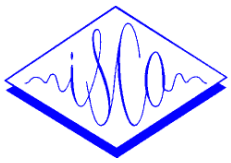

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

MULTIPLE APPROACHES TO ROBUST SPEECH RECOGNITION

*Richard M. Stern, Fu-Hua Liu, Yoshiaki Ohshima, Thomas M. Sullivan, Alejandro Acero**

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

This paper compares several different approaches to robust speech recognition. We review CMU's ongoing research in the use of acoustical pre-processing to achieve robust speech recognition, including the first evaluation of pre-processing in the context of the DARPA standard ATIS domain for spoken language systems. We also describe and compare the effectiveness of three complementary methods of signal processing for robust speech recognition: acoustical pre-processing, microphone array processing, and the use of physiologically-motivated models of peripheral signal processing. Recognition error rates are presented using these three approaches in isolation and in combination with each other for the speaker-independent continuous alphanumeric census speech recognition task.

1. INTRODUCTION

The need for speech recognition systems and spoken language systems to be robust with respect to their acoustical environment has become more widely appreciated in recent years (*e.g.* [1]).

Results of several studies have demonstrated that even automatic speech recognition systems that are designed to be speaker independent can perform very poorly when they are tested using a different type of microphone or acoustical environment from the one with which they were trained (*e.g.* [2, 3]), even in a relatively quiet office environment. The use of microphones other than a "close-talking" headset also tends to severely degrades speech recognition performance. Even in a relatively quiet office environment there is significant additive noise from fans, door slams, and other conversations, as well as reverberations arising from surface reflections in the room. Applications such as speech recognition over telephones, in automobiles, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

The CMU speech group is committed to the development of speech recognition systems that are robust with respect to environmental variation, just as it has been an early proponent of speaker-independent recognition. While most of our work presented to date has described new *acoustical pre-processing* algorithms (*e.g.* [2, 4, 5], we have always regarded pre-processing as one of several approaches that must be developed in concert to achieve robust recognition.

In this paper we first review some of our most recent algorithms and results using acoustical pre-processing. We then describe and compare the effectiveness of three complementary methods of signal processing for robust speech recognition: acoustical pre-processing, microphone array processing, and the use of physiologically-motivated models of peripheral signal processing.

2. ACOUSTICAL PRE-PROCESSING

We have found that two major factors degrading the performance of speech recognition systems using desktop microphones in normal office environments are additive noise and unknown linear filtering. We showed in [2, 6] that simultaneous *joint* compensation for the effects of additive noise and linear filtering is needed to achieve maximal robustness with respect to acoustical differences between the training and testing environments of a speech recognition system. We described in [2, 6] two algorithms that perform such joint compensation, based on additive corrections to the cepstral coefficients of the speech waveform. The more effective and adaptive of these algorithms, *Codeword-Dependent Cepstral Normalization* (CDCN) [2], uses EM techniques to compute ML estimates of the additive noise and linear filtering that corrupt "clean" speech signals. The CDCN algorithm adapts automatically to new testing environments, using structural knowledge about the nature of the degradations to the speech signal to achieve good recognition accuracy.

More recently we developed several new algorithms which combine the environmental independence of CDCN with greater computational simplicity. One such algorithm is the *Blind SNR-Dependent Cepstral Normalization* (BSDCN) algorithm [5]. The BSDCN algorithm estimates compensation parameters by computing average cepstra at each signal-to-noise ratio (SNR) in the training and testing environments. A correspondence is established between the SNRs in the training and testing environments by use of traditional nonlinear warping technique on histograms of SNRs from each of the two environments. The compensation vectors of BSDCN are the differences between average cepstra in the training testing environment for each SNR pair that is matched by the warping algorithm.

