ENVIRONMENTAL ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION USING PHYSIOLOGICALLY-MOTIVATED SIGNAL PROCESSING

(*)*Yoshiaki Ohshima and* (**)*Richard M. Stern, Jr.*
(*)Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato, Kanagawa 242, Japan
(**)Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA

ABSTRACT

This paper examines methods by which speech recognition systems can be made more environmentally robust by analyzing the performance of Seneff's model of auditory periphery [7]. The purpose of the paper is threefold. First, we document the extent to which the Seneff model reduces the degradation in speech recognition accuracy caused by additive noise and/or linear filtering. Second, we examine the extent to which individual components of the nonlinear neural transduction (NT) stage of the Seneff model contribute to recognition accuracy by evaluating the recognition accuracy with individual components of the model removed from the processing. Third, we determine the extent to which the robustness provided by the Seneff model is complementary to and independent of the improvement in recognition accuracy already provided by existing successful acoustical pre-processing algorithms such as *codeword-dependent cepstral normalization* (CDCN) [1]. Experimental techniques are proposed in the course of investigating the above issues. The results of speech recognition experiments using CMU's SPHINX [4] system under real and simulated degradation are reported.

1. INTRODUCTION

As speech recognition nears practical use in various application areas, the acoustical robustness of the speech recognizer has become an important issue to be addressed. In fact, it is well known that a speech recognition system often fails to maintain reasonable performance as the acoustical conditions of its operating environment depart from the ones that were used for the training. This phenomenon is observed not only for applications in which the system is used in very noisy environments, but also for those with more subtle changes in the acoustical environment that cause no problems for human listeners. Signal processing algorithms must be developed that can eliminate such effects in an effective and efficient manner.

Three major types of signal processing technique have been used to enable speech recognition systems to achieve environmental robustness: acoustical pre-processing, such as codeword-dependent cepstral normalization (CDCN) that compensates for the effects of additive noise and linear filtering; physiologically-motivated processing, such as the computational models of the auditory periphery (*e.g.* [3,5,7]); and microphone-array processing, which includes adaptive processing techniques and the recent correlation-based algorithm [9].

In a previous study [8], we used Seneff's auditory model as an example of physiologically-motivated processing and evaluated the baseline performance of its 40 channel mean-rate and synchrony outputs for speech with or without degradations. We reported that the Seneff model provides a substantial amount of environmental robustness, while the most effective front-end processing technique among those examined was provided by the cepstral front-end in conjunction with CDCN.

Although the effects of physiologically-based auditory modelling and acoustical pre-processing have been analyzed in isolation, it is not known whether a combination of the two types of procedure can produce a greater degree of improvement in the recognition accuracy than can already by achieved by the use of either technique alone.

In this paper, we further investigate the effectiveness of physiologically-motivated processing. First, we search for the feature parameters of the Seneff model, which are more relevant to speech recognition than its raw outputs. Next, we evaluate the significance of the individual neural transduction (NT) components in terms of the robustness achieved. Finally, we propose a few ways of combining the Seneff model with CDCN in an attempt to determine whether the types of robustness provided by these two approaches are complementary.

The remainder of the paper is organized as follows. In Section 2, we describe the proposed experimental procedures, and in Section 3 report on the results of speech recognition experiments. In the last section, we summarize our findings and offer some concluding remarks.

2. EXPERIMENTAL METHODS

2.1. Software and Data Resources

We used a modified version of CMU's SPHINX-I speech recognition system. The LPC-based front end as described in [4] was replaced by a candidate physiologically-motivated front-end, which included several components to represent the BPF bank, non-linear half-wave rectification, short-term adaptation and rapid AGC stages, synchrony fall-off, and the generalized synchrony detector (GSD), as described in Seneff [7].

We also used the census database, which contains 1018 multi-speaker continuous alphanumeric utterances that are either random sequences or spelled-out addresses [1]. Utterances were recorded simultaneously in stereo, using the close-talking Sennheiser HMD-414 (CLSTK) microphone and the omnidirectional desktop Crown PZM6FS (CRPZM) microphone. The close-talking microphone speech was generally free from noise and is therefore regarded as "clean." In contrast, the PZM microphone speech was corrupted by additive noise, and also exhibited a different spectral tilt from the speech recorded by the close-talking microphone. This tilt is assumed to be a consequence of the combined effects of the linear filtering due to the microphone placement, room acoustics, and the differences in the transfer function of the microphones.

