Exploring the parameter space of Cochlear Implant Processors for consonant and vowel recognition rates using normal hearing listeners

D. Sen, W. Li, D. Chung & P. Lam

School of Electrical Engineering & Telecommunications
University of New South Wales
d.sen@unsw.edu.au

Abstract

This paper explores two specific parameters in cochlear implant processors in terms of their effect on consonant and vowel recognition rates. The two parameters in our study are the compression factor and the number of electrodes in an *M* of *N* SPEAK processor. In particular, the aim is to evaluate the values of these parameters required to match the performance of acoustic re-synthesis models to that of cochlear implant recipients.

1.Introduction

While there have been other studies to evaluate speech intelligibility as a function of various parameters in cochlear implant processors, those studies have mostly focused on the isolated effect of each parameter or strategy rather than a joint study of the combined effects. The purpose of this work is not only to perform a joint study of the compression factor and the *M* of *N* parameters in the Spectral Peak (SPEAK) strategy, but also, in the process, to validate a re-synthesis model that strives to emulate the function of the cochlear implant for normal hearing listeners. Fu and Shannon (Fu & Shannon, 1998; Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995) measured recognition rate as a function of compression factor in both implant patients and normal hearing listeners. Loizou et. al. (Loizou, Graham, Dickins, Dorman & Poroy, 1997) compared Continuous Interleaved Sampling (CIS) and SPEAK strategies in terms of speech recognition rates amongst implant patients.

Cochlear implant recipients are often difficult to access in engineering affiliated institutions and even when they are available, testing on them is often not possible due to ethical and medical requirements. Further, the fact that the recipients often have diverse auditory pathology also makes subjective responses somewhat dubious. It is thus much more desirable to use normal listening subjects and provide them with an auditory stimulus that would be indicative of what the recipients would hear given the same auditory stimuli.

Cochlear implants convert acoustic stimuli into a set of electrical signals that stimulates the neurons in cochlea. In order to emulate this "electric hearing" in normal hearing listeners, a re-synthesis model tries to re-synthesise an acoustic waveform from these electrical outputs of the cochlear implant. Ideally, the re-synthesised acoustic stimuli will produce the same perception in a normal hearing listener as would be heard by a normal hearing listener who was implanted with a cochlear implant and subjected to the original stimulus waveform. While these assumptions are unrealistic, it is expected that a good model will provide accurate indications of average recipient perception when a large number of recipients and normal hearing listeners are used for the comparison.

The re-synthesised waveform can subsequently be used to carry out subjective tests on normal hearing listeners and the responses would be indicative of recipient responses. To validate the re-synthesis model, the responses are compared to subjective responses of actual cochlear implant recipients. In our work, the tests measure the recognition rates of consonants and vowels of CVC words.

2. The re-synthesis model.

Figure 1 shows the block diagram of our model that converts an input acoustic digital signal $s[n]$ into 22 electrical signals $e_i[n]$ and subsequently back to an acoustic signal $y[n]$.

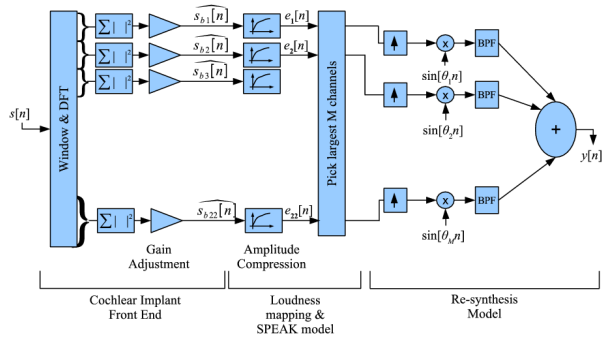

Figure 1. The re-synthesis model.

2.1.Filterbank and envelope computation

The first stage in the model represents a typical front-end of a cochlear implant processor which tries to

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

emulate the cochlear peripheral processing. The implementation uses a uniform DFT filter-bank followed by appropriate frequency binning and energy computation. The frequency binning emulates the frequency resolution of a normal cochlea. The windowed DFT, binning and energy computation effectively results in the computation of 22 outputs, representing the envelope of the output of a non-uniform filter-bank. The input signal is sampled at 16 kHz, the window size is 128 points (8 ms) with 96 point overlap representing a 2 ms time integration in the calculation of the envelopes. Each channel can thus be viewed as the full-wave-rectified and low-pass-filtered (LPF) output of a band-pass filter, where the cut-off frequency of the LPF is approximately 500 Hz.

