The performance of different synthesis signals in acoustic models of cochlear implants

Trudie Strydom and Johan J. Hanekom[a)]
Department of Electrical, Electronic, and Computer Engineering, University of Pretoria, Pretoria 0002, South Africa

(Received 23 November 2009; revised 15 October 2010; accepted 26 October 2010)

Synthesis (carrier) signals in acoustic models embody assumptions about perception of auditory electric stimulation. This study compared speech intelligibility of consonants and vowels processed through a set of nine acoustic models that used Spectral Peak (SPEAK) and Advanced Combination Encoder (ACE)-like speech processing, using synthesis signals which were representative of signals used previously in acoustic models as well as two new ones. Performance of the synthesis signals was determined in terms of correspondence with cochlear implant (CI) listener results for 12 attributes of phoneme perception (consonant and vowel recognition; F1, F2, and duration information transmission for vowels; voicing, manner, place of articulation, affrication, burst, nasality, and amplitude envelope information transmission for consonants) using four measures of performance. Modulated synthesis signals produced the best correspondence with CI consonant intelligibility, while sinusoids, narrow noise bands, and varying noise bands produced the best correspondence with CI vowel intelligibility. The signals that performed best overall (in terms of correspondence with both vowel and consonant attributes) were modulated and unmodulated noise bands of varying bandwidth that corresponded to a linearly varying excitation width of 0.4 mm at the apical to 8 mm at the basal channels.
 [DOI: 10.1121/1.3518760]

PACS number(s): 43.66.Ts, 43.71.Ky, 43.71.Es, 43.66.Lj [MAH] Pages: 920–933

I. INTRODUCTION

Acoustic models are used to investigate aspects of importance for speech intelligibility in general, but specifically for cochlear implant (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, 2005). Although acoustic models have shown relatively good correspondence with best CI listener results in quiet 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 (NH) 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.*, 1997). 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 center 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 study investigated the performance of nine different synthesis signals in terms of correspondence to a selected set of CI listener results.

Dorman *et al.* (1997) studied noise bands and sinusoids in quiet and found hardly any differences between results obtained with these signals. They studied speech intelligibility of Iowa vowels (Tyler *et al.*, 1986), a subset of Hillenbrand's vowels (Hillenbrand *et al.*, 1995), Iowa consonants (Tyler *et al.*, 1986), and Hearing in Noise Test (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 multitalker 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.*,

[a)]Author to whom correspondence should be addressed. Electronic mail: johan.hanekom@up.ac.za

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 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; Boothroyd *et al.*, 1996; Fu and Nogaki, 2005). 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 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 signal-to-noise ratio (SNR) the 6-ERB condition produced a significantly lower score of 68% than 90% for the 3-ERB condition. At −3 dB SNR, these scores dropped to 35% and 72% for the 6-ERB and 3-ERB conditions, 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 250–8000 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 modeling are described in the subsequent texts.

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 center 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 difference limen (DL) and pitch scaling experiments on both NH listeners (using the amplitude-modulated noise bands) and CI listeners, and manipulated the modulation depth and smoothing factor (see Fig. 2) of the modulator signals for the NH 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, Central Institute for the Deaf (CID) and Speech in Noise (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 that 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 modeled the rate of stimulation. Although their study did not consider speech intelligibility, by studying frequency discrimination, interaural 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 and Carlyon (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 study of Deeks and Carlyon verified that their signals gave excitation patterns similar to what is expected from electrical stimulation, using Patterson *et al.*'s (1995) model. 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 3-channel and 6-channel models. At channel 6 the scores were 83% and 71%, and for channel 3 the scores were 45% and 34% at the rates of 140 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 study addresses this issue by investigating vowel and consonant intelligibility for nine different synthesis signals originating from three different sources. First, previously used synthesis signals such as pure tones and noise bands of different widths (Boothroyd *et al.*, 1996; Dorman *et al.*, 1997; Whitmal III *et al.*, 2007), modulated noise bands (Blamey *et al.*, 1984b), and Filtered harmonic complexes (Deeks and Carlyon, 2004) were included. Second, transposed tones (Oxenham *et al.*, 2004), which had previously been used in a psychoacoustic study, were used. Third, new synthesis signals were developed by building on concepts from existing signals. The study compared results from these experiments to CI listener results from a previous study (Pretorius *et al.*, 2006) that used the same speech material to analyze similarities and differences between acoustic model and CI results. The study of Pretorius *et al.* used listeners using either the SPEAK or the ACE speech processing strategy (Pretorius *et al.*, 2006) and, therefore,

the present study 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.

II. METHODS

A. Signal processing

Since the aim of the study 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 more explicitly include implant parameters that may influence the effective number of channels. This includes (1) the use of realistic implant parameters in the model (e.g., using actual interelectrode distances) and (2) modeling current spread through the use of different synthesis filter widths. This approach was followed in the present study, as expanded on below.

The generic signal processing steps are illustrated in Fig. 1. The filtering into contiguous channels was performed using a fast Fourier transform (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 Fig. 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 6 of 20 strategy, whilst listeners using the ACE strategy typically used an 8 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

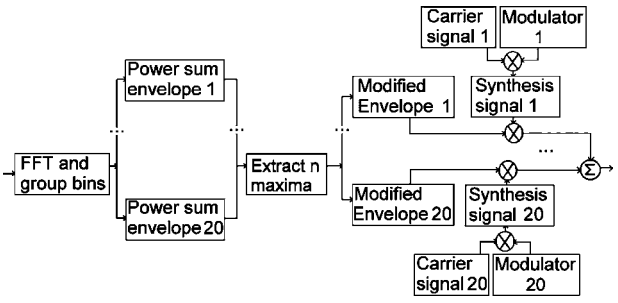

FIG. 1. Signal processing steps. The term modified envelope refers to some channels being set to zero in the SPEAK and ACE strategy when they are not among those containing the spectral peaks. The modulator block is only applicable to the modulated signals.

the design of the synthesis signal. Some aspects that were common to the nine synthesis signals are described below, while Secs. II A 1 and II A 2 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, 2003, 2005), 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 center 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 I.