Experimental comparisons of CDCN and BSDCN. Figure 1 compares the error rate obtained when the original discrete-HMM SPHINX system is trained using the DARPA standard HMD-414 closetalking microphone (CLSTLK), and tested using either the CLSTLK microphone or the omnidirectional desktop Crown PZM-6FS microphone (PZM6FS). The census database was used, which contains simultaneous recordings of speech from the CLSTLK and PZM6FS microphones in the context of a speaker-independent continuous-speech alphanumeric task with perplexity 65, recorded in an office environment [2]. Our previous analyses of this database indicate that the PZM6FS data are corrupted by linear filtering as well as by additive noise. The horizontal dotted lines indicate the recognition accuracy obtained when the system is tested on the microphone with which it was trained, with no processing. The intersection of the upper curve with the upper horizontal line indicates that with CDCN compensation, SPHINX can recognize speech using the PZM6FS microphone just as well when trained on the CLSTLK microphone as when trained using the PZM6FS.

*Present address: Telefónica Investigación y Desarrollo, Emilio Vargas 6, Madrid 28043, Spain

In the lower panel of Fig. 1 we summarize recent results obtained using data from the February, 1992, ATIS-domain robust-speech evaluation. For this evaluation, the semicontinuous-HMM SPHINX-II recognizer was trained using the CLSTLK microphone, and tested using both the CLSTLK microphone and the unidirectional Crown PCC-160 microphone (PCC160). In each case, the system was not provided with explicit knowledge of the identity of the environment within which it is operating. Using the CDCN algorithm causes the error rate to increase from 15.1% to only 20.4% as the testing microphone is changed from the CLSTLK to the PCC160 microphone. In contrast, the error rate increases from 12.2% to 38.8% when one switches from the CLSTLK to the PCC160 microphone without CDCN. (Formal results were not obtained using BSDCN on the ATIS task with the CLSTLK microphone, but the error rate should be comparable to that obtained with no processing.)

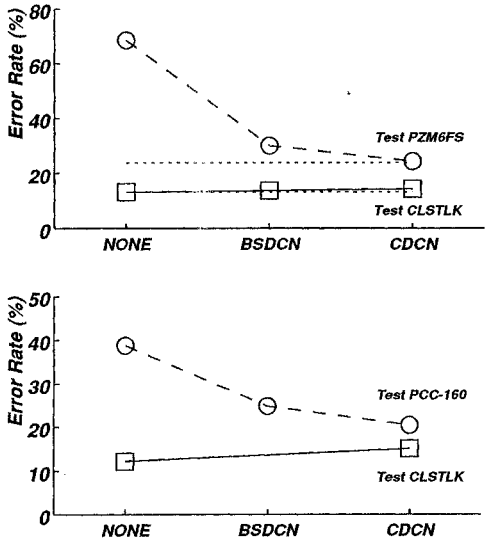

Figure 1: Comparison of error rates obtained on the census task (upper panel) and the DARPA ATIS task (lower panel) with no pre-processing, the BSDCN algorithm, and the CDCN algorithm. The system was trained using the CLSTLK microphone in all cases. For the census task SPHINX was tested using either the CLSTLK microphone (solid curve) or the PZM6FS microphone (broken curve). For the DARPA ATIS task SPHINX-II was tested using either the CLSTLK microphone (solid curve) or the cardiod desktop Crown PCC160 microphone (broken curve).

The BSDCN algorithm is much simpler, and it compensates speech approximately 80 times faster than the CDCN algorithm, but it produces error rates that are about 20% worse than those of CDCN for both the census and ATIS tasks. We also found that recognition accuracy using CDCN converges with only 2 seconds of speech in the test environment for best performance, while BSDCN requires about 70 seconds of speech [5]. We believe that the CDCN algorithm converges more rapidly because it imposes more structure on the compensation process (from knowledge of how speech is likely to be degraded), while the BSDCN algorithm is purely data driven.

