EURASIP Journal on Applied Signal Processing 2005:18, 2979–2990

The Effects of Noise on Speech Recognition in Cochlear Implant Subjects: Predictions and Analysis Using Acoustic Models

Jeremiah J. Remus

Department of Electrical & Computer Engineering, Pratt School of Engineering, Duke University, P.O. Box 90291, Durham, NC 27708-0291, USA
Email: jeremiah.remus@duke.edu

Leslie M. Collins

Department of Electrical & Computer Engineering, Pratt School of Engineering, Duke University, P.O. Box 90291, Durham, NC 27708-0291, USA
Email: lcollins@ee.duke.edu

Received 1 May 2004; Revised 30 September 2004

Cochlear implants can provide partial restoration of hearing, even with limited spectral resolution and loss of fine temporal structure, to severely deafened individuals. Studies have indicated that background noise has significant deleterious effects on the speech recognition performance of cochlear implant patients. This study investigates the effects of noise on speech recognition using acoustic models of two cochlear implant speech processors and several predictive signal-processing-based analyses. The results of a listening test for vowel and consonant recognition in noise are presented and analyzed using the rate of phonemic feature transmission for each acoustic model. Three methods for predicting patterns of consonant and vowel confusion that are based on signal processing techniques calculating a quantitative difference between speech tokens are developed and tested using the listening test results. Results of the listening test and confusion predictions are discussed in terms of comparisons between acoustic models and confusion prediction performance.

Keywords and phrases: speech perception, confusion prediction, acoustic model, cochlear implant.

1. INTRODUCTION

The purpose of a cochlear implant is to restore some degree of hearing to a severely deafened individual. Among individuals receiving cochlear implants, speech recognition performance varies, but studies have shown that a high level of speech understanding is achievable by individuals with successful implantations. The speech recognition performance of individuals with cochlear implants is measured through listening tests conducted in controlled laboratory settings, which are not representative of the typical conditions in which the devices are used by the individuals in daily life. Numerous studies have indicated that a cochlear implant patient's ability to understand speech effectively is particularly susceptible to noise [1, 2, 3]. This is likely due to a variety of factors, such as limited spectral resolution, loss of fine temporal structure, and impaired sound-localization abilities.

The manner and extent to which noise affects cochlear implantee's speech recognition can depend on individual characteristics of the patient, the cochlear implant device, and the structure of the noise and speech signals. Not all of these relationships are well understood. It is generally presumed that increasing the level of noise will have a negative effect on speech recognition. However, the magnitude and manner in which speech recognition is affected is more ambiguous. Particular speech processing strategies may be more resistant to the effects of certain types of noise, or noise in general. Other devices parameters, such as the number of channels, number of stimulation levels, and compression mapping algorithms, have also been shown to influence how speech recognition will be affected by noise [4, 5, 6]. The effects of noise also depend on the type of speech materials and the linguistic knowledge of the listener. With all of these interdependent factors, the relationship between noise and speech recognition is quite complex and requires careful study.

The goals of this study were to analyze and predict the effects of noise on speech processed by two acoustic models of cochlear implant speech processors. The listening test was conducted to examine the effects of noise on speech

recognition scores using a complete range of noise levels. Information transmission analysis was performed to illustrate the results of the listening test and to verify assumptions regarding the acoustic models. The confusion prediction methods were developed to investigate whether a signal processing algorithm would predict patterns of token confusion similar to those seen in the listening test. The use of the similarities and differences between speech tokens for prediction of speech recognition and intelligibility has a basis in previous studies. Müsch and Buus [7, 8] used statistical decision theory to predict speech intelligibility by calculating the correlation between variations of orthogonal templates of speech tokens. A mathematical model developed by Svirsky [9] used the ratio of frequency-channel amplitudes to locate phonemes in a multidimensional perceptual space. A study by Leijon [10] used hidden Markov models to approximate the rate of information transmitted through a given acoustic environment, such as a person with a hearing aid.

The motivation for estimating trends in token confusions and overall confusion rate, based solely on information in the processed speech signal, is to enable preliminary analysis of speech materials prior to conducting listening tests. Additionally, a method that estimates token confusions and overall confusion rate would have applications in the development of speech processing methods and noise mitigation techniques. Sets of processed speech tokens that are readily distinguishable by the confusion prediction method should also be readily distinguishable by cochlear implantees, if the prediction method is well conceived and robust.

The rest of this paper is organized as follows. Section 2 discusses the listening test conducted in this study. The experimental methods using normal-hearing subjects and the information transmission analysis of vowel and consonant confusions are detailed. Results, in the form of speech recognition scores and information transmission analyses, are provided and discussed. Section 3 describes the methods and results of the vowel and consonant confusion predictions developed using signal processing techniques. The methods of speech signal representation and prediction metric calculation are described, and potential variations are addressed. Results are presented to gauge the overall accuracy of the investigated confusion prediction methods for vowels and consonants processed with each of the two acoustic models.

2. LISTENING TEST

The listening test measured normal-hearing subjects' abilities to recognize noisy vowel and consonant tokens processed by two acoustic models. Using acoustic models to test normal-hearing subjects for cochlear implant research is a widely used and well-accepted method for collecting experimental data. Normal-hearing subjects provide a number of advantages: they are more numerous and easier to recruit, the experimental setups tend to be less involved, and there are not subject variables, such as experience with cochlear implant device, type of implanted device, cause of deafness, and quality of implantation, that affect individual patient's performance. Results of listening tests using normal-hearing subjects are often only indicative of trends in cochlear implant patient's performance; absolute levels of performance tend to disagree [1, 11]. There are several sources of discrepancies between the performance of cochlear implant subjects and normal-hearing subjects using acoustic models, such as experience with the device, acclimation to spectrally quantized speech, and the idealistic rate of speech information transmission through the acoustic model. However, acoustic models are still an essential tool for cochlear implant research. Their use is validated by numerous studies where cochlear implant patient's results were successfully verified and by the flexibility they provide in testing potential speech processing strategies [12, 13].

Subjects

Twelve normal-hearing subjects were recruited to participate in a listening test using two acoustic models for vowel and consonant materials in noise. Prior to the listening tests, subjects' audiograms were measured to evaluate thresholds at 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz to confirm normal hearing, defined in this study as thresholds within two standard deviations of the subject group's mean. Subjects were paid for their participation. The protocol and implementation of this experiment were approved by the Duke University Institutional Review Board (IRB).