In all experiments, the system was trained on the CLSTK microphone speech and tested on both microphones. Testing data were provided from a disjoint subset of the corpus. White noise was used in experiment 2.3.

2.2. Feature extraction of the mean-rate and synchrony parameters

We performed principal component analysis [2] on the 40-channel mean-rate and synchrony outputs. The basis function of the Karhuenen-Loève transform, which diagonalized the corresponding covariance matrix of the CLSTK training data, was used to transform both CLSTK and CRPZM testing data so as to reduce correlation among channel outputs.

For the principal components of the mean-rate and synchrony outputs, we tried a number of different dimensions, from the original 40 down to 2, and searched for the minimum feature vector size necessary for successful classification performance by using a composite report of all the NIST benchmark tests for statistical significance.

2.3. Evaluation of the significance of individual NT components

The general procedure for this evaluation is to remove an individual component of the NT model from the sequence of front-end processing steps, and then to train and test the SPHINX system to observe the impact of eliminating that component. Short-term adaptation, automatic gain control (AGC), and low-pass filtering (LPF) for the effect of synchrony fall-off were the candidates for elimination.

2.4. Techniques for combining acoustical pre-processing and auditory modeling

We employed two approaches to combine acoustical pre-processing algorithms and Seneff's model: a *waveform-domain approach* and a *parameter-domain approach* [6].

Waveform-domain approach. Figure 1 shows a block diagram of the processing. The acoustical pre-processing methods proposed by Acero all operate in the cepstral domain. In the waveform-domain approach, speech is resynthesized from the sequences of cepstral coefficients normalized by pre-processing techniques such as CDCN. The synthesis filter is excited by the residual error signal saved in the LPC analysis procedure. The resynthesized speech waveform is then processed by the Seneff auditory model in the usual fashion.

To handle occasional problems involving instability of the resynthesis filter, we reflected poles outside the unit circle into the inside, in order to regain stability with no audible degradation while preserving the spectral envelope.

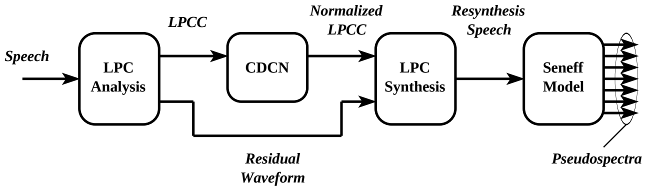

Figure 1: Waveform-domain integration of acoustical pre-processing and physiologically-motivated signal processing.

Parameter-domain approach. An alternative to the waveform-domain approach is to combine auditory modelling and acoustical pre-processing in the parameter domain. In this case, speech is first processed by the auditory model. The output features from the auditory processor (AP) are first converted to a form that is more like a cepstral representation of the incoming speech. At this point, cepstral normalization algorithms such as CDCN can be applied to the derived cepstra of the auditory model outputs. In Figure 2, cepstral parameters are derived by all-pole fitting of the Seneff model auditory spectra. Another way to derive cepstral parameters is to apply the IDCT directly to the Seneff model outputs. The resulting feature vectors resemble Bark auditory cepstral coefficients (BACC).

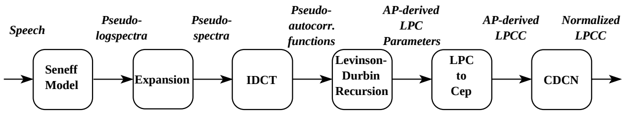

Figure 2: Derivation of cepstral parameters for the parameter-domain integration of the Seneff model and CDCN. The magnitude of pseudospectra from the Seneff model are expanded to compensate for the compressive nature of the model. The resulting spectra are fitted by all-pole modelling and the LPCC parameters are computed.

