the Spectral Maxima Sound Processor (SMSP), was developed specifically for use with the University of Melbourne's advanced cochlear implant (McDermott, 1989). The success of this project led to the modification of the SMSP to make it suitable for use with the Nucleus implant, known as the Mini System 22. Over the past few years the design of the SMSP has been refined further and it has been evaluated with a number of implantees (McKay, McDermott, Vandali & Clark, 1991; McDermott, McKay & Vandali, 1992).

It is well known that the variability among implantees with any particular device is extremely high, with expected performance scores in the total population ranging from zero to 100% on some types of speech tests. To successfully compare two or more speech processors, therefore, it is necessary to either directly take account of this variability by using the same subjects for each processor, or if this is not possible, a very large random sample of implantees with each processor would be necessary in order that the sample mean scores would be a reasonable estimate of the population mean scores. In addition, comparing data from independently performed studies poses the additional difficulty of unknown effects of evaluation procedures, analysis methodologies, test materials, etc.

Results of speech perception tests have been reported between five hearing prostheses (Tyler, Moore & Kuk, 1989; Tyler, Tye-Murray, Moore & McCabe, 1989). In one test, two groups of better-performing implantees using the Ineraid and Nucleus implants, respectively, were tested with English words (containing three or four phonemes) and sentences. The Ineraid prosthesis generated monopolar stimulation at four intracochlear locations based on the amplitudes in corresponding band-pass filters (Edington, 1980), whereas the Nucleus system used a predecessor of the MSP speech processor (the WSP In) that activated two electrodes selected according to the estimated frequencies of the first two speech formants (Blamey, Dowell, Clark & Seligman, 1987).
For both processors, the average score for words was 12%, and the average score for words in sentences was 33%, approximately, although the authors pointed out that any quantitative comparison should be cautious for the reasons discussed above.

In a major study reported recently by Waltzman, Cohen and Fisher (1992), 30 implantees received the Nucleus device, 30 the Ineraid, and 22 received a single-channel 3M-Vienna implant (Hochmair-Desoyer, Hochmair & Stiglbrunner, 1985). Initially the Nucleus implantees used the WSP III processor and obtained levels of performance not significantly different from those using the Ineraid. Both multielectrode devices performed significantly better than the single-channel device. Later 27 Nucleus implantees received the MSP and after 3 months obtained significantly higher scores than the Ineraid users in some tests; the mean score obtained for monosyllabic words was 25% and for sentences the mean score was 58%, approximately.

Another research group has developed an experimental speech processor and tested it with a group of Ineraid implantees selected as excellent performers (Wilson, Finley, Lawson, Wolford, Eddington & Rabinowitz, 1991). The continuous interleaved sampling strategy presents the amplitude envelope of signals in five or six band-pass filters as pulsatile stimulation at a relatively high rate (typically over 800 Hz per channel) on corresponding electrodes. This strategy enabled the subjects to obtain significantly higher scores on several speech perception tests than with their usual Ineraid processor. For example, average scores for monosyllables increased from approximately 43% to 63%, whereas average scores for words in sentences increased from 87% to over 95%, approximately.

Skinner, Holden et al (1991) have compared the MSP with the WSP III processor in a group of Nucleus implantees not selected on the basis of performance. Average scores for monosyllabic words improved from 13% to 29% with the MSP, and for words in sentences from 52% to 70%, approximately.

Finally, a recent report (McKay, McDermott, Vandali & Clark, 1992) compared the performance of the SMSP with that of the MSP in a group of implantees also not selected on the basis of performance. Average scores for monosyllabic words improved from 40% to 57% with the SMSP, and for words in sentences from 81% to 94%. Although, as noted above, it would be misleading to compare the results from these various studies directly, it is noteworthy that excellent speech perception can be restored to at least some multichannel cochlear implantees by advanced speech processors including the continuous interleaved sampling approach and SMSP.

This paper reports on an analysis of the performance of a group of implantees using the SMSP and MSP processors, undertaken to identify which aspects of speech perception were contributing to the improvements, in order to investigate the effects of differences between the strategies. The first experiment reported below provided data from five users of the two processors for a detailed analysis of phoneme feature perception. Some of the raw response data obtained in the previous study (McKay, McDermott, Vandali & Clark, 1992) were included in this analysis along with results obtained from a fifth subject. In Experiment II, the processors' performance in conveying information concerning a speaker's identity and speech intonation patterns was assessed.

When comparing the performance of two speech processing schemes in the same subjects, it is important to consider possible confounding of the results by two effects. First, if the duration of the study is extensive, it is possible that improvements could be partly or wholly due to experience rather than to the new processor. The second potential problem, which arises particularly when the new speech processing scheme is very different, is that a period of adjustment and learning is required before the new strategy can be assessed fairly. Consequently, a test protocol was devised for each subject to provide an acceptable compromise between the effects of time and experience. This is described in the Method I section, and a discussion of the possible confounding effects of learning can be found in the Discussion I section.

Functional Description of the MSP

A detailed description of the MSP speech processor and the MULTIPEAK coding strategy may be found in Skinner, Holden et al (1991). Briefly, the strategy involves estimation of the fundamental frequency (f0) and the frequencies (f1, f2) and amplitudes (a1, a2) of the first two formants of speech signals. In addition, amplitudes are estimated in the frequency ranges 2.0 to 2.8 kHz (band 3), 2.8 to 4.0 kHz (band 4), and 4.0 to 6.0 kHz (band 5). The encoding strategy generates four stimulus pulses per estimated voice pitch period (1/f0) when voiced speech is detected. Two electrodes are selected based on f1 and f2, respectively, whereas the remaining two are basal electrodes permanently allocated to bands 3 and 4. When no voicing is detected, four electrodes are stimulated nonperiodically at a rate averaging 200 to 300 Hz. In the original version of MULTIPEAK,
F2 was used to select one electrode, and the remaining three were those allocated to Bands 3, 4 and 5. In a newer version, which was used in the study reported here, the selected electrodes were those allocated to F1, F2, Band 3 or 4, and Band 5. One of Bands 3 or 4 is selected to maximize the instantaneous spatial separation from the F2 electrode. The current levels and widths of the stimulus pulses are related nonlinearly to the amplitudes of the corresponding signals. The levels on each electrode are constrained to lie between empirically determined threshold and maximum comfortable values.