Speech materials

Vowel and consonant tokens were taken from the Revised Cochlear Implant Test Battery [14]. The vowel tokens used in the listening test were {had, hawed, head, heard, heed, hid, hood, hud, who'd}. The consonants tested were {b, d, f, g, j, k, m, n, p, s, sh, t, v, z} presented in /aCa/ context. The listening test was conducted at nine signal-to-noise ratios: quiet, +10 dB, +8 dB, +6 dB, +4 dB, +2 dB, +1 dB, 0 dB, and −2 dB. Pilot studies and previous studies in the literature [3, 5, 15, 16] indicated that this range of SNRs would provide a survey of speech recognition ability over the range of scores from nearly perfect correct identification to performance on par with random guessing. Speech-shaped noise, that is, random noise with a frequency spectrum that matches the average long-term spectrum of speech, is added to the speech signal prior to acoustic model processing.

Signal processing

This experiment made use of two acoustic models implemented by Throckmorton and Collins [17], based on acoustic models developed in [18, 19]. The models will be referred to as the 8F model and the 6/20F model, named for the number of presentation and analysis channels. A block diagram of the general processing common to both acoustic models is shown in Figure 1. With each model, the incoming speech is prefiltered using a first-order highpass filter with a 1 kHz cutoff frequency, to equalize the spectrum of the incoming signal. It is then passed through a 6th-order antialiasing Butterworth lowpass filter with an 11 kHz cutoff. Next, the filterbank separates the speech signal into M channels using 6th-order Chebyshev filters with no passband overlap.

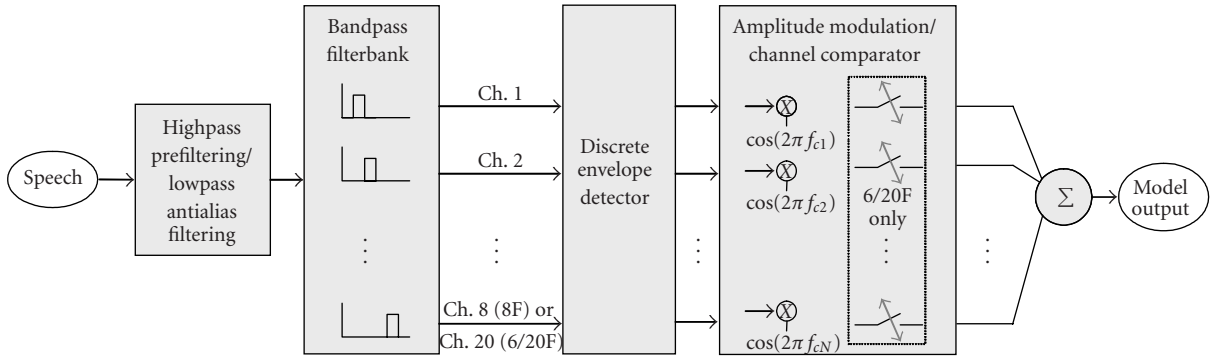

Figure 1: Block diagram of acoustic model. Temporal resolution is equivalent in both models, with channel envelopes discretized over 2-millisecond windows. In each 2-millisecond window, the 8F model presents speech information from 150 Hz to 6450 Hz divided amongst eight channels, whereas the 6/20F model presents six channels, each with narrower bandwidth, chosen from twenty channels spanning 250 Hz to 10823 Hz.

Each channel is full-wave rectified and lowpass filtered using an 8th-order Chebyshev with 400 Hz cutoff to extract the signal envelope for each frequency channel. The envelope is discretized over the processing window of length L using the root-mean-square value.

The numbers of channels and channel cutoff frequencies for the two acoustic models used in this study were chosen to mimic two popular cochlear implant speech processors. For the 8F model, the prefiltered speech is filtered into eight logarithmically spaced frequency channels covering 150 Hz to 6450 Hz. For the 6/20F model, the prefiltered speech is filtered into twenty frequency channels covering 250 Hz to 10823 Hz, with linearly spaced cutoff frequencies up to 1.5 kHz and logarithmically spaced cutoff frequencies for higher filters. The discrete envelope for both models is calculated over a two-millisecond window, corresponding to 44 samples for speech recorded at a sampling frequency of 22050 Hz.

The model output is assembled by determining a set of presentation channels, the set of frequency channels to be presented in the current processing window, then amplitude modulating each presentation channel with a separate sine-wave carrier and summing the set of modulated presentation channels. In each processing window, a set of N ($N \leq M$) channels is chosen to be presented. All eight frequency channels are presented ($N = M = 8$) with the 8F model. With the 6/20F model, only the six channels with the largest amplitude in each processing window are presented ($N = 6$, $M = 20$). The carrier frequency for each presentation channel corresponds to the midpoint on the cochlea between the physical locations of the channel bandpass cutoff frequencies. The discrete envelopes of the presentation channels are amplitude modulated with sinusoidal carriers at the calculated carrier frequencies, summed, and stored as the model output.

Procedure

The listening tests were conducted in a double-walled sound-insulated booth, separate from the computer, experimenter, and sources of background noise, with stimuli stored on disk and presented through headphones. Subjects recorded their responses using the computer mouse and graphical user interface to select what they had heard from the set of tokens. Subjects were trained prior to the tests on the same speech materials processed through the acoustic models to provide experience with the processed speech and mitigate learning effects. Feedback was provided during training.

Testing began in quiet and advanced to increasingly noisy conditions with two repetitions of a randomly ordered vowel or consonant token set for training, followed by five repetitions of the same randomly ordered token set for testing. The order of presentation of test stimuli and acoustic models were randomly assigned and balanced among subjects to neutralize any effects of experience with the previous model or test stimulus in the pooled results. Equal numbers of test materials were presented for each test condition, defined by the specific acoustic model and signal-to-noise ratio.

Results

The subjects' responses from the vowel and consonant tests at each SNR for each acoustic model were pooled for all twelve subjects. The results are plotted for all noise levels in Figure 2. Statistical significance, indicated by asterisks, was determined using the arcsine transform [20] to calculate the 95% confidence intervals. The error bars in Figure 2 indicate one standard deviation, which were also calculated using the arcsine transform. The vowel recognition scores show that the 6/20F model significantly outperforms the 8F model at all noise levels. An approximately equivalent level of performance was achieved with both acoustic models on the consonant recognition test, with differences between scores at most SNRs not statistically significant. Vowel recognition is heavily dependent on the localization of formant frequencies, so it is reasonable that subjects using the 6/20F model, with 20 spectral channels, perform better on vowel recognition.

