VOWEL AND CONSONANT CONFUSION IN NOISE BY COCHLEAR IMPLANT SUBJECTS: PREDICTING PERFORMANCE USING SIGNAL PROCESSING TECHNIQUES

Jeremiah J. Remus and Leslie M. Collins

Department of Electrical and Computer Engineering
Duke University, Durham, NC 27708

ABSTRACT

Cochlear implants are able to restore some degree of hearing to deafened individuals; however implant users are particularly susceptible to background noise. The effect of noise can be assessed using vowel and consonant confusions measured in listening experiments. This paper presents three signal processing methods developed to predict patterns in vowel and consonant confusion in noise for cochlear implant users. Prediction performance is tested using the results of a listening experiment conducted with acoustic models of two cochlear implant speech processors and normal hearing subjects. Confusion prediction is based on prediction metrics calculated using each method's unique representation of the speech tokens.

1. INTRODUCTION

Cochlear implants have been used successfully to restore some degree of hearing in severely deafened patients. Studies by numerous investigators have shown a high level of word and sentence recognition by cochlear implant patients in controlled listening experiments [1-3]. However, listening experiments and patient testimony indicate that cochlear implant patients are particularly susceptible to the deleterious effects of background noise, which cochlear implant patients must contend with in many everyday situations.

Listening experiments are conducted using a variety of speech materials, such as vowel and consonant sounds, monosyllabic words, and sentences. Closed set tests, where subjects must select their response from limited options, allow for the analysis of confusions between tokens. Results of a closed-set test can be represented in a confusion matrix, containing the number of occurrences of each possible token given/responded combination, with the correct identification of tokens presented along the diagonal. The results can then be analyzed using information transmission analysis developed by Miller and Nicely [4], or other methods of classifying patterns of confusions.

Acoustic models of cochlear implant speech processors have been employed by many investigators to conduct listening experiments using normal-hearing subjects [e.g. 5,6]. Acoustic models provide control over the experimental parameters, allowing for fixed conditions across subjects. The individual performance of cochlear implant patients can be affected by a variety of personal factors – age, duration of deafness, duration of implantation, etiology of deafness, electrode insertion depth, etc. Using acoustic models with normal hearing subjects enables mitigation of these external factors, varying only the parameter of interest during testing.

The use of normal hearing subjects in listening experiments with acoustic models for investigations of cochlear implant speech recognition is a widely used and well accepted experimental method. Many developments in cochlear implants have been made using acoustic models; for example, saturation of pitch as a function of pulse rate [7] and improved performance using a simple spectral mapping speech processor versus a feature extraction algorithm [8]. However, results of listening experiments using normal hearing subjects are often only indicative of trends in cochlear implant patient performance; absolute levels of performance tend to disagree.

In this study, signal processing techniques were developed to measure similarities between speech tokens processed by acoustic models for the purpose of predicting patterns of vowel and consonant confusions. Prediction results were compared to confusion matrices generated from normal hearing subjects tested for vowel and consonant recognition in noise using two acoustic models.

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 experiments. 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 implant patients, if the prediction method is well developed and robust.

2. LISTENING EXPERIMENT

Twelve normal hearings subjects were tested for vowel and consonant recognition in noise with speech tokens processed using two different acoustic models. The two acoustic models used in this experiment [9] are identified throughout

as 8F and 6/20F, and imitate the CIS [10] and SPEAK [11] processing strategies.

Each subject's vowel and consonant recognition ability was tested in quiet and at eight signal-to-noise (SNR) levels: +10 dB, +8 dB, +6 dB, +4 dB, +2 dB, +1 dB, 0 dB, and –2 dB. Both the 8F and 6/20F implant models were used at each SNR level, for a total of 18 test conditions. The vowel tokens used in the listening experiment 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/ format.

Independently interchanging the order of test material and acoustic model generates four possible experiment sequences that are divided equally among the subjects to neutralize any effects of experience with the previous model or token set. The experiment began with two repetitions of the randomly ordered token set for training, followed by five repetitions of the randomly ordered token set for testing. Subjects were tested at all noise levels with one token set / acoustic model combination before proceeding to the next set of token materials. Feedback was given during training, and unlimited repeats were allowed in both training and testing. The experiment started in quiet and proceeded through the increasing noise levels, ending at -2 dB.