3. RESULTS

Feature Extraction. Figure 3 shows the results of feature extraction of the mean-rate, in which the recognition accuracy gradually dropped as the number of principal components was reduced. When the CLSTK microphone was used for both training and testing with the mean-rate principal component system, we found no statistically significant difference in the recognition performance for any number of principal components from the full set of 40 down to 4. There was also a small fluctuation among the best performing systems. For instance, the 6-component system was reported to be better than the 4-, 5-, 40-component systems in a matched-pairs test.

In the CRPZM microphone testing, performance degradation started when the 4-component system was used, and systems with fewer components than four were clearly inferior to the system with five or more principal components. We also found that the mean-rate principal components had an advantage over their raw-feature counterparts in that they improved the cross-microphone performance from 32.3% to 52.2%, while there was no performance degradation in the same microphone testing.

Owing to the limited space, we omit the plot for the GSD, which shows a very similar trend to the mean-rate case. We observed that the dimension was successfully reduced from 40 to 4 with no performance degradation in the CLSTK testing. In the CRPZM testing, however, the 10-component system was the most compact. The best performance of GSD principal components in the cross-microphone testing was 63.5%, using the 12-component system, as compared with the GSD baseline performance of 48.7%.

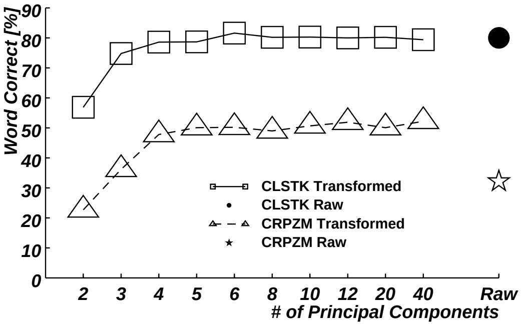

Figure 3: Dependence of the recognition accuracy on the number of principal components derived from the mean-rate outputs of the Seneff model. The upper curves were obtained by testing with the CLSTK microphone (boxes), and the lower curves were obtained by testing with the CRPZM microphone (triangles). Also plotted are the baseline recognition rates using the raw features of mean-rate outputs (a bullet and a star).

Evaluation of NT Components. The four panels of Figure 4 show the results obtained by disengaging the short-term adaptation. When the mean-rate parameters were used, we observed a significant drop in the recognition performance when the adaptation stage was eliminated, except for testing in the two least noisy conditions with the CLSTK microphone. We recognized two characteristic trends in the performance regardless of whether the mean-rate or the GSD outputs are used.

First, for SNRs of +20 dB or greater, the omission of short-term adaptation does not cause significant degradation in recognition performance when the CLSTK microphone is used for training and testing. Second, eliminating the adaptation stage results in an unacceptable performance degradation for all SNRs when the CRPZM microphone is used for testing. In summary, short-term adaptation is an important element in noisy conditions, but its value is not obvious for clean speech.

In Figure 5, it can be seen that removing the AGC did not adversely affect recognition accuracy for the most part when the mean-rate outputs were used, and that the trend was similarly true for the GSD outputs, except for the testing with the lowest SNR. The results suggest that the AGC could be omitted from the current implementation if only the mean-rate parameters are required.

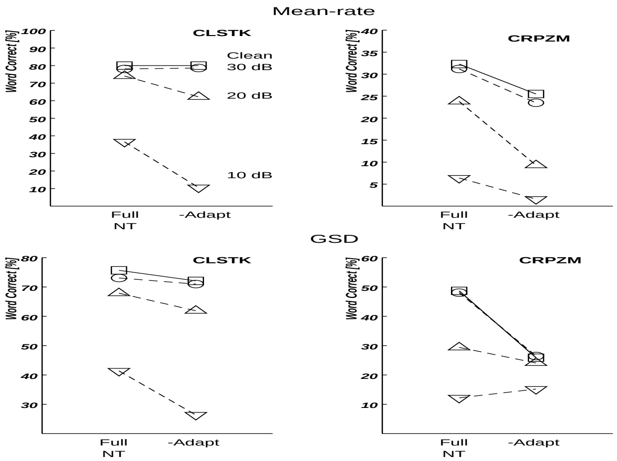

Figure 4: Evaluation of the NT components using the mean-rate and GSD outputs of the Seneff model, training with the CLSTK microphone and testing with the CLSTK and CRPZM microphones. Full NT denotes the original Seneff model, in which all the NT functions are enabled, while -Adapt denotes that the short-term adaptation NT component is disengaged.