2.2.Amplitude compression and SPEAK simulation.

The second stage shows an amplitude compression block followed by a block that picks the *M* largest outputs out of the *N* channels. In this second block, we systematically varied the parameter *M*, to investigate the "*M* of *N*" SPEAK strategy where only the *M* most salient electrodes are chosen to excite the nerves in the cochlea. No attempt is made to time-interleave the electrodes as is typically done in cochlear implants. The ACE strategy employs a higher rate of stimulation to the implant electrodes than SPEAK, but otherwise the two strategies are quite similar. Since both the recipient data and our model used Nucleus devices (from Cochlear Ltd.) we had 22 electrodes from which to choose (i.e *N*=22).

The amplitude compression block can be viewed as an attempt to ensure correct loudness perception by the recipients. This is a difficult prospect given that the acoustic dynamic range of normal hearing is about 110 dB while the current dynamic range between the smallest level (T-level) and maximum comfort level (C-level) in the implant electrode is only about 8 dB. To alleviate the problem somewhat, implants are designed to map only about 30 dB of the acoustic dynamic range where speech is mostly found (40-70 dB SPL say). This can be achieved if a mapping of the form:

$$e[n]=\left(s_b[n]\right)^k \quad (1)$$

is made between the electrode $e[n]$ and the signal at the output of the bandpass filters, $s_b[n]$. The exponent k, is thus approximately given by 8/30 = 0.27.

An alternative perspective on computing k can be found by considering loudness growth functions. In normal hearing, the loudness L as a function of acoustic pressure P is given by Stevens' law (Stevens, 1955; Shannon et al., 1995):

$$L \propto P^{k_1}, \quad (2)$$

where, k_1 is approximately 0.6. A 30 dB dynamic range is thus mapped to a 30x0.6 = 18 dB loudness range. Fu and Shannon (Fu & Shannon, 1998) found a similar loudness "growth" relationship for implant recipients as a function of electrode current level, *I*:

$$L \propto I^{k_2}, \quad (3)$$

where k_2 was found to be approximately 2.72. The 18 dB loudness range thus maps to 18/2.72=6.62 dB current range. The relationship between pressure to electrode current is thus given by:

$$I \propto P^{k_3}, \quad (4)$$

with k_3=6.62/30=0.22. This value is fairly close to the 0.27 found earlier for the exponent in Equation 1. The second interpretation is depicted in Fig. 2.

While these analyses provide the optimal mapping between the acoustic signal and the electrode signal in cochlear implants, the question that is pertinent to the re-synthesis model is whether the electrode signal will need to be expanded further by a factor of 0.6 to account for the fact that the normal cochlea (of the listener) will compress the re-synthesised signal by a factor of 0.6. This would mean that the optimal mapping to be used in the electrodes for the re-synthesis model will need an exponent of 0.22/0.6=1/2.72=0.37 and not the 0.22-0.27 predicted earlier. To investigate this further, we systematically varied the value of *k,* to find the value that will best match recipient responses.

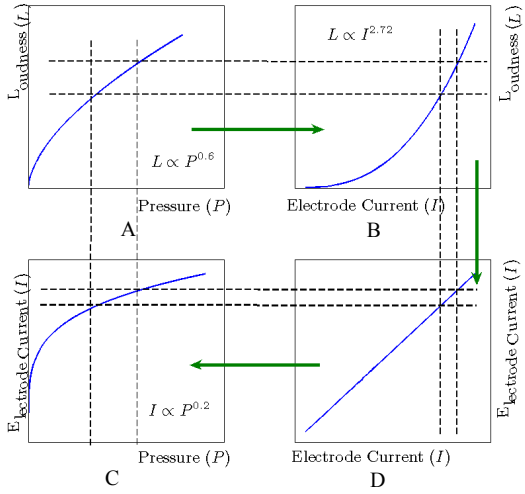

Figure 2. Loudness growth functions. **A** shows loudness as a function of pressure as given by Stevens' Law. **B** shows the loudness as a function of electrode current as found by (Fu, 1998) and **C** shows how the relationship between electrode current and pressure can be derived from **A** and **B**.

