CHAPTER 5
MODELLING THE ELECTRICAL INTERFACE: EFFECTS OF SIMULTANEOUS STIMULATION AND COMPRESSION FUNCTION

In Chapter 4 electrical field interaction was studied. Simultaneous stimulation was assumed, but envelope extraction was performed as in non-simultaneous strategies. In this chapter an investigation into the effects of simultaneous stimulation was made, while remaining closer to actual signal-processing steps of SAS. This required some adaptations to the model described in Chapter 3. The experiment in Chapter 4 suggested that the compression function could affect speech intelligibility. An experiment regarding the effects of compression function was therefore also conducted. These two experiments explored aspects related to the electrical interface.

5.1 MODELLING SIMULTANEOUS STIMULATION

5.1.1 Introduction

The SAS strategy is available in the Clarion and Med-El implants. No envelope extraction is used; the signal is filtered into contiguous frequency channels and compressed to fit the dynamic range of the CI listener. SAS differs from CIS strategies in that SAS does not extract envelopes during the initial processing stages, nor does it use interleaved pulsatile stimulation; SAS rather uses simultaneous analogue stimulation of all electrodes. This strategy therefore preserves all fine-structure information of the filtered signal, but channel interactions are a concern in this strategy, since all electrodes are stimulated simultaneously. Speech intelligibility for the SAS strategy is similar to that obtained with interleaved strategies. For example, Friesen *et al.* (2001) found no significant differences between speech intelligibility in listeners using CIS and SAS for all speech material. The Stollwerck *et al.* study (2001) with 50 listeners also showed similar intelligibility scores for CIS and SAS listeners, with 75% of the listeners preferring CIS. In a study on strategy preferences, Zwolan *et al.* (2005) found intelligibility scores that were similar for quiet listening conditions, but that were significantly higher using the CIS strategy, when listening in noisy conditions. Most listeners preferred the CIS strategy over the SAS strategy. A block diagram of the SAS strategy is shown in Figure 5.1.

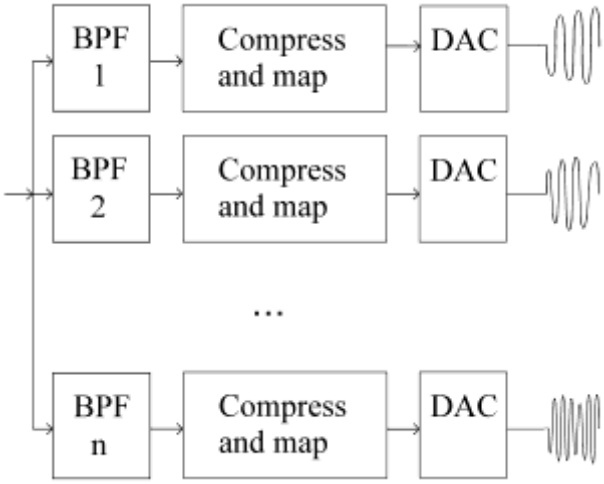

Figure 5.1 Block diagram for the SAS strategy. BPF denotes the band-pass filters. DAC denotes the digital-to-analogue converter.

Although this strategy does not present modelling challenges regarding modelling of non-simultaneous stimulation, it is especially important to include the electrical layer, as well as some assumptions pertaining to the electrophysiological layer, as will be shown. The electrical layer with its modelling of the electrical field interaction is modelled with less uncertainty about the values of effective current decay, since no temporal current decay effects need to be considered or assumed. In the experiment described in Chapter 4, the effects of non-simultaneous stimulation were ignored.

5.1.2 Methods

5.1.2.1 Assumptions

Assumptions for this model were the same as those described for the SPREAD model in Chapter 4. The signal envelope was not extracted in the initial signal-processing stages in SAS processing as in other processing strategies. The SAS model therefore modelled the effects of analogue stimulating currents, which could result in either strengthening or weakening of delivered currents, while still including insertion depths and reduced dynamic ranges. This will be discussed in more detail in the next section.

5.1.2.2 Signal processing for the SAS model

Figure 5.2 illustrates the signal-processing steps for the SAS model. The different stages of signal processing shown here are explained below. Figure 5.3 shows the outputs of the signal-processing steps shown in Figure 5.2.

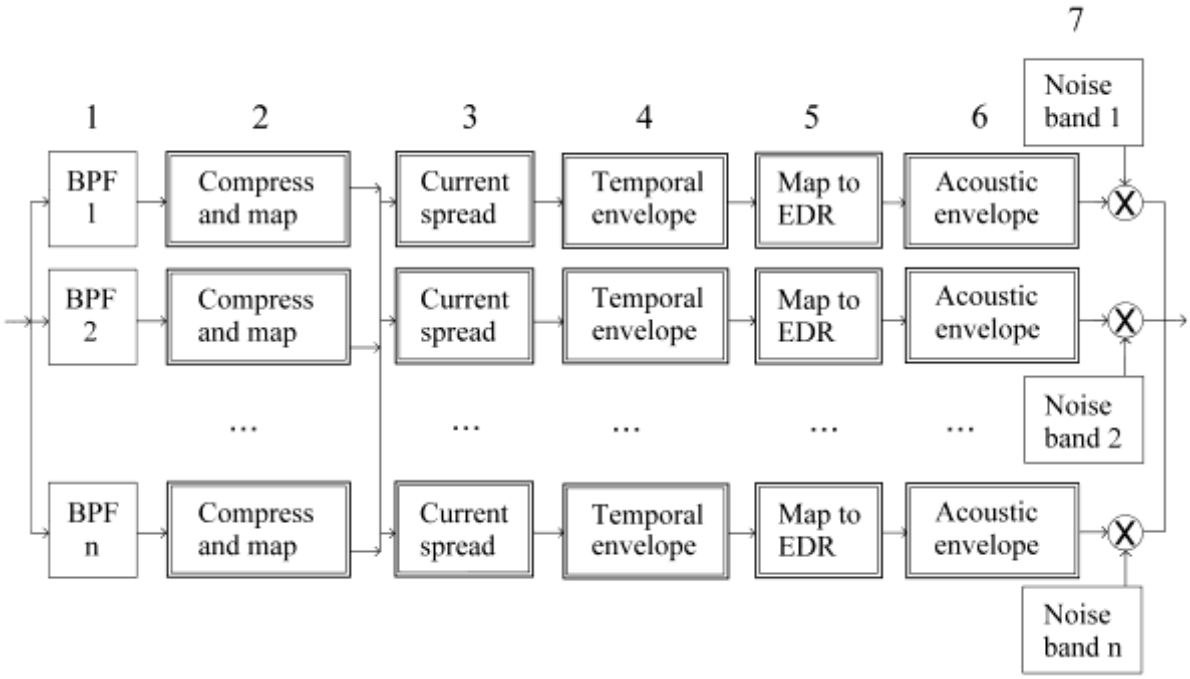

Figure 5.2 Signal processing for the SAS model. BPF denotes the band-pass filters. EDR denotes the electrical dynamic range.

5.1.2.2.1 Step 1: Filtering

Filtering for this model was the same as that described for the SPREAD model, but no envelopes were extracted.

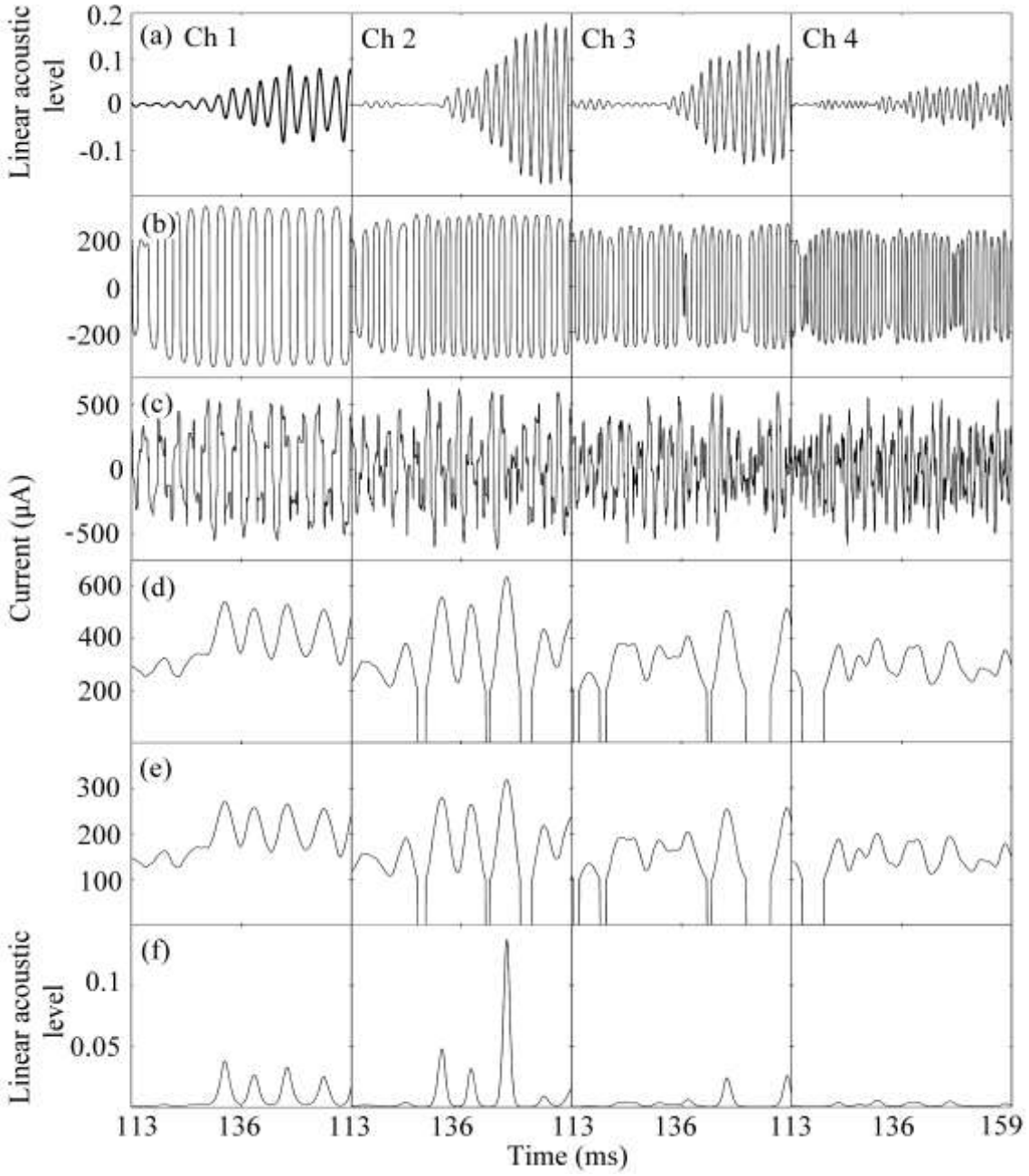

Figure 5.3 Outputs of signal-processing steps in the SAS model using an input dynamic range of 60 dB and electrical dynamic range of 11 dB. (a) Band-pass filtered signal. (b) Signal compressed using logarithmic compression function. (c) Signal with effects of spread from neighbouring channels. (d) Temporal envelope, as a model of temporal integration. (e) Temporal envelope downscaled to original comfort level and electrical dynamic range of 11 DB. (f) Acoustic envelope after inverse loudness mapping function is applied.

5.1.2.2.2 Step 2: Compression

Compression of the acoustic intensities differed somewhat from that of the SPREAD model, since both positive and negative values had to be considered. The signal was full-wave rectified, while keeping track of the negative values. The acoustic comfort and threshold levels were calculated as for the SPREAD model and the signal was compressed using Equation 3.2. Values in the signal that were below the acoustic threshold (determined by the input dynamic range) were set to 0. Finally, the signal was manipulated to reverse all the values that were initially negative, so that these values became negative again. The output of this step is shown in Figure 5.3b.

5.1.2.2.3 Step 3: Current spread

The effects of current spread were determined in the same way as for the SPREAD model, using Equations 3.5 and 3.6. Outputs of this step are shown in Figure 5.3c. Note that a different pattern of current spread effects emerges here, owing to the signal values that could be positive or negative.

5.1.2.2.4 Step 4: Temporal envelope

At this stage an additional step, namely extracting the temporal envelope of the electrical signal, was included in the SAS model. This was done for two reasons:

- Firstly, as fluctuations in the signal at a rate higher than the typical frequencies used in the synthesis signal should be avoided, the extraction of the envelope is necessary from a signal-processing perspective.

- Secondly, if a temporal integration period of 6 - 7 ms is assumed (McKay *et al.*, 2001), the use of a half-wave rectifier and low-pass filter (third order Butterworth with cut-off frequency of 160 Hz) can be justified. The output of this step is shown in Figure 5.3d.

5.1.2.2.5 Step 5: Interpreting the effective current effects (acoustic envelope)

This step was the same as that used in the SPREAD model, i.e., downscaling the electrical envelope and applying the inverse compression function. The processing steps for determining the inverse were described in Chapter 3, using Equation 3.9. The outputs of this step are shown in Figure 5.3e and 5.3f (loudness perception).

5.1.2.2.6 Step 6: Synthesis signals

Noise bands were used as synthesis signals, similar to the approach in Chapter 4.

5.1.2.2.7 Modulation of synthesis signals by envelope outputs

This step was the same as the modulation step used in the SPREAD model described in Chapter 4.

5.1.2.3 Experimental methods

5.1.2.3.1 Listeners

The same listeners who took part in the SPREAD model experiment were used.

5.1.2.3.2 Speech material

The same speech material that was used in the SPREAD model was used for this experiment, except that sentences were not included.

5.1.2.3.3 Experiments

Vowel and consonant intelligibility were tested, for seven and 16 electrodes, for SNRs of +10 dB and +5 dB, for a total of four conditions.

5.1.2.3.4 Procedure

The procedure was the same as that followed in the SPREAD model.

5.1.3 Results

Speech intelligibility results using SAS processing are shown in Figure 5.4 for consonant and vowel intelligibility. The results for the SPREAD and STANDARD model, discussed in Chapter 3, are also shown for comparison.

Consonant intelligibility. The SAS model results generally appear similar to the SPREAD model results, although there appears to be a substantial decrease in score at 16 channels for the SAS model for some aspects. Statistical analysis was performed on the consonant intelligibility scores using a two-way repeated measures ANOVA, followed by post-hoc paired t-tests where significant effects were found. Significant differences for each model are indicated by the same character as the symbol used for the graph. One symbol indicates

a significant difference at the 0.05 level. Two symbols indicate significant differences at the 0.001 level. For example, the symbol ⟷ indicates a significant difference (at the 0.001 level) in scores for the SPREAD model.