At each SNR, results of the vowel and consonant test were pooled across subjects and tallied in confusion matrices, with

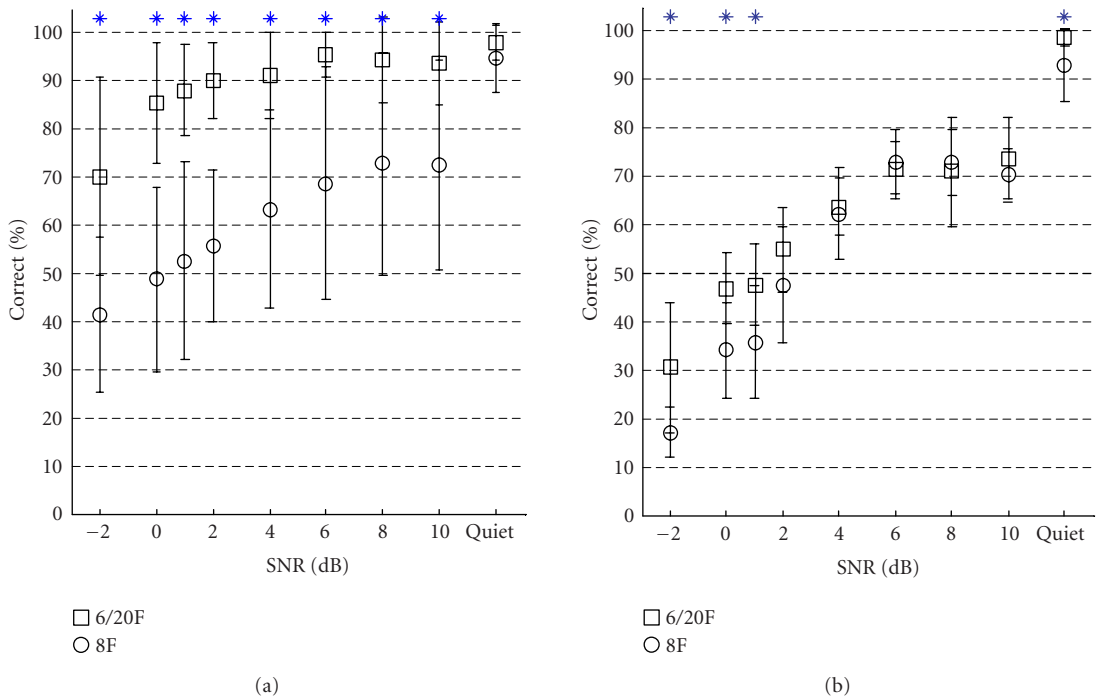

Figure 2: (a) Vowel token recognition scores. (b) Consonant token recognition scores.

rows corresponding to the actual token played, and columns indicating the token chosen by the subject. An example confusion matrix is shown in Table 1. Correct responses lie along the diagonal of the confusion matrix. The confusion matrices gathered from the vowel and consonant test can be analyzed based on the arrangement and frequency of incorrect responses. One such method of analysis is information transmission analysis, developed by Miller and Nicely in [21]. In each set of tokens presented, it is intuitive that some incorrect responses will occur more frequently than others, due to common phonetic features of the tokens. The Miller and Nicely method groups tokens based on the common phonetic features and calculates information transmission using the mean logarithmic probability (MLP) and mutual information $T(x;y)$ which can be considered the transmission from x to y in bits per stimulus. In the equations below, p_i is the probability of confusion, N is the number of entries in the matrix, n_i is the sum of the ith row, n_j is the sum of the jth column, and n_{ij} is a value from the confusion matrix resulting from grouping tokens with common phonetic features:

$$\begin{aligned} \mathrm{MLP}(x) &= -\sum_i p_i \log p_i, \\ T(x;y) &= \mathrm{MLP}(x) + \mathrm{MLP}(y) - \mathrm{MLP}(xy) \\ &= -\sum_{i,j} \frac{n_{ij}}{N} \log_2 \frac{n_i n_j}{N n_{ij}}. \end{aligned} \tag{1}$$

The consonant tokens were classified using the five features in Miller and Nicely—voicing, nasality, affrication, duration, and place. Information transmission analysis was also applied to vowels, classified by the first formant frequency, the second formant frequency, and duration. The feature classification matrices are shown in Table 2. Information transmission analysis calculates the transmission rate of these individual features, providing a summary of the distribution of incorrect responses, which contains useful information unavailable from a simple token recognition score.

Figure 3 shows the consonant feature percent transmission, with percent correct recognition or "score" from Figure 2 included, for the 6/20F model and 8F model. The plots exhibit some deviation from the expected monotonic result; however, this is likely due to sample variability and variations in the random samples of additive noise used to process the tokens. It appears that increasing levels of noise deleteriously affect all consonant features for both acoustic models. It is interesting to note that consonant recognition scores for the 6/20F model and 8F model are nearly identical, but feature transmission levels are quite different. The differences in the two acoustic models result in two distinct sets of information that result in approximately the same level of consonant recognition. A previous study by Fu et al. [3] performed information transmission analyses on consonant data for 8-of-8 and 6-of-20 models and calculated closely grouped feature transmission rates at each SNR for both models, resembling the 8F results shown here. Both Fu et al. models as well as the 8F model in this study have similar model bandwidths, and it is possible that the inclusion of higher frequencies in the 6/20F model and their effect on channel location and selection of presentation channels results in the observed spread of feature transmission rates. Further comments on these results are presented in the discussion.

Table 1: Example confusion matrix for 8F vowels at +1 dB SNR. Responses are pooled from all test subjects.

8F acoustic model, SNR = 1 dB										
		Responded								
		had	hawed	head	heard	heed	hid	hood	hud	who'd
Played	had	29	10	12	3	0	0	1	5	0
	hawed	0	53	0	1	1	0	0	4	1
	head	9	2	19	5	3	14	5	2	1
	heard	0	2	4	34	1	4	9	3	3
	heed	2	0	1	6	31	0	7	0	13
	hid	2	2	15	2	6	26	2	3	2
	hood	0	2	4	6	4	2	26	4	12
	hud	1	19	1	2	0	0	3	31	3
	who'd	1	1	1	7	2	1	12	0	35

Table 2: Information transmission analysis classification matrices for (a) consonants and (b) vowels. The numbers in each column indicate which tokens are grouped together for analysis of each of the features. For some features, multiple groups are defined.