The results in (Fu & Shannon, 1998) indicates that recipient and normal hearing listener recognition rates were almost identical when they both have a compressive mapping of 0.2. This is in contradiction to the 0.37 that we calculated for the synthesis model. Fu's experiments were, however, restricted to four-channel processing (both implants and re-synthesis model), made no attempt to emulate a SPEAK/CIS/ACE type strategy in the re-synthesis model and used noise bands instead of sinusoids in the re-synthesis procedure.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

2.3.Signal re-synthesis

The final stage of our re-synthesis model involves modulating the M sinusoids at different frequencies before bandpass filtering and summation to create the re-synthesised signal, $y[n]$. The model is quite similar to that used by Loizou (Loizou, Dorman & Fitzke, 2000; Loizou, Dorman & Poroy, 2000; Spahr, Dorman & Loizou, 2002; Loizou, Dorman & Tu, 1999) and is only different from Shannon's re-synthesis model (Fu & Shannon, 1998; Shannon et al., 1995) in that Shannon uses filtered noise instead of sinusoids in the final stages of the model. Dorman, however, found no significant difference between the use of sinusoids or filtered noise for the modulation signal (Dorman, Loizou & Rainey, 1997).

3. Methods & stimuli

3.1.Subjects

Six subjects between the ages of 20 and 29 participated in the test. All of the subjects were screened for normal hearing (within 10 dB HL) by carrying out informal tests at four frequencies between 0 and 4 kHz. Four of the subjects were native speakers of Australian English, one of Canadian English and one had English as their second language.

3.2.Equipment

All tests including the screening tests were carried out using DT770-Pro closed headphones, with stimuli played mono-aurally. The subjects were seated in a sound-insulated anechoic chamber (Acoustic Systems) with computer systems and audio playback equipment situated outside the chamber. TDT (system-II) hardware and software was used for the screening test. An external Multiface (RME) card was used for playback along with software written in-house for the actual CVC testing.

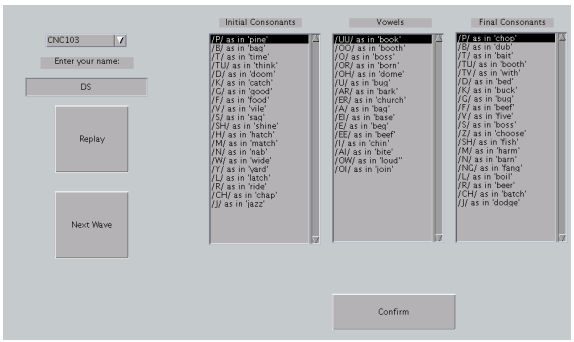

Figure 3: Subjective testing interface

3.3.Speech material

The test stimuli consisted of 928 male CVC words divided into 30 balanced word-lists. There were 20 initial consonants, 16 middle-vowels, and 20 final consonants in the material. The consonants and vowels are shown in Table 1. Thus, while there was not a large difference in the number of vowels and consonants to choose from, the perceptual space between units may have been smaller for the vowels than the consonants. The entire set of test material were from a set used by the University of Melbourne for testing on cochlear implant recipients.

3.4.Procedure

The tests were conducted over a six month period on multiple subjects and re-synthesis models. Each subject was screened with an informal listening test to ensure normal hearing thresholds and put through a training process to identify consonants and vowels in a non-noise environment. For every permutation of k and M, subjects were asked to listen to the single CVC words in individual word-lists and identify the initial consonant, middle-vowel and final-consonant. Each subject used at least three word-lists for each permutation. The subjects spent up to an hour training on the clean (undistorted) sounds before attempting the test on the synthesised material. The presentation level was 60 dBA. The listeners were allowed to re-play the words as many times as they wished before they performed an N-Alternative Forced Choice task where they selected a single entry from each of three lists corresponding to the consonants or vowels. Each list had entries such as "/P/ as in pine", "/UU/ as in book" and "/D/ as in bed" for the initial consonants, vowels and final consonants respectively as shown in the graphical interface in Fig 3. The word-lists were counter-balanced between subjects.