Confusion matrices were compiled from results pooled across all subjects and noise levels. Pooling across noise levels was justified by information transmission analysis, which indicated that the additive noise proportionally affected all token features. Hence, for noisier test conditions, the same patterns of confusions are expected, but larger in magnitude. The resulting confusion matrices, separated by test material and acoustic model, serve as the reference for measuring the confusion predictions using the signal processing methods.

3. CONFUSION PREDICTION METHODS

Three signal processing methods were developed for predicting trends in vowel and consonant confusion matrices. Confusion predictions were based on prediction metrics, calculated as some measure of similarity or distance between two speech tokens. The prediction methods utilize the processed speech tokens with no additive noise, assuming confusions are predominantly dictated by the token itself. The process that was used for calculating these prediction metrics for each of the three methods is detailed below.

3.1. Token envelope correlation

Token envelope correlation (TEC) uses the discrete envelope of the speech signal, an element of the acoustic model, for calculation of the prediction metric. As equation 1 shows, the prediction metric is the correlation coefficient between the two speech tokens' discrete envelopes.

$$M_{i,j} = \frac{\mathbf{x}_i^{\mathrm{T}} \mathbf{s}_j}{\sqrt{\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i}\sqrt{\mathbf{s}_j^{\mathrm{T}} \mathbf{s}_j}} \quad \text{for the } i\text{th, } j\text{th tokens} \qquad (1)$$

To address the issue of different token lengths, as well as the fact that correlation is dependent on the alignment of the two signals, the discrete envelopes are aligned, according to the least cost mapping determined by dynamic time warping [12], prior to calculating the correlation. Dynamic time warping was carried out using 512-sample windows of the processed speech signal with 50% overlap. Token envelope correlation was included as a prediction method since calculation of the prediction metric is based strictly on the temporal information contained in the discrete envelope, thus testing the adequacy of using only temporal information for token discrimination.

3.2. Dynamic time warping

Dynamic time warping (DTW) calculates the prediction metric using the same Mel-cepstrum coefficients used to align the discrete envelopes in TEC [12]. DTW finds the least cost mapping through a cost matrix, where each entry is the Euclidean distance (eq. 2) between the Mel-cepstrum coefficients at two points in the token.

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

The prediction metric M used in DTW is the value of the least cost mapping (eq. 3) through the cost matrix.

$$D_{i+1,j+1} = d(x_{i+1}, y_{j+1}) + \min(D_{i,j}, D_{i+1,j}, D_{i,j+1}) \qquad (3)$$

$$M_{i,j} = D(x_I, y_J) \qquad (4)$$

Dynamic time warping was included in this study to provide a prediction metric calculated using windowed spectral information, to contrast the strictly temporal information basis of the prediction metric for TEC.

3.3. Hidden Markov models

Hidden Markov models (HMM) use the Mel-cepstrum coefficients for calculating the prediction metric but adopt a statistical representation of the token [13]. In this study, HMMs were trained using 100 sample speech tokens, recorded by the first author, in order to develop the statistics of each token's HMM. Sets of HMMs were produced with numbers of states ranging from two to four, and continuous observation probability functions assembled using two to six Gaussian mixtures.

Two different methods were developed for calculating the prediction metric using the HMMs. Each method uses the log likelihood of each token's HMM producing a target observation for the prediction metric. In the first method the target observation is a real speech token. In the second method, the target observation is produced by sampling the HMM of the desired token. This process is averaged over 100 trials to account for different realizations of the stochastic model. The results reported here use the first method for calculating the prediction metric. Additionally, the HMMs used have three states, and observation probabilities were generated using six Gaussian mixtures.

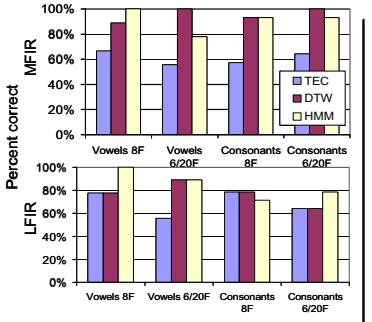

Fig. 1. Near prediction of confusions. Correct prediction of MFIR tokens (top) and LFIR tokens (bottom).