(a)

Consonants	Voicing	Nasality	Affrication	Duration	Place
b	1	0	0	0	0
d	1	0	0	0	1
f	0	0	1	0	0
g	1	0	0	0	4
j	1	0	0	0	3
k	0	0	0	0	4
m	1	1	0	0	0
n	1	1	0	0	1
p	0	0	0	0	0
s	0	0	1	1	2
sh	0	0	1	1	3
t	0	0	0	0	1
v	1	0	1	0	0
z	1	0	1	1	2

(b)

Vowels	Duration	F1	F2
had	2	2	1
hawed	1	2	0
head	1	1	1
heard	1	1	0
heed	2	0	1
hid	0	1	1
hood	0	1	0
hud	0	2	0
who'd	0	0	0

The patterns of feature transmission are much more consistent between the two acoustic models for vowels, as shown in Figure 4. The significantly higher vowel recognition scores at all noise levels using the 6/20F model translate to greater transmission of all vowel features at all noise levels. Hence, the better performance of the 6/20F model is not due to more effective transmission of any one feature.

3. CONFUSION PREDICTIONS

Several signal processing techniques were developed in the context of this research to measure similarities between processed speech tokens for the purpose of predicting patterns of vowel and consonant confusions. The use of the similarities and differences between speech tokens has a basis in previous studies predicting speech intelligibility [7, 8], and investigating the perception of speech tokens presented through an impaired auditory system [10] and processed by a cochlear implant [9].

The three prediction methods that are developed in this study use two different signal representations and three different signal processing methods. The first method is token envelope correlation (TEC), which calculates the correlation between the discrete envelopes of each pair of tokens. The second method is dynamic time warping (DTW) using the cepstrum representation of the speech token. The third prediction method uses the cepstrum representation and hidden Markov models (HMMs). These three methods provide for comparison a method using only the temporal information (TEC), a deterministic measure of distance between the speech cepstrums (DTW), and a probabilistic distance measure using a statistical model of the cepstrum (HMM).

Dynamic time warping

For DTW [22], the (ith, jth) entry in the prediction metric matrix is the value of the minimum-cost mapping through a cost matrix of Euclidean distances between the cepstrum coefficients of the ith given token and the jth response token. To calculate the (ith, jth) entry in the prediction metric matrix, the cepstrum coefficients are computed from energy-normalized speech tokens. A cost matrix is constructed from the cepstrums of the two tokens. Each row of the cost matrix

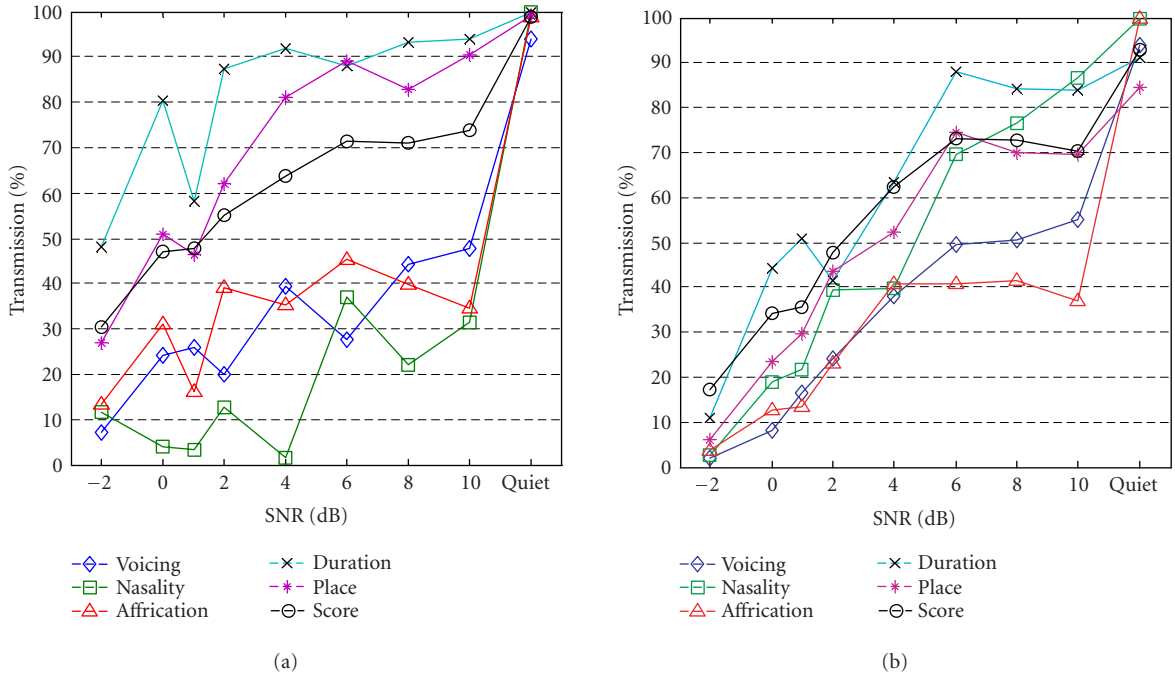

Figure 3: (a) 6/20F consonant information transmission analysis. (b) 8F consonant information transmission analysis.

specifies a vector of cepstrum coefficients calculated during one window of the given signal, each column corresponds to a vector of cepstrum coefficients calculated during one window of the response signal, and the entry in the cost matrix is a measure of distance between the two vectors. In this project, the coefficient vector differences were quantified using the Euclidean distance $d_2(x, y)$,

$$d_2(x,y) = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2}. \tag{2}$$

The minimum-cost path is defined as the contiguous sequence of cost matrix entries from (1,1) to (N, M), where N is the length of the given token cepstrum and M is the length of the response token cepstrum, such that the sum of the sequence entries is minimized. To reduce the complexity of searching for the minimum-cost path, sequence steps are restricted to three cases: horizontal $(n, m+1)$, vertical $(n+1, m)$, and diagonal $(n+1, m+1)$. Additionally, since the shortest path from (1,1) to (N, M) will be nearly diagonal, the cost matrix entry is multiplied with a weighting parameter in the case of a diagonal step, to prevent the shortest path from becoming the default minimum-cost path. The value for the weighting parameter, equal to 1.5 in this study, can be increased or decreased resulting in a lesser or greater propensity for diagonal steps.

Next, the cumulative minimum-cost matrix D_{ij} containing the sum of the entries for the minimum-cost path from (1,1) to any point (n, m) in the cost matrix is calculated. Given the restrictions on sequence-step-size, sequence step direction, and weighting parameter, the cumulative cost matrix is calculated as

$$D_{n+1,m+1} = \begin{cases} 1.5 \cdot d_{n+1,m+1} + \min\left(D_{n,m}, D_{n+1,m}, D_{n,m+1}\right) \\ \qquad \text{if } \min\left(D_{n,m}, D_{n+1,m}, D_{n,m+1}\right) = D_{n,m}, \\ d_{n+1,m+1} + \min\left(D_{n,m}, D_{n+1,m}, D_{n,m+1}\right) \\ \qquad \text{if } \min\left(D_{n,m}, D_{n+1,m}, D_{n,m+1}\right) \neq D_{n,m}. \end{cases} \tag{3}$$

The value of the minimum-cost path from (1,1) to (N, M) is $D_{N,M}$. The final value of the prediction metric is the minimum cost $D_{N,M}$ divided by the number of steps in the path to normalize values for different token lengths. Diagonal steps are counted as two steps when determining the path length.