Initial Consonant	Vowel	Final Consonant
'**P**ine'	'b**OO**k'	'cho**P**'
'**B**ag'	'b**OO**th'	'du**B**'
'**T**ime'	'b**O**ss'	'bai**T**'
'**T**hink'	'b**O**rn'	'boo**TH**'
'**D**oom'	'd**O**me'	'wi**TH**'
'**C**atch'	'b**U**g'	'be**D**'
'**G**ood'	'b**A**rk'	'bu**CK**'
'**F**ood'	'ch**UR**ch'	'bu**G**'
'**V**ile'	'b**A**g'	'bee**F**'
'**S**ag'	'b**A**se'	'fi**VE**'
'**B**hine'	'b**E**g'	'bo**SS**'
'**B**atch'	'b**EE**f'	'choo**SE**'
'**M**atch'	'ch**I**n'	'fi**SH**'
'**N**ab'	'b**I**te'	'ha**RM**'
'**W**ide'	'l**OU**d'	'ba**RN**'
'**Y**ard'	'j**OI**n'	'fa**NG**'
'**L**atch'		'boi**L**'
'**R**ide'		'bee**R**'
'**C**hap'		'ba**TCH**'
'**J**azz'		'do**DGE**'

Table 1: Initial consonant, middle vowel and final consonant representatives in the test stimuli.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

4.Results

Figures 4-6 show the results when no compression is used (k=1) and all channels are used for synthesis (M=22). Darker tones in the figures indicate higher rates of recognition. The results are plotted as confusion matrices with the matrix on the left representing actual recipient responses. The recipient data were from the University of Melbourne (on the same speech material as this test) and comprised of responses from six patients implanted with Cochlear Ltd prostheses. Each of the patients were post-lingually implanted with 1-6 years of implant experience and ranged between ages of 48-75 years. The presentation level was 60dBA. Five of the recipients used the SPEAK strategy while the last used the ACE strategy at 260 Hz/channel. The difference between ACE and SPEAK is mainly in the high rate used in ACE and since the 260 Hz/channel is comparable to what is used in SPEAK, we deemed the last recipient as using the equivalent of the SPEAK strategy.

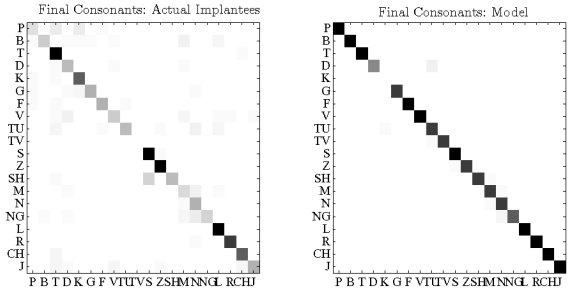

Figure 4: Confusion matrix of responses for final consonants (k=1 and M=22). Recipient accuracy 64.68 % and model accuracy 90.51%.

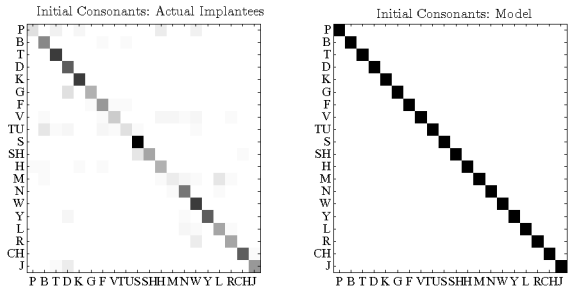

Figure 5: Confusion matrix of responses for initial consonants (k=1 and M=22). Recipient accuracy 69.30 % and model accuracy 99.67%.

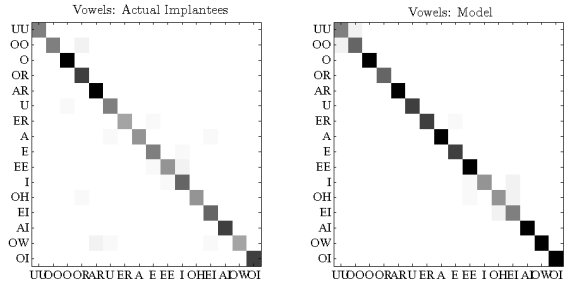

Figure 6: Confusion matrix of responses for vowels (k=1 and M=22). Recipient accuracy 81.31 % and model accuracy 90.00%.

Figures 7-8 show the results at the other extreme of performance when k=0.2 and M=5.

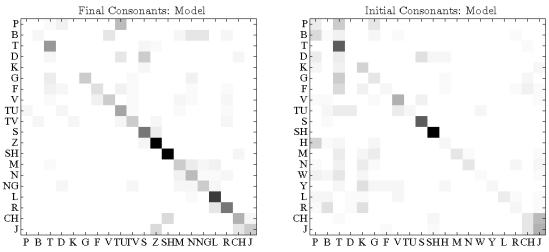

Figure 7: Confusion matrix of responses for final and initial consonants when k=0.2 and M=5. Model accuracy is 47.13% and 33.47% for final and initial consonants respectively.