3. MICROPHONE ARRAYS AND ACOUSTICAL PRE-PROCESSING

Despite the encouraging results that we have achieved using acoustical pre-processing, we believe that further improvements in recognition accuracy can be obtained in difficult environments by combining acoustical pre-processing with other complementary types of signal processing. The use of microphone arrays is motivated by a desire to improve the effective SNR of speech as it is input to the recognition system. For example, the headset-mounted CLSTLK microphone produces a higher SNR than the PZM6FS microphone under normal circumstances because it picks up a relatively small amount of additive noise. Furthermore, the speech signal is not degraded by room reverberation when the CLSTLK microphone is used.

In error analyses of the original census data we found that a large fraction of errors observed while training and testing using the PZM6FS were caused by the confusion of silence or noise segments with weak phonetic events. Microphone arrays can, in principle, produce directionally-sensitive gain patterns that can be adjusted to produce maximal sensitivity in the direction of the speaker and reduced sensitivity in the direction of competing sound sources. By increasing the SNR to the input of the speech recognition system, microphone arrays can produce a substantial decrease in error rate by reducing the number of confusions between weak phonetic events and noise.

Several different types of array-processing strategies have been applied to automatic speech recognition. The simplest approach is that of the delay-and-sum beamformer, in which delays are inserted in each channel to compensate for differences in travel time between the desired sound source and the various sensors (*e.g.* [7, 8]). A second option is to use an adaptation algorithm based on minimizing mean square energy such as the Frost or Griffiths-Jim algorithm [9]. These algorithms can provide nulls in the direction of noise sources as well as more sharply focused beam patterns, but they assume that the desired signal is statistically independent of all sources of degradation. Consequently, these algorithms can provide good improvement in SNR when signal degradations are caused by additive independent noise sources, but they do not perform well in reverberant environments when the distortion is at least in part a delayed version of the desired speech signal [10, 11]. (This problem can be avoided by only adapting during non-speech segments [12]). A third type of approach to microphone array processing involves the cross-correlation of outputs from different sensors, based on the processing of the human binaural system (*e.g.* [13]). While cross-correlation is an efficient way to identify the direction of a strong signal source, the nonlinear nature of the cross-correlation operation renders it inappropriate as a means to directly process waveforms.

Pilot evaluation of the Flanagan array. In order to obtain a better understanding of the ability of array processing to provide further improvements in recognition accuracy we conducted a pilot evaluation of the 23-microphone array developed by Flanagan and his colleagues at AT&T Bell Laboratories [7, 8]. The microphones of this one-dimensional delay-and-sum beamformer are unevenly spaced in order to provide a beamwidth that is approximately constant over the range of frequencies of interest. First-order gradient microphones are used, which develop a null response in the vertical plane. We compared the recognition accuracy on the census task obtained using the Flanagan array with the accuracy observed using the CLSTLK and PZM6FS microphones, with particular interest in determining the extent to which array processing provides an improvement in recognition accuracy that is complementary to the improvement in accuracy provided by acoustical pre-processing algorithms such as the CDCN algorithm.

14 utterances from the census database were obtained from each of five male speakers in a sparsely-furnished laboratory at the Rutgers University CAIP Center with hard walls and floors. The reverberation time of this room was informally estimated to be between 500 and 750 ms. Simultaneous recordings were made of each ut-

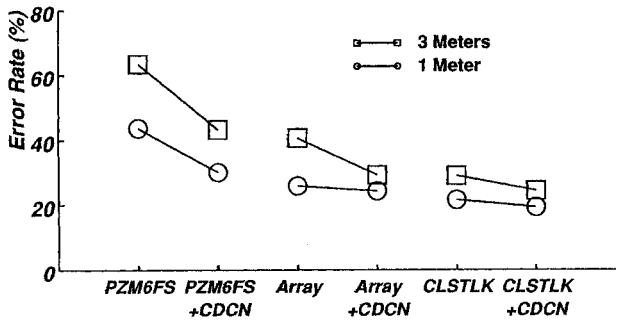

Figure 2: Comparison of error rates obtained on a portion of the census task using the omnidirectional desktop PZM6FS, the 23-microphone array developed by Flanagan, and the CLSTLK microphone, each with and without CDCN. Data were obtained from simultaneous recordings using the three microphones at distances of 1 and 3 meters (for the PZM6FS and the array).

terance using three microphones: the CLSTLK microphone, the PZM6FS, and the Flanagan array with input lowpass-filtered at 8 kHz. Recordings were made with the speaker seated at distances of 1, 2, and 3 meters from the PZM6FS and Flanagan array microphones, wearing the CLSTLK microphone in the usual fashion at all times.