Token envelope correlation

For TEC, the (ith, jth) entry in the prediction metric matrix is the normalized inner product of the discrete envelopes of two processed speech tokens that have been temporally aligned using dynamic time warping. The discrete envelope was originally calculated as a step in the acoustic model processing. The discrete envelope used in TEC is similar to the discrete envelope calculated in the acoustic model, with a lower cutoff frequency on the envelope extraction filter.

The cepstrums of the ith processed given token and the jth processed response token are used in the DTW procedure to calculate the minimum-cost path for the two tokens. The minimum-cost path is then used to temporally align the two discrete envelopes, addressing the issue of different token

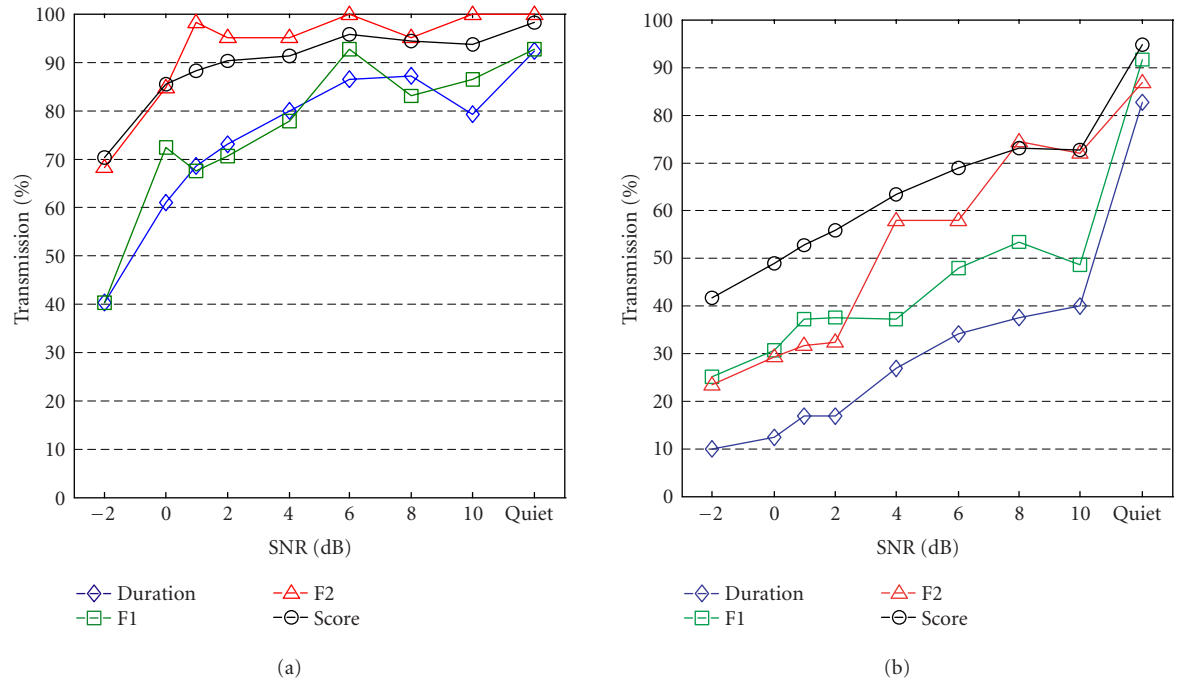

Figure 4: (a) 6/20F vowel information transmission analysis. (b) 8F vowel information transmission analysis.

lengths in a more elegant manner than simple zero padding. Using DTW to align the signals injects flexibility in the alignment to account for potential listener ambiguity regarding the starting point and pace of the speech token.

After alignment of the given token and response token, the final value of the prediction metric can be calculated as

$$M_{i,j} = \frac{\mathbf{x}_i{}^T \mathbf{y}_j}{\sqrt{\mathbf{x}_i{}^T \mathbf{x}_i}\sqrt{\mathbf{y}_j{}^T \mathbf{y}_j}}, \quad (4)$$

where $\mathbf{x}_i$ is the discrete envelope of the ith given token, $\mathbf{y}_j$ is the discrete envelope of the jth response token, and $M_{i,j}$ is the (ith, jth) entry in the prediction metric matrix.

Hidden Markov models

The third prediction method is based on hidden Markov models (HMMs) [22, 23]. Using HMMs, the (ith, jth) entry in the prediction metric matrix is the log-likelihood that the cepstrum of the ith given token is the observation produced by the HMM for the cepstrum of the j^{th} response token. To calculate the (ith, jth) entry in the prediction metric matrix using HMMs, a continuous-observation HMM was trained for each speech token using a training set of 100 tokens. All training data were collected from a single male speaker in quiet. HMMs were trained for different numbers of states Q and numbers of Gaussian mixtures M, with Q ranging from two to six and M ranging from two to four. Training was performed using the expectation-modification method to iteratively determine the parameters that locally maximize the probability of the observation sequence. The state transition matrix and Gaussian mixture matrix were initialized using random values. A k-means algorithm was used to initialize the state-observation probability distributions. The probability of an observation was determined using the forward algorithm [23] to calculate $P(O_1 O_2 \cdots O_T, q_T = S_i \mid \lambda)$, where O_i is the ith element in the observation sequence, $q_T = S_i$ indicates that the model is in the ith state at time T, and λ are the HMM parameters.

Prediction performance

The accuracy of each prediction method was verified using the vowel and consonant confusion matrices generated in the listening test as basis for comparison. The confusion matrices at each of the eight noise levels and in quiet were pooled to produce a general pattern of confusions independent of any specific noise level. Combining the confusion matrices across noise levels was justified by information transmission analyses, which indicated that increasing the amount of additive noise most significantly affected the rate of confusions rather than the pattern of confusions.