Functional Description of the SMSP

In contrast to the MSP, the SMSP was not designed to extract speech features. Instead, as the term “spectral maxima” suggests, the processor examines the spectrum of incoming signals and selects those components with the largest amplitudes (McDermott, McKay & Vandali, 1992). The spectrum is estimated using a bank of band-pass filters, in which the first eight have center frequencies spaced linearly from 250 Hz to 1.65 kHz, and the remaining eight are spaced logarithmically to 5.4 kHz. The amplitude in each filter output is obtained from a rectifier via a low-pass filter rolling off from 200 Hz. The digitized spectral amplitudes are equalized by a microprocessor so as to match the overall frequency response of the SMSP and microphone approximately to that of normal hearing. The six largest amplitudes are selected periodically, and corresponding stimulus current levels and pulse widths are calculated using an algorithm analogous to that in the MSP. In the implant, 16 electrodes are allocated to the channels of the filterbank in tonotopic order. The duration of each stimulation period, containing up to six pulses when input signal levels are high, is typically 4 ms. The period is extended automatically when the data transmission time to the implant, determined partly by the subject’s pulse width requirements, makes this necessary.

Experiment I: Phoneme Feature Analysis

Subjects

Table 1 summarizes some details of the implantees who took part in Experiment I. All subjects were users of the Mini System 22 cochlear implant, but there was considerable variability in their experience with the implant and with the MSP speech processor. Before implantation, all were profoundly deaf, and obtained no significant speech perception benefit from hearing aid use. There were no special selection criteria; a willingness and ability to try a new speech processor on a take-home basis was sufficient. All subjects had between 16 and 20 electrode pairs activated by their MSP, of which the most-apical 16 were used with their SMSP.

Method I, Test Protocols

As Subjects 1 and 2 were experienced MSP users, their speech perception was evaluated over 2 to 3 weeks before the changeover to the SMSP, and then monitored for 3 months. Each of these subjects showed a rapid improvement in scores in the first 5 weeks of SMSP use, and a gradual improvement in subsequent weeks. Therefore, the data obtained during the first 5 weeks were excluded from the study.

Subject 3 was the only subject with no previous MSP experience, although he had more than 3 years of experience using the WSP III processor. To equalize his experience with the unfamiliar processors, he was given continuous use of the MSP for 12 weeks, followed by 12 weeks with the SMSP. Subsequently he used the MSP for a further 8 weeks, followed by the SMSP for 8 weeks. Data were gathered during these last two periods.

Subject 4 was implanted only 3 months before the beginning of the study, and hence was likely to be in the rapid-learning phase of implant use. It was decided therefore to monitor her performance over time with the MSP before changing to the SMSP. She was monitored for 3 months, in the last 7 weeks of which the data showed no significant change and were used in the comparison. She was then given

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
<th>Duration of Profound Deafness</th>
<th>Etiology</th>
<th>Implant Experience/MSP Experience</th>
<th>Number of Active Electrodes (MSP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
<td>17 years</td>
<td>Accidental injury</td>
<td>22 months/11 months</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1 year</td>
<td>Progressive unknown</td>
<td>45 months/16 months</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>73</td>
<td>30 years</td>
<td>Otosclerosis</td>
<td>38 months/none</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td>2 years</td>
<td>Ototoxic drug</td>
<td>3 months/3 months</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>51</td>
<td>38 years</td>
<td>Meningitis/ototoxic drug</td>
<td>12 months/12 months</td>
<td>20</td>
</tr>
</tbody>
</table>
the SMSP for 14 weeks, with data being gathered over the last 8 weeks.

Subject 5's performance using the MSP was monitored for 5 weeks before she started using the SMSP. Over the following 8 weeks, her performance with the SMSP was monitored, and data were collected during the last 4 weeks. She then agreed to return to the MSP for another 6 weeks, and data were collected in the last 2 weeks of that time.

Test Materials

Experiment I was based on CNC word tests and vowel and consonant confusion tests, all presented at a level of 70 dB SPL using electrical stimulation alone from recordings made by Australian adult speakers.

The CNC word lists comprised sets of monosyllables, each having a Consonant/vowel-Nucleus/Consonant structure (Peterson & Lehiste, 1962). Ten lists, each containing 50 words, were recorded by a male speaker for use in testing Subjects 1, 3, and 4; and CNC material for Subjects 2 and 5 was recorded by a female speaker.

For the vowel confusion tests, 4 different tokens for each of 11 steady vowels in /hVd/ format were recorded by another male speaker. In each test, all 44 tokens were presented in random order to the subject in a closed-set identification task. At the commencement of each test, the subjects were permitted to listen to the complete set of items but were given no feedback during the test about the correctness of their responses.

The consonant confusion test had a similar format, comprising 12 consonants in an /aCa/ context. In the case of Subject 2, the speaker for the consonant confusion test was female, but for the other four subjects the speaker was male.

There were two reasons for testing subjects with materials recorded by different speakers. First, Subject 2 had taken part in speech perception research previously, and it was considered that her familiarity with some of the male speaker's recordings would adversely influence the results of the comparison. The use of the female speaker's recordings avoided this difficulty. The second reason was to verify that the improvements observed with the SMSP were not specific to a particular speaker's vocal characteristics. Accordingly, the CNC word lists recorded by the female speaker were used in testing Subject 5 also. Note also that two different male speakers recorded the CNC and the vowel and consonant tokens, respectively.