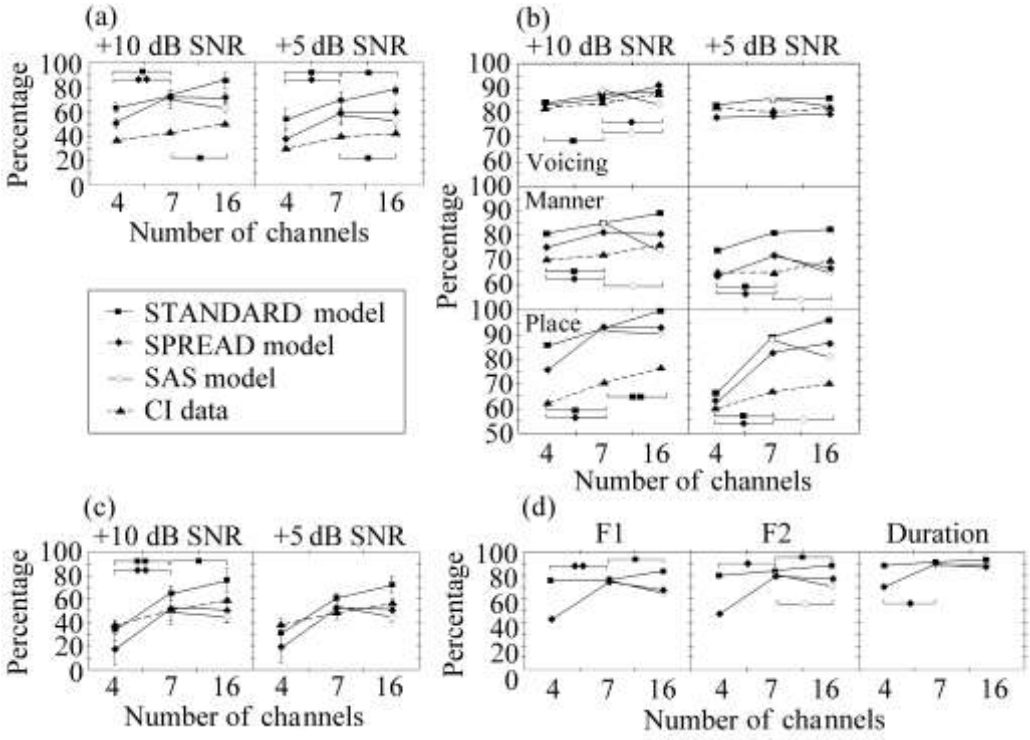

Figure 5.4 Speech intelligibility results for the SAS model. (a) Consonant recognition. (b) Consonant feature percentage correct. (c) Vowel recognition. (d) Vowel feature percentage correct at +10 dB SNR and +5 dB SNR. The CI data are from the Friesen *et al.* study (2001). Error bars on (a) and (c) indicate +-1SD. Significant differences are indicated using the same notation as in Figure 4.3.

A two-way repeated measures ANOVA on the SAS model results indicated a significant main effect of noise level ($F(1,20)=27.29$, $p<0.001$) and significant effect of number of channels ($F(1,20)=6.17$, $p<0.05$). Although there was a significant drop in the score averaged over the two noise levels from seven to 16 channels, this was not reflected in any of the individual noise levels, as indicated on Figure 5.4a.

The SPREAD and SAS model results were compared using a two-tailed paired t-test. There was a significant difference in scores between the model results ($p<0.05$). Results at the two noise levels were pooled. Paired t-tests revealed that there was no significant difference at seven channels between the two models ($p=0.41$). At 16 channels there was a significant difference between the model results ($p<0.001$), with the average for the SPREAD model at 66.2% versus 57.8% for the SAS model.

A feature analysis of consonant intelligibility was performed using the method described in Miller and Nicely (1955). The percentage correct scores for the different features was calculated to allow comparison with the Friesen *et al.* (2001) scores. These scores for voicing, manner and place of articulation for the three models are displayed in Figure 5.4b. Scores from implant listeners from the Friesen *et al.* study are displayed for comparison. The categories for voicing, manner of articulation and place of articulation are displayed in Table 3 on page 79.

Figure 5.4b shows that consonant feature transmission also appears to drop or remain constant at 16 channels.

Comparison of models. Two-tailed paired t-tests were used to determine if there were differences between model results. There were significant differences between the voicing cue score at +5 dB SNR of the SPREAD and SAS model at seven channels ($p<0.05$), for the manner cue at +10 dB SNR at 16 channels ($p<0.05$) and for the place cue at +5 dB SNR at 16 channels ($p<0.05$).

In conclusion, consonant intelligibility showed an asymptote at seven channels, and the manner of articulation and place of articulation features also displayed the asymptote.

Vowel intelligibility. Results for vowel intelligibility are displayed Figure 5.4c and d. The results for the SPREAD and SAS model are noticeably lower than the results for the STANDARD model. The vowel intelligibility scores also do not appear to drop as the SNR becomes poorer for the SPREAD and SAS models. Statistical analysis was performed on the vowel intelligibility scores using a two-way repeated measures ANOVA, followed by paired t-tests. Similar to the consonant intelligibility scores, an analysis, using post-hoc

paired t-tests, was also performed to determine if the results for the SAS model differed at seven and 16 channels. Significant differences between scores are indicated on Figure 5.4c, using the symbols as discussed for consonant intelligibility.

A two-way repeated measures ANOVA was performed on the SAS model results. It indicated a non-significant effect of noise level ($F(1,20)=2.15$, $p=0.16$), significant main effect of number of electrodes ($F(1,20)=4.8$, $p<0.05$) and non-significant interaction ($F(1,20)=2.3$, $p=0.14$). Pooling data across both noise levels showed average scores at seven and 16 electrodes of 50.1% and 44.8% respectively.

The SAS and SPREAD model results were compared using a paired t-test, by pooling the data for +10 dB SNR and +5 dB SNR and for seven and 16 electrodes. There was no significant difference between results obtained with the SAS and SPREAD models ($p=0.14$).

The vowel features F1, F2 and duration were analysed, using the method described by Van Wieringen and Wouters (1999). Categories for the cues F1, F2 and duration are shown in Table 3 in Chapter 4. The vowel features F1, F2 and duration were analysed, using the method described by Van Wieringen and Wouters (1999). Categories for the cues F1, F2 and duration are shown in Table 3 in Chapter 4. Paired t-tests were performed for each cue to determine if there were significant differences between scores at 4 and 7 channels, and between scores at 7 and 16 channels. Differences are indicated in the same way as with consonant features. The percentage correct for F1, F2 and duration cues for the STANDARD, SPREAD and SAS model are displayed in Figure 5.4d. The figure indicates that the SPREAD and SAS models display asymptote at 7 channels for F1 and F2 transmission. Only the SAS and STANDARD models showed asymptote at 7 channels for the duration cue. Outputs of the different signal processing blocks were shown in Figure 5.3. Power spectral densities for selected vowels are displayed in Figure 5.5.

A comparison between the transmission of F1, F2 and duration for the SPREAD and SAS model results using paired t-tests revealed no significant difference between the models at 7 channels for the transmission of F1 ($p=0.79$), F2 ($p=0.73$) and duration cues ($p=0.95$). At 16 channels there was no significant difference between the transmission of F1 ($p=0.44$)

and F2 ($p=0.79$) for the two models, but there was a significant difference between the transmission of duration cues for the two models ($p<0.05$).

5.1.4 Discussion

Consonants. For consonants there was an asymptote in speech intelligibility at seven channels in the SPREAD and SAS models. The SAS model results were significantly lower than the SPREAD model results at 16 channels for consonant intelligibility, significantly higher than the SPREAD model results for voicing at +5 dB SNR, significantly lower than the SPREAD model for manner at +10 dB SNR and significantly lower than the SPREAD model results at +5 dB SNR for place of articulation. It appears that the features which rely on temporal cues (Xu *et al.*, 2005) are affected more by the SAS model than by the SPREAD model, if the severe drops in voicing and manner are considered, especially at +10 dB SNR (Figure 5.4b). Place of articulation is believed to rely more on spectral cues (Xu and Zheng, 2007; Xu *et al.*, 2005). At +5 dB SNR, place of articulation cues also suffered in the SAS model.

The manner of articulation cue showed an asymptote at seven channels for both the SPREAD and SAS models, but not for the STANDARD model. This suggests that current spread, reduced dynamic range and insertion depth affect the transmission of this cue. The SAS model distorts the fine structure of the signal through current spread effects, after which an envelope is extracted, whereas the SPREAD envelope distorts the signal after the envelope is extracted. It is therefore possible that the SAS model causes more fine structure temporal envelope distortions, as illustrated in Figure 5.3 when panels (a) and (f) are compared. These temporal envelope distortions appear to be more detrimental to the manner of articulation cue perception than to the other cues. The SAS and SPREAD model results differ significantly at 16 channels for consonants. This suggests that the SAS model causes more severe temporal envelope damage than the SPREAD model.

Place of articulation cues are believed to be mostly spectral in nature (Xu and Zheng, 2007), relying not only on the spectral content of the consonant itself, but also on the successful coding of the vowel formant movements of the vowel following it (Miller and Nicely, 1955). This is confirmed by the significant increase in transmission of the place of

articulation cue in the STANDARD model from four to seven and from seven to 16 channels. The place of articulation cue asymptotes or decreases in the SAS model from seven to 16 channels. Spectral information is influenced by current spread through the alteration of relative intensities between channels. Figure 5.5 illustrates that SAS processing also damages spectral cues, but to a smaller extent than the SPREAD model. The SAS model appears to preserve the relative magnitude of the signals in the different channels. The place of articulation cue therefore also contributes to the observed asymptote in intelligibility for consonants at seven channels. The observation that both manner of articulation cues, which are mostly temporal in nature, and place of articulation cues, which are mostly spectral in nature (Xu and Zheng, 2007; Xu *et al.*, 2005), are affected in the SPREAD and SAS models at 16 channels, indicates that both temporal and spectral cues are affected by the models. It is interesting to note that place of articulation cues for the SAS model at 16 channels are significantly lower than for the SPREAD model at +5 dB SNR. The spectral information appears to be better preserved in the SAS model as evidenced in Figure 5.5, which would suggest that place of articulation cues could be better preserved in the SAS model. This is not the case, so the temporal distortions discussed in the previous paragraph for the SAS model also appear to affect the place of articulation cue, but more so at +5 dB SNR. This confirms that both spectral and temporal cues contribute to place of articulation transmission (Xu and Zheng, 2007). In this case, it appears that the damage to temporal cues caused by the SAS model was tolerated at +10 dB SNR, but that the added noise, combined with this damage, caused a severe drop in place of articulation at +5 dB SNR.

It therefore appears that a potential of 16 clearly distinguishable spectral channels are reduced to only seven distinguishable channels of information when current spread, dynamic range and insertion depth effects are considered. Figure 5.5 shows that the SAS model appears to cause more damage to temporal cues than the SPREAD model, but maintains spectral cues somewhat better.

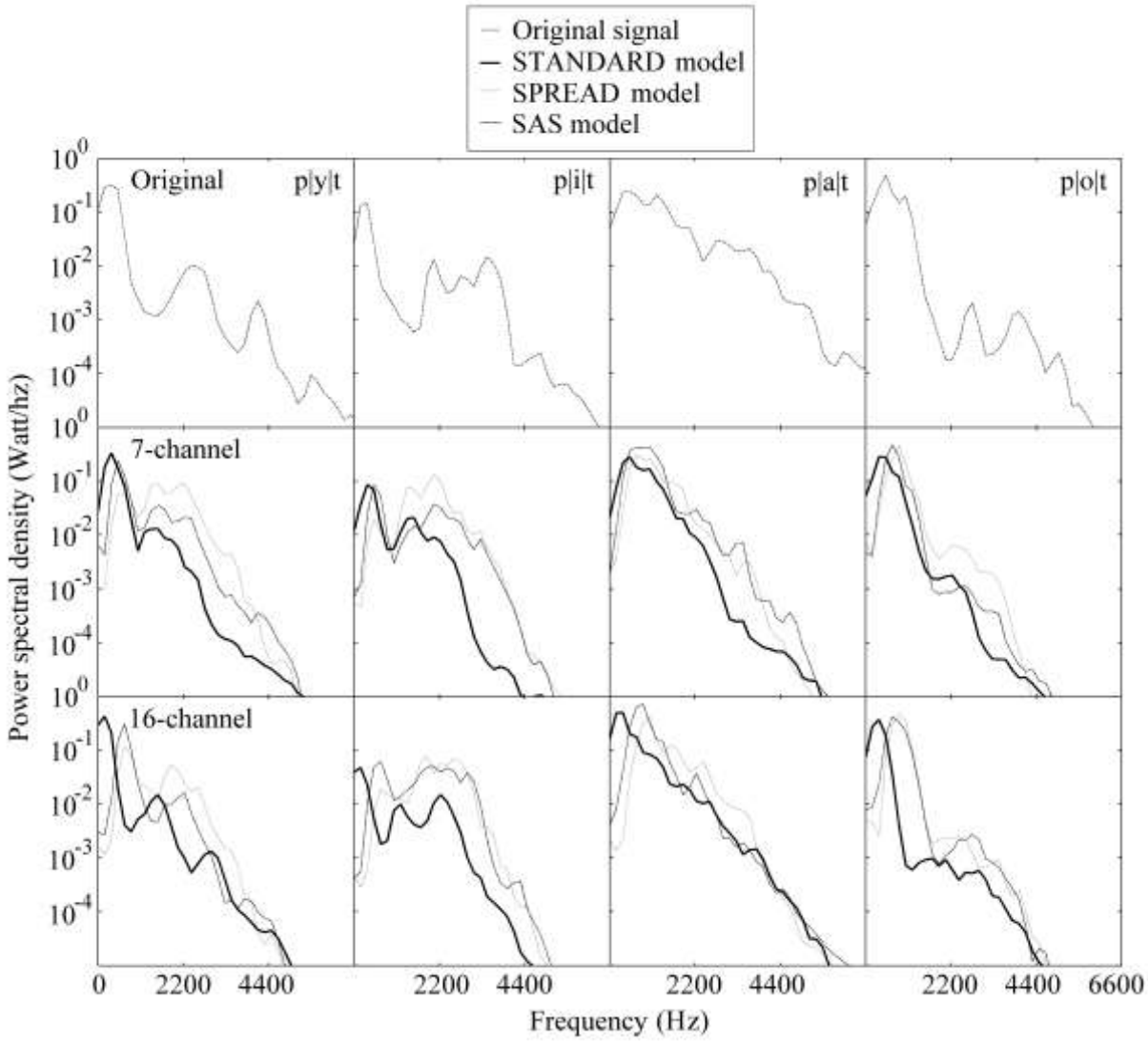

Figure 5.5. Power spectral densities for vowels processed using the SAS, SPREAD and STANDARD models.

Vowel intelligibility. The observed asymptote in vowel intelligibility at seven channels in the SAS model may be explained by the transmission of F1 and F2 cues, both of which show asymptote at seven channels. The transmission of F1 and F2 cues can be influenced by insertion depth effects, filtering effects and current decay effects, as discussed in Chapter 4. The SAS and SPREAD models appear to have the same problems resulting from current decay, which manifest as shifts in the first formant, or merging of formants, owing to border-type effects, as discussed in Chapter 4. The SAS model differs from the SPREAD model in some respects, though.

The SPREAD model, owing to the extraction of envelopes in the initial signal processing, causes increases in current in all channels, although in different measures. The result is typically a set of elevated current levels. Some of the channels are "boosted", owing to border or merged formant effects, and generally there is lower peak-to-trough contrast. This was illustrated in Chapter 4. The SAS model, on the other hand, can cause either increases or decreases in current level, owing to the analogue-stimulation strategy. This was illustrated in Figure 5.3.

The results appear to be less predictable for the SAS model, depending on the relative phases of nearby channels. However, these differences did not appear to cause differences in intelligibility, as illustrated by the non-significant differences between the SAS and SPREAD models for vowel and vowel feature identification.

5.2 MODELLING THE COMPRESSION FUNCTION

5.2.1 Introduction

Linear and non-linear dynamic range compression respectively decreases or distorts the difference in intensity levels between channels, as shown in Figure 4.2 in Chapter 4. This could influence the perceptual effects of current spread. Loizou *et al.* (2000b) studied the effects of linear compression of the dynamic range using an acoustic model. They found that all speech material was affected by dynamic range compression, with vowels affected most and consonant place of articulation also affected significantly. At a 12 dB dynamic range, vowel intelligibility fell to about 55% correct (versus 75% correct for no compression), and consonant intelligibility fell to 65% correct (versus 80% in the no-compression condition). They hypothesised that the poor vowel recognition and place of articulation identification were the result of reduced spectral contrast. Fu and Shannon (1998a) studied the effects of different power-law compression functions in CI listeners using a four-channel CIS processor and normal-hearing listeners. They found similar patterns of effects in both groups of listeners, but with normal-hearing listeners having optimal recognition using linear mapping functions and CI listeners having optimal recognition using an exponent of 0.25, which presumably restored normal loudness

growth. Normal-hearing listeners performed better than implant listeners for all speech material.

5.2.2 Methods

5.2.2.1 Assumptions

The assumptions for the compression experiment were the same as those for the SPREAD model described in Chapter 4. Furthermore, it was assumed that the perception of the electrical intensity was related to the compression function that was used. For example, if a power-law compression with compression factor of 0.07 was used for the compression phase, the inverse of the same function was used to determine acoustic intensity in the perceptual layer. The reason for this was that only effects of the compression function on current decay were to be investigated, without adding any confounding effects of normal loudness growth or the lack thereof.

5.2.2.2 Signal processing

Signal processing was the same as for the SPREAD model described in Chapter 4. The compression functions used are described in Equation 3.1 and 3.2 in Chapter 3. The functions used to convert back to acoustic intensity are described in Equation 3.9 and 3.10 in Chapter 3.

5.2.2.3 Experimental methods

5.2.2.3.1 Listeners and speech material

The listeners were the same as for the SPREAD model. The same vowels and consonants were used as those of the SPREAD model.

5.2.2.3.2 Experiments

Experiments were conducted at +10 dB SNR for 16 channels for each of the three compression functions used.

5.2.2.3.3 Procedure

The conditions for the three compression functions were randomised across listeners to eliminate the effects of learning in the average results. The other procedures were as for the experiment described in Chapter 4.

5.2.3 Results

Results shown in Figure 5.4 are for a logarithmic compression in the SPREAD and SAS models. One of the theories of the present experiment was that logarithmic compression could influence the effects of current spread. To explore this assumption, speech intelligibility was measured at an SNR of +10 dB, at 16 channels, for three different compression functions. Results are shown in Figure 5.6. The aim was to explore both more compressive and less compressive functions, so a power-law compression with an exponent of 0.07 and a linear compression were studied in addition to the logarithmic compression.

Figure 5.6a shows the shapes of these compression functions, with the acoustic intensity plotted on a logarithmic scale, using c=0.07, c=1 and a logarithmic compression function. It is clear that the more compressive function (c=0.07) reduces the contrast between higher intensities, while the linear compression function (c=1) effectively increases the contrast between higher intensities as compared to the logarithmic function. Similarly, the more compressive function effectively increases the contrast between the low intensities, whereas the linear compression effectively decreases the contrast between the low intensities relative to the logarithmic function.

Single factor ANOVAs performed on each of the aspects indicated no significant effects of compression function on vowel or consonant intelligibility, or on any of the features of consonant intelligibility and vowel intelligibility (consonant recognition: $F(2,17)=0.50$, $p=0.61$, vowel recognition: $F(2,17)=0.21$, $p=0.82$, voicing: $F(2,17)=0.31$, $p=0.74$, manner: $F(2,17)=0.04$, $p=0.97$, place: $F(2,17)=1.60$, $p=0.24$, F1: $F(2,17)=0.52$, $p=0.60$, F2: $F(2,17)=0.40$, $p=0.67$, duration: $F(2,17)=0.63$, $p=0.55$).

However, when the recognition and feature transmission scores for individual vowels and consonants were compared, there were significant differences, even though the average

scores did not show such differences. Figure 5.6e to i display individual scores for vowels or consonants, but only those in which significant individual differences ($p<0.05$) were found. No significant differences were found in the consonant features voicing and manner of articulation, both of which are believed to be mostly temporal cues. There were also no significant differences in individual vowel duration cues.

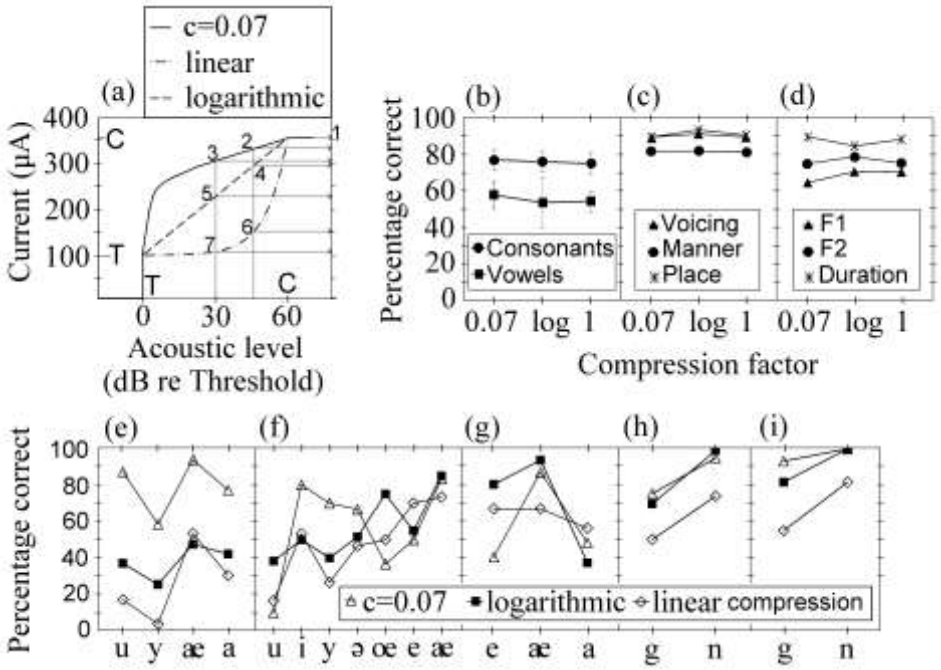

Figure 5.6 (a) Compression functions used to compress a 60 dB input dynamic range to an 11 dB electrical dynamic range. (b) Consonant and vowel intelligibility at 16 channels at +10 dB SNR, for the three different compression factors. (c) Consonant feature percentage correct for the three compression functions. (d) Vowel feature percentage correct for the three compression functions. (e) Individual vowel scores. (f) Individual vowel F1 scores. (g) Individual vowel F2 scores. (h) Individual consonant scores. (i) Individual consonant place of articulation scores. (e) to (i) only show results where significant differences were found.

5.2.4 Discussion

The results were not significantly different for any of the compression functions, although there were significant differences for individual vowel and consonant intelligibility. Figure

5.6e shows that the more compressive mapping (c=0.07) provided superior intelligibility for the vowels |u|, |y|, |ae| and |a|. With the exception of |y| and |æ| (F1, panel f) and |æ| and |a| (F2, panel g), this does not appear to be primarily attributable to F1 and F2 transmission. Studying panels (e) to (i) suggests that the more compressive mapping in general yields most benefit for individual vowel and consonant intelligibility. This is confirmed by the slightly better (although not significantly so) scores for vowel and consonant recognition, shown in Figure 5.6b. It appears as if average F1 and F2 transmission is lower for the more compressive function (although not significantly so), as hypothesised. Surprisingly this did not affect average vowel intelligibility. Figure 5.6f and g indicate that the effects of compression are not as simple as suggested by studying signal envelope profiles and power spectral densities (Figure 4.2 and Figure 4.8 in Chapter 4). It appears as if duration cues are conveyed slightly better, although not significantly so, with the more compressive mapping (panel d), which could be the aspect that facilitated the slightly better vowel intelligibility, despite poorer F1 and F2 transmission. Contrary to the theory, it appears as if more compressive mapping could enhance speech intelligibility, although it appears to exacerbate current spread effects in high-intensity channels. This is possibly facilitated by the suppression of noise by increasing the contrast between low-intensity and high-intensity channels. The more compressive function does provide speech material which sounds less noisy, but this aspect was not tested with listeners. The reduction in contrast between formant peaks appears to be less detrimental to intelligibility than was theorised.

Vowels which appeared to benefit from more compressive mapping were those which had low first formant and high second formant frequencies, i.e. large peak separation (|y| and |i|), but also some with smaller peak separation. It appeared as if the large peak separation protected the formants from the effects of current spread, while the suppression of noise aspect aided in increased intelligibility. For the smaller peak separation, the more compressive mapping appeared to facilitate the merging of peaks, which then boosted F2 transmission, but not F1 transmission (|ə|). For the vowel |æ|, both formants appeared to benefit from the more compressive function. Some other mechanism appeared to influence F1 and F2 transmission for this vowel, which also has low peak separation.

Bear in mind that the inverse function of each function was applied to model loudness perception. It is known that the loudness growth function for electrical stimulation is logarithmic (e.g. Zeng and Shannon, 1992). By applying the inverse of the compression function, this aspect is ignored. A conclusion that more compressive mapping could provide superior speech intelligibility is therefore probably presumptuous. The interaction between mapping, loudness perception (of electrical stimulation) and current decay needs further investigation before such conclusions can be drawn.

5.3 CONCLUSION

Two experiments described in this chapter illustrated the issues related to modelling simultaneous stimulation and compression function.

The SAS experiment illustrated that some mechanism is needed to ensure that fluctuations in the processed signal are not faster than those available in the synthesis signals. A half-wave rectifier and low-pass filter were used as a model of temporal integration in this experiment. It appeared that SAS processing, as modelled in this chapter, was more detrimental to temporal cues such as manner of articulation and duration, than the SPREAD model processing, described in the previous chapter. Spectral cues were also distorted, although the PSDs suggested that such distortions were less in the SAS model than in the SPREAD model.

The second experiment illustrated how the compression function used could influence the observed effects of current decay. Although no significant differences were found between average scores for vowel and consonant intelligibility and feature transmission scores for the different compression functions, there were differences in individual phoneme and feature transmission scores.

The modelling assumption of using the inverse of the compression function could conceal effects of loudness perception of electrical stimulation. Modelling the perception of loudness therefore requires a separate assumption, for example a logarithmic mapping function, to ensure that sensible conclusions may be drawn from such acoustic models.

CHAPTER 6
MODELLING THE PERCEPTUAL LAYER: EFFECTS OF DIFFERENT SYNTHESIS SIGNALS

This chapter describes an experiment that studies the correspondence of different synthesis signals' results with cochlear implant results. The work described in this chapter was accepted for publication in the Journal of the Acoustical Society of America (Strydom and Hanekom, 2011b).

6.1 INTRODUCTION

Acoustic models are used to investigate aspects of importance for speech intelligibility in general, but also specifically for CI listeners. The models typically focus on one or two controlled parameters, such as the number of channels needed for optimal speech intelligibility (Shannon *et al.*, 1995; Dorman *et al.*, 1998; Friesen *et al.*, 2001) or insertion depth effects (Baskent and Shannon, 2003; Baskent and Shannon, 2005). Although acoustic models have shown relatively good correspondence with best CI listener results in quiet listening conditions for about four channels, there are several aspects where acoustic models still differ from the outcomes achieved by CI listeners. One example is the saturation in speech intelligibility for CI listeners at about eight channels, whereas an increase in performance is observed in normal-hearing listeners (listening to sounds processed by an acoustic model) for up to 20 channels (Friesen *et al.*, 2001). As the aim of most studies using acoustic models has been to draw conclusions on the implications of the specific experimental outcomes for listening through a CI, acoustic model results may be seen as benchmarks for CI listener results and may be used to direct CI design. Consequently, it is necessary to find among the various approaches in the design of acoustic models, those that most accurately correspond to CI listener results.

Most of the published acoustic models use signal-processing steps that correspond to those used in modern-day implants, i.e. filtering the speech signal into contiguous frequency channels (the analysis filters), extracting the temporal envelope in each channel by half-wave or full-wave rectification, low-pass filtering at about 160-400 Hz and modulating a carrier signal with these envelopes (Shannon *et al.*, 1995; Dorman *et al.*, 1997b). Noise bands with filter cut-offs matched to the analysis filter cut-offs are the carrier signals

(synthesis signals) which have most commonly been used, while sinusoids that are generated with frequencies matched to the centre frequencies of the analysis filter bands have also been popular. Modulated noise bands (Blamey *et al.*, 1984b) and filtered harmonic complexes (Deeks and Carlyon, 2004) have been used to model low-rate stimulation. The present experiment investigated the performance of nine different synthesis signals in terms of correspondence to a selected set of CI listener results.

Dorman *et al.* (1997b) studied noise-bands and sinusoids in quiet listening conditions and found hardly any differences between results obtained with these signals. They studied speech intelligibility of Iowa vowels (Tyler, Preece and Tye-Murray, 1986), a subset of Hillenbrand's vowels (Hillenbrand, Getty, Clark and Wheeler, 1995), Iowa consonants (Tyler *et al.*, 1986) and HINT sentences without added noise (Nilsson *et al.*, 1994). For most of the speech material and speech features there was no significant difference between the scores obtained with the noise bands and sinusoids. The exceptions were the multi-talker vowels (Hillenbrand *et al.*, 1995), where the sinusoids produced scores that were slightly (<10%), but significantly higher than those of the noise bands, and consonant place of articulation, where the noise band processor gave higher scores than the sinusoid processor. The scores for all speech material were quite high at about 90% or better, which is substantially higher than average scores of 70% and less obtained by CI listeners (Friesen *et al.*, 2001; Pretorius *et al.*, 2006), although some individual CI listeners obtained good scores of about 80-90% for consonant recognition in these studies. Whitmal III *et al.* (2007) focused mainly on consonant intelligibility and intelligibility of words in sentences using different types of synthesis signals, including sinusoids and noise bands. The sinusoids produced better consonant intelligibility than the noise bands when listening in noisy conditions, but the outcomes in quiet listening conditions were not significantly different, with both at around 60%, much closer to implant listener results than earlier studies. The intelligibility for words in sentences was significantly better for the sinusoids at around 85% than for the noise-vocoder at around 75%.

Parameters of noise bands were manipulated in several studies to produce different groups of synthesis signals to model speech intelligibility of CI listeners (Baer and Moore, 1993; Fu and Nogaki, 2005; Boothroyd *et al.*, 1996). Spectral smearing, or varying amounts of filter overlap, was achieved by broadening the filter widths or by adjusting the filter slopes. Baer and Moore (1993) used equivalent rectangular bandwidths (ERB) to study key word recognition in sentences at three noise levels, simulating the broadened auditory filters of hearing-impaired listeners. Filter widths varied from lower to higher frequencies, with a 3-ERB condition having bandwidths of 318 Hz at 750 Hz, 561 Hz at 1500 Hz and 1044 Hz at 3000 Hz. Negligible differences in recognition were found in quiet listening conditions between 3-ERB and 6-ERB conditions (with the latter filters twice as wide as in the 3-ERB condition), with all scores more than 95%, but at 0 dB SNR the 6-ERB condition produced a significantly lower score of 68% than the 90% for the 3-ERB condition. At -3 dB SNR, these scores dropped to 35% and 72% for the 6-ERB and 3-ERB condition respectively. Fu and Nogaki (2005), using HINT sentences (Nilsson *et al.*, 1994), varied the slopes of the filters used for the noise bands to change the amount of spectral smearing. They found that results using -6 dB/octave noise bands with four simulated channels gave the closest results to implant user results, with 50% HINT sentence recognition at +10 dB SNR. Boothroyd *et al.* (1996) used smearing bandwidths of 250 Hz to 8000 Hz to study spectral smearing using vowels, consonants and isolated consonant-vowel-consonant words. At a smearing bandwidth of 250 Hz, they found small but significant changes in intelligibility for vowels and consonants (both still at more than 90%) relative to the no-smearing condition. Recognition decreased to around 15% when the smearing bandwidth was increased to 8000 Hz. Vowels were slightly more susceptible to the effects of smearing than consonants. Vowel and consonant recognition dropped to 55% and 65% respectively at a smearing bandwidth of 1000 Hz. Different approaches to modelling are described in the next three paragraphs.

An early acoustic model by Blamey *et al.* (1984b) incorporated the effect of stimulation rate into their model by using modulated noise bands as synthesis signals. The modulation rate represented the rate of stimulation, with the centre frequency of the noise bands representing place of stimulation. The width of the noise bands was presumably intended

to model current spread, although the authors did not state this explicitly. They performed pitch DL and pitch-scaling experiments on both normal-hearing listeners (using the amplitude-modulated noise bands) and CI listeners, and manipulated the modulation depth and smoothing factor (see Figure 6.2) of the modulator signals for the normal-hearing listeners to get best correspondence with the CI data. Their model results using these signals (Blamey *et al.*, 1984a) showed good correspondence with CI listener results for a wide variety of sound material, including initial and final consonants, vowels, CID and SPIN sentences, and speaker identification. The processing scheme which was used was F0 F1 F2 processing.

Oxenham *et al.* (2004) studied pitch psychoacoustics of transposed signals, which consisted of sine-wave carrier signals which typically represented place of stimulation (frequencies of more than 4 kHz), modulated by half-wave rectified sinusoids of a much lower frequency 320 Hz), which modelled rate of stimulation. Although their study did not consider speech intelligibility, by studying frequency discrimination, inter-aural time discrimination, F0 discrimination and pitch matching it was shown that mismatching rate and place of stimulation was detrimental to pitch perception. They also showed that the transposed tones at low rates of stimulation gave temporal nerve response patterns similar to what is found in the auditory nerve (Meddis and O'Mard, 1997).

Deeks *et al.* (2004) studied the effect of rate of stimulation on speech intelligibility using an acoustic model. Their model used filtered harmonic complexes as synthesis signals, which consisted of complexes of overtones of some fundamental tone (which represented the stimulation rate) to model the perception of electrical stimulation at a specific rate at a specific tonotopic place. They combined all overtones of the chosen fundamental tone in a given frequency band to find the synthesis signal for that frequency band. The Deeks study verified that their signals gave excitation patterns similar to what is expected from electrical stimulation, using Patterson's model (1995). Results from the study showed that a rate of 140 pps gives significantly higher identification of key words in sentences than a rate of 80 pps for both three and six channels. At six channels the scores were 83% and 71%, and for three channels the scores were 45% and 34% for rates of 140 pps and 80 pps respectively.

Taken together, these outcomes provide a clear motivation for the importance of careful selection of synthesis signals in creating an acoustic model, since the different signals yielded vastly different results. The present experiment addresses this issue by investigating vowel and consonant intelligibility for nine different synthesis signals originating from three different sources. Firstly, previously used synthesis signals such as pure tones and noise bands of different widths (Boothroyd *et al.*, 1996; Dorman *et al.*, 1997b; Whitmal III *et al.*, 2007), modulated noise bands (Blamey *et al.*, 1984b) and filtered harmonic complexes (Deeks and Carlyon, 2004) were included. Secondly, transposed tones (Oxenham *et al.*, 2004) which had previously been used in a psychoacoustic study, were used. Thirdly, new synthesis signals were developed by building on concepts from existing signals. The experiment compared results from these experiments to CI listener results from a previous study (Pretorius *et al.*, 2006), that used the same speech material to analyse similarities and differences between acoustic model and CI results. The Pretorius *et al.* study used listeners using either the SPEAK or the ACE speech processing strategy (Pretorius *et al.*, 2006), and therefore the present experiment used SPEAK and ACE-like processing (Skinner *et al.*, 2002). The objective was to determine which signals were the best models of CI speech intelligibility as determined by a set of performance measures.

6.2 METHODS

6.2.1 Signal processing

Since the aim of the experiment was to compare results with CI listener results, similar to the approach of Verschuur (2007), CI signal processing was followed closely without adding too much processing detail.

The observed reduced spectral resolution in CI listeners may be approached in two different ways in an acoustic model. As CI listeners have been shown to have at most four to eight spectral information channels available (e.g. Friesen *et al.*, 2001; Fu and Nogaki, 2005), the first approach would be to use a reduced number of channels in the model (typically four; see for example Fu and Shannon, (1998)), disregarding possible causes of the reduction in the number of channels.

The alternative approach would be to include implant parameters that may influence the effective number of channels more explicitly. This includes (i) the use of realistic implant parameters in the model (e.g. using actual inter-electrode distances), and (ii) modelling current spread through the use of different synthesis filter widths. This approach was followed in the present experiment, as expanded on below.

The generic signal-processing steps are illustrated in Figure 6.1. The filtering into contiguous channels was performed using an FFT, similar to the processing in the Nucleus CIs. FFT bins are combined by adding the power in relevant bins to arrive at analysis filter outputs.

SPEAK (or ACE)-type processing was used, with either six or eight strongest channels out of 20 extracted in each time window. The signal-processing block that selected these six or eight maxima in Figure 6.1 set the values in the remaining channels to 0. In the set of CI listener results that was used for comparison (Pretorius *et al.*, 2006), listeners using SPEAK processing typically used a six of 20 strategy, whilst listeners using the ACE strategy typically used an eight of 20 strategy.

In the final step, the extracted speech signal envelopes in each frequency band were modulated by the synthesis signal of each frequency band. Up to the point where the maxima are extracted, the signal processing for all nine variations in the acoustic model was the same. The nine variations differed in the design of the synthesis signal. Some aspects that were common to the nine synthesis signals are described below, while the next two sections describe the aspects that were different.

An insertion depth of 23 mm was assumed. This assumption was made to ensure that the low-rate modulators' frequency (250 Hz) would be lower than the lowest frequency of carrier signal used (722 Hz). This insertion depth could affect speech intelligibility substantially, especially if the analysis filters were not matched to the synthesis filters (Baskent and Shannon, 2005; Baskent and Shannon, 2003), but it was also a realistic value for CI implant depths. Average insertion depths of 25 mm (Baumann and Nobbe, 2006), 21.75 mm (Boex *et al.*, 2006) and 28.8 mm (Baskent and Shannon, 2005) were found in implant users, with an average insertion depth across the 16 listeners in these studies of

23.6 mm. The synthesis filter centre frequencies corresponded to simulated electrode positions, with the electrodes spaced at 0.75 mm, as in the Nucleus CI. Moreover, the average range of analysis frequencies was used, with analysis filter cut-offs as indicated in Table 4.

Effects of current spread are indirectly included through the use of different filter widths, an approach followed in several studies (e.g. Baer and Moore, 1993; Boothroyd *et al.*, 1996; Blamey *et al.*, 1984b). Bingabr *et al.* (2008) used both filter widths and filter slopes to model current spread, whereas Fu and Nogaki (2005) used filter slopes to model current spread. Bipolar stimulation excites a narrower population of nerve fibres than monopolar stimulation (e.g. Hanekom, 2001; Kral *et al.*, 1998). Typical values for the spread of excitation at the -3 dB point in electrical stimulation is 0.4 mm for bipolar stimulation using electrodes separated by 0.75 mm and 0.8 mm for monopolar stimulation (Kral *et al.*, 1998). These values were used as a guide for filter widths in some of the synthesis signals.

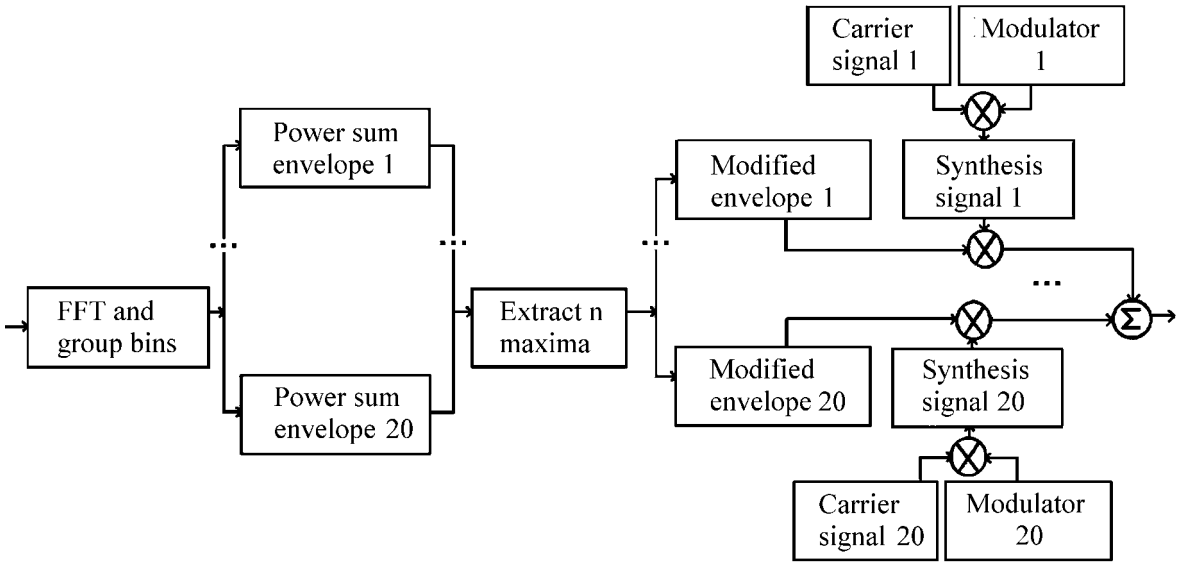

Figure 6.1. Signal-processing steps. FFT denotes the fast Fourier transform. The term modified envelope refers to some channel intensities being set to zero in the SPEAK and ACE strategy when these channels are not among those containing the spectral peaks. The modulator block is only applicable to the modulated signals.

It is acknowledged that many more aspects that were not included in the model could influence speech intelligibility, including input dynamic range (Zeng *et al.*, 2002), signal bandwidth, amplitude compression function and pulse duration (Loizou *et al.*, 2000d).

When constructing the signals, informal listening confirmed that all the signals had at least a monotone rising pitch when moving from apical to basal channels. The intention was to avoid pitch reversals which could affect speech intelligibility severely (Throckmorton and Collins, 2002).

The following two sections describe the aspects that uniquely identified the nine different synthesis signals. The signals that were used were grouped into a modulated signal group and an unmodulated signal group, as synthesis signals used in previous acoustic models were of these two types.

6.2.1.1 Modulated synthesis signals

Dual pitch percepts are reported by CI listeners, indicating that both rate and place of stimulation play a role in the perception of pitch (McKay and Carlyon, 1999). These effects are perceived up to rates of about 300-800 pps. The default stimulation rate in SPEAK processing is 250 pps, which would typically influence the perception of pitch. The similarity of amplitude modulated (AM) pulse trains, which also give a dual pitch percept up to an AM rate of about 300 Hz (McKay and Carlyon, 1999), presents AM pulse trains as a reasonable choice for synthesis signals for acoustic models of low-rate stimulation.

AMN: Amplitude modulated noise. This signal was constructed by modulating a carrier signal (representing place pitch) with a smoothed rectangular pulse (Blamey *et al.*, 1984b). The carrier signal used in the AMN synthesis signal was wide-band noise with a width of 40% of the analysis filter centre frequency, similar to the Blamey study. For the first channel, this width is 289 Hz (40% of 722 Hz). The width increases to 2476 Hz (40% of 6190 Hz) for channel 20. A duty cycle of 0.5, smoothing parameter of 0.1 and modulation index of 1 are used. The shape of the synthesis signal and its constituent signals are displayed in Figure 6.2. With the exception of the modulator, the amplitudes were

normalised to a maximum of 0.5 for all signals. The filter cut-off frequencies for the wide-band noise are given in Table 4.

AMS: Short amplitude modulated noise signal. This signal has not been used previously in an acoustic model. It has a modulator pulse width which is much shorter than that of AMN, to correspond to the typical pulse width that is used in implants with a pulse rate of 250 pps. The combined anodic and cathodic phase of a bi-phasic pulse would be 667 μs for a strategy where six maxima are extracted. The carrier signal, as model of place of stimulation, has a spread of excitation of 8 mm for this synthesis signal (corresponding to a noise bandwidth of 1000 Hz in the most apical channel, widening towards 7000 Hz at the most basal channel), which is wider than for the AMN signal, but the same as the bandwidth used in the wide noise band (WN) signal, which is discussed later. The synthesis signals for channel 1 and channel 9 are shown in Figures 6.3a and 6.3c respectively.

TT: Transposed tones. Transposed tones were used, based on the concepts used in a study by Oxenham *et al.* (2004). The rate of stimulation was modelled by the modulating envelope, which was a half-wave rectified sinusoid of frequency 250 Hz. The half-wave rectified sinusoid was low-pass filtered to avoid spectral spread of energy. The low-pass filter used in the present experiment was somewhat different from that used in the Oxenham study, namely a fourth order Butterworth filter with a low-pass cut-off of 3000 Hz. Place of stimulation was modelled by sine-wave carriers, with frequencies at the centre of the synthesis filter bands (Table I). One other adjustment was needed to ensure a monotonically rising pitch for the resulting signals, when moving from apical to basal channels. The sine wave carrier phases were adjusted within each modulator pulse to ensure that each pulse started with the same phase of the sine-wave carrier. This may be seen as a model for locking the phase of the elicited action potential to the phase of the electrical stimulus, which should be valid for low-rate stimulation, as the action potentials are phase-locked to the stimulus for low stimulation rates (van den Honert and Stypulkowski, 1987b). Another approach would be to use multiples of the modulating wave for the carrier, as was done by McKay *et al.* (1999). The present approach was chosen to ensure that the filter centre frequencies remain the same for all conditions for all

synthesis signals. The TT synthesis signals for channel 1 and channel 9 are shown in Figure 6.3.

Table 4. Analysis and synthesis filters of different signals.

Filter	**Filter -3 dB pass band (Hz)**
Analysis filters: All signals	440–565, 565–690, 690–815, 815–940, 940–1065, 1065–1190, 1190–1315, 1315–1440, 1440–1690, 1690–1940, 1940–2190, 2190–2565, 2565–2940, 2940–3440, 3440–3940, 3940–4565, 4565–5315, 5315–6190, 6190–7190, 7190–7999
Synthesis signal filters: AMN	595–881, 666–997, 751–1126, 847–1269, 952–1427, 1069–1603, 1199–1797, 1343–2013, 1503–2253, 1680–2519, 1877–2814, 2095–3141, 2337–3504, 2605–3906, 2903–4353, 3233–4848, 3599 -5397, 4005–6006, 4455–6682, 4955–7431
Synthesis signal filters: AMS, WN, FHC	354–1363, 409–1528, 469–1710, 536–1913, 610–2138, 693–2387, 785–2663, 886–2970, 999–3310, 1124–3687, 1262–4106, 1416–4570, 1586–5085, 1775–5656, 1985–6289, 2218–6992, 2476–7771, 2762–8635, 3080–9594, 3432–10668
Synthesis signal filters: NN	678–769, 769–868, 868–979, 979–1102, 1102–1238, 1238–1389, 1389 - 1557, 1557–1743, 1743–1949, 1949–2177, 2177–2431, 2431–2712, 2712–3024, 3024–3370, 3370–3754, 3754–4180, 4180–4652, 4652–5176, 5176–5757, 5757–6401
Synthesis signal filters: VN, MVN	699–747, 765–872, 837–1015, 915–1177, 999–1363, 1089–1574, 1187–1816, 1292–2091, 1405–2404, 1528–2762, 1660–3170, 1802–3635, 1956–4165, 2122–4770, 2301–5459, 2494–6245, 2702–7141, 2927–8163, 3170–9328, 3432–10668
Synthesis signal filters: SS, TT (centre frequencies only)	703, 797, 902, 1019, 1148, 1292, 1451, 1627, 1823, 2040, 2281, 2548, 2844, 3173, 3537, 3942, 4390, 4887, 5439, 6051

FHC: Filtered harmonic complexes. This is not a modulated signal, but it is included as a signal which has a pattern reminiscent of a modulated signal, as shown in Figure 6.3**.** This signal was constructed based on concepts used in a study by Deeks and Carlyon (2004). A rate of 250 pps was modelled in the FHC synthesis signal by using harmonic complexes with an F0 of 125 Hz summed in alternating phase, which corresponds to a pulse rate of 250 pps (Deeks and Carlyon, 2004). Harmonics (overtones) of 125 Hz were found within a filter band corresponding to an excitation range of 8 mm and were summed in alternating phase to construct the synthesis signal for each filter. The width of 8 mm, which corresponded to around 1000 Hz, 2500 Hz and 7000 Hz respectively in the lowest frequency, mid-frequency and highest frequency regions, differed from the 2 mm width used in the Deeks *et al.* study. In that study, F0s of 40 and 70 Hz were used, analysis filters below 1089 Hz were discarded and the synthesis filter cut-offs were matched to the analysis filter cut-offs, as it proved to give best intelligibility. As resolved harmonics provide the normal-hearing listener with place of excitation cues, the use of analysis filters above 1089 Hz, combined with a filter width of 2 mm in the Deeks study, ensured that harmonics of the fundamental frequency were not resolved. The wider filter width of the present experiment ensured that harmonics of the fundamental frequency were not resolved. The synthesis signals for channels 1 and 9 are shown in Figure 6.3. The harmonics used for channels 1, 9 and 17 are harmonics 3 - 10, 8 - 26 and 20 - 62 respectively. As an example, harmonic 8 is the lowest harmonic used for channel 9, and is 1000 Hz (8 x 125 Hz).

MVN: **Modulated noise bands with varying width.** The signal was constructed in an attempt to improve correspondence with CI listener vowel intelligibility results for modulated signals. Analysis of the first set of modulated synthesis signals showed that the TT signals and AMN signals provided best correspondence with CI listener results. The areas of concern were the low vowel recognition and poor vowel feature transmission scores when compared to CI listener results. The use of sinusoids as place carrier signals did not allow any adaptations to the typical TT signals. The place carrier signal was therefore modelled as noise bands with varying widths, with modulators similar to the original AMN signal. By using narrower noise bands in the low-frequency region, it was

hypothesised that better vowel intelligibility would be realised, while maintaining correspondence with consonant intelligibility results. The design of this signal is determined by the varying spread of excitation in apical and basal regions of the cochlea for electrical stimulation, which may be attributed to the narrower cochlear duct in the apical region (Kral *et al.*, 1998), and also possibly the spiral shape with the spiral radius smaller in the apical region than in the basal region (Hanekom, 2001). The first parameter for such a signal is the width of the pass band for the first filter. This was chosen to correspond with the values for bipolar stimulation as reported by Kral *et al.* (1998), namely 0.4 mm at the -3 dB point in the apical region. This width was adjusted to become wider in the basal region, reaching a width of 8 mm, i.e. 4 mm on either side of the electrode. Although experimental spread data show a widening of the filters in the basal region (Kral *et al.*, 1998), with an increase in width of about 0.4 mm over a distance of 2.2 mm in cat cochlea, the increase for the MVN signal was not so much determined by this experimental spread data, but rather by the observation that consonant recognition appears to be better modelled by the WN signal, rather than narrow noise bands. By retaining the relatively narrow widths in the low-frequency region, it was hypothesised that vowel intelligibility would not suffer, and that the broadening of filters in the high-frequency regions would lower consonant recognition. It was hypothesised that a signal such as this would better model speech intelligibility for both consonants and vowels. The equation for calculating the filter width is given in Equation 6.1,

$$width(i) = 0.4 + 7.6[(i-1)/19], \qquad \textbf{(6.1)}$$

where i denotes the number of the filter, and i=1 is the most apical filter. Figure 6.3 shows the typical synthesis signal for channel 1 and channel 9.

6.2.1.2 Unmodulated synthesis signals

Unmodulated synthesis signals exclude modelling of the stimulation rate. In Figure 6.1 this would imply that the modulator block is absent (or may be replaced by a constant signal with an amplitude of 1). The rationale for excluding the effects of rate explicitly in these synthesis signals is that rate of stimulation does not affect pitch above stimulation rates of

about 800 pps. Note that the unmodulated carrier signal is still modulated by the modified envelope of the speech signal, as shown in Figure 6.1.

WN: Wide noise bands. An unmodulated noise band is used as synthesis signal. The BM of length 35 mm was divided into four roughly equal portions of 8 mm each, following results from the Fu and Nogaki study (2005), where CI listeners were found effectively to have four channels of information. A choice of 8 mm corresponds to a smearing width of 1000 Hz at the most apical electrode and a smearing width of 2700 Hz at the most basal electrode, which should yield vowel and consonant recognition scores of between 50% and 30%, according to the Boothroyd study (1996). The synthesis signal for channel 1 is shown in Figure 6.3. The synthesis filters were designed using third order Butterworth filters with a width of 8 mm at the -3 dB point.

NN: Narrow noise bands with filter widths of 0.75 mm, corresponding to electrodes spaced 0.75 mm apart. The width of excitation for bipolar electrical stimulation is 0.4 mm at the -3 dB point for bipolar stimulation and 0.8 mm at the -3 dB point for monopolar stimulation (Kral *et al.*, 1998). The design of these filters specifies a width of 0.75 mm at the -3 dB point, which corresponds to the excitation width for monopolar stimulation. The typical synthesis signal for channel 1 is shown in Figure 6.3.

SS: Sinusoids (Dorman *et al.*, 1997b). Sinusoids were constructed with frequencies equal to the centre frequencies of the analysis filters, and with rms level the same as that in the original envelope. The synthesis signal for channel 1 is shown in Figure 6.3.

VN: Noise bands with varying width. These signals were used to simulate differential spread of excitation in the apical and basal regions. They were identical to MVN, except that no modulator was used. The synthesis signal for channel 1 is shown in Figure 6.3.

6.2.2 Listeners

Seven Afrikaans-speaking listeners with normal hearing, aged between 18 and 30 years, took part in the experiment. All had normal hearing as determined by a hearing screening test, with all subjects having thresholds better than 20 dB at frequencies ranging from 250 Hz to 8000 Hz.

6.2.3 Speech material

Fifteen medial consonants (b d g p t k m n f s v z j r l x), spoken by a male voice were presented in an a/Consonant/a context. Twelve medial vowels (ɑ ɑ: œ æ ɛ ɛ: u i y ə ɔ e:), spoken by a male voice in the context p/Vowel/t, were presented to the same listeners. The speech material and speaker were the same as those used in the Pretorius *et al.* study (2006). The original speech material was processed by the acoustic model, and nine different versions were created using the nine synthesis signals.

6.2.4 Procedure

Experiments were conducted in a double-walled sound booth. Processed speech material was presented in sound field using a PC with an external sound card (M-Audio Fasttrack Pro) and a Yamaha MS101 II loudspeaker. Listeners could adjust the volume to comfortable levels (found to range between around 60 and 70 dB SPL). Listeners were seated 1 m from the loudspeaker and faced the loudspeaker, which was at ear level. Consonants and vowels were presented to listeners in random order using customised software (Geurts and Wouters, 2000) without any practice session. Twelve repetitions of each vowel or consonant were presented. The software played processed consonant or vowel material, and the listener had to select the correct consonant or vowel by clicking on the appropriate button on the screen. Consonants or vowels which were processed using each of the synthesis signals each represented one condition. The material was presented one condition at a time. Vowels and consonants for all the conditions, except for VN and MVN, were presented in random order to the listeners to ensure that learning effects would not affect results. Vowels and consonants for VN and MVN were presented about a month later, with the conditions using these signals once again randomised. Chance performance level for the vowel test was 8.3%, and the 95% confidence level was at 12.48% correct. Chance performance level for the consonant test was 6.7%, with the 95% confidence level at 11.1% correct. No feedback was given.

A control experiment using six representative synthesis signals (SS, NN and VN for vowels and AMN, TT and WN for consonants) was conducted with three of the listeners.

Both vowel and consonant intelligibility were tested to determine if learning effects may have played a role in the intelligibility of the phonemes processed by the acoustic model.

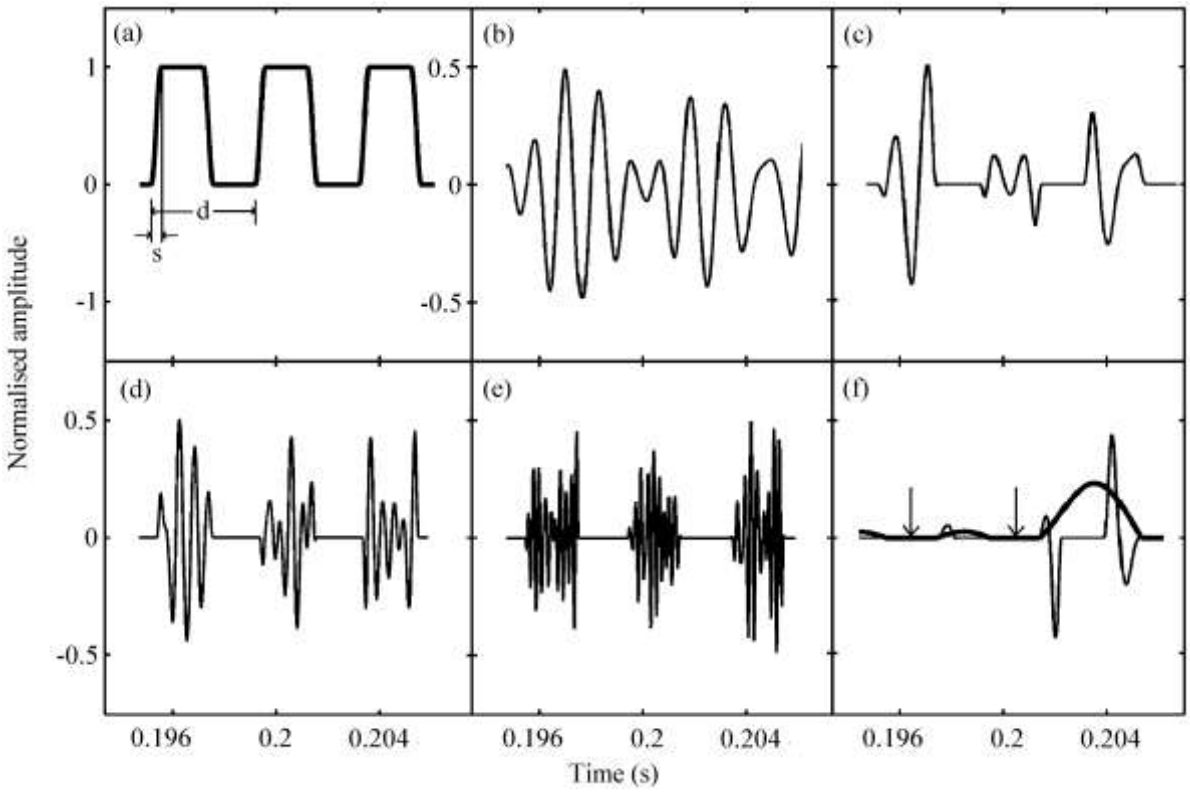

Figure 6.2. Modulated wide-band noise synthesis signal (synthesis signal AMN). Only a brief time segment is shown. Signal amplitudes were normalised to a maximum of 0.5. (a) Modulator signal corresponding to the stimulation pulse rate. Smoothing parameter = s/d (0.1 for this signal). (b) Wide-band noise centred around 722 Hz for channel 1. Filter width is 289 Hz. (c) Synthesis signal for channel 1, being the product of (a) and (b). (d) Synthesis signal for channel 9 (wide-band noise centred at 1843 Hz). (e) Synthesis signal for channel 17 (wide-band noise centred at 4410 Hz). (f) An example of the output of channel 1 for a particular input speech signal: the extracted envelope of the speech signal in channel 1 (shown in bold) was modulated by the synthesis signal in panel (c). Note how the SPEAK (and ACE) strategies set some speech envelope values to zero, as indicated by the arrows in Figure 6.2f.

Twelve repetitions of each processed phoneme were presented in random order to the listeners (four repetitions of each phoneme synthesised using three synthesis signals). This was repeated four times, so that there were four consecutive sets of twelve repetitions. Each set was seen as representing a learning event. The objective was to establish whether learning occurred over the period of presentation of these four sets of repetitions. This control experiment was conducted several months after the original experiment. Thus, learning effects from the original experiment would be minimal. Loudness was fixed at 65 dB SPL during this control experiment.

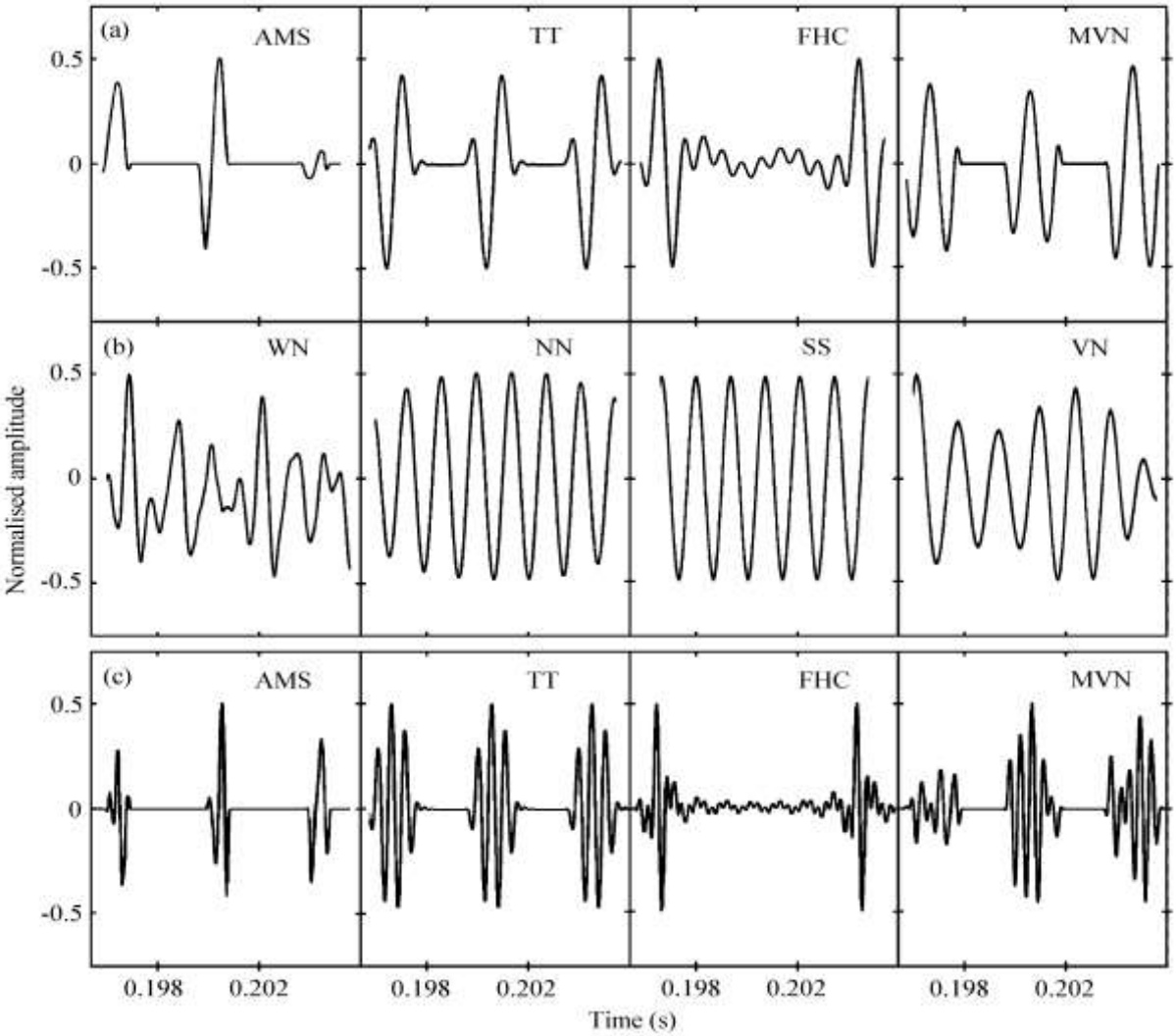

Figure 6.3. (a) Modulated synthesis signals for channel 1. (b) Unmodulated synthesis signals for channel 1. (c) Modulated synthesis signals for channel 9. All signals were normalised to give maximum amplitudes of 0.5.

6.2.5 Performance measures

Analysis of the confusion matrices for consonants into the features voicing, manner of articulation, place of articulation, affrication, burst, nasality and amplitude envelope was carried out using information transmission analysis as described in Miller and Nicely (1955). Analysis of the confusion matrices for vowels was carried out in a similar manner, studying the features F1 and F2 and duration of each vowel, with categories described by Van Wieringen and Wouters (1999). To allow statistical analysis, feature information transmission scores were obtained from information transmission analysis of the confusion matrices of each individual.

In acoustic modelling studies, quantitative comparisons between results obtained with the acoustic model and results of CI listeners listening to the same speech material are typically made using comparisons between feature information transmission scores (e.g. Friesen *et al.*, 2001; Fu and Shannon, 1998). Discussion of the differences between model and CI results has often been qualitative, for example highlighting that information transmission of a particular feature differs between CI and normal-hearing listeners. The present experiment, however, compared different synthesis signals, and therefore had to determine which synthesis signal results were closest to those of CI listeners. Four different performance measures were used to compare the confusion matrices obtained with acoustic models with that of CI listeners to determine which synthesis signal best modelled CI listener perception. Each performance measure used emphasised different aspects of performance. Therefore, the measured performance of a synthesis signal was expected to be related to the specific performance measure employed. As performance measures have not been used before when comparing acoustic model outcomes to CI results, part of the objective of the present experiment was to comment on the suitability of possible performance measures.

The first measure of performance (Eq. 2) was a sum of squares of differences in information-transmission scores. The squares of differences between CI and normal-hearing results were obtained by using differences in average scores for each of the attributes considered to characterise phoneme intelligibility. Information-transmission scores for the three vowel features F1, F2 and duration, as well as percentage correct vowel

recognition, were used as four attributes that characterise vowel intelligibility, while the consonant features voicing, manner, place of articulation, amplitude envelope, affrication, burst, nasality and percentage correct consonant recognition were used as eight attributes that characterise consonant intelligibility.

The square of differences (*SD*)

$$SD(i, j) = (IT(i, j) - IT_{CI}(i))^2, \tag{6.2a}$$

and means of these squares of differences (*MSD*)

$$MSD(k, j) = \frac{1}{n_k} \sum_{i=1}^{k_1} SD(i, j), \tag{6.2b}$$

were obtained, with *IT(i,j)* the average information transmission score (or percentage correct in the case of vowel and consonant recognition) measured for the speech attribute *i* using the synthesis signal *j*, $IT_{CI}(i)$ the average information transmission score measured for CI listeners for phoneme attribute *i* and *SD(i,j)* the square of differences for attribute *i* using synthesis signal *j*. *MSD(k,j)* is the mean of the squares of differences for lumped measure *k* (for example all four vowel attributes) for synthesis signal *j*, where the summation is over all the relevant phoneme attributes for the specific lumped measure and n_k denotes the number of these attributes for the lumped measure *k* (n_k = 4 for vowels and 8 for consonants). *SD* and *MSD* are then transformed to values between 0 and 1 to ensure that good performance would be represented by values close to 1,

$$NSD(i, j) = 1 - \frac{SD(i, j)}{NF_1}, \tag{6.2c}$$

$$NMSD(k, j) = 1 - \frac{MSD(k, j)}{NF_2}, \tag{6.2d}$$

where *NSD(i,j)* is a normalised performance measure for each phoneme attribute *i* when the synthesis signal is *j*. *NMSD(k,j)* is the normalised mean performance measure for

synthesis signal *j*, for the lumped measure *k*. NF_1 and NF_2 are normalisation factors, found from the maximum of all *SD* and *MSD* values respectively, to ensure that the *NSD* and *NMSD* scores are normalised to a maximum of 1, with higher values of *NSD* and *NMSD* indicating better performance.

A second performance measure was the concordance index, as described in Brusco (2004). This performance as expressed by the concordance index will be denoted by PCI. The concordance index gives an indication of how well the rows of one confusion matrix follow the trends of the same rows in a second confusion matrix. A particular row in a confusion matrix shows the fraction of correct classifications of a particular phoneme and (off-diagonally) the confusions with all other phonemes in the set. Thus, this index considers to which extent the confusions in two different confusion matrices correspond. When two confusion matrices are identical or when the same rows of the two matrices are linearly related, the concordance index is 1.

A third performance measure was Pearson's correlation coefficients (PCC) between the diagonal confusion matrix elements obtained from normal-hearing listeners listening to each version of the acoustic model and the diagonal confusion matrix elements of CI listener results, in each case summed over listeners. This coefficient gives an indication of the correspondence between individual phoneme recognition scores for the group of CI listeners and group of normal-hearing listener results, the average of which are usually reported as the vowel and consonant recognition scores.

The fourth performance measure was the Pearson's correlation coefficient found from the correlation between off-diagonal matrix elements for each acoustic model and CI listener results (denoted as PCC-O hereafter). Diagonal elements were removed from the summed confusion matrices and the remaining matrices were then compared using correlation analysis. This coefficient may be seen as a measure of how well the phoneme confusions for CI listeners correlate with those of normal-hearing listeners listening to a particular version of the acoustic model.

To arrive at a lumped measure, the four performance measures were then ranked for each phoneme attribute considered, by assigning values 2 to 10 to each synthesis signal (nine

synthesis signals), with 10 indicating the best performance and 2 indicating worst performance. These rank values were then summed and normalised to a maximum of 1, for vowels, consonants and vowels and consonants combined, to typify overall performance of each synthesis signal.

Finally, the most prevalent confusions for CI listeners as well as for normal-hearing listeners listening to each version of the acoustic model were examined to determine whether similar confusions were present in both.

6.3 RESULTS

The primary objective of the experiment was to determine which synthesis signal gave the best performance in terms of correspondence with CI listener results. Where the term "performance" is used, this denotes correspondence with CI data using the four different performance measures (Figures 6.5 and 6.6), whereas the term "intelligibility" refers to the phoneme intelligibility scores (Figure 6.4) obtained with a particular synthesis signal, with high intelligibility indicated by high scores (high percentage correct or high percentage information transmission). High intelligibility is not necessarily related to good performance of a synthesis signal. For synthesis signals NN and SS for example, performance for consonant attributes is generally poor (Figures 6.5b and 6.6b), although their intelligibility is high (Figure 6.4).

Figure 6.4 shows the consonant and vowel recognition scores, as well as the feature information transmission scores obtained with the different synthesis signals. Data from CI listeners for the Pretorius *et al.* study (2006) are also displayed.

Figure 6.5 shows the normalised square differences (defined in Eq. 2c) for the individual phoneme attributes for the nine synthesis signals. The performance indices using the four measures of performance are shown in Figure 6.6. Figures 6.6a and 6.6b show that the performance measures generally display mixed trends. The trend of the concordance index (PCI) appears to differ generally from the trends of the other three measures. Performance indices for vowels appear to be generally higher than those of consonants, except when measured by the concordance index, which was typically lower for vowels than for

consonants. The SS and AMS synthesis signals are the poorest performers in predicting consonant attributes, while the AMS signal performs poorest for vowel attributes.

Figure 6.6c shows the best overall rank scores for vowel performance to be similar to those of consonant performance, but the signals that performed best for vowel attributes were different from those that performed best for consonant attributes. The four best performing synthesis signals for predicting vowel attributes (as judged from the rank scores) are SS, VN, NN and MVN in that order. Similarly, the four best synthesis signals for predicting consonant performance are MVN, AMN, TT and VN. Considering prediction performance of vowels and consonants together, the best synthesis signals were VN, MVN, AMN and TT, with NN very close to TT.

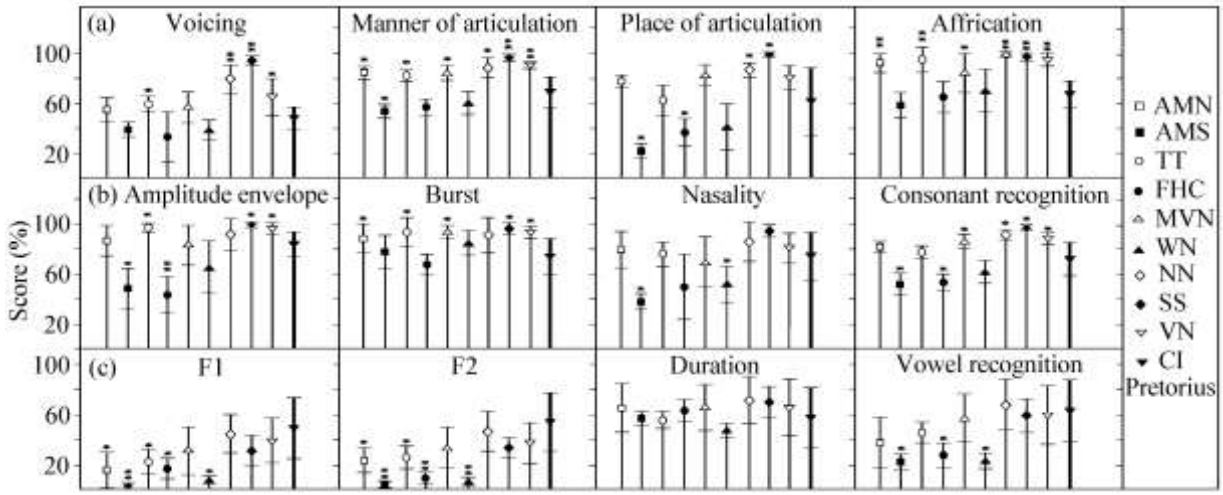

Figure 6.4. (a) and (b) Consonant feature information transmission scores and consonant recognition percentage correct. (c) Vowel feature information transmission scores and vowel recognition percentage correct. Error bars indicate +-1SD. Results from CI listener study are indicated using bold lines. * indicates significant difference from the CI listener results (Pretorius *et al.*, 2006) at the 0.05 level, whereas ** indicates significant difference at the 0.001 level.

The three best performing synthesis signals' results were compared with CI listener results for each consonant and vowel attribute using one-way ANOVAs. The observation that intelligibility of consonant attributes appeared to benefit from synthesis signals with

narrow spread of excitation (Figure 6.4a and b, synthesis signals SS, NN and VN), prompted a comparison of the SS signal results with those of the NN and VN signals for all consonant and vowel attributes, to determine whether these differences were significant. Comparison with the VN results was expected to show up sensitivities to simulated spread of excitation in different cochlear regions. Synthesis signals VN and MVN generally had good performance for both consonant and vowel attributes. They differed in one aspect only, namely the use of a modulator signal. Tables 4 and 5 show the results of the one-way ANOVAs between best-performing signals and CI results, and between the synthesis signal groupings MVN and VN, VN and SS, VN and NN, and NN and SS.

Table 6 shows that the results for SS, NN and VN all differ non-significantly from CI listener results for all vowel attributes. The degrees of freedom are shown at the top of each section, with the values of F and p shown in the table. Table 5 shows a mixed pattern of differences for consonant attributes, with the AMN signal appearing to differ significantly for only two attributes. MVN results differ non-significantly from VN results for all attributes. The results for NN, VN and SS differ non-significantly for all vowel attributes, and results for VN and NN differ non-significantly for all phoneme attributes. There were significant differences between NN and SS and between VN and SS for voicing, manner and place of articulation. VN, but not NN, differed significantly from SS for nasality.

A comparison between the most prevalent phoneme confusions predicted using the synthesis signals and the most prevalent confusions for CI listeners shows that these generally differ. The five most prevalent vowel confusions for CI listeners were |y| with |i|, |u| with |œ|, |ɛ| with |ə|, œ with | ə| and |e:| with |ɛ:|, while for consonants CI listeners mostly confused |t| with |d|, |l| with |n|, |w| with |b|, |p| with |b| and |n| with |j|. None of the synthesis signals showed exactly these same confusion patterns for either vowels or consonants, although some synthesis signals had one or two of these confusions in their five most prevalent confusions, with the MVN synthesis signal faring the best.

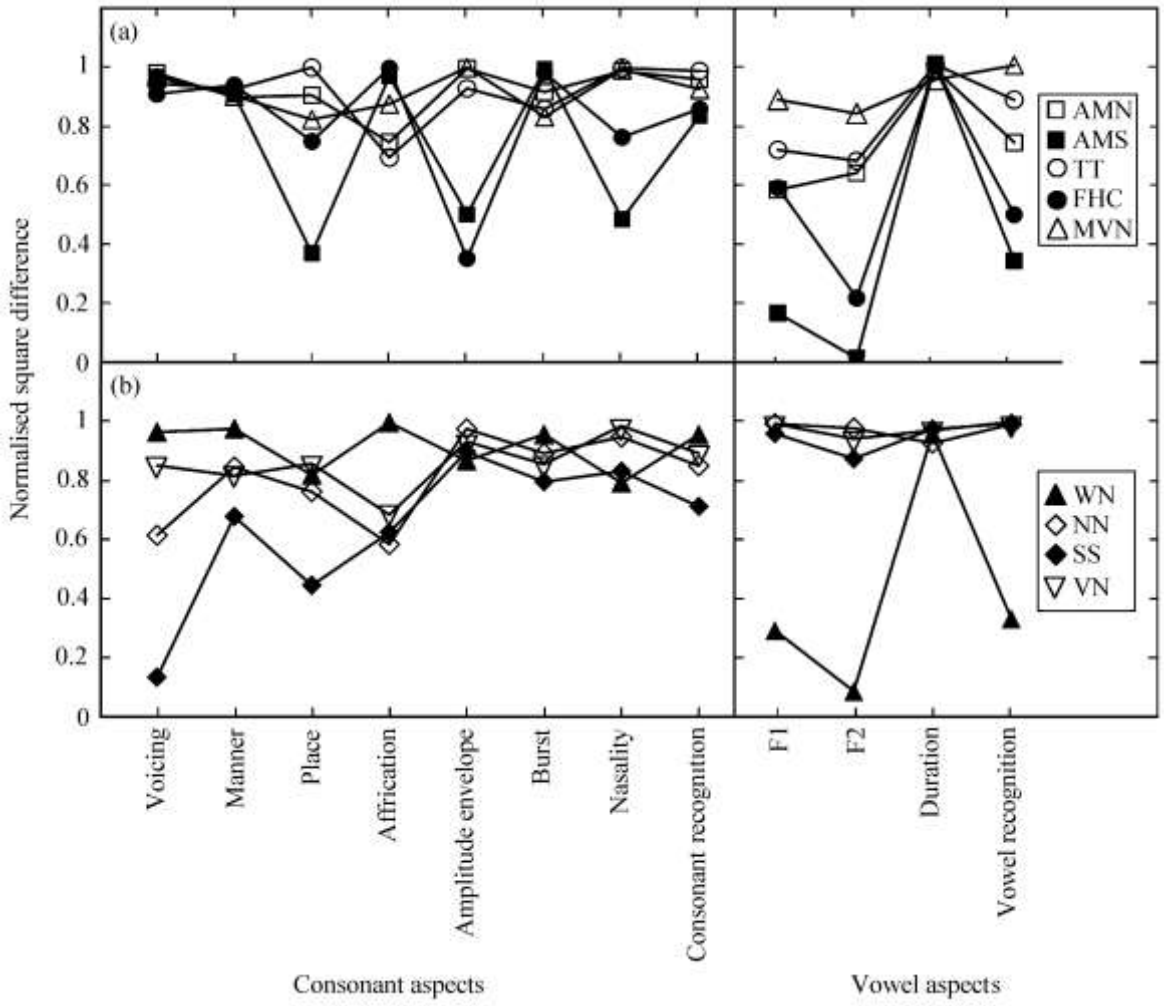

Figure 6.5. Performance of different synthesis signals for individual attributes using normalised square difference scores (NSD in Eq. 2c). (a) Normalised square difference for modulated synthesis signals. (b) Normalised square difference for unmodulated synthesis signals.

Learning effects. An analysis of the original 12 repetitions was done by dividing the 12 repetitions into three sets (or learning events) of four repetitions each. A two-way ANOVA (factors synthesis signal and learning event) was performed to determine if any learning effects could be observed which might possibly affect interpretation of results. However, no effects of learning were observed for vowels (main effect of synthesis signal, $F(8,188)=31.46$, $p<0.001$; no main effect of learning event, $F(2,188)=0.47$, $p=0.62$) or consonants (main effect of synthesis signal, $F(8,188)=112.50$, $p<0.001$; no main effect of

learning event, F(2,188)=2.50, p=0.09). The control experiment that was performed several months later using four learning events of four repetitions each for three listeners, confirmed that no significant learning effects were observed for either vowels or consonants (two-way ANOVA for consonants: no main effect of learning event, F(3,35)=0.31, p=0.82; two-way ANOVA for vowels: no main effect of learning event, F(3,35)=0.52, p=0.67). The control experiment results (for the six selected synthesis signals, three for vowels and three for consonants; three listeners; loudness level fixed at 65 dB SPL) were also compared to the original results for these synthesis signals for the same three listeners (who had originally listened at their comfortable listening levels), using a two-way ANOVA (factors synthesis signal and listening level). This comparison indicated no significant main effect of level for either vowels (F(1,17)=0.61, p=0.45) or consonants (F(1,17)=2.66, p=0.13), which confirms that the comfortable listening levels did not yield results different from results obtained at a fixed loudness level of 65 dB SPL.

6.4 DISCUSSION

Learning effects. Although learning effects may play a role in results, as illustrated by Rosen *et al.* (1999), acoustic modelling studies have in general not been consistent in their approach to possible learning effects. Many acoustic model studies provided no training, but relied on randomisation of test conditions to eliminate learning effects (Baskent, 2006; Baskent and Shannon, 2007; Verschuur, 2009; Deeks and Carlyon, 2004; Fu and Nogaki, 2005). Other studies relied on the experience of the listeners (Loizou *et al.*, 2000a), some used moderate training of around one hour or less (Bingabr *et al.*, 2008; Stickney, Zeng, Litovsky and Assmann, 2004; Loizou *et al.*, 2000b; Green, Faulkner and Rosen, 2004), whereas still others allowed extensive training of three hours or more or used some measure to ensure that performance had stabilised (Xu and Zheng, 2007; Souza and Boike, 2006; Throckmorton, Selin Kucukoglu, Remus and Collins, 2006). The analysis of the original experiment results into three sets of learning events, as well as analysis of the control experiment results, confirmed that learning effects were not important during the present experiment, probably because of the extensive experience of the group of listeners combined with the random presentation of signals. Similarly, the use of comfortable listening levels as opposed to fixed loudness levels did not affect results.

Performance and intelligibility. The terms performance and intelligibility were defined earlier (see the Results section). Figures 6.5 and 6.6 show a general trend of modulated synthesis signals (MVN, AMN and TT) leading to better performance for consonant attributes, and narrow-spread unmodulated synthesis signals (SS, NN and VN) giving better performance for vowel attributes. Two distinct aspects that influence results may be identified in the set of synthesis signals. The first is the modulation or absence of modulation in the synthesis signal, coupled with the ability of the synthesis signal to sample the speech envelope effectively. The second aspect is the modelled spread of excitation of the synthesis signal. In this respect a distinction must be made between the spread of excitation of the carrier signal in the case of modulated signals, and the spread of excitation of the synthesis signal, which results from modulating the carrier signal. The modulation of any signal effectively broadens its spectrum, since the modulation adds high-frequency components to the synthesis signal spectrum, as exemplified by the increase in excitation width in channel 2 from 0.8 mm (carrier signal) to 2mm (synthesis signal) for the MVN signal. Figure 6.6a, 6b and 6c show signals ordered from the smallest spread of excitation on the left to the largest spread of excitation, based on the filter width of the synthesis signal in channel 1.

Vowel performance and intelligibility. Vowel performance was best for the SS, NN and VN signals. These signals also had the best intelligibility. The non-significant difference of the NN signal results from the SS signal results (Table 6) suggests that the typical spread associated with monopolar stimulation (of which NN is a good model) in the apical region does not affect vowel intelligibility. The SS and NN signals model relatively narrow spread of excitation in all regions of the cochlea, which explains their good intelligibility results. This may be compared to the findings of Dorman *et al.* (1997b), who found no difference between results obtained using sine-wave processors and noise-band processors for all vowel material, except for multi-talker vowels, where the sine-wave processor gave slightly better intelligibility. The analysis filters and synthesis filters were matched in that study. The best intelligibility results obtained in the present experiment were still relatively low and may be explained by the modelled insertion depth of 23 mm. Baskent and Shannon (2003) found decreases in vowel intelligibility of about 20% for insertion depths

of 25 mm with a compression of 5 mm in a noise-carrier simulation with normal-hearing listeners. Baskent *et al.* (2005) found decreases of 17% in vowel intelligibility for CI listeners when insertion depths were reduced from 28 mm to 24 mm.

Although both the VN and MVN signals have relatively narrow filters for their carrier signals in the apical region (widths of less than 2.8 mm up to 1300 Hz), both have exaggerated filter widths widening to 4.8 mm at channel 12 (2568 Hz) and to 8 mm at channel 20 (6071 Hz). Their intelligibility for all vowel attributes differed non-significantly from that of the NN signal and the SS signal, which both have narrower spread of excitation in all but the first channel. This suggests that vowel intelligibility is tolerant of relatively wide spread of excitation in higher-frequency channels, at least for SPEAK and ACE-like processing.

Consonant performance and intelligibility. The best performing synthesis signals for consonant attributes were the MVN, AMN and TT signals (Figures 6.6b and 6.6c). The AMN and MVN signals have widening spread of excitation towards the higher-frequency regions, as indicated in Table 3. The synthesis signals that produced the best consonant intelligibility were SS, NN and VN, with the SS signal having significantly better intelligibility than NN and VN for most attributes of consonant intelligibility, as shown in Figure 6.4 and Table 5. The Whitmall III *et al.* study (2007) showed a similar trend, with the sinusoids yielding better scores than noise-carriers in both quiet listening conditions and noise. Table 5 shows some interesting trends. Voicing, manner, place of articulation and consonant recognition were sensitive to spread of excitation: both the NN and VN synthesis signals had significantly lower scores than the SS signal. The VN results for nasality were significantly lower than those of SS, whereas the NN results were not, suggesting that nasality transmission does not tolerate wide higher-frequency excitation widths. Affrication, amplitude envelope and burst transmission appeared less sensitive to spread of excitation, as illustrated by the non-significant differences between the NN and VN signal results and SS signal results.

Two hypotheses can be formulated to explain the performance of the MVN, AMN and TT synthesis signals. The first relates to spread of excitation. Both the MVN and AMN

synthesis signals have carrier signal filter widths widening towards the basal region. The AMN carrier signal width widens from 2.3 mm in the apical region to 3 mm in the basal region, whereas the MVN carrier signal width changes from 0.4 to 8 mm from apex to base. Both are modulated signals. The increasing excitation widths of the synthesis signal carriers towards the basal region may therefore be the key to the good performance, as they may be seen as models of the current spread increasing towards the basal region of the cochlea. The filter widths of both of these signals' carriers at the basal end are however exaggerated relative to the excitation width of 0.8 mm found in the Kral *et al.* study (1998), which suggests that there may be other aspects which could cause some additional widening of the cochlear filters for CI listeners. Severe hearing loss in the high-frequency regions, versus residual hearing in the low-frequency regions (von Ilberg, Frankfurt and für Hals-Nasen-Ohrenheilkunde, 1999; Gantz and Turner, 2003) for some listeners, and trends of increasing thresholds towards the higher frequencies for listeners with hearing loss (e.g. Baskent, 2006) suggest that degeneration of peripheral axonal processes of nerve fibres may be more severe in the basal region than in the apical region, leading to wider auditory filters in the basal region. This, in combination with increasing current spread towards the base, may explain why exaggerated excitation widths in the synthesis signal gives good correspondence with CI results.

The second hypothesis is that modulation type effects broaden the spectrum, without the need for an unrealistic amount of current spread. The extent of this broadening is determined by the modulation depth and smoothing factor (described in Figure 6.2) of the modulating signal. For example, the NN signal has a spread of excitation of 0.75 mm at the -3 dB point for channel 2. At this channel, the MVN signal has a similar spread (0.8 mm) in its carrier signal, but its synthesis signal has a spread of excitation of 2 mm at the -3 dB point. This effect of modulation could conceivably provide spread of excitation approaching that of AMN, with a carrier signal modelling much smaller current spread of 0.8 mm – the typical monopolar excitation width – in channel 2. This could explain why modulated signals such as TT, AMN and MVN provide good performance for consonant attributes, even though the spread of excitation of their carrier signals differs substantially. It appears that the modulation in these signals provides the widening filters, without the

need for unrealistic amounts of current spread. In the case of TT, there is no spread of excitation in the carrier signal, but the excitation width of its synthesis signal is 3.5 mm in channel 2. The presence of modulation could also be used to study the effects of temporal sampling rate, as discussed next.

Comparison of MVN and VN. The scores obtained with the MVN synthesis signal, which is a modulated version of VN, differed non-significantly from the VN scores for all attributes, as shown in Tables 4 and 5. This suggests that consonant and vowel intelligibility are not affected by a low rate of sampling (down to rates of 250 Hz) of the speech signal, at least in quiet listening conditions for SPEAK and ACE-like processing. Studies with CI listeners yield mixed results, reporting both no effect of stimulation rate (e.g. Fu and Shannon, 2000a; Holden, Skinner, Holden and Demorest, 2002) and significant effects of stimulation rate (e.g. Kiefer *et al.*, 1997; Loizou *et al.*, 2000d; Buechner *et al.*, 2006; Frijns *et al.*, 2003). The increase in intelligibility with higher stimulation rates may possibly be attributed to the improved stochastic firing of the neurons when using higher stimulation rates (Rubinstein and Hong, 2003), rather than to the improved sampling ability associated with such stimulation rates.

Comparison of vowel and consonant performance. Generally vowel results using the synthesis signals were closer to CI results than consonant results. SS, VN and NN all differ non-significantly from CI results for the four attributes of vowel intelligibility studied, as shown in Table 6. The occurrence of significant differences between the results for the SS, NN and VN group for some consonant attributes indicate that consonant intelligibility is more sensitive to reduced spectral selectivity than vowel intelligibility for SPEAK and ACE-like processing. Figure 6.6c shows the best performing signals for consonant attributes to be moderate performers for vowel attributes, and vice versa. This illustrates that no synthesis signal (among those considered in this experiment) models the perception of phonemes optimally for both consonants and vowels.

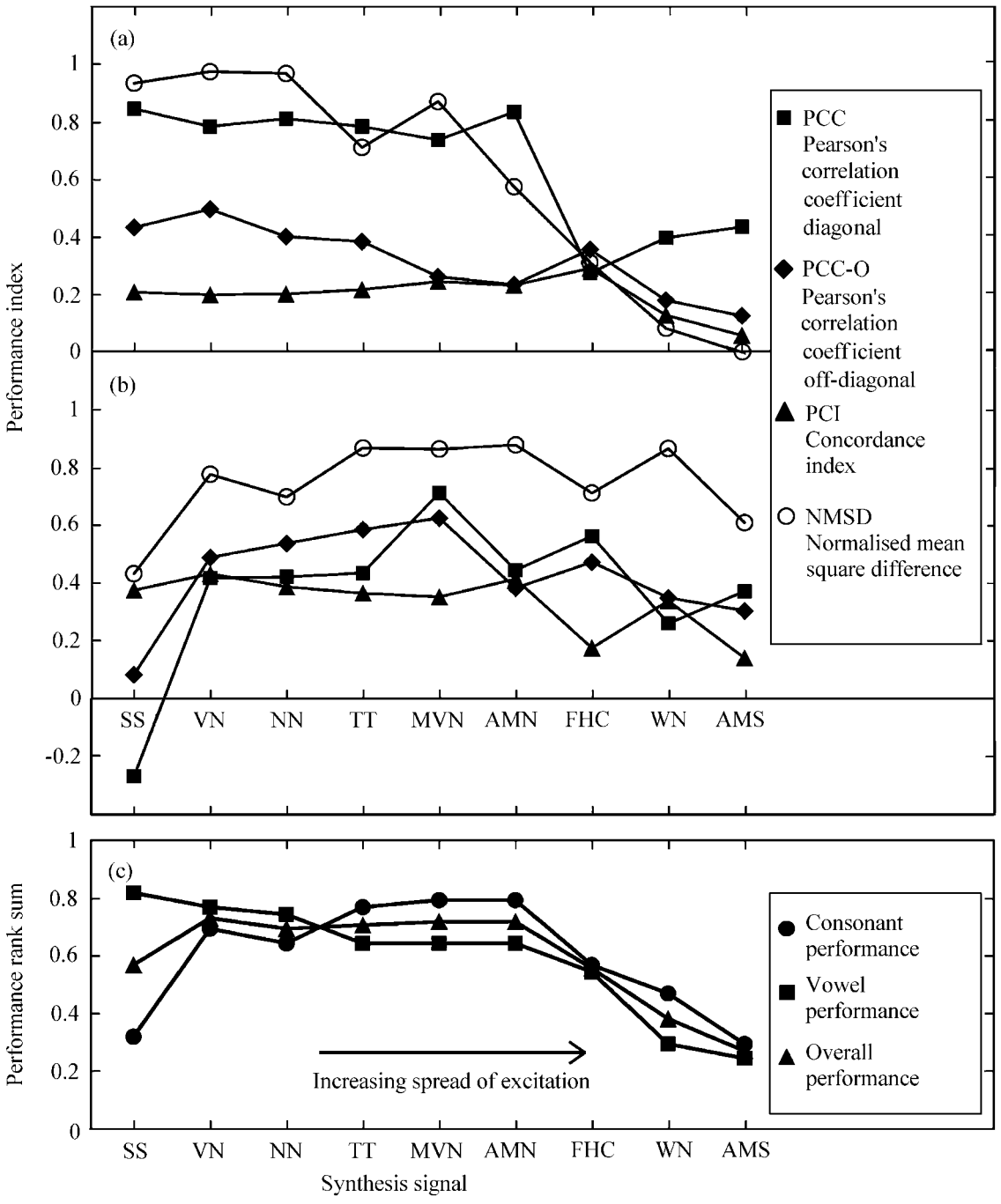

Figure 6.6. Lumped performance measures. (a) Performance indices for vowels. (b) Performance indices for consonants. (c) Normalised performance rank sums of four performance measures.

Table 5. Results from one-way ANOVAs, comparing best performing signal results for consonants with those of CI listeners (left panel), and comparing synthesis signal results (right panel). Significant differences at the 0.05 level are marked with *, whereas significant differences at the 0.001 level are marked with **.

Consonant attributes								
Speech attribute	MVN-CI F(1,13)	TT-CI F(1,13)	AMN-CI F(1,13)		MVN-VN F(1,13)	NN-SS F(1,13)	NN-VN F(1,13)	VN-SS F(1,13)
Voicing	F=2.35 p=0.15	F=7.28 p<0.05*	F=1.76 p=0.21		F=1.95 p=0.19	F=10.25 p<0.01*	F=2.70 p=0.13	F=23.38 p<0.001*
Manner	F=7.91 p<0.05*	F=6.79 p<0.05*	F=9.11 p<0.05*		F=4.53 p=0.06	F=6.46 p<0.05*	F=0.27 p=0.61	F=18.41 p<0.001 **
Place	F=3.93 p=0.07	F=0.00 p=0.99	F=2.20 p=0.16		F=0.20 p=0.66	F=27.24 p<0.001 **	F=1.49 p=0.25	F=22.71 p<0.001 **
Affrica-tion	F=5.73 p<0.05*	F=21.10 p<0.001 *	F=22.81 p<0.001*		F=2.65 p=0.13	F=0.64 p=0.44	F=2.55 p=0.14	F=0.86 p=0.37
Amp. env.	F=0.01 p=0.95	F=9.66 p<0.01*	F=0.22 p=0.65		F=4.48 p=0.06	F=2.44 p=0.14	F=0.86 p=0.37	F=2.22 p=0.16
Burst	F=12.00 p<0.005 *	F=6.75 p<0.05*	F=3.94 p=0.07		F=0.25 p=0.63	F=1.03 p=0.33	F=0.19 p=0.67	F=1.43 p=0.25
Nasality	F=0.14 p=0.71	F=0.04 p=0.84	F=0.37 p=0.56		F=1.26 p=0.28	F=1.83 p=0.20	F=0.43 P=0.52	F=7.18 p<0.05*
Cons recog.	F=5.25 p<0.05*	F=0.89 p=0.36	F=3.10 p=0.10		F=0.73 p=0.41	F=15.87 p<0.005 *	F=1.22 p=0.29	F=26.91 p<0.001 **

Table 6. Results from one-way ANOVAs, comparing best performing signal results for vowels with those of CI listeners (left panel), and comparing synthesis signal results (right panel). Significant differences at the 0.05 level are marked with *, whereas significant differences at the 0.001 level are marked with **.

Vowel attributes								
Speech attribute	SS-CI F(1,11)	VN-CI F(1,11)	NN-CI F(1,11)		MVN-VN F(1,13)	NN-SS F(1,13)	NN-VN F(1,13)	VN-SS F(1,13)
F1	F=0.72 p=0.42	F=0.10 p=0.75	F=0.14 p=0.17		F=1.73 p=0.21	F=0.56 p=0.47	F=0.00 p=0.99	F=0.42 p=0.53
F2	F=2.29 p=0.16	F=1.07 p=0.33	F=0.55 p=0.47		F=1.13 P=0.31	F=2.28 p=0.16	F=0.47 p=0.51	F=0.45 p=0.51
Dura-Tion	F=0.44 p=0.52	F=0.57 p=0.47	F=1.41 p=0.26		F=0.17 P=0.69	F=0.45 p=0.52	F=0.40 p=0.54	F=0.01 p=0.94
Vowel recog.	F=0.00 p=0.95	F=0.01 p=0.92	F=0.66 p=0.44		F=0.04 p=0.84	F=1.21 p=0.29	F=0.55 p=0.47	F=0.04 p=0.84

Performance measures. When acoustic model results are used to model speech intelligibility for vowels and consonants, confusion matrices are usually analysed using information transmission analysis, and statistical significance of differences determined using an ANOVA. If an acoustic model is used to study changes in feature information transmission scores using different signal-processing schemes or other experimental manipulations, NMSD is the most appropriate measure of performance, since it is based on feature information transmission percentages (Eq. 2).

PCC, on the other hand, reflects the relationship between individual scores for phonemes, the average of which yields consonant recognition scores. The FHC signal, for example, has a PCC of 0.6, indicating moderate correlation between CI and normal-hearing listener results for consonant attributes (Figures 6.6b and 6.6c), but has a low intelligibility score for consonant recognition of 53% (Figure 6.4). This indicates that, although relative scores

between the different consonant tokens follow a trend similar to those of CI listeners (indicated by the PCC of 0.6), the actual values are on average lower than those of CI listeners, as indicated by the difference in average scores (53% versus 72%).

Whereas PCC does not consider confusions, PCC-O and PCI both do. While PCC-O is sensitive to the magnitude of deviations from the comparison matrix, it reflects the correlation between individual confusions. Although PCI appears to be the more suitable measure, as it reflects similarity in confusion patterns between two matrices, it assigns 0, -1 or 1 to indicate differences (equal, smaller than or larger than respectively) between corresponding pairs of elements in the two matrices that are compared, and consequently does not reflect the magnitude of these differences. NMSD goes further than any of these measures and reflects feature-based grouping of phoneme confusions (using feature information transmission analysis), making this measure the most appropriate for the present task.

The correspondence between many of these measures for the best performers (with the exception of AMN) is an indication that the best performing synthesis signals perform well from the different viewpoints reflected by the different performance measures. The PCC, PCC-O and concordance index reflect specific confusions occurring for individual phoneme tokens, but do not consider groupings of errors (e.g., phonemes with similar F2 confused, irrespective of F1). This may explain some of the differences between PCI, PCC-O and NMSD trends in general.

Selection of the most appropriate synthesis signal. The present experiment showed that a number of adjustments to an acoustic model could improve correspondence with CI data, which may improve the utility of acoustic models. These adjustments are (i) the careful choice of simulated insertion depth, with the accompanying simulated positioning of electrodes for the synthesis filters, and (ii) the use of an appropriate synthesis signal. If a study involves only vowel intelligibility, the noise-bands with widths of 0.75 mm (NN), sinusoids (SS) and varying noise bands (VN) give good correspondence to CI results. For studies where only consonant intelligibility is measured, the MVN, AMN or TT signals may be used.

For studies where both consonant and vowel intelligibility needs to be measured, the VN, MVN, AMN, TT and NN signals appear best (in that order). Considering the importance of the NMSD measure when using information transmission analysis, the AMN and TT signals are not recommended because of their poor performance for vowel NMSD (Figure 6.6a). Similarly, NN is not recommended because of its poor performance for consonant NMSD (Figure 6.6b). MVN and VN both have satisfactory performance for both vowel and consonant NMSD. Figure 6.4 shows that MVN and VN results differ non-significantly from CI listener results for all vowel attributes. MVN results differ significantly from CI listener results for four consonant attributes (Table 5 and Figure 6.4). VN results also differ from CI consonant results for these four attributes (but more significantly so for affrication and manner of articulation), as well as for voicing and amplitude envelope. Although the VN signal is easier to construct, it does appear that MVN gives better correspondence with CI data when looking at the pattern of statistical differences shown in Figure 6.4.

Implications for CI listeners. Even though some signals were identified as better performers than others, each of the signals had difficulty in modelling some aspects of speech intelligibility. For example, the AMN signal did not model affrication well (Figure 6.5a), but had good performance for consonant attributes and also phoneme attributes taken together (Figure 6.6b and 6c). The prevalent confusions in CI listener results did not correspond well with any of the prevalent confusions of the synthesis signal results. This emphasises that acoustic models can predict confusion categories (as measured through information transmission analysis, as confirmed in this article) when the synthesis signal is judiciously chosen, but they generally do not predict specific confusions. This is generally true and is a fundamental limitation of acoustic models. This does not negate the utility of acoustic models in directing designs or interpreting CI findings, provided that these limitations are acknowledged. Specifically, lack of correspondence between acoustic model outcomes and CI results for particular attributes may be an indication of a modelling deficiency of some aspect of CI perception, which may lead to misinterpretation of results. Also, of course, although there are observed confusion trends among CI listeners, specific confusions vary greatly among these listeners. Models should rightly predict trends in feature information transmission, and not specific confusions.

Other aspects of the present acoustic model (which is representative of acoustic models generally found in literature) may need further development to improve correspondence with CI data for various experimental conditions and performance measures. Finally, correspondence with CI listener data for a wider range of environments (performance should be tested in noise), processing algorithms (e.g. CIS processing) and speech material must be tested to extend the applicability of the present experimental results.

6.5 CONCLUSION

- With the correct modelling choices, acoustic models may predict average trends of phoneme perception observed in CI users. Trends in categories of phoneme confusions may be modelled correctly, but, irrespective of synthesis signal used, acoustic models generally do not predict specific phoneme confusions found in CI listener results. Although this appears to be a fundamental limitation of acoustic models, this does not negate their value.

- Correspondence with CI listener results, using acoustic models of CIs, may be improved for a variety of performance measures by appropriate choice of synthesis signal. The choice of the synthesis signal depends also on the speech material tested, since vowel performance and consonant performance are not predicted best by the same synthesis signal.

- Synthesis signals that give best correspondence with CI results are those that model narrow spread of excitation (best correspondence with vowel perception of CI users) and those that use modulated signals (best correspondence with CI user consonant perception). Synthesis signals VN, MVN and AMN provide the best performance when both vowels and consonants are tested in acoustic simulation studies. Based on a qualitative evaluation of the different performance measures, the MVN signal is recommended.

- The choice of performance measure influences the observed correspondence between CI listener data and normal-hearing listener acoustic model results. The information

transmission analysis-based NMSD performance measure appears to be the most useful choice of performance measure.