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

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.*, 2000c).

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).

Sections II A 1 and II A 2 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.

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 the 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

TABLE I. Band pass filters: Analysis filters and synthesis filters.

Filter	Filter −3 dB pass band
Analysis filters	440–565 Hz, 565–690 Hz, 690–815 Hz, 815–940 Hz, 940–1065 Hz, 1065–1190 Hz, 1190–1315 Hz, 1315–1440 Hz, 1440–1690 Hz, 1690–1940 Hz, 1940–2190 Hz, 2190–2565 Hz, 2565–2940 Hz, 2940–3440 Hz, 3440–3940 Hz, 3940–4565 Hz, 4565–5315 Hz, 5315–6190 Hz, 6190–7190 Hz, 7190–7999 Hz
Synthesis signal filters: AMN	595–881 Hz, 666–997 Hz, 751–1126 Hz, 847–1269 Hz, 952–1427 Hz, 1069–1603 Hz, 1199–1797 Hz, 1343–2013 Hz, 1503–2253 Hz, 1680–2519 Hz, 1877–2814 Hz, 2095–3141 Hz, 2337–3504 Hz, 2605–3906 Hz, 2903–4353 Hz, 3233–4848 Hz, 3599–5397 Hz, 4005–6006 Hz, 4455–6682 Hz, 4955–7431 Hz
Synthesis signal filters: AMS, WN, FHC	354–1363 Hz, 409–1528 Hz, 469–1710 Hz, 536–1913 Hz, 610–2138 Hz, 693–2387 Hz, 785–2663 Hz, 886–2970 Hz, 999–3310 Hz, 1124–3687 Hz, 1262–4106 Hz, 1416–4570 Hz, 1586–5085 Hz, 1775–5656 Hz, 1985–6289 Hz, 2218–6992 Hz, 2476–7771 Hz, 2762–8635 Hz, 3080–9594 Hz, 3432–10 668 Hz
Synthesis signal filters: NN	678–769 Hz, 769–868 Hz, 868–979 Hz, 979–1102 Hz, 1102–1238 Hz, 1238–1389 Hz, 1389–1557 Hz, 1557–1743 Hz, 1743–1949 Hz, 1949–2177 Hz, 2177–2431 Hz, 2431–2712 Hz, 2712–3024 Hz, 3024–3370 Hz, 3370–3754 Hz, 3754–4180 Hz, 4180–4652 Hz, 4652–5176 Hz, 5176–5757 Hz, 5757–6401 Hz
Synthesis signal filters: VN, MVN	699–747 Hz, 765–872 Hz, 837–1015 Hz, 915–1177 Hz, 999–1363 Hz, 1089–1574 Hz, 1187–1816 Hz, 1292–2091 Hz, 1405–2404 Hz, 1528–2762 Hz, 1660–3170 Hz, 1802–3635 Hz, 1956–4165 Hz, 2122–4770 Hz, 2301–5459 Hz, 2494–6245 Hz, 2702–7141 Hz, 2927–8163 Hz, 3170–9328 Hz, 3432–10 668 Hz
Synthesis signal filters: SS, TT (center frequencies only)	703 Hz, 797 Hz, 902 Hz, 1019 Hz, 1148 Hz, 1292 Hz, 1451 Hz, 1627 Hz, 1823 Hz, 2040 Hz, 2281 Hz, 2548 Hz, 2844 Hz, 3173 Hz, 3537 Hz, 3942 Hz, 4390 Hz, 4887 Hz, 5439 Hz, 6051 Hz

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.

a. Amplitude modulated noise (AMN). 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 center 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 Fig. 2. With the exception of the modulator, the amplitudes were normalized to a maximum of 0.5 for all signals. The filter cut-off frequencies for the wide-band noise are given in Table I.

b. Short amplitude modulated noise signal (AMS). 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 toward 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 (WN) band signal, which is discussed later. The synthesis signals for channel 1 and channel 9 are shown in Figs. 3(a) and (c), respectively.

c. Transposed tones (TT). TT were used, based on the concepts used in a study by Oxenham *et al.* (2004). The rate of stimulation was modeled 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 study 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 modeled by sine-wave carriers, with frequencies at the center 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

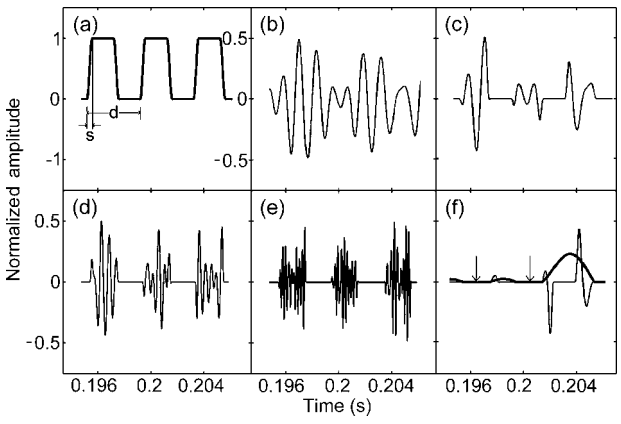

FIG. 2. Modulated wide-band noise synthesis signal (synthesis signal AMN). Only a brief time segment is shown. Signal amplitudes were normalized to a maximum of 0.5. (a) Modulator signal corresponding to the stimulation pulse rate. Smoothing parameter = s/d (0.1 for this signal). (b) Wide-band noise centered 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 centered at 1843 Hz). (e) Synthesis signal for channel 17 (wide-band noise centered 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 Fig. 2(f).