Results of Experiment I

CNC word test

From the results of each list of CNC words tested, three sets of confusion matrices (for vowels, initial consonants and final consonants) were constructed. The matrices were analyzed to obtain percentage correct scores for a set of phoneme features. Using a method analogous to that described by Miller and Nicely (1955), the phonemes were grouped so that a small number of values could be assigned to each feature. For vowels, the features examined were duration (with two possible values), the frequencies of the first and second formants (three values each), and the frequency transition directions of the first two formants (three values each). For consonants the features were manner of articulation (five values), place of articulation (seven values), and voicing (two

<table>
<thead>
<tr>
<th>Feature</th>
<th>MSP* Mean Score</th>
<th>Range</th>
<th>SMSP Mean Score</th>
<th>Range</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOWELS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>82.3 (70.7-93.3)</td>
<td></td>
<td>89.0 (78.7-97.0)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>F1</td>
<td>83.8 (70.0-93.0)</td>
<td></td>
<td>87.6 (70.0-96.9)</td>
<td></td>
<td>&lt;0.002</td>
</tr>
<tr>
<td>F2</td>
<td>83.7 (65.3-95.7)</td>
<td></td>
<td>89.3 (74.7-98.0)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>F1-trans</td>
<td>87.9 (76.0-95.0)</td>
<td></td>
<td>92.7 (84.0-98.0)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>F2-trans</td>
<td>89.6 (77.3-96.9)</td>
<td></td>
<td>92.8 (80.0-98.9)</td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>INITIAL CONSONANTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voicing</td>
<td>90.2 (79.3-97.3)</td>
<td></td>
<td>92.8 (84.0-99.9)</td>
<td></td>
<td>&lt;0.02</td>
</tr>
<tr>
<td>Place</td>
<td>65.3 (47.0-76.0)</td>
<td></td>
<td>74.6 (52.3-87.7)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Manner</td>
<td>76.2 (57.3-88.0)</td>
<td></td>
<td>83.2 (63.7-93.0)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>FINAL CONSONANTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voicing</td>
<td>84.0 (72.0-95.0)</td>
<td></td>
<td>91.3 (84.0-96.3)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Place</td>
<td>67.9 (54.3-82.0)</td>
<td></td>
<td>78.4 (63.0-90.6)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Manner</td>
<td>75.0 (57.0-90.3)</td>
<td></td>
<td>82.4 (64.3-95.7)</td>
<td></td>
<td>&lt;0.0005</td>
</tr>
</tbody>
</table>

* The mean and range of the subject averages for each feature are given for each processor along with the significance of the processor factor obtained by two-way analysis of variance.
values). Average scores were calculated for each processor and each subject. Table 2 shows the mean and range of these subject averages for each phoneme feature, along with the significance of the processor effect on a two-way analysis of variance test for repeated measures within subjects.

Before carrying out the statistical analyses on these data, the average results for the male speaker (Subjects 1, 3, and 4) were compared with those for the female speaker (Subjects 2 and 5). As it was evident that the increases in each score with the SMSP had a similar pattern across phonemes for each speaker, any effects of speaker differences were disregarded.

For all five vowel features, the results show a significant increase in scores for the SMSP (duration: \( F(1, 40) = 22.59, p < 0.0005 \); first formant frequency: \( F(1, 40) = 10.83, p < 0.002 \); second formant frequency: \( F(1, 40) = 25.19, p < 0.0005 \); first formant transitions: \( F(1, 40) = 31.08, p < 0.0005 \); and second formant transitions: \( F(1, 40) = 14.34, p < 0.001 \)).

For initial consonants, the results show significantly higher SMSP scores for manner of articulation \( F(1, 40) = 22.59, p < 0.0005 \), place of articulation \( F(1, 40) = 54.44, p < 0.0005 \), and voicing \( F(1, 40) = 5.97, p < 0.02 \). Similarly, for final consonants, the results show significant improvements for manner of articulation \( F(1, 40) = 22.11, p < 0.0005 \), place of articulation \( F(1, 40) = 33.58, p < 0.0005 \), and voicing \( F(1, 40) = 31.24, p < 0.0005 \).

The raw percentage correct scores reported above do not reflect the relative accuracy of perception of different phoneme features, however. For example, the voicing feature, with only two permissible values, will have a score due to chance alone of 50% correct, but this same score would be significantly high for the place of articulation feature, which has seven values. Information transmission analysis (described by Miller & Nicely, 1955), is an alternate means for quantifying phoneme feature perception. It measures the information content of the subjects’ responses as a percentage of the information present in the stimulus set. It takes into account both the distribution of features in the test materials and the number of values that are permissible for each feature.

To calculate the percentage of information transfer for each processor, the confusion matrices for the five subjects were added together. The combined confusion matrices for vowels, initial consonants and final consonants are displayed in Figures 1, 2 and 3 respectively. Each matrix contains responses from 25 word lists, with Subjects 1 to 5 contributing 6, 3, 6, 7, and 3 lists respectively. The resultant information transfer scores for vowels and consonants are shown in Figures 4 and 5. Large gains in information transfer are evident for all phoneme features except the initial consonant voicing feature where the gain is small.

Vowel and Consonant Confusion Tests

The perception of phoneme features in closed set confusion tests using each processor, was analysed using a method similar to that described above for the CNC word test results. The same features were examined, with the exception of the two vowel formant frequency transition features, which do not occur in steady vowels. Fewer values were used for the consonant manner and place of articulation features, reflecting the smaller number of consonants in the test material.

Table 3 shows the mean and range of the average scores for each subject, when their responses were converted to percentages correct for each vowel or consonant feature, along with the significance of the processor effect on a two-way analysis of variance test.