The first test of confusion prediction performance gauged the ability to predict the most frequent incorrect responses (MFIRs). The prediction of MFIRs was measured in terms of successful near predictions, defined as the case where one token in the set of MFIRs matches one token in the predicted set of MFIRs. Sets of two tokens were used for vowel near predictions (25% of possible incorrect responses), three tokens for consonants (23% of possible incorrect responses). For example, if the two MFIRs for "head" were "hid" and "had," then either "hid" or "had" would have to be one of the two predicted MFIRs for a successful near prediction. Measuring prediction performance using near

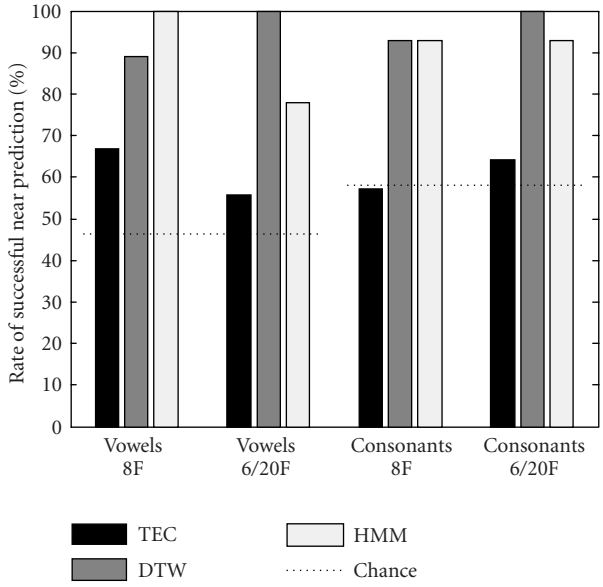

Figure 5: Most frequent incorrect response (MFIR) near predictions for each combination of speech material (vowel, consonant) and acoustic model (8F, 6/20F). Chance scores are included for comparison.

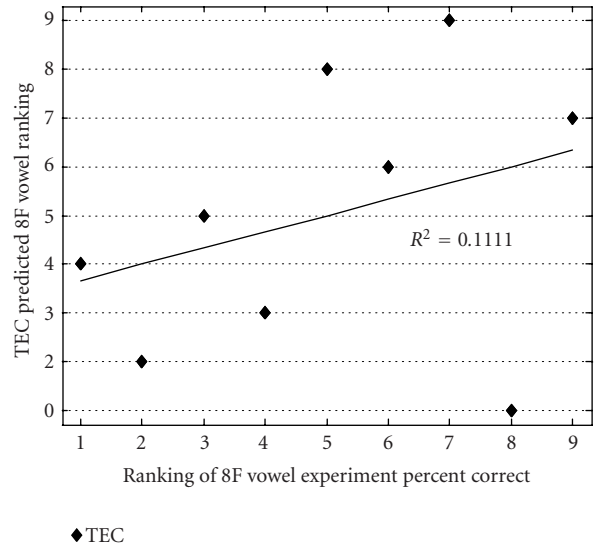

Figure 6: Scatter plot for 8F vowel predicted rankings using TEC versus actual recognition rankings. Includes regression line and R^2 value, corresponding to top-left value in Table 3a.

predictions satisfies the objective of predicting patterns in the confusions, rather than strictly requiring that the predicted MFIR was indeed the most frequent incorrect response. The purpose of measuring near predictions is to test whether the methods are distributing the correct tokens to the extremes of the confusion response spectrum.

Figure 5 shows the percentages for successful near prediction of the MFIR tokens for each acoustic model and token set. Percentages of successful near prediction were calculated out of possible nine trials for vowels ($N = 9$) and fourteen trials for consonants ($N = 14$). Near-perfect performance is achieved using DTW. The HMM method performs at a similarly high level. The TEC method consistently underperforms the two methods utilizing the cepstrum coefficients for confusion prediction. Chance performance is also shown for comparison.

The second test of confusion prediction performance analyzed the ability of each method to discern how frequently each individual token will be confused, as represented by the main diagonal of the confusion matrices. Rather than predicting the absolute rate of confusion, which would be dependent on noise level, the test evaluates the accuracy of a predicted ranking of the tokens from least to most recognized, or most often to least-often confused.

To calculate the predicted ranking of the individual-token recognition rates, the off-diagonal values in each row of the prediction metric matrix were averaged and ranked, as a means of evaluating each token's uniqueness. The more separation between the played token and the set of incorrect responses, where separation is measured by the prediction metrics, the less likely it is that an incorrect response will occur.

The fit of the predicted token recognition rankings to the actual recognition rankings was represented using linear regression. The coefficient of determination R^2 [24] was calculated for the linear regression of a scatter plot with one set of rankings plotted on the ordinate and another on the abscissa. R^2 values were calculated for two different sets of scatter plots. The first set of scatter plots was created by plotting the predicted recognition rankings and token length rankings against the true recognition rankings. A ranking of token lengths was included to investigate any potential effects of token length on either the calculation of the prediction metrics or the listening test results. Figure 6 displays an example scatter plot for TEC predicted 8F vowels rankings including the regression line and R^2 value. Each token is represented by one point on the chart. The x-axis value is determined by token rank in terms of recognition rate in the listening test, and the y-axis value is determined by the token's predicted recognition ranking using TEC. Similar scatter plots (not shown) were created for the other prediction methods. All of the R^2 values with listening test rankings on the x-axis are shown in Table 3a. A second set of scatter plots was created by assigning token length rankings to the x-axis, rather than listening test rankings, and using predicted rankings and listening test rankings for the y-axis values (Table 3b).

Table 3 shows the R^2 values for the two different methods of plotting. With the percent correct plotted on the x-axis, the HMM is shown to perform very well for vowel recognition rankings with either acoustic model. DTW and HMM perform similarly on 8F consonants, but not at the level of HMM on vowels. HMM performance is weaker for 6/20F consonants than for 8F consonants. Predicted recognition rankings for any material set using TEC do not appear promising.

Table 3: Summary of coefficient of determination for linear fittings. R^2 values calculated for percent correct along x-axis (a) and length along x-axis (b).

(a)

	Percent correct along x-axis			
Method	8F vow.	6/20F vow.	8F cons.	6/20F cons.
TEC	0.1111	0.3403	0.0073	0.003
DTW	0.0336	0.0278	0.4204	0.4493
HMM	0.6944	0.5136	0.4261	0.2668
Length	0.3463	0.0544	0.0257	0.0333

(b)

	Length along x-axis			
Method	8F vow.	6/20F vow.	8F cons.	6/20F cons.
TEC	0.0711	0.0044	0.0146	0.3443
DTW	0.16	0.444	0.01	0.0136
HMM	0.64	0.5378	0.09	0.2759
Correct (%)	0.34	0.0544	0.0257	0.033

Investigating the potential relationship between token length and predicted recognition rankings leads to the observation that HMM predicted rankings for vowels with both acoustic models and DTW predicted rankings for 6/20F vowels appear to correspond to token length. The true recognition ranking also appears related to length for 8F vowels. The relationship between HMM predicted rankings and token length can potentially be explained by the structure of the HMM. The state transition probabilities are adapted to expect tokens of a certain length; longer or shorter tokens can cause state transitions that are forced early or delayed. This would affect the calculated log-likelihood values, and could result in artifacts of token length in the predicted recognition rankings.