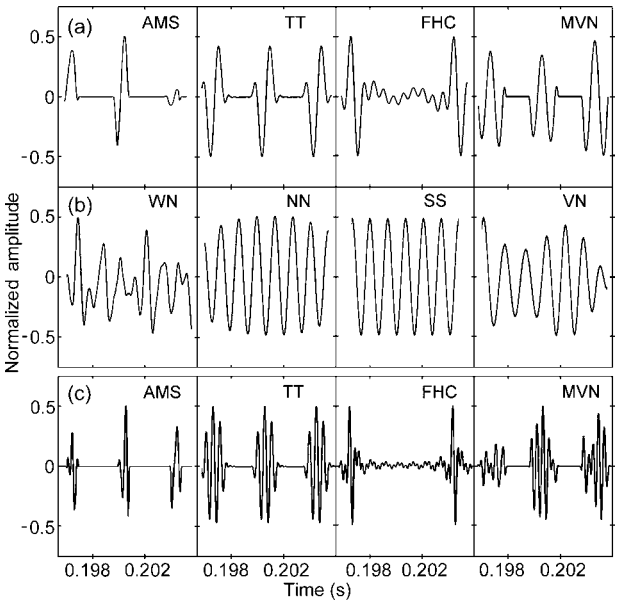

FIG. 3. (a) Modulated synthesis signals for channel 1. (b) Unmodulated synthesis signals for channel 1. (c) Modulated synthesis signals for channel 9. All signals were normalized to give maximum amplitudes of 0.5.

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, 1987). Another approach would be to use multiples of the modulating wave for the carrier, as was done by McKay and Carlyon (1999). The present approach was chosen to ensure that the filter center frequencies remain the same for all conditions for all synthesis signals. The TT synthesis signals for channels 1 and 9 are shown in Fig. 3.

d. Filtered harmonic complexes (FHC). 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 Fig. 3. This signal was constructed based on concepts used in a study by Deeks and Carlyon (2004). A rate of 250 pps was modeled 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, 2500, and 7000 Hz, respectively, in the lowest frequency, mid-frequency, and highest frequency regions, differed from the 2 mm width used in the study of Deeks and Carlyon. 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 NH listener with place of excitation cues, the use of analysis filters above 1089 Hz, combined with a filter width of 2 mm in the study of Deeks and Carlyon, ensured that harmonics of the fundamental frequency were not resolved. The wider filter width of the present study ensured that harmonics of the fundamental frequency were not resolved. The synthesis signals for channels 1 and 9 are shown in Fig. 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 (8×125 Hz).

e. Modulated noise bands with varying width (MVN). 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 modeled as noise bands with varying widths (VN), with modulators similar to the original AMN signal. By using narrower noise bands in the low-frequency region, it was hypothesized that better vowel intelligibility would be realized, whilst 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 the 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 modeled by the WN signal, rather than narrow noise bands. By retaining the relatively narrow widths in the low-frequency region, it was hypothesized that vowel intelligibility would not suffer, and that the broadening of filters in the high-frequency regions would lower consonant recognition. It was hypothesized 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 Eq. (1),

$$\text{width}\,(i) = 0.4 + 7.6\,[(i-1)/19], \tag{1}$$

where i denotes the number of the filter and $i = 1$ is the most apical filter. Figure 3 shows the typical synthesis signal for channels 1 and 9.

2. Unmodulated synthesis signals

Unmodulated synthesis signals exclude modeling of the stimulation rate. In Fig. 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 Fig. 1.

a. Wide noise bands (WN). An unmodulated noise band is used as synthesis signal. The basilar membrane of length 35 mm was divided into four roughly equal portions of 8 mm each, following results from the study of Fu and Nogaki (2005), where CI listeners were found to have effectively 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 *et al.* (1996) study. The synthesis signal for channel 1 is shown in Fig. 3. The synthesis filters were designed using third order Butterworth filters with a width of 8 mm at the −3 dB point.

b. Narrow noise bands (NN). Narrow noise bands with filter widths 0.75 mm, corresponding to electrodes spaced 0.75 mm, were used. 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 Fig. 3.

c. Sinusoidal signals (SS). (Dorman et al., 1997). SS were constructed with frequencies equal to the center frequencies of the analysis filters, and with root-mean-square (rms) level the same as that in the original envelope. The synthesis signal for channel 1 is shown in Fig. 3.

d. Noise bands with varying width (VN). 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 Fig. 3.

B. Listeners

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

C. 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 study by Pretorius *et al.* (2006). The original speech material was processed by the acoustic model, and nine different versions were created using the nine synthesis signals.