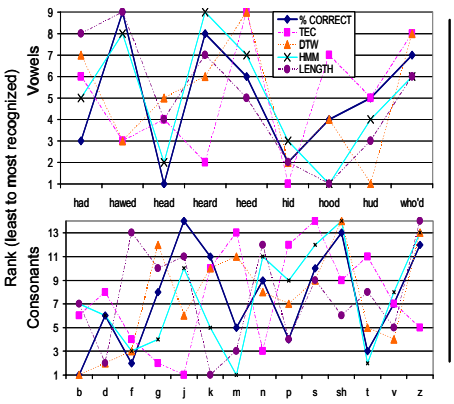

Fig. 2. Token recognition rankings.

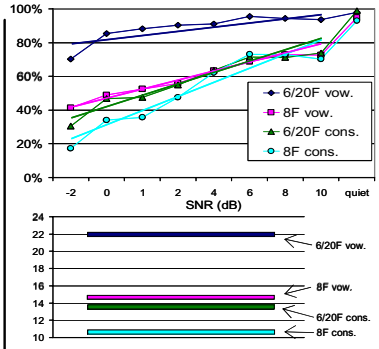

Fig. 3. (top) Trends in listening experiment results. (bottom) Average confusion distance computed using DTW.

4. PREDICTION PERFORMANCE

The confusion prediction methods were evaluated based on their ability to predict the most frequent incorrect responses (MFIRs) and least frequent incorrect responses (LFIRs) for each token presented. MFIR prediction could have important implications in designing speech processors and noise mitigation algorithms by identifying potentially problematic token confusions. Since for each token played there is usually only one or two dominant MFIRs, and a large number of less likely responses, the prediction of MFIR tokens is expected to be more robust, or at least more accurately measured, than the LFIR token predictions.

Three tests were used to gauge prediction performance. The first test measured successful near prediction of MFIRs and LFIRs. Successful near prediction is defined as the case where one token in the set of MFIRs or LFIRs matches one token in the predicted MFIRs or LFIRs. For example, if the two MFIRs for "head" are "hid" and "had", then either "hid" or "had" would have to be one of the predicted MFIRs for a successful near prediction. Sets of two tokens were used for vowel near predictions (25% of possible incorrect responses), three tokens for consonants (23% of possible incorrect responses). Measuring prediction performance using near predictions follows with predicting patterns in the confusions, rather than strictly requiring that the predicted most or least frequent incorrect response was indeed the most or least 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.

The second test measured each prediction method's ability to rank the tokens in terms of frequency of correct identification, corresponding to the values along the diagonal of the confusion matrix. Rather than predicting the exact frequency of identification, which would be dependent on noise level, the prediction methods produced a ranking of the tokens from least to most recognized (most often to least often confused). The predicted ranking was determined by comparing differences in the prediction metrics to evaluate each token's uniqueness.

The third test was a prediction of the differences in token correct identification, measured in the listening experiment as percent correct, for the different acoustic models and token sets. The level of token correct identification was predicted by averaging the differences in the prediction metrics, used in the second test, for all tokens. Greater average difference would indicate more separation between tokens, which translates into less confusability, and hence better recognition.

Results of the three prediction tests are shown in figures 1, 2 and 3. DTW and HMM perform the near prediction task (figure 1) better than TEC. This result is shown most clearly among the more legitimate MFIR near predictions, and is consistent across token sets and acoustic models.

Results shown in figure 2 are for the 8F model only, since the pattern of results was consistent for both acoustic models. Token recognition ranking performance of the three prediction methods was compared by calculating the Euclidean distance between the predicted recognition rankings and the true rankings. HMM performed well on vowels, ranking all but two tokens to within one spot of their true ranks. DTW and HMM performed similarly on consonants, but not at the level of HMM for vowels. A ranking of token length was included to expose any effects of token length on prediction method performance or potential relationship to results of the listening experiment. Token length did not factor into the recognition ranking performance; each method's recognition rankings appear to be a function primarily of the signal content, rather than signal length.

DTW was the only method successful at the third task of predicting the differences in token correct identification for the different acoustic models and token sets, and hence is the only result shown. The failure of TEC at the third task supports the conclusion that the strictly temporal representation lacks sufficient distinguishing characteristics. Based on performance in the first two tasks, the failure of HMMs was unexpected; however, follow-up work indicates HMM performance on the third task may improve with larger sets of training data. The plot of DTW average confusion distance shown in figure 3 conforms to the overall trends seen in the listening experiment results.