The third task tested whether the performance gap seen in the listening test between the token sets with different materials and acoustic models was forecast by any of the prediction methods. DTW was the only method that appeared to have any success predicting the differences in token correct identification for the different acoustic models and token sets. The token identification trend lines for vowels and consonants are shown in Figure 7a. The overall level of token recognition for any combination of token set and acoustic model was predicted with DTW by averaging the off-diagonal prediction metrics. The average confusion distance is plotted as a constant versus SNR in Figure 7b since the metric is not specific to the performance at any particular noise level, and indicates that the pattern of the trends of recognition levels is reasonably well predicted.

Predicted trends for TEC and HMM are not shown, but did not accurately indicate the trends in the listening test. The failure of TEC at the third task supports the conclusion that the strictly temporal representation lacks sufficient distinguishing characteristics. Since the measure for this task is essentially an average of the token recognition rankings calculated in the second task, another measure of prediction performance for which TEC scored poorly, the poor performance using TEC for this task is not surprising. However, the HMM prediction metric performed very well on the first two tasks. Based on that performance, the failure of HMMs was unexpected, especially with the accuracy of the predicted trends using DTW.

4. DISCUSSION

Information transmission analysis using the method developed by Miller and Nicely [21] calculates how effectively the two acoustic models transmitted the features of vowels and consonants. The increased spectral resolution of the 6/20F model, credited for the better performance of the 6/20F model for vowel token recognition, also appeared in the information transmission results, with proportionally greater transmission of both the F1 and F2 features. The results of the consonant feature analyses are more difficult to classify. A reasonable hypothesis would be that the 8F model should more effectively transmit broadband features, since it has a continuous frequency spectrum with greater bandwidth than the 6/20F model. The 6/20F model should better transmit frequency-specific consonant features due to greater frequency resolution. However, many outcomes from the consonant feature transmission analysis disagree with this hypothesis. Affrication, a broadband feature, is transmitted with similar efficiency by both acoustic models. Voicing is relatively narrowband and suspected to be more effectively transmitted by the 6/20F model; however, it is also transmitted with similar efficiency by both acoustic models. The 6/20F model transmits place and duration more effectively than the 8F model. Duration is essentially a temporal feature, and differences between the acoustic models should not

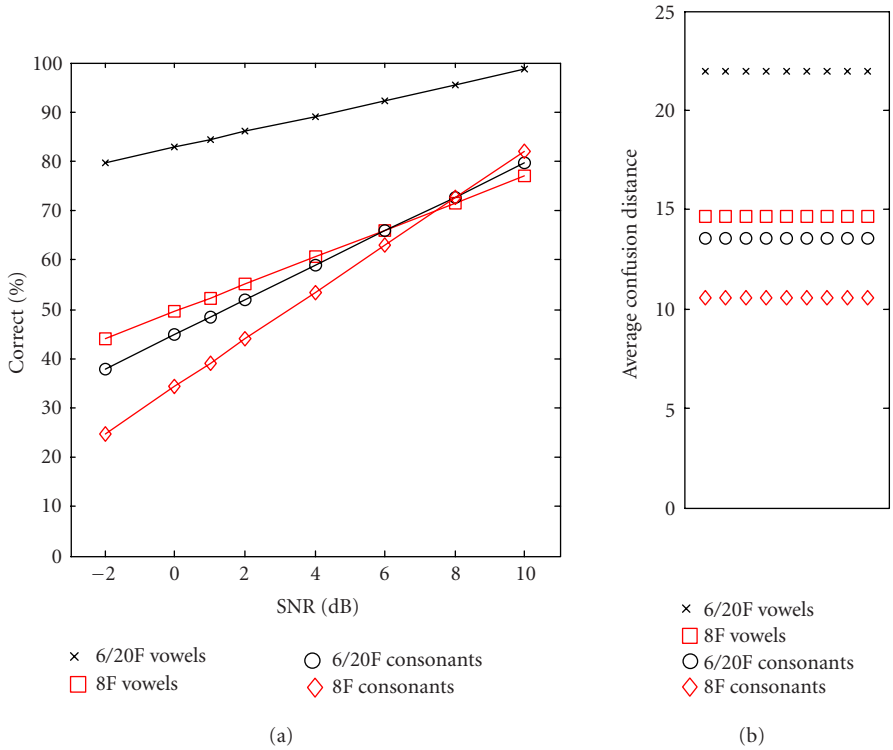

Figure 7: (a) Trends in results of the listening tests separated by model and test material. (b) Trends predicted by DTW using average confusion distance.

affect transmission of this feature. An acoustic description of the effect of place is very complex and difficult to describe in general terms for all tokens. Place can appear as a broadband or narrowband feature in different regions of the frequency spectrum. The 8F model was more efficient at transmitting the nasality feature. With regards to the 6/20F model, it is possible that six spectral channels do not provide sufficient information to maximally transmit some of the consonant features, whereas for vowels, only a few channels are required for transmission of the formants.

This examination of the information transmission analysis results can benefit from the observation that the vowel and consonant features are not independent of each other. For example, the vowel feature duration, a temporal feature, was much more effectively transmitted, with a difference of approximately 50% transmission across all noise levels, by the 6/20F model than by the 8F model; however, the two models have the same temporal resolution. The increased spectral resolution would have legitimately increased transmission of the formant features, resulting in a reduced number of incorrect responses in the listening test, which would in turn raise the calculated transmission of duration information as a side effect. It is expected that some of the calculated percent transmission of features for consonants may also reflect strong or weak performance of other features, or could potentially be influenced by unclassified features. Analysis of the feature classification matrices could help explain potential relationships between the calculated values for feature transmission.

The results of the confusion predictions indicate that analysis of the differences between tokens can provide insight to the token confusions. The three tasks used in this study to analyze the performance of the confusion predictions investigate prediction of trends along the rows, the diagonal, and the overall separation of prediction metrics, providing a multifaceted view of the accuracy of the overall token confusion pattern. The two methods utilizing the cepstrum coefficients for representing the speech token outperformed the method using strictly temporal information in all three tests. The experiment setup and speech-shaped noise characteristics, either of which could potentially affect patterns of token confusion, were not considered in the prediction metric calculations. Expanding the prediction methods to include such additional factors could improve the accuracy of confusion pattern prediction.