D. Procedure

Experiments were conducted in a double-walled sound booth. Processed speech material was presented in sound field using a personal computer (PC) with an external sound card (M-Audio Fasttrack Pro, Avid Technology Inc, MA) and a Yamaha MS101 II loudspeaker (Yamaha Electronics Corporation, CA). Listeners could adjust the volume to comfortable levels [found to range between around 60 and 70 dB sound pressure level (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 customized 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 randomized. 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 study 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. Twelve repetitions of each processed phoneme were presented in random order to the listeners (four repetitions of each phoneme synthesized using three synthesis signals). This was repeated four times, so that there were four consecutive sets of 12 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 study was conducted several months after the original experiment. Thus, learning effects from the original study would be minimal. Loudness was fixed at 65 dB SPL during this control study.

E. 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 modeling 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., Fu and Shannon, 1998; Friesen *et al.*, 2001). 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 NH listeners. The present study, 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 modeled CI listener perception. Each performance measure used emphasized 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 study 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 NH results were obtained by using differences in average scores for each of the attributes considered to characterize 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 characterize 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 characterize consonant intelligibility. The square of differences (SD),

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

and means of these squares of differences (MSD),

$$\mathrm{MSD}(k,j) = \frac{1}{n_k}\sum_{i=1}^{k_1}\mathrm{SD}(i,j) \tag{2b}$$

were obtained, with $\mathrm{IT}(i,j)$ 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, $\mathrm{IT}_{\mathrm{CI}}(i)$ the average information transmission score measured for CI listeners for phoneme attribute i and $\mathrm{SD}(i,j)$ the square of differences for attribute i using synthesis signal j. $\mathrm{MSD}(k,j)$ 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,

$$\mathrm{NSD}(i,j) = 1 - \frac{\mathrm{SD}(i,j)}{\mathrm{NF}_1}, \tag{2c}$$

$$\mathrm{NMSD}(k,j) = 1 - \frac{\mathrm{MSD}(k,j)}{\mathrm{NF}_2}, \tag{2d}$$

where $\mathrm{NSD}(i,j)$ is a normalized performance measure for each phoneme attribute i when the synthesis signal is j. $\mathrm{NMSD}(k,j)$ is the normalized mean performance measure for synthesis signal j, for the lumped measure k. NF_1 and NF_2 are normalization factors, found from the maximum of all SD and MSD values, respectively, to ensure that the normalized square difference (NSD) and normalized means of these squares of differences (NMSD) scores are normalized 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 by 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 NH 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 NH listener results, the average of which are usually reported as the vowel and consonant recognition scores.

The fourth performance measure was the PCC 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 NH 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 2–10 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 normalized 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 NH listeners listening to each version of the acoustic model were examined to determine whether similar confusions were present in both.

III. RESULTS

The primary objective of the study was to determine which synthesis signal gave the best performance in terms of

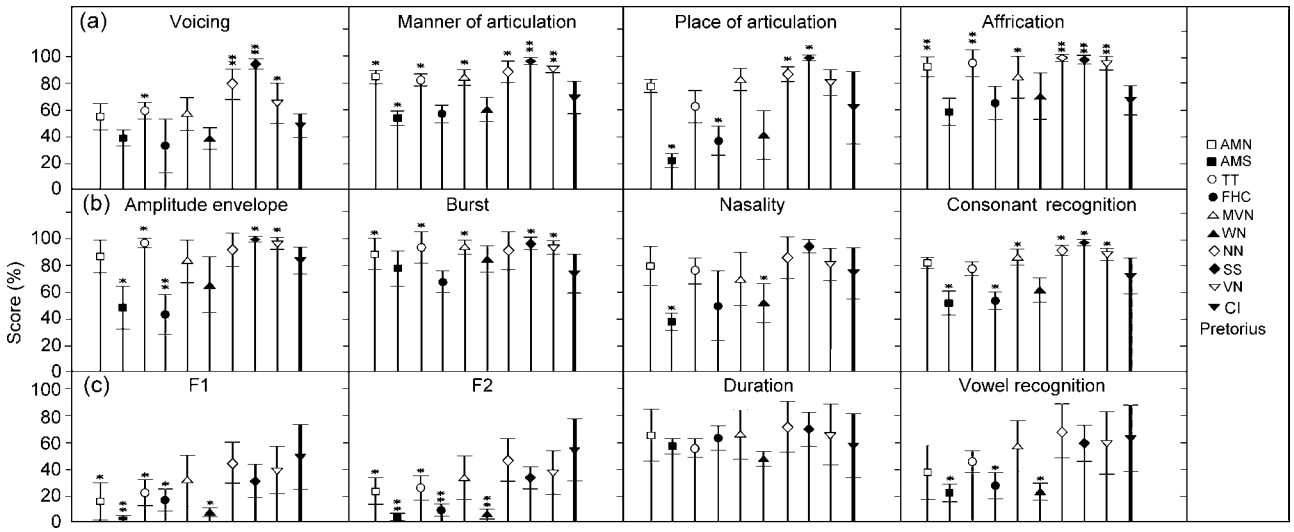

FIG. 4. (a) and (b) Consonant feature information transmission scores and consonant recognition percentage correct. (c) Vowel feature information transmission scores and vowel recognition percentage correct. Error bars indicate ±1 SD. Results from CI listener study are indicated using bold lines. * indicates significant difference from the CI listener results (Pretorius *et al.*, 2006) at the 0.05 level, where ** indicates significant difference at the 0.001 level.

correspondence with CI listener results. Where the term "performance" is used, this denotes correspondence with CI data using the four different performance measures (Figs. 5 and 6), whereas the term "intelligibility" refers to the phoneme intelligibility scores (Fig. 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 [Figs. 5(b) and 6(b)], although their intelligibility is high (Fig. 4).

Figure 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.*'s study (2006) are also displayed.

Figure 5 shows the NSDs [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 Fig. 6. Figures 6(a) and (b) 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(c) 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.