Figure 2 summarizes the error rates obtained from these speech samples at two distances (1 and 3 meters) and three microphones (CLSTLK, PZM6FS, and the Flanagan array), with and without the CDCN algorithm. Error rates using the CLSTLK microphone differed somewhat for the two distances because different speech samples were obtained at each distance and because the sample size is small. (It is also possible that speakers may have spoken less naturally when the recording microphone is placed at the 3-meter distance.) The SPHINX system had been previously trained on speech obtained using the CLSTLK microphone. As expected, the worst results were obtained using the PZM6FS microphone, while the lowest error rate was obtained for speech recorded using the CLSTLK. More interestingly, the results in Fig. 2 show that both the Flanagan array and the CDCN algorithm are effective in reducing the error rate, and that in fact the error rate at each distance obtained with the combination of the two is very close to the error rate obtained with the CLSTLK microphone and no environmental compensation. The complementary nature of the improvement of the Flanagan array and the CDCN algorithm is indicated by the fact that adding CDCN to the array improves the error rate (comparing columns 3 and 4 of Fig. 2), and that converting to the array even when CDCN is already employed also improves performance (comparing columns 2 and 4 of Fig. 2). We believe that this is observed because the array improves the SNR under all circumstances, and the CDCN algorithm compensates for mismatches between the training environment (using the CLSTLK microphone) and the testing environment (using the array).

4. PHYSIOLOGICALLY-MOTIVATED FRONT ENDS AND ACOUSTICAL PRE-PROCESSING

In recent years there has also been an increased interest in the use of peripheral signal processing schemes that are motivated by human auditory physiology and perception, and a number of such schemes have been proposed (*e.g.* [14, 15, 16, 17]). Most auditory models include a set of linear bandpass filters with bandwidth that increases nonlinearly with center frequency, a nonlinear rectification stage that frequently includes short-term adaptation and lateral suppression, and, in some cases, a more central display based on short-term temporal information. Recent evaluations indicate that with "clean" speech, such approaches tend to provide recognition accuracy that is comparable to that obtained with conventional LPC-based or DFT-based signal processing schemes, but that these auditory models can provide greater robustness with respect to environmental changes when the quality of the incoming speech (or the extent to which it resembles speech used in training the system) decreases [18, 19]. Despite the apparent utility of such processing schemes, no one has a deep-level understanding of why they work as well as they do, and in fact different researchers choose to emphasize rather different aspects of the peripheral auditory system's response to sound in their work. We estimate that the number of arithmetic operations of some of the currently-popular auditory models ranges from 35 to 600 times the number of operations required for the LPC-based processing used in the CMU speech recognition systems.

Pilot evalution of the Seneff auditory model. We recently completed a series of pilot evaluations using an implementation of the Seneff auditory model [17] and the census database. While almost all evaluations of physiologically-motivated front ends to date have been performed using artificially-added white Gaussian noise, we have been interested in the extent to which auditory models are helpful in recognizing speech that has been degraded by reverberation or other types of linear filtering. As in the case of microphone arrays, we are also interested in determining the extent to which improvements in robustness provided by auditory modelling complement those that we already enjoy by the use of acoustical pre-processing algorithms such as CDCN.

We compared error rates obtained using the standard 12 LPC-based cepstral coefficients normally input to the SPHINX system, with and without CDCN, with those obtained using an implementation of the 40-channel mean-rate output of the Seneff model [17], and with the 40-channel outputs of Seneff's Generalized Synchrony Detectors (GSDs). The system was evaluated using the original testing database from the census task with the CLSTLK and PZM6FS microphones, and also with white Gaussian noise artificially added at SNRs of +10, +20, and +30 dB, measured using the global SNR method described in [15].