For vowel duration, there was no significant difference between processors \( F(1, 24) = 2.72, p = 0.1 \). This result is not surprising, considering the near perfect scores obtained by all subjects for this feature using either processor, and indicates that the duration cue is very salient in the context of the confusion test. For first formant frequency (F1), there was a significant increase in scores for the SMSP \( F(1, 24) = 9.36, p < 0.005 \), as for the second formant frequency (F2) \( F(1, 24) = 60.63, p < 0.0005 \).

For consonant voicing there was no significant overall difference between processors \( F(1, 27) = 3.97, p = 0.06 \). There was a significant improvement with the SMSP for consonant manner of articulation \( F(1, 27) = 6.90, p < 0.014 \) and place of articulation \( F(1, 27) = 38.33, p < 0.0005 \).

Combined confusion matrices (Figs. 6 and 7) were constructed for each processor. Subjects 1 to 5 contributed \( (3, 2, 6, 2, 3) \) test sessions to the MSP vowel matrix and \( (3, 2, 6, 4, 3) \) test sessions to the SMSP vowel matrix. Similarly, the five subjects contributed \( (3, 2, 6, 3, 3) \) test sessions to the MSP consonant matrix, and \( (3, 4, 6, 4, 3) \) test sessions to the SMSP consonant matrix. The matrices were subjected to information transmission analysis, and the results with each processor for vowel and consonant confusions are displayed in Figures 8 and 9. Increases can be seen for the transfer of information about all features except consonant voicing, with vowel second formant frequency and place and manner of articulation of consonants showing the largest improvements.
subjects therefore it is not expected that learning factors would have any significant effect on the data. For Subject 4 it is possible that some of the improvements recorded were due to experience alone, as the average elapsed time between the two periods of data collection was approximately 14 weeks (7 to 10 months after implantation). However, when a composite phoneme score is calculated for this subject based on data published previously (McKay, McDermott, Vandali & Clark, 1992), the increase over this relatively short time is 13.8%. In contrast, an average increase of 8.6% over the much longer interval of 2 to 9 months after implantation was reported by Tye-Murray et al (1992). Furthermore, the assessments made during the period of monitoring indicate that increasing experience alone did not contribute significantly to this subject’s improvements. Although a subsequent period of MSP use would have been desirable to confirm the magnitude of the changes, this subject refused to return to the MSP for enough time for this purpose, claiming that the SMSP was clearly superior. For Subject 5, the sequence of evaluations suggests that any learning that occurred over the duration of the study would have had reduced the size of the improvements recorded for the SMSP.

The quantitative analysis described above on the results of the CNC word tests and confusion tests details the improvement in both vowel and consonant feature discrimination with the use of the SMSP by these implantees. The data presented in the combined confusion matrices of Figures 1, 2, 3, 6, and 7 confirm the findings of the analyses. They also provide more specific details about where improvements were most evident, and how they might relate to differences in processor design.

The vowel matrices (Figs. 1 and 6) show clear evidence of a reduction in errors of formant frequency perception using the SMSP. For example, Figure 6 shows there is a reduction in /i/ for /u/ confusions in the long duration vowels and a reduction in confusions among the short vowels /æ/, /æ/ and /u/, which are distinguished mostly by the second formant frequencies. As shown in Figure 1, there is a reduction in confusions of the long vowel /æ/ with /u/ (first formant error) and /a/ (second formant error). In the case of the diphthongs, the main improvement in formant perception with the SMSP is seen with /æ/ and /a/. Both these diphthongs have a rising second formant and falling first formant, and the confusions with the MSP suggest that the rising second formant is not perceived as well as with the SMSP, being confused with a steady or falling second formant in some cases.

The observation that the SMSP conveys more information about the vowel formants and their

### Figure 1. Pooled vowel confusion matrices for five subjects from the CNC word tests using the MSP and SMSP processors. A total of 25 word lists were used for each processor (1, 250 vowels).

#### Discussion of Experiment I

Before examining the detailed results, the potentially confounding effects of learning over time require consideration. Improvements in perception over time since implantation have been documented by Tye-Murray, Tyler, Woodworth & Gantz (1992), who studied a group of 13 Nucleus and 14 Ineraid cochlear implantees. They calculated a phoneme composite score by averaging the results of closed-set consonant and vowel recognition tests and phoneme scores derived from open-set monosyllabic word tests. This score increased by 8.6% (averaged across subjects) in the first 9 months of implant use (2 to 9 months), and by a further 4.4% in the second 9 months, with only small nonsignificant changes from 18 months to 3 years.

Subjects 1 and 2 were experienced users of the implant (2 and 4 years, respectively) and of the MSP (11 and 16 months, respectively). Both had been involved in research since receiving the MSP, and their monitored speech perception performance had not changed significantly over at least the last 6 months of that time, in agreement with the findings of Tye-Murray et al (1992). Subject 3 was given equal periods of experience with both processors. For these periods of experience with both processors. For these

---

<table>
<thead>
<tr>
<th>RESPONSES - MSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>79  3  3  16  4</td>
</tr>
<tr>
<td>1  3  45  2  3  7</td>
</tr>
<tr>
<td>9  25  21  3  1  3  8</td>
</tr>
<tr>
<td>1  2  56  1  1  1</td>
</tr>
<tr>
<td>21  1  1  1  1  1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RESPONSES - SMSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>90  1  1  10  1</td>
</tr>
<tr>
<td>1  49  3  5  1</td>
</tr>
<tr>
<td>3  44  1  9  8</td>
</tr>
<tr>
<td>1  3  36  1</td>
</tr>
<tr>
<td>2  1  63  1</td>
</tr>
<tr>
<td>1  4  1  1  10  1</td>
</tr>
</tbody>
</table>

---

---

---

---

---

---

---
transitions may be explicable both in terms of its generally higher stimulation rate (250 Hz per electrode, typically), and in the way it selects spectral components to generate patterns of electrical stimulation.