The three best performing synthesis signals' results were compared with CI listener results for each consonant and vowel attribute using one-way analysis of variances (ANOVAs). The observation that intelligibility of consonant attributes appeared to benefit from synthesis signals with narrow spread of excitation [Figs. 4(a) 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

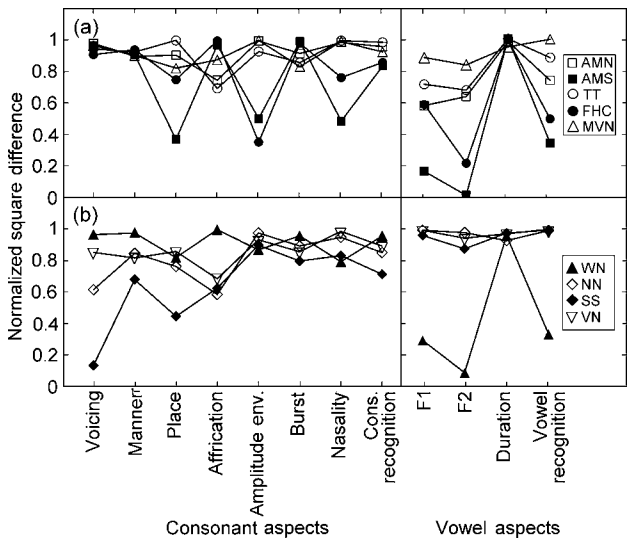

FIG. 5. Performance of different synthesis signals for individual attributes using NSD scores [in Eq. (2c)]. (a) NSD for modulated synthesis signals. (b) NSD for unmodulated synthesis signals.

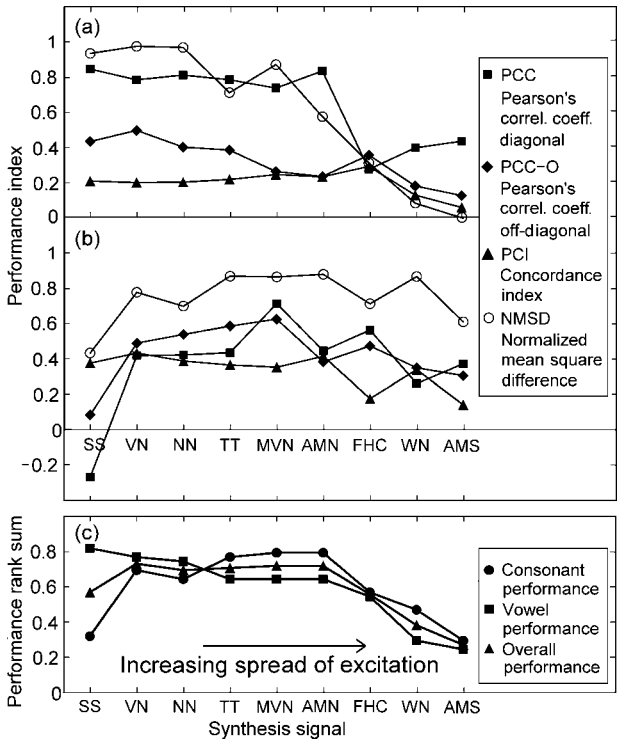

FIG. 6. Lumped performance measures. (a) Performance indices for vowels. (b) Performance indices for consonants. (c) Normalized performance rank sums of four performance measures.

only, namely the use of a modulator signal. Table II shows 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 II 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. The table 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.

A. 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 may possibly affect interpretation of results. However, no effects of learning was 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 study 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 study 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.

IV. DISCUSSION

A. Learning effects

Although learning effects may play a role in results, as illustrated by Rosen *et al.* (1999), acoustic modeling studies have in general not been consistent in their approach to possible learning effects. Many acoustic model studies provided no training, but relied on randomization of test conditions to eliminate learning effects (Deeks and Carlyon, 2004; Fu and Nogaki, 2005; Baskent, 2006; Baskent and Shannon, 2007; Verschuur, 2009). Other studies relied on the experience of the listeners (Loizou *et al.*, 2000a), some used moderate training of around 1 h or less (Loizou *et al.*, 2000b; Green *et al.*, 2004; Stickney *et al.*, 2004; Bingabr *et al.*, 2008), whereas still others allowed extensive training of 3 h or more, or used some measure to ensure that performance had stabilized (Souza and Boike, 2006; Throckmorton *et al.*, 2006; Xu and Zheng, 2007). The analysis of the original study results into three sets of learning events, as well as analysis of the control study results, confirmed that learning effects were not important during the present study, 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.

TABLE II. Results from one-way ANOVAs, comparing best performing signal results with those of CI listeners (left panel), and comparing synthesis signal results (right panel).