Figure 3 summarizes the results of these comparisons, with error rate plotted as a function of SNR using each of the three peripheral signal processing schemes. The upper panel describes recognition error rates obtained with the system both trained and tested using the CLSTLK microphone, and the lower panel describes error rates obtained with the system trained using the CLSTLK microphone but tested using the PZM6FS microphone. When the system is trained and tested using the CLSTLK microphone, best performance is obtained using conventional LPC-based signal processing for "clean" speech. As the SNR is decreased, however, error rates obtained using either the mean rate or GSD outputs of the Seneff model degrade more gradually confirming similar findings from previous studies. The results in the lower panel of Fig. 3, demonstrate that the mean rate and GSD outputs of the Seneff model provide lower error rates than conventional LPC cepstra when the system is trained using the CLSTLK microphone and tested using the PZM6FS. Nevertheless, the combination of conventional LPC-based processing and the CDCN algorithm produced performance that equaled or bettered the best performance obtained with the auditory model for each test condition, as indicated by the lower set of dashed curves in each panel of Fig. 3. We have also explored several ways of combining CDCN with the outputs of the auditory model, but so far we have not been able to obtain a better error rate than the error rate observed using CDCN alone. Because the auditory model is nonlinear and not easily ported from one site to another, these comparisons should all

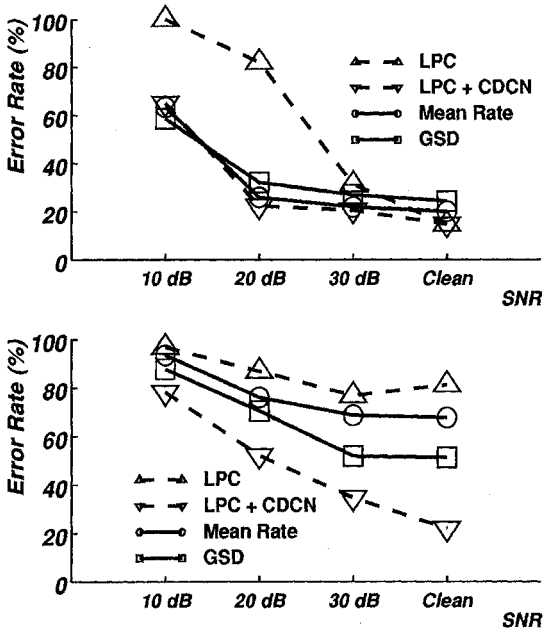

Figure 3: Pilot data comparing error rates obtained on the census task using the conventional LPC-based processing of SPHINX with results obtained using the mean rate and synchrony outputs of the Seneff auditory model. SPHINX was trained on the CLSTLK microphone in all cases, and tested using either the CLSTLK microphone (upper panel) or the Crown PZM6FS microphone (lower panel). White noise was artificially added to the speech signals and data are plotted as a function of global SNR.

be regarded as preliminary. It is quite possible that performance using the auditory model could further improve if greater attention were paid to tuning it to more closely match the characteristics of SPHINX.

5. SUMMARY AND CONCLUSIONS

In this paper we summarize our current research in acoustical pre-processing for robust speech recognition, and we describe our first attempts to integrate pre-processing with other approaches to robust speech recognition. We found that the CDCN and BSDCN algorithms provide significant error reduction for the DARPA ATIS task using the unidirectional desktop PCC160 microphone. We demonstrated that the CDCN algorithm and the Flanagan delay-and-sum microphone array can provide complementary benefits to speech recognition in reverberant environments. Although the Seneff auditory model improves the recognition accuracy of the CMU speech system in reverberant as well as in noisy environments, results were no better than those obtained with CDCN alone.