Figure 10 shows the spectrogram of the word "wish" along with the output of both processors for the same utterance. It can be seen in this figure that, as well as operating at a higher rate, the SMSP represents spectral features, such as vowel formants, by groups of pulses on nearby electrodes that overlap in time. The MSP, in contrast, selects only one electrode per formant in each stimulation period and, consequently, only one or a few pulses are presented on each electrode during a rapid change of formant frequency. This is illustrated by the rapidly rising F2 frequency during the "wish" utterance. The SMSP's behavior may thus improve implantees' chances of identifying formant frequencies during transitions, and perceiving the continuity of moving formants in diphthongs.

The SMSP's tendency to activate several adjacent electrodes to represent F1 and F2 in vowel sounds may also permit implantees to extract more precise information regarding the bandwidth and center frequency of formants than is possible using the MSP, which selects only one electrode to represent each of the first two formants. For example, a shift of the spatial centroid of stimulation on adjacent electrodes with the SMSP may be perceived as a change of formant frequency, whereas this change may be too small to be represented by a change of electrode selection in the MSP.

The results for vowel duration are particularly interesting when the results obtained in the confusion tests are compared with those obtained in the CNC word tests. In the confusion tests, duration errors are infrequent with either processor. In the mono-syllabic words, duration is less well perceived, or less utilized by subjects as a cue to vowel identity. This may be due partly to the greater variability in vowel duration caused by variation in the surrounding consonants. For example, vowels are longer when the final consonant is voiced. It may also be partly due to the introduction of a further class of vowels (the diphthongs) not present in the confusion study.

The duration errors among the steady vowels in CNC words are more frequent but show the same pattern as the errors in the confusion tests, with more errors confusing short vowels for longer ones.
Figure 4. A comparison of the percentage of information transfer derived from the pooled response data from the CNC word tests with the MSP and SMSP processors. The graph shows total vowel information transfer, and transfer of information about the features: duration, first and second formant frequencies, and formant transitions.

than vice versa. There is a moderate improvement with the SMSP on both tests for this duration error, evidenced particularly by the improvement in the /i/ for /I/ confusion.

The major vowel duration improvement for CNC words, however, is seen in the confusions between short vowels and diphthongs. Short vowels are confused with diphthongs twice as often with the MSP (8.7% of short vowels compared with 4.1% with the SMSP). Conversely, there are twice the errors with the MSP confusing diphthongs with short vowels (12.8% of diphthongs compared with 6.1%). A probable reason for both these types of confusions is that the second part or target of the diphthong is being confused with part of the following consonant or pre-articulation for this consonant. In support of this proposition it can be seen that there are very few confusions (2 with the MSP and 4 with the SMSP) between diphthongs and the vowels /I/ and /u/ which are the target segments of all diphthongs in these words. The continuity of diphthongs may be better perceived with the SMSP for the same reasons as discussed above for vowel formant perception.

Consonant voicing was another phoneme feature that showed a different pattern of improvement for CNC words than for the confusion test. Vowel duration information is a vital cue for final consonant voicing perception. It is not surprising, then, that the increase in vowel duration accuracy is associated with a large increase in final consonant voicing information for CNC words in spite of there being only a small voicing information advantage for initial consonants, and none for consonants in the /aCa/ environment (Figs. 5 and 9). The data suggest that subjects using the SMSP are either better at utilising vowel duration cues to determine whether the final consonant is voiced, or better at perceiving the voicing of the final consonant, making the vowel duration clearer.

An example from the final consonant confusion matrices (Fig. 3) is the nasals (/m/, /n/, and /ŋ/), which are often confused with unvoiced plosives in the case of the MSP but only rarely with the SMSP. Unvoiced plosives in the final position are characterized by a sudden cutoff of voicing in the preceding vowel as the articulators close. The MSP is designed to detect the voicing cutoff and switch to the “unvoiced” mode in which aperiodic stimulation at a generally higher rate is generated, and different electrodes are selected. For nasals in the final position, the intensity of the voicing preceding or during the
nasal may occasionally fall below the voicing detector’s threshold, causing the processor to switch prematurely to the “unvoiced” mode. This may make the nasal sound more like a plosive, causing a manner of articulation error as well as a voicing error. In contrast, the SMSP does not change its pattern of stimulation for unvoiced sounds.

The manner of articulation of consonants is conveyed both by spectral shape (e.g., presence of high frequency noise in fricatives, absence of energy in the F2 region for nasality) and temporal cues (e.g., silent periods before plosives or affricates, and other amplitude envelope cues). There are many other examples of improvements in manner of articulation perception with the SMSP, which are evident in the confusion matrices. For example, voiced plosives are confused with nasals or glides in all matrices and the incidence of these errors with the SMSP is reduced to one-third of the incidence with the MSP, as are the errors confusing affricates with fricatives or plosives.

The SMSP may provide a more detailed repre-

Table 3. Analysis of phoneme feature perception for closed-set vowel and consonant confusion tests for five subjects.

<table>
<thead>
<tr>
<th>Feature</th>
<th>MSP*</th>
<th>SMSP</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Score (Range)</td>
<td>Mean Score (Range)</td>
<td></td>
</tr>
<tr>
<td>VOWELS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duration</td>
<td>97.4 (92.8–100)</td>
<td>98.7 (95.0–100)</td>
<td>NS*</td>
</tr>
<tr>
<td>F1</td>
<td>94.4 (89.4–98.8)</td>
<td>97.0 (93.2–100)</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>F2</td>
<td>84.1 (76.1–98.6)</td>
<td>95.2 (86.3–100)</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>CONSONANTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voicing</td>
<td>94.4 (89.6–98.9)</td>
<td>92.8 (85.4–98.9)</td>
<td>NS</td>
</tr>
<tr>
<td>Place</td>
<td>70.8 (58.3–81.9)</td>
<td>81.9 (62.5–93.7)</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Manner</td>
<td>87.7 (75.7–96.8)</td>
<td>91.5 (85.4–99.3)</td>
<td>&lt;0.02</td>
</tr>
</tbody>
</table>

* The mean and range of the subject averages for each feature are given for each processor along with the significance of the processor factor obtained by two-way analysis of variance.
* NS denotes not significant at p = 0.05 level.
Figure 6. Pooled confusion matrices for five subjects from the vowel confusion tests using the MSP and SMSP processors. The total number of test sessions was 16 for the MSP and 18 for the SMSP (total number of stimuli/row equal to 64 and 72 respectively).