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	$F=2.35$	$F=7.28$	$F=1.76$	$F=1.95$	$F=10.25$	$F=2.70$	$F=23.38$
	$p=0.15$	$p<0.05$*	$p=0.21$	$p=0.19$	$p<0.01$*	$p=0.13$	$p<0.001$*
Manner	$F=7.91$	$F=6.79$	$F=9.11$	$F=4.53$	$F=6.46$	$F=0.27$	$F=18.41$
	$p<0.05$*	$p<0.05$*	$p<0.05$*	$p=0.06$	$p<0.05$*	$p=0.61$	$p<0.001$**
Place	$F=3.93$	$F=0.00$	$F=2.20$	$F=0.20$	$F=27.24$	$F=1.49$	$F=22.71$
	$p=0.07$	$p=0.99$	$p=0.16$	$p=0.66$	$p<0.001$**	$p=0.25$	$p<0.001$**
Affrication	$F=5.73$	$F=21.10$	$F=22.81$	$F=2.65$	$F=0.64$	$F=2.55$	$F=0.86$
	$p<0.05$*	$p<0.001$*	$p<0.001$*	$p=0.13$	$p=0.44$	$p=0.14$	$p=0.37$
Amplitude envelope	$F=0.01$	$F=9.66$	$F=0.22$	$F=4.48$	$F=2.44$	$F=0.86$	$F=2.22$
	$p=0.95$	$p<0.01$*	$p=0.65$	$p=0.06$	$p=0.14$	$p=0.37$	$p=0.16$
Burst	$F=12.00$	$F=6.75$	$F=3.94$	$F=0.25$	$F=1.03$	$F=0.19$	$F=1.43$
	$p<0.005$*	$p<0.05$*	$p=0.07$	$p=0.63$	$p=0.33$	$p=0.67$	$p=0.25$
Nasality	$F=0.14$	$F=0.04$	$F=0.37$	$F=1.26$	$F=1.83$	$F=0.43$	$F=7.18$
	$p=0.71$	$p=0.84$	$p=0.56$	$p=0.28$	$p=0.20$	$p=0.52$	$p<0.05$*
Consonant recognition	$F=5.25$	$F=0.89$	$F=3.10$	$F=0.73$	$F=15.87$	$F=1.22$	$F=26.91$
	$p<0.05$*	$p=0.36$	$p=0.10$	$p=0.41$	$p<0.005$*	$p=0.29$	$p<0.001$**

Vowel attributes							
	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	$F=0.72$	$F=0.10$	$F=0.14$	$F=1.73$	$F=0.56$	$F=0.00$	$F=0.42$
	$p=0.42$	$p=0.75$	$p=0.17$	$p=0.21$	$p=0.47$	$p=0.99$	$p=0.53$
F2	$F=2.29$	$F=1.07$	$F=0.55$	$F=1.13$	$F=2.28$	$F=0.47$	$F=0.45$
	$p=0.16$	$p=0.33$	$p=0.47$	$p=0.31$	$p=0.16$	$p=0.51$	$p=0.51$
Duration	$F=0.44$	$F=0.57$	$F=1.41$	$F=0.17$	$F=0.45$	$F=0.40$	$F=0.01$
	$p=0.52$	$p=0.47$	$p=0.26$	$p=0.69$	$p=0.52$	$p=0.54$	$p=0.94$
Vowel recognition	$F=0.00$	$F=0.01$	$F=0.66$	$F=0.04$	$F=1.21$	$F=0.55$	$F=0.04$
	$p=0.95$	$p=0.92$	$p=0.44$	$p=0.84$	$p=0.29$	$p=0.47$	$p=0.84$

*Significant differences at the 0.05 level.
**Significant differences at the 0.001 level.

B. Performance and intelligibility

The terms performance and intelligibility were defined earlier (see Sec. III). Figures 5 and 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 modeled 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 (carrier signal) to 2 mm (synthesis signal) for the MVN signal. Figures 6(a)–(c) 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.

C. 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 II) 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.* (1997), 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 study were still relatively low and may be explained by the modeled 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 NH listeners. Baskent *et al.* (2005) found decreases of 17% in vowel intelligibility for CI listeners when insertion depths were reduced from 28 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.

D. Consonant performance and intelligibility

The best performing synthesis signals for consonant attributes were the MVN, AMN, and TT signals [Figs. 6(b) and (c)]. The AMN and MVN signals have widening spread of excitation toward the higher-frequency regions, as indicated in Table I. 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 Fig. 4 and Table II. The Whitmall III *et al.*'s study (2007) showed a similar trend, with the sinusoids yielding better scores than noise–carriers in both quiet and noise. Table II 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 toward 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 toward the basal region may therefore be the key to the good performance, as they may be seen as models of the current spread increasing toward 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.*'s (1998) study, 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 *et al.*, 1999; Gantz and Turner, 2003) for some listeners, and trends of increasing thresholds toward the higher frequencies for listeners with hearing loss (e.g., Baskent, 2006) suggest that degeneration of peripheral axonal processes of nerve fibers 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 toward 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 Fig. 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 modeling 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 differ 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 effects of temporal sampling rate, as discussed in the following sections.

E. 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 Table II. 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 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, 2000; Holden *et al.*, 2002) and significant effects of stimulation rate (e.g., Kiefer *et al.*, 1997; Loizou *et al.*, 2000c; Frijns *et al.*, 2003; Buechner *et al.*, 2006). 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.

F. 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 II.

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(c) 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 study) models the perception of phonemes optimally for both consonants and vowels.

G. Performance measures

When acoustic model results are used to model speech intelligibility for vowels and consonants, confusion matrices are usually analyzed 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 yield consonant recognition scores. The FHC signal, for example, has a PCC of 0.6, indicating moderate correlation between CI and NH listener results for consonant attributes [Figs. 6(b) and (c)], but has a low intelligibility score for consonant recognition of 53% (Fig. 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.

H. Selection of the most appropriate synthesis signal

The present study 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 (1) the careful choice of simulated insertion depth, with the accompanying simulated positioning of electrodes for the synthesis filters and (2) the use of an appropriate synthesis signal. If a study involves only vowel intelligibility, the noise bands with widths of 0.75 mm NN, 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 due to their poor performance for vowel NMSD [Fig. 6(a)]. Similarly, NN is not recommended due to its poor performance for consonant NMSD [Fig. 6(b)]. MVN and VN both have satisfactory performance for both vowel and consonant NMSD. Figure 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 II and Fig. 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 Fig. 4.