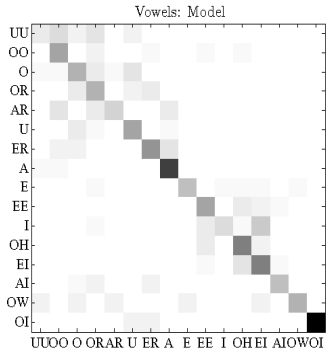

Figure 8: Confusion matrix of responses for vowels when k=0.2 and M=5. Model accuracy is 48.44%.

Figure 9 is a three-dimensional plot showing the overall results for initial consonants, final consonants and vowels as a function of the compressional exponent k and the number of electrodes in the M of N paradigm.

5. Discussion & dissemination

As the values of [k,M] are varied from [1.0, 22] to [0.2, 5] all three recognition rates of initial, final consonants and vowels spiral down from close to 100% to below 50%. The spiral shape is interesting as it suggests that the compression factor k and M can be traded off against each other to an extent. The greatest range of accuracy can be observed for the initial consonants.

Figure 10 shows the recognition rates as a function of the compression factor k. Recipient data is shown using dashed horizontal lines. Even with the perturbation due to the change in M, it can be observed that the best match between acoustic hearing and electric hearing for vowels is achieved with a compression factor of approximately 0.37 – exactly as predicted analytically in Section 2. The match for initial and final consonants seem to occur at a compression factor value of between 0.2 and 0.3. It may be tempting to suggest that this is a validation for Fu and Shannon's observed results (Fu & Shannon,

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

1998). However, we suggest that the reason for the mismatch at a compression factor of 0.37 for consonants is due to the fact, that in actual implants, only signals at 60 dBA with a dynamic range of only 30-40 dB are represented. Since consonants typically have lower levels than vowels, this fact affects the consonants more than vowels, causing the lower recognition rates for consonants. The fact that there is match at a lower compression factor is not indicative of the real amplitude mapping, but rather the fact that the synthesis model currently allows the full dynamic range to be represented unlike cochlear implants. This hypothesis is easily tested and left for future work.

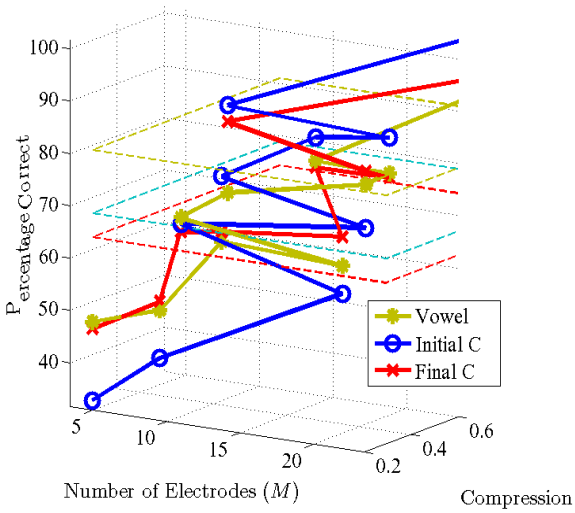

Figure 9: Plot showing the overall results as a function of the exponent k, and the number of electrodes M. The dotted planes show recipient scores.

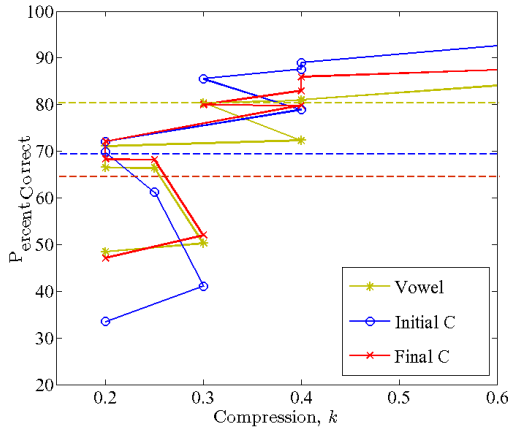

Figure 10: Recognition accuracy as a function of compression factor k. Horizontal dashed lines indicate recipient accuracy rates.