Figure 7. Pooled confusion matrices for five subjects from the consonant confusion tests using the MSP and SMSP processors. The total number of test sessions was 17 for the MSP and 20 for the SMSP (total number of stimuli/row equal to 68 and 80 respectively).

The large improvement in perception of place of articulation is very pleasing, as this feature is particularly difficult for cochlear implantees to perceive. Previous studies using a variety of test materials and analytic techniques have shown that place information is the worst, or close to the worst, transmitted consonant feature (Blamey, Dowell, Brown et al, 1987; Dorman, Soli, Dankowski et al, 1990; Hochmair-Desoyer, Hochmair & Stiglbbrunner, 1985; Preece & Tyler, 1992; von Wallenberg & Battmer, 1991). Place information is conveyed mostly by the spectral shape and intensity of the consonant sound, and by the formant frequency transitions in the vowels surrounding the consonant. The consonant confusion matrices (Figs. 2, 3, and 7) show a marked improvement with the SMSP for place of articulation accuracy, particularly with the plosives /t/and /k/, to a lesser extent their voiced versions /d/and /g/, and the nasals in the final position. These improvements with the
Figure 8. A comparison of the percentage of information transfer derived from the pooled response data from the vowel confusion tests with the MSP and SMSP processors. The graph shows total vowel information transfer, and transfer of information about the features: duration, and first and second formant frequencies.

SMSP may arise both from better definition of vowel formant transitions, and better information about high frequency spectral shape, as discussed above.

**Experiment II: Identification of Speaker and Intonation Pattern**

One of the major differences between the two processing strategies described in this report is in the use of rate of stimulation to encode features of speech. As described earlier, the MSP directly encodes the estimated fundamental frequency (F0) as the stimulation rate for voiced sounds, and in addition encodes the absence of voicing by changing to a typically higher-rate, aperiodic stimulation and an alternate electrode selection algorithm. In contrast, the SMSP strategy does not involve a voicing decision or a measurement of F0, the rate of stimulation being constant. There is, nevertheless, representation of F0 in the amplitude modulation of the stimuli in some channels. Figure 11 illustrates this modulation in the SMSP output during the vowel segment of the word “hard” as spoken by a male speaker.

The fundamental frequency of speech normally provides a voice pitch percept that can help to convey suprasegmental aspects of speech such as emphasis within an utterance or whether it is a question or statement. These intonation changes may also convey information about the speaker such as their emotional state. There are, however, a multitude of other cues available for speaker identification and intonation change, which may make these tasks possible if F0 information is removed, and the importance for implantees of F0 information is unknown.

It is therefore of interest to investigate whether different representations of F0 (rate of stimulation in the MSP and stimulus amplitude modulation in the SMSP) affect the performance of subjects on speaker identification and intonation pattern recognition tasks, and whether any such differences can be related to the differences between the speech processing strategies. These issues were investigated in Experiment II.

**Method II**

**Speaker Identification**

Four Australian speakers (two male and two female) were chosen on the basis of their fundamental frequency, each being up to approximately one
Figure 9. A comparison of the percentage of information transfer derived from the pooled response data from the consonant confusion tests with the MSP and SMSP processors. The graph shows total consonant information transfer, and transfer of information about the features: manner and place of articulation, and voicing.

standard deviation above or below the mean fundamental frequency of their sex. The average fundamental frequencies for Speakers 1 to 4 were 91, 126, 189 and 201 Hz respectively. The last two speakers exhibited a large natural variation in their F0's, causing their ranges to overlap significantly. Each speaker had recorded a 12-sentence passage from a scouting manual (Millar, O’Kane & Bryant, 1989). These recordings were used to construct four new recordings of the passage with each passage containing three sentences from each speaker, distributed in a random fashion. In this recording, each speaker retained his or her natural speaking loudness variation (up to 15 dB within each sentence), but the range of voice level and most common word level were roughly equal across speakers. No attempt was made to equalize levels on a word-by-word or sentence-by-sentence basis. There were, however, differences in the individual patterns of loudness variation; for example Speaker 3 had noticeably more frequent variations, due to the way she emphasized words. In addition to the test recordings, a training segment was constructed using one sentence repeated by each speaker in turn.

Three implantees took part in this study (Subjects 1, 3, and 4). In each test session, the subject first listened to the training segment until they were satisfied that they could remember what each speaker sounded like. Throughout all segments of the test, the subject had a typed copy of the speech material, so that any difficulty understanding the sentences would not interfere with their task of identifying the speaker. In the test segment of the session, the subject listened to one of the four prepared passages, and after each of the twelve sentences was required to respond with one of the four speaker numbers. No feedback was given regarding the correctness of the responses. In each test session a different recording was used to prevent any learning effect due to familiarity with the speaker order.

