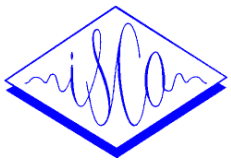

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SIGNAL PROCESSING FOR ROBUST SPEECH RECOGNITION

*Richard M. Stern, Fu-Hua Liu, Pedro J. Moreno, and Alejandro Acero**

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

ABSTRACT

This paper describes several new cepstral-based compensation procedures that render the SPHINX-II system more robust with respect to acoustical environment. The first algorithm, phone-dependent cepstral compensation, is similar in concept to the previously-described MFCDCN method, except that cepstral compensation vectors are selected according to the current phonetic hypothesis, rather than on the basis of SNR or VQ codeword identity. We also describe two procedures to accomplish adaptation of the VQ codebook for new environments. Use of the various compensation algorithms in consort produces a reduction of error rates for SPHINX-II by as much as 40 percent relative to the rate achieved with cepstral mean normalization alone.

1. INTRODUCTION

A continuing problem with current speech recognition technology is the lack of robustness with respect to environmental variability. For example, the use of microphones other than the ARPA standard Sennheiser HMD-414 "close-talking" headset (CLSTLK) severely degrades the performance of systems like the original SPHINX system, even in a relatively quiet office environment [*e.g.* 1, 2]. Applications such as speech recognition in automobiles, over telephones, on a factory floor, or outdoors demand an even greater degree of environmental robustness.

In this paper we describe and compare the performance of a series of cepstrum-based procedures that enable the CMU SPHINX-II [3] speech recognition system to maintain a high level of recognition accuracy over a wide variety of acoustical environments. In previous years we described the performance of cepstral mapping procedures such as the CDCN algorithm, which is effective but fairly computationally costly [2]. More recently we discussed the use of cepstral highpass-filtering algorithms, such as the popular RASTA and cepstral-mean-normalization algorithms (CMN) [4]. These algorithms are very simple to implement but somewhat limited in effectiveness. CMN, in which the mean of the cepstral vectors is subtracted on a frame-by-frame basis before recognition, is now a component of routine baseline processing for the CMU SPHINX-II system and for many other systems.

In this paper we describe several new procedures that when used in consort can provide as much as an additional 40 percent improvement over baseline processing with CMN. These techniques include phone-dependent cepstral compensation, environmental interpolation of compensation vectors, and codebook adaptation. In Sec. 2 we describe the various compensation procedures in detail, and we examine their effect on recognition accuracy in Sec. 3.

* Currently at Microsoft Corporation

2. ENVIRONMENTAL COMPENSATION ALGORITHMS

We begin this section by reviewing the previously-described MFCDCN algorithm. We then discuss blind environment selection and environmental interpolation as they apply to MFCDCN. Finally, the complementary procedures of phone-dependent cepstral normalization and codebook adaptation are described.

2.1. Multiple Fixed Codeword-Dependent Cepstral Normalization (MFCDCN)

Multiple fixed codeword-dependent cepstral normalization (MFCDCN) provides additive cepstral compensation vectors that depend on signal-to-noise ratio (SNR) and that also vary from codeword to codeword of the vector-quantized (VQ) representation of the incoming speech at each SNR [4]. At low SNRs these vectors primarily compensate for effects of additive noise. At higher SNRs, the algorithm compensates for linear filtering, while at intermediate SNRs, they compensate for both of these effects. Environmental independence is provided by computing compensation vectors for a number of different environments and selecting the compensation environment that results in minimal residual VQ distortion.