5. DISCUSSION AND CONCLUSIONS

The results of the prediction performance tests indicate that signal processing techniques utilizing the Mel-cepstrum representation of the speech waveform can forecast trends in token confusion. The fitness of the cepstral-based predictors (DTW and HMM) versus the strictly temporal predictor (TEC) concurs with conclusions gathered from the listening experiment, where subjects performed better using the model that emphasized spectral resolution over the inclusion of all spectral information. Further analysis of the TEC results may reveal subsets of tokens for which confusion predictions are more accurate. In general, the cepstral representation of the tokens appears better suited for confusion prediction than the temporal envelope. Consideration of other factors influencing subject responses, such as noise characteristics and experiment setup, might allow for more accurate prediction of confusions.

Prediction metrics generated by TEC and DTW are symmetric, a property not entirely consistent with the listening experiment confusion matrices. It is assumed that the asymmetries in the confusion matrices are due to the other contributory factors mentioned previously. However, even the symmetric prediction metrics should be capable of reasonably accurate confusion predictions.

The prediction methods were implemented such that the calculated prediction metrics were not biased by differences in token length. However, the listening experiment token rankings shown in figure 2 indicate that differences in token length are a potential distinguishing feature for some tokens. Performance of the prediction methods used here may improve if the influence of token length is indeed present and were removed.

Development of a robust method to forecast confusions using only the processed speech signals would allow for *a priori* analysis of new speech processing and noise mitigation schemes. Further work is needed to improve the prediction performance reported in this paper, potentially through variations in the prediction methods or manner of token representation, as well as to include the additional contributory factors mentioned here.

6. REFERENCES

[1] Q.J. Fu, R.V. Shannon, and X. Wang, "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *J. Acoust. Soc. Am*., pp. 3586-3596, Vol.104, 1998

[2] B.S. Wilson, C.C. Finley, D.T. Lawson, R.D. Wolford, and M. Zerbi, "Design and evaluation of a continuous interleaved sampling (CIS) processing strategy for multichannel cochlear implants," *J. Rehabil. Res. Dev*., pp. 110-116, Vol.30, 1993

[3] B.L. Fetterman and E.H. Domico, "Speech recognition in background noise of cochlear implant patients," *Otolaryngol. Head Neck Surg*., pp. 257-263, Vol.126, 2002

[4] G.A. Miller and P.E. Nicely, "An Analysis of Perceptual Confusion Among Some English Consonants," *J. Acoust. Soc. Am*., pp. 338-352, Vol.27, 1955

[5] P.J. Blamey, R.C. Dowell, Y.C. Tong, A.M. Brown, S.M. Luscombe, and G.M. Clark, "Speech processing studies using an acoustic model of a multiple-channel cochlear implant," *J. Acoust. Soc. Am*., pp. 104-110, Vol.76, 1984

[6] M.F. Dorman, P.C. Loizou, A.J. Spahr, and E. Maloff, "A comparison of the speech understanding provided by acoustic models of fixed-channel and channel-picking signal processors for cochlear implants," *J. Speech Lang. Hear. Res.*, pp. 783-788, Vol.45, 2002

[7] P.J. Blamey, R.C. Dowell, Y.C. Tong, and G.M. Clark, "An acoustic model of a multiple-channel cochlear implant," *J. Acoust. Soc. Am*., pp. 97-103, Vol. 76, 1984

[8] Y.C. Tong, J.M. Harrison, J. Huigen, and G.M. Clark, "Comparison of who speech processing schemes using normal-hearing subjects," *Acta Otolaryngol. Suppl*., pp. 135-139, Vol.469, 1990

[9] C.S. Throckmorton and L.M. Collins, "The effect of channel interactions on speech recognition in cochlear implant subjects: predictions from an acoustic model," *J. Acoust. Soc. Am.*, pp.285-296, Vol.112, 2002

[10] B.S. Wilson, C.C. Finley, D.T. Lawson, R.D. Wolford, D.K. Eddington, and W.M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature*, pp.236-238, Vol.352, 1991

[11] M.W. Skinner, G.M. Clark, L.A. Whitford, P.M. Seligman, S.J. Staller, D.B. Shipp, J.K. Shallop, et. al., "Evaluation of a new spectral peak coding strategy for the Nucleus 22 Channel Cochlear Implant System," *Am. J. Otol.,* pp. 215-227, Vol.15 Suppl.2, 1994

[12] Deller, J.R., Proakis, J.G., and Hansen, J.H.L., *Discrete-Time Processing of Speech Signals*, Macmillan, 1993

[13] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. of the IEEE*, pp. 257-286, Vol.77, Feb. 1989