intonation Pattern Identification

Two Australian speakers, one male and one female, recorded the material for this test. Tokens of two isolated voiced phonemes (/a/ and /m/) were recorded, spoken naturally with rising, falling, or steady voice pitch. Four different tokens for each speaker, phoneme, and intonation pattern were selected to make four lists of twelve tokens. Each set of twelve tokens contained four of each intonation pattern (rising, steady or falling pitch) in a random order, all from the same speaker using the same phoneme. Within each list, the average loudness (70 dB SPL) and length of utterance (approximately 1 sec) were approximately equal for each token, with
Figure 10. The spectrogram of the word “wish” (a), and the resultant outputs of the MSP processor (b), and the SMSP processor (c). The vertical axis of the spectrogram represents frequency, plotted on a linear scale. The vertical axes in (b) and (c) represent electrode position, with the more-basal (higher pitch) electrodes at the top. In the processor output graphs, the vertical bars each represent one biphasic pulse of electrical stimulation with height proportional to the stimulation level. The horizontal axes represent time (75 ms/div).

any variation in these quantities being similar for each intonation pattern. No attempt was made to remove naturally occurring coincident cues to intonation changes. For example, the loudness change
within each token was highly correlated to the pitch change (loudness increasing with pitch). The variation in F0 for rising or falling pitch was approximately 30% for the male speaker, and close to one octave for the female speaker. There was an associated change of 3 to 5 dB in the loudness levels. The male speaker's F0 range encompassed 90 to 160 Hz overall, whereas the female speaker's range was 180 to 425 Hz, approximately. Three different randomizations of each list were made to prevent learning effects due to familiarization with the token order. A training segment was constructed for each speaker and phoneme, containing six identified tokens (two of each intonation pattern).

Subjects 1, 2, 3 and 4 took part in this experiment. In each session, the subject listened to four lists (each with a different speaker/phoneme), and was asked to respond after each token by saying if the pitch was rising, falling, or steady. Before each list, they were able to listen to the training segment for as many times as they wished. No feedback was given during the test regarding the correctness of their responses.

Results of Experiment II

Speaker Identification

Speaker confusion matrices were constructed for each subject. These matrices were very similar for each subject, showing the same pattern and little variation in overall accuracy. They were therefore combined to give the matrices for each processor shown in Figure 12. Each element of these matrices represents the average across the three subjects of the percentage of responses obtained for each stimulus. Each row represents the responses from a total of 30-sentence presentations by each speaker.

It is evident from Figure 12 that there is no significant difference between the processors in ability to convey speaker identity, at least in the context of the chosen test. Two features common to both matrices are noteworthy. First, there are virtually no confusions with either processor between male and female speakers (Speakers 1 or 2 with Speakers 3 or 4). Subjects could also easily differentiate the two male speakers (96.7% accuracy with the MSP and 94.5% with the SMSP). In contrast, there was a great deal of difficulty distinguishing the two female speakers (54.15% accuracy with the MSP and 54.2% with the SMSP). The scores for each processor are within one standard deviation of variability predicted by the binomial theorem, and thus are not significantly different from each other in either of the above cases. In addition the scores for the female speakers are well within one standard deviation of the theoretical chance score of 50%.

Intonation Pattern Identification

Figure 13 shows the combined-subject confusion matrices for the intonation pattern test, the elements of which are the percentage of responses averaged across the four subjects. As with the speaker identification test, all subjects showed very similar patterns of response. The results for the female speaker are presented separately from the results for the male speaker, and each row contains the percentages of responses from a total of 112 presentations of each intonation pattern for each speaker sex.

Each intonation pattern was identified with an accuracy well above the chance level with both processors. Subjects obtained higher scores with the female speaker in every case, this result being consistent with the greater pitch change for the female speaker. The overall accuracy (combining the two speakers) was not significantly different for the two processors (74.2% correct for the MSP and 73.8% correct for the SMSP). However, when the scores for the female or male speaker are examined separately, it is evident that the subjects performed more poorly.
Figure 12. Confusion matrices for the speaker identification test. The stimuli are sentences spoken by four different speakers (numbered 1 to 4). The matrix elements are averages over the three participants (Subjects 1, 3 and 4) of the percentages of responses obtained for each speaker.

with the SMSP for the male speaker (59.9% correct compared with 71.9% with the MSP), and better with the SMSP for the female speaker (87.7% correct compared with 76.6% correct with the MSP). The differences between the processors for each speaker were significantly more than the variability predicted from binomial theory (p < 0.05). The advantage of the female speaker over the male speaker was thus much greater with the SMSP (an increase of 27.8% for the female speaker compared to an increase of 4.7% with the MSP).

Discussion of Experiment II

The two tests administered in Experiment II investigate aspects of speech perception that are often taken for granted in normal-hearing listeners. The speakers and intonation patterns used in the tests can be easily identified by normal-hearing subjects. The results show that these tasks can be more difficult for implantees, and that improvements to speech processing schemes that aim to facilitate these aspects of speech perception would be beneficial.

Two aspects of the speaker identification results are worthy of discussion. First, there was no difficulty with either processor in distinguishing between male speakers or distinguishing between the male and female speakers. If subjects were using mainly voice pitch to tell the speakers apart, then it would follow that the amplitude modulations at F0 in the SMSP provide a voice pitch percept analogous to that provided by the F0 rate in the MSP. Qualitative comments from users of both processors support the presence of such a voice pitch, and its use in identifying the speakers in the test and also in everyday situations.

Secondly, neither processor performed consistently in enabling differentiation between the female speakers. This result is not surprising if subjects were relying heavily on F0 information, because the F0 ranges of the two females overlapped considerably. Apart from this, there are other reasons why each processor may not enable accurate differentiation between female speakers. In a system with pulse rates as low as that of the SMSP, a woman’s fundamental frequency is too high to be accurately represented by amplitude modulation. Instead, the actual rate of stimulation may provide the dominant temporal pitch cue, providing a higher pitch percept than the amplitude modulated stimulation from male voices, but not providing differentiation between the females. In the case of the MSP, errors in the estimation of F0 may occur more often at higher F0 frequencies because of technical limitations in the processor’s circuitry.