Figure 11 shows accuracy as a function of the number of electrodes chosen in the M of N strategy. We see that a larger number of electrodes are required to match recipient scores. With M=14, it seems both vowel and consonant scores can be matched with our synthesis model.

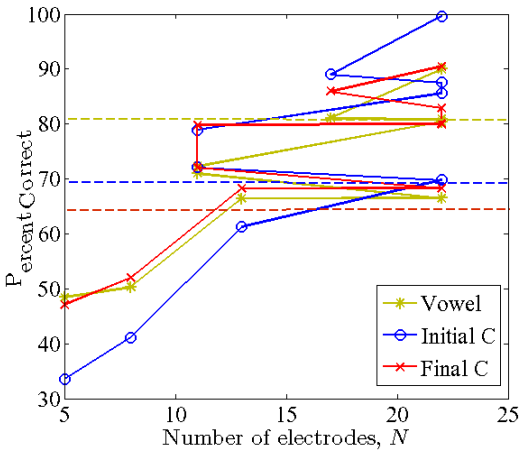

Figure 11: Recognition accuracy as a function of the number of channels used in the M of N strategy to simulate SPEAK. Horizontal dashed lines indicate recipient accuracy rates.

Interestingly, Fu and Shannon's data (Fu & Shannon, 1998) show a lower accuracy for vowels than consonants (for a constant compression factor), which is also reflected in our normal-hearing data down to a compression factor of 0.3. However, the recipient data show higher vowel recognition rates than consonants. This is intriguing and could be due to the fact that only four channels were used in Fu's study. The trend of vowels scoring higher accuracy scores than consonants is also found by Loizou for SPEAK strategies (Loizou et al., 1997). We attribute this discrepancy to the same reason as why a compression factor of k=0.37 does not seem to match recipient data – entirely due to the synthesis model producing better results for consonants than actual implants – as it is providing a bigger acoustic dynamic range to the listener. If this is accepted as the likely explanation of both Fu's and our synthesis models, the only discrepancy that remains is for the four channel implant data in Fu's study.

6. Conclusions

The paper has validated the ability of an acoustic re-synthesis model to replicate recognisability of vowels and consonants in implant recipients. An analytical estimate of the optimum compression factor required to achieve this was validated experimentally for vowels. The overall results show that when values of N=14 and a compression ratio of about k=0.3-0.37 is used in the model, normal hearing listener data are a close match to actual recipient data. This was an extremely satisfying result as it validated our re-synthesis model,

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

thus providing the authors a viable way of testing algorithmic changes to cochlear implant processors without the requirement for a large pool of recipients (that are unavailable to the authors) for subjective testing.

7. Acknowledgements

We would like to acknowledge the assistance of Hugh McDermott, Colette McKay and Catherine Sucher from the University of Melbourne for providing us with recipient CVC recognition scores as well as the corresponding database of CVC stimuli. We would also like to thank Cochlear Ltd., for providing us with a Matlab toolbox of the Nucleus speech processor. We would also like to thank our subjects – for their patience and trust.

8. References

Fu, Q. & Shannon, R.V. (1998), Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners, *JASA*, *104,* 5, pp 2570-2577.

Dorman, M.F., Loizou, P.C. & Rainey, D. (1997), Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs, *JASA*, *102,* 4, pp 2403-2411.

Loizou, P.C., Dorman, M. & Fitzke, J. (2000), The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants, *Ear and Hearing*, *21*, 1, pp 25-31.

Loizou, P.C., Dorman, M. & Poroy, O. (2000), Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution, *JASA*, *108,* 5, pp 2377-2387.

Loizou, P.C., Dorman, M. & Tu, Z. (1999), On the number of channels needed to understand speech, JASA, *106*, 4, pp 2097-2103.

Loizou, P., Graham, S., Dickins, J., Dorman, M. & Poroy, O. (1997), Comparing the performance of the SPEAK strategy (Spectra 22) and the CIS strategy (MED-EL) in quiet and in noise, *Conference on implantable auditory prosthesis*, Asilomar, Monterey, CA.

Shannon, R.V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995)., Speech recognition with primarily temporal cues, *Science, 270*, pp 303–304.

Spahr, A., Dorman, M. & Loizou, P. (2002)., Effects on performance of partial misalignments of spectral information in acoustic simulations of cochlear implants," *JASA*, *112*, 5, pp 2356.

Stevens, S.S. (1955), The measurement of loudness, *JASA*, 27, pp 815-829**.**