Not considering the effects of the noise characteristics and experiment setup also resulted in symmetric prediction metrics matrices calculated using DTW and TEC. This is not entirely consistent with the results of the listening test, however the results presented in this study using DTW indicate that symmetry does not prohibit prediction of trends in token confusion. The procedure for calculating the prediction metric with each prediction method included steps to normalize the outcome for tokens of different lengths, to emphasize the differences within the speech signals and minimize any effect of differences in token length. However, Table 3 indicates that token length may have been used by the listening

test participants to distinguish the speech tokens. Reinserting some effect of token length in the calculation of the prediction metrics or removing token length as a factor in the listening test may also improve confusion prediction accuracy.

In summary, this study presented results of a listening test in noise using materials processed through two acoustic models mimicking the type of speech information presented by cochlear implant speech processors. Information transmission analyses indicate different rates of transmission for the consonant features, likely due to differences in spectral resolution, number of channels, and model frequency bandwidth, despite similar speech recognition scores. The development of signal processing methods to robustly and accurately predict token confusions would allow for preliminary analysis of speech materials to evaluate prospective speech processing and noise mitigation schemes prior to running listening tests. Results presented in this study indicate that measures of differences between speech tokens calculated using signal processing techniques can forecast token confusions. Future work to improve the accuracy of the confusion predictions should include refining the prediction methods to consider additional factors contributing to token confusions, such as speech-shaped noise characteristics, experiment setup, and token length.

ACKNOWLEDGMENTS

This work was supported by NSF Grant NSF-BES-00-85370. We would like to thank the three anonymous reviewers for comments and suggestions. We would also like to thank the subjects, who participated in this experiment, as well as Dr. Chris van den Honert at Cochlear Corporation and Doctors Robert Shannon and Sigfrid Soli at House Ear Institute for supplying speech materials.

REFERENCES

[1] M. F. Dorman, P. C. Loizou, and J. Fitzke, "The identification of speech in noise by cochlear implant patients and normal-hearing listeners using 6-channel signal processors," *Ear & Hearing*, vol. 19, no. 6, pp. 481–484, 1998.

[2] B. L. Fetterman and E. H. Domico, "Speech recognition in background noise of cochlear implant patients," *Otolaryngology—Head and Neck Surgery*, vol. 126, no. 3, pp. 257–263, 2002.

[3] Q.-J. Fu, R. V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3586–3596, 1998.

[4] D. Baskent and R. V. Shannon, "Speech recognition under conditions of frequency-place compression and expansion," *Journal of the Acoustical Society of America*, vol. 113, no. 4, pp. 2064–2076, 2003.

[5] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, "Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants," *Journal of the Acoustical Society of America*, vol. 110, no. 2, pp. 1150–1163, 2001.

[6] P. C. Loizou, M. F. Dorman, O. Poroy, and T. Spahr, "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," *Journal of the Acoustical Society of America*, vol. 108, no. 5, pp. 2377–2387, 2000.

[7] H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility. I. Model structure," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2896–2909, 2001.

[8] H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance," *Journal of the Acoustical Society of America*, vol. 109, no. 6, pp. 2910–2920, 2001.

[9] M. A. Svirsky, "Mathematical modeling of vowel perception by users of analog multichannel cochlear implants: temporal and channel-amplitude cues," *Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1521–1529, 2000.

[10] A. Leijon, "Estimation of sensory information transmission using a hidden Markov model of speech stimuli," *Acustica—Acta Acustica*, vol. 88, no. 3, pp. 423–432, 2001.

[11] Q.-J. Fu, J. J. Galvin, and X. Wang, "Recognition of time-distorted sentences by normal-hearing and cochlear-implant listeners," *Journal of the Acoustical Society of America*, vol. 109, no. 1, pp. 379–384, 2001.

[12] Q.-J. Fu and R. V. Shannon, "Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners," *Journal of the Acoustical Society of America*, vol. 107, no. 1, pp. 589–597, 2000.

[13] Y. C. Tong, J. M. Harrison, J. Huigen, and G. M. Clark, "Comparison of two speech processing schemes using normal-hearing subjects," *Acta Otolaryngology Supplement*, vol. 469, pp. 135–139, 1990.

[14] Cochlear Corporation and the University of Iowa, Cochlear Corporation/the University of Iowa Revised Cochlear Implant Test Battery, Englewood, Colo, USA, 1995.

[15] M. F. Dorman, P. C. Loizou, J. Fitzke, and Z. Tu, "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6-20 channels," *Journal of the Acoustical Society of America*, vol. 104, no. 6, pp. 3583–3585, 1998.

[16] P. C. Loizou, M. F. Dorman, Z. Tu, and J. Fitzke, "Recognition of sentences in nosie by normal-hearing listeners using simulations of SPEAK-type cochlear implant signal processors," *Annals of Otology, Rhinology, and Laryngology Supplement*, vol. 185, pp. 67–68, December 2000.

[17] C. S. Throckmorton and L. M. Collins, "The effect of channel interactions on speech recognition in cochlear implant subjects: predictions from an acoustic model," *Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 285–296, 2002.

[18] P. J. Blamey, R. C. Dowell, Y. C. Tong, and G. M. Clark, "An acoustic model of a multiple-channel cochlear implant," *Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 97–103, 1984.

[19] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 2403–2411, 1998.

[20] A. R. Thornton and M. J. M. Raffin, "Speech-discrimination scores modeled as a binomial variable," *Journal of Speech and Hearing Research*, vol. 21, no. 3, pp. 507–518, 1978.

[21] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[22] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, NY, USA, 1993.
[23] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
[24] J. L. Devore, *Probability and Statistics for Engineering and the Sciences*, Duxbury Press, Belmont, Calif, USA, 1995.

Jeremiah J. Remus received the B.S. degree in electrical engineering from the University of Idaho in 2002 and the M.S. degree in electrical engineering from Duke University in 2004. He is currently working towards the Ph.D. degree in the Department of Electrical & Computer Engineering at Duke University. His research interests include statistical signal processing with applications in speech perception and auditory prostheses.

Leslie M. Collins was born in Raleigh, NC. She received the B.S.E.E. degree from the University of Kentucky, Lexington, and the M.S.E.E. and Ph.D. degrees in electrical engineering, both from the University of Michigan, Ann Arbor. She was a Senior Engineer with the Westinghouse Research and Development Center, Pittsburgh, Pa, from 1986 to 1990. In 1995, she became an Assistant Professor in the Department of Electrical & Computer Engineering (ECE), Duke University, Durham, NC, and has been an Associate Professor in ECE since 2002. Her current research interests include incorporating physics-based models into statistical signal processing algorithms, and she is pursuing applications in subsurface sensing, as well as enhancing speech understanding by hearing impaired individuals. She is a Member of the Tau Beta Pi, Eta Kappa Nu, and Sigma Xi honor societies.