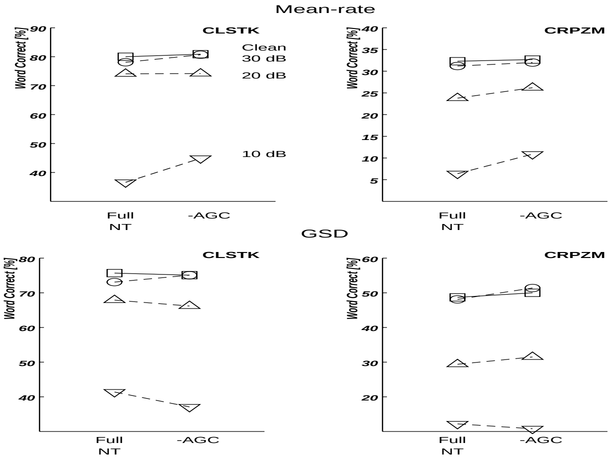

Figure 5: Evaluation of the NT components using the mean-rate and GSD outputs of the Seneff model, training with the CLSTK microphone and testing with the CLSTK and CRPZM microphones. Full NT denotes the original Seneff model, in which all the NT functions are enabled, while -AGC denotes that the AGC NT component is disengaged.

The four panels of Figure 6 show the results from two sets of experiments involving the LPF. We noticed that eliminating the LPF produces no degradation in the recognition accuracy when the CLSTK microphone is used for testing, for either the mean-rate or the GSD outputs. On the other hand, when the CRPZM microphone is used for testing, there is a 2% drop in recognition accuracy when the mean-rate outputs are used, and a greater difference when the GSD outputs are used.We found that better results were often obtained by not using the AGC, although the performance gain may be marginal. The modified mean-rate front end without the use of the LPF performed just as well as the original full NT version did. On the other hand, the lack of short-term adaptation had a severe negative impact. In order to determine the extent to which short-term adaptation component is the single dominant component of the NT model after the rectifier, we disabled both the AGC and LPF, leaving in place only the rectifier and the short-term adaptation. As with some of the previous results, there was no significant degradation in performance when the CLSTK microphone was used for testing. However, eliminating both the AGC and LPF produced a significant degradation in the recognition accuracy both when the CRPZM was used for testing and when the CLSTK microphone was used for testing with noise added at an SNR of +20 dB or below.

For these reasons it was concluded that short-term adaptation by itself is not sufficient to provide good recognition accuracy in adverse conditions.

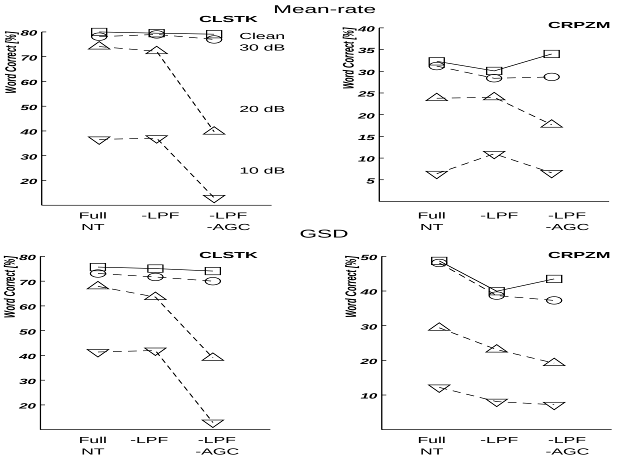

Figure 6: Evaluation of the NT components using the mean-rate and GSD outputs of the Seneff model, training with the CLSTK microphone and testing with the CLSTK and CRPZM microphones. Full NT denotes the original Seneff model, in which all the NT functions are enabled, while -AGC and -LPF respectively denote that the AGC and synchrony fall-off NT components are disengaged.