ACKNOWLEDGMENTS

This research is sponsored by the Defense Advanced Research Projects Agency, DoD, through ARPA Order 7239, and monitored by the Space and Naval Warfare Systems Command under contract N00039-91-C-0158. Views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or of the United States Government. We thank Hsiao-Wuen Hon, Xuedong Huang, Kai-Fu Lee, Raj Reddy, Eric Thayer, Bob Weide, and the rest of the speech group for their contributions to this work. We also thank Jim Flanagan, Joe French, Greg Legowski, and A. C. Surendran for their assistance in obtaining the experimental data using the array microphone, and Stephanie Seneff for providing source code for her auditory model. The graduate studies of Tom Sullivan and Yoshiaki Ohshima have been supported by Motorola and IBM Japan, respectively.

REFERENCES

1. Juang, B. H., "Speech Recognition in Adverse Envoronments", *Comp. Speech and Lang.*, Vol. 5, 1991, pp. 275-294.
2. Acero, A. and Stern, R. M., "Environmental Robustness in Automatic Speech Recognition", *ICASSP-90*, April 1990, pp. 849-852.
3. Erell, A. and Weintraub, M., "Estimation Using Log-Spectral-Distance Criterion for Noise-Robust Speech Recognition", *ICASSP-90*, April 1990, pp. 853-856.
4. Acero, A. and Stern, R. M., "Robust Speech Recognition by Normalization of the Acoustic Space", *ICASSP-91*, May 1991, pp. 893-896.
5. Liu, F.-H., Acero, A., and Stern, R. M., "Efficient Joint Compensation of Speech for the Effects of Additive Noise and Linear Filtering", *ICASSP-92*, March 1992, pp. 865-868.
6. Acero, A. and Stern, R. M., "Acoustical Pre-Processing for Robust Spoken Language Systems", *ICSLP-90*, November 1990, pp. 1121-1124.
7. Flanagan, J. L., Johnston, J. D., Zahn, R., and Elko, G.W., "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", *J. Acoust. Soc. Amer.*, Vol. 78, Nov. 1985, pp. 1508-1518.
8. Flanagan, J. L., Berkeley, D. A., Elko, G. W., West, J. E., and Sondhi, M. M., "Autodirective microphone systems", *Acustica*, Vol. 73, February 1991, pp. 58-71.
9. Widrow, B., and Stearns, S. D., *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1985.
10. Peterson, P. M., "Adaptive Array Processing for Multiple Microphone Hearing Aids". RLE TR No. 541, Res. Lab. of Electronics, MIT, Cambridge, MA
11. Alvarado, V. M., Silverman, H. F., "Experimental Results Showing the Effects of Optimal Spacing Between Elements of a Linear Microphone Array", *ICASSP-90*, April 1990, pp. 837-840.
12. Van Compernolle, D., "Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", *ICASSP-90*, April 1990, pp. 833-836.
13. Lyon, R. F., "A Computational Model of Binaural Localization and Separation", *ICASSP-83*, 1983, pp. 1148-1151.
14. Cohen, J. R., "Application of an Auditory Model to Speech Recognition", *J. Acoust. Soc. Amer.*, Vol. 85, No. 6, June 1989, pp. 2623-2629.
15. Ghitza, O., "Auditory Nerve Representation as a Front-End for Speech Recognition in a Noisy Environment", *Comp. Speech and Lang.*, Vol. 1, 1986, pp. 109-130.
16. Lyon, R. F., "A Computational Model of Filtering, Detection, and Compression in the Cochlea", *ICASSP-82*, May 1982, pp. 1282-1285.
17. Seneff, S., "A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing", *Journal of Phonetics*, Vol. 16, No. 1, January 1988, pp. 55-76.
18. Hunt, M., "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", *ICASSP*, May 1989.
19. Meng, H., and Zue, V. W., "A Comparative Study of Acoustic Representations of Speech for Vowel Classification Using Multi-Layer Perceptrons", *ICSLP-90*, November 1990, pp. 1053-1056.