I. Implications for CI listeners

Even though some signals were identified as better performers than others, each of the signals had difficulty in modeling some aspects of speech intelligibility. For example, the AMN signal did not model affrication well [Fig. 5(a)], but had good performance for consonant attributes and also phoneme attributes taken together [Figs. 6(b) and (c)]. The prevalent confusions in CI listener results did not correspond well with any of the prevalent confusions of the synthesis signal results. This emphasizes 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 modeling 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 study results.

V. CONCLUSIONS

(1) With the correct modeling choices, acoustic models may predict average trends of phoneme perception observed in CI users. Trends in categories of phoneme confusions may be modeled 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.
(2) 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.
(3) 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.
(4) The choice of performance measure influences the observed correspondence between CI listener data and NH listener acoustic model results. The information transmission analysis-based NMSD performance measure appears to be the most useful choice of performance measure.

ACKNOWLEDGMENTS

This study was made possible in part by grants from the National Research Foundation of South Africa. The authors would also like to thank listeners for their commitment and time for the experiments.

Baer, T., and Moore, B. C. J. (**1993**). "Effects of spectral smearing on the intelligibility of sentences in noise," J. Acoust. Soc. Am. **94**, 1229–1241.
Baskent, D. (**2006**). "Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels," J. Acoust. Soc. Am. **120**, 2908–2925.
Baskent, D., and Shannon, R. V. (**2003**). "Speech recognition under conditions of frequency-place compression and expansion," J. Acoust. Soc. Am. **113**, 2064–2076.
Baskent, D., and Shannon, R. V. (**2005**). "Interactions between cochlear implant electrode insertion depth and frequency-place mapping," J. Acoust. Soc. Am. **117**, 1405–1416.
Baskent, D., and Shannon, R. V. (**2007**). "Combined effects of frequency compression–expansion and shift on speech recognition," Ear Hear. **28**, 277–289.
Baumann, U., and Nobbe, A. (**2006**). "The cochlear implant electrode–pitch function," Hear. Res. **213**, 34–42.
Bingabr, M., Espinoza-Varas, B., and Loizou, P. C. (**2008**). "Simulating the effect of spread of excitation in cochlear implants," Hear. Res. **241**, 73–79.
Blamey, P. J., Dowell, R. C., Tong, Y. C., Brown, A. M., Luscombe, S. M., and Clark, G. M. (**1984a**). "Speech processing studies using an acoustic model of a multiple-channel cochlear implant," J. Acoust. Soc. Am. **76**, 104–110.
Blamey, P. J., Dowell, R. C., Tong, Y. C., and Clark, G. M. (**1984b**). "An acoustic model of a multiple-channel cochlear implant," J. Acoust. Soc. Am. **76**, 97–103.
Boex, C., Baud, L., Cosendai, G., Sigrist, A., Kos, M. I., and Pellizone, M. (**2006**). "Acoustic to electric pitch comparisons in cochlear implant subjects with residual hearing," J. Assoc. Res. Otol. **7**, 110–124.
Boothroyd, A., Mulhearn, B., Gong, J., and Ostroff, J. (**1996**). "Effects of spectral smearing on phoneme and word recognition," J. Acoust. Soc. Am. **100**, 1807–1818.
Brusco, M. J. (**2004**). "On the concordance among empirical confusion matrices for visual and tactual letter recognition," Percept. Psychophys. **3**, 392–397.
Buechner, A., Frohne-Buechner, C., Gaertner, L., Lesinski-Schiedat, A., Battmer, R. D., and Lenarz, T. (**2006**). "Evaluation of Advanced Bionics high resolution mode," Int. J. Audiol. **45**, 407–416.
Deeks, J. M., and Carlyon, R. P. (**2004**). "Simulations of cochlear implant hearing using filtered harmonic complexes: Implications for concurrent speech segregation," J. Acoust. Soc. Am. **115**, 1736–1746.
Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (**1998**). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," J. Acoust. Soc. Am. **104**, 3583–3585.
Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.
Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.
Frijns, J. H. M., Klop, W. M. C., Bonnet, R. M., and Briaire, J. J. (**2003**). "Optimizing the number of electrodes with high-rate stimulation of the Clarion CII cochlear implant," Acta Oto-Laryngol. **123**, 138–142.
Fu, Q. J., and Nogaki, G. (**2005**). "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," J. Assoc. Res. Otol. **6**, 19–27.
Fu, Q. J., and Shannon, R. V. (**1998**). "Effects of amplitude nonlinearity on phoneme recognition by cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **104**, 2570–2577.
Fu, Q. J., and Shannon, R. V. (**2000**). "Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners," J. Acoust. Soc. Am. **107**, 589–587.
Gantz, B., and Turner, C. W. (**2003**). "Combining acoustic and electric hearing," Laryngoscope **113**, 1726–1730.
Geurts, L., and Wouters, J. (**2000**). "A concept for a research tool for experiments with cochlear implant users," J. Acoust. Soc. Am. **108**, 2949–2956.
Green, T., Faulkner, A., and Rosen, S. (**2004**). "Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants," J. Acoust. Soc. Am. **116**, 2298–2310.
Hanekom, T. (**2001**). "Three-dimensional spiraling finite element model of the electrically stimulated cochlea," Ear Hear. **22**, 300–315.
Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Holden, L. K., Skinner, M. W., Holden, T. A., and Demorest, M. E. (**2002**). "Effects of stimulation rate with the Nucleus 24 ACE speech coding strategy," Ear Hear. **23**, 463.
Kiefer, J., Ilberg, C., Rupprecht, V., Hubnet-Egener, J., Baumgartner, W., Gstottner, W., Forgasi, K., and Stephan, K. (**1997**). "Optimized speech