Another possible explanation for the similarity of the results with the two processors is that subjects were using only features other than voice pitch to distinguish speakers. Such additional information, which is coded similarly in both processors, would include formant patterns, and idiosyncratic temporal and loudness patterns. However, it is not likely that subjects were relying solely on these other features, first because the two female speakers in this test were not accurately identified in spite of having very different loudness variation patterns, and secondly because it is contrary to the subjects’ comments about how they performed the test.

The results of the intonation pattern test show that the different representation of F0 in the SMSP has no overall disadvantage for detection of intonation change. As with the speaker identification task, the subjects may have used another cue, such as loudness change, which is similarly coded in both processors, to identify the intonation patterns. It is unlikely that loudness was the only cue used, however, as this would be inconsistent with the differences seen between the two processors when the speaker sex is separated. The more likely case is that, as well as the (presumably equal) loudness cue, there was an associated pitch change cue for each processor.

In the case of the MSP, it is unlikely that subjects did not perceive some pitch change as a result of the
rate changing, as rate pitch is a well-documented phenomenon (Tong & Clark, 1985). In the case of the SMSP, a similar pitch change would be associated with the change in frequency of the stimulus amplitude modulation, at least for male speakers. The results for the male speaker would then imply that the rate change in the MSP is more effective at conveying voice pitch change than the change in frequency of amplitude modulations in the SMSP when the fundamental frequency is in the male range.

For the female speaker, the high FO is no longer represented in amplitude modulations in the output of the SMSP. However comparison of the results suggests that the SMSP provides better cues to FO change than the MSP when the fundamental frequency is high. An alternative source of voice pitch information with the SMSP could be the small change in formant frequencies due to the shift in underlying harmonic structure as the fundamental frequency changes. This shift is evident especially in the area of the first formant. In the MSP, only one electrode is selected for F1, and the changes in F1 frequency with FO variation may often be too small to alter the electrode selection, and hence will have no effect on the spatial stimulation pattern. In contrast, the SMSP stimulates several adjacent electrodes in the F1 region, and the relative levels of stimulation on these electrodes change with small F1 variations, causing a corresponding change in the spatial centroid of the stimulation. This effect can be seen in Figure 11, where F1 during the word “hard” is represented by stimulation over three electrodes. The slight shift of F1 frequency during the vowel, is evident as a shift in the centroid of stimulation of these three electrodes, whereas the MSP would have selected one constant electrode to represent F1 throughout the vowel. It is possible that subjects were perceiving these formant shifts with the SMSP for the female speaker, whose FO changes were greater than those of the male speaker, and this extra information enabled them to perceive the intonation change better than with the rate change in the MSP.

Conclusions

The results obtained in Experiment I demonstrate that the processing strategy implemented in the Spectral Maxima Sound Processor has provided significantly better performance in understanding speech to a group
of cochlear implanters. The performance of these implanters using the SMSP has been compared directly with that provided by use of the MSP with the MULTIPEAK strategy. Analysis of phoneme feature detection has shown that the improvements in perception are greatest for vowel formants and consonant manner and place of articulation. Possible reasons for these improvements can be found in the higher rate of stimulation with the SMSP, the use of several nearby electrodes to represent broad spectral peaks including vowel formants, and the continuous mapping of frequency to place of stimulation. The improvements achieved with the SMSP are of sufficient magnitude to be of considerable value to these implanters in everyday circumstances.

Experiment II was aimed at determining whether the use of a constant rate, rather than F0 rate, produced any differences in the ability to identify speakers and intonation patterns. There were not any overall differences in level of performance that could be attributable to the different treatment of the fundamental frequency. In the case of identification of intonation patterns, however, there were differences evident when the speaker sex was separated, with the SMSP performing worse for intonation changes of a male voice, and better for the larger intonation changes in the female voice. In the case of the SMSP, F0 information may be provided by amplitude modulations at the F0 frequency for male speakers, and intonation changes may be detected by either changes in the modulation frequency or, for female voices, by small changes in the vowel formant position caused by shifts in the underlying harmonic structure. Subjects have commented that, subjectively, the SMSP is no less effective than the MSP in enabling them to distinguish or recognize speakers in their everyday lives.

Most implanters who have taken the SMSP away from the laboratory for use at home have expressed a strong preference for the processor over the MSP, in spite of its larger size and weight. Most of these subjects have also reported that music sounds more realistic and enjoyable than it sounded with the MSP. Mental sounds has improved noticeably.

References
Wilson BS, Finley CC, Lawson DT, Woldford RD, Eddington DK

Acknowledgments: Financial support for this project has been provided by the National Health and Medical Research Council of Australia Program Grant “Studies to develop sensory prostheses for deaf children and adults”, the U.S. National Institutes of Health Grant “Improved cochlear implants: psychophysics and engineering,” the U.S. National Institutes of Health Contract “Speech processors for auditory prostheses”, the Australian Research Council’s special research center “The Human Communication Research Centre”, and the Australian Bionic Ear and Hearing Research Institute. We also appreciate the substantial assistance provided by our colleagues, especially Mr. Andrew Vandali, Prof. Graeme Clark, Ms. Lesley Whitford, Dr. Peter Blamey, Mr. Mark Harrison, and all the persons with cochlear implants who have taken part in the studies.

Received June 8, 1992; accepted May 20, 1993.
Address reprint requests to Dr. Colette McKay, Australian Bionic Ear & Hearing Research Institute, 384-388 Albert Street, East Melbourne, VIC 3002, Australia.
Author/s:
McKay, Colette M.; McDermott, H. J.

Title:
Perceptual performance of subjects with cochlear implants using the spectral maxima sound processor (SMSP) and the mini speech processor (MSP)

Date:
1993

Citation:

Persistent Link:
http://hdl.handle.net/11343/27336