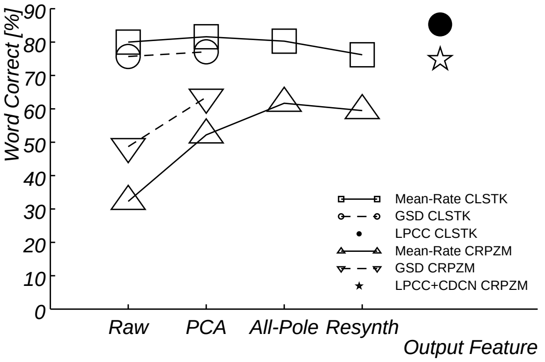

Figure 7: Combination of the Seneff model with CDCN. Plotted are the recognition accuracies for the waveform-domain approach (Resynth) and the parameter-domain approach (All-Pole). Also plotted are the accuracies based on the principal component analysis (PCA) and the raw output features (Raw) of the Seneff model.

Combination of the Seneff model with CDCN. Figure 7 summarizes the results of a series of experiments using the waveform-domain and parameter-domain approaches. We observed that both approaches produced better results than those based on the mean-rate principal components, and that the results were comparable with those based on the synchrony principal components. Nevertheless, we found that the best-performing front-end is still the conventional cepstral parameters with CDCN processing.

Specifically, in the waveform-domain approach, the recognition accuracy based on the mean-rate parameters for the resynthesized CLSTK microphone data was 76.2%, while the accuracy for the resynthesized CRPZM was 59.5%. In the parameter-domain approach, the derived cepstral parameters of the mean-rate output were normalized by applying CDCN and marked 80.1% for the CLSTK microphone testing and 61.7% for the CRPZM testing.

4. SUMMARY

We found that both the mean-rate and synchrony outputs of the Seneff model provide better recognition accuracy than conventional signal processing using the LPC cepstral coefficients, when speech is subjected to additive noise and linear filtering. Although there are 40 frequency-specific outputs in the Seneff model, we found that no loss in recognition accuracy is incurred if classification decisions are made on the basis of five principal components of the mean-rate outputs and 10 principal components of the synchrony outputs.

We found that short-term adaptation was the most important component of the neural transduction stage of the Seneff model.

We developed several ways of combining auditory processing with environmental normalization techniques such as CDCN in both waveform and cepstral domains. We showed that the recognition accuracy provided by physiologically-motivated signal processing can be further improved by combining this with environmentally normalized cepstral processing. Both approaches improved the recognition accuracy based on the mean-rate parameters up to about 60% in the CRPZM testing, but neither outperformed the application of CDCN to conventional cepstral processing.

REFERENCES

1. Acero, A., “Acoustical and Environmental Robustness in Automatic Speech Recognition,” *Ph.D. Thesis*, Carnegie Mellon University, 1990.

2. Duda, R.O. and Hart, P.E., “Pattern Classification and Scene Analysis,” John Wiley and Sons, 1973.

3. Ghitza, O., “Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment,” *Computer Speech and Language*, Vol. 2, No.1, 1987, pp. 109-130.

4. Lee, K.-F., “Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System,” *Ph.D. Thesis*, Carnegie Mellon University, 1988.

5. Lyon, R. F.,”A Computational Model of Filtering, Detection, and Compression in the Cochlea,” *ICASSP-82*, 1982, pp. 1282-1285.

6. Ohshima Y., “Environmental Robustness in Speech Recognition Using Physiologically-Motivated Signal Processing”, *Ph.D. Thesis,* Carnegie Mellon University, 1993.

7. Seneff, S., “A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing,” *J. Phonetics*, Vol. 16, 1988, pp. 55-76.

8. Stern, R.M., Liu, F.-H., Ohshima, Y., Sullivan, T.M., and Acero, A., “Multiple Approaches to Robust Speech Recognition,” *DARPA Speech and Natural Language Workshop*, February, 1992, pp. 274-279.

9. Sullivan, T.M. and Stern, R.M., “Multi-Microphone Correlation-Based Processing for Robust Speech Recognition,” *ICASSP-93*, 1993, pp. II-91-94.