understanding with the CIS speech coding strategy in cochlear implants: The effect of variations in stimulus rate and number of channels," *Vth International Cochlear Implant Conference*, New York, NY, pp. 1009–1020.
Kral, A., Hartmann, R., Mortazavi, D., and Klinke, R. (**1998**). "Spatial resolution of cochlear implants: The electrical field and excitation of auditory afferents," Hear. Res. **121**, 11–28.
Loizou, P. C., Dorman, M., and Fitzke, J. (**2000a**). "The effect of reduced dynamic range on speech understanding: Implications for patients with cochlear implants," Ear Hear. **21**, 25–31.
Loizou, P. C., Dorman, M., Poroy, O., and Spahr, T. (**2000b**). "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," J. Acoust. Soc. Am. **108**, 2377–2386.
Loizou, P. C., Poroy, O., and Dorman, M. (**2000c**). "The effect of parametric variations of cochlear implant processors on speech understanding," J. Acoust. Soc. Am. **108**, 790–802.
McKay, C. M., and Carlyon, R. P. (**1999**). "Dual temporal pitch percepts from acoustic and electric amplitude-modulated pulse trains," J. Acoust. Soc. Am. **105**, 347–357.
Meddis, R., and O'Mard, L. (**1997**). "A unitary model of pitch perception," J. Acoust. Soc. Am. **102**, 1811–1820.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Oxenham, A. J., Bernstein, J. G. W., and Penagos, H. (**2004**). "Correct tonotopic representation is necessary for complex pitch perception," Proc. Natl. Acad. Sci. U.S.A. **101**, 1421–1425.
Patterson, R. D., Allerhand, M. H., and Giguère, C. (**1995**). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," J. Acoust. Soc. Am. **98**, 1890–1894.
Pretorius, L. L., Hanekom, J. J., Van Wieringen, A., and Wouters, J. (**2006**). "′n Analitiese tegniek om die foneem-herkenningsvermoë van Suid-Afrikaanse kogleêre inplantingsgebruikers te bepaal" ("Analytical technique to determine the phoneme-recognition ability of South African cochlear implant users"), Die Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie **25**, 195–207.
Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629.
Rubinstein, J. T., and Hong, R. (**2003**). "Signal coding in cochlear implants: Exploiting stochastic effects of electrical stimulation," Ann. Otol. Rhinol. Laryngol. Suppl. **191**, 14–19.
Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Skinner, M. W., Holden, L. K., Whitford, L. A., Plant, K. L., Psarros, C., and Holden, T. A. (**2002**). "Speech recognition with the Nucleus 24 SPEAK, ACE, and CIS speech coding strategies in newly implanted adults," Ear Hear. **23**, 207–223.
Souza, P. E., and Boike, K. T. (**2006**). "Combining temporal-envelope cues across channels: Effects of age and hearing loss," J. Speech. Lang. Hear. Res. **49**, 138–149.
Stickney, G. S., Zeng, F. G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Throckmorton, C. S., and Collins, L. M. (**2002**). "The effect of channel interactions on speech recognition in cochlear implant subjects: Predictions from an acoustic model," J. Acoust. Soc. Am. **112**, 285–296.
Throckmorton, C. S., Selin Kucukoglu, M., Remus, J. J., and Collins, L. M. (**2006**). "Acoustic model investigation of a multiple carrier frequency algorithm for encoding fine frequency structure: Implications for cochlear implants," Hear. Res. **218**, 30–42.
Tyler, R. S., Preece, J., and Tye-Murray, M. (**1986**). "The Iowa audiovisual speech perception laser videodisc," in *Laser Videodisc and Laboratory Report* (Department of Otolaryngology, Head and Neck Surgery, University of Iowa Hospital and Clinics, Iowa City, IA).
van den Honert, C., and Stypulkowski, P. H. (**1987**). "Temporal response patterns of single auditory nerve fibers elicited by periodic electrical stimuli," Hear. Res. **29**, 207–222.
van Wieringen, A., and Wouters, J. (**1999**). "Natural vowel and consonant recognition by Laura cochlear implantees," Ear Hear. **20**, 89–103.
Verschuur, C. (**2007**). "Acoustic models of consonant recognition in cochlear implant users," Doctoral thesis, Faculty of Engineering, Science and Mathematics, University of Southampton.
Verschuur, C. (**2009**). "Modeling the effect of channel number and interaction on consonant recognition in a cochlear implant peak-picking strategy," J. Acoust. Soc. Am. **125**, 1723–1736.
von Ilberg, C., Frankfurt, U., and für Hals-Nasen-Ohrenheilkunde, K. (**1999**). "Electric-acoustic stimulation of the auditory system new technology for severe hearing loss," Logo **61**, 334–340.
Whitmal, N. A. III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.
Xu, L., and Zheng, Y. (**2007**). "Spectral and temporal cues for phoneme recognition in noise," J. Acoust. Soc. Am. **122**, 1758–1764.
Zeng, F. G., Grant, G., Niparko, J., Galvin, J., Shannon, R., Opie, J., and Segel, P. (**2002**). "Speech dynamic range and its effect on cochlear implant performance," J. Acoust. Soc. Am. **111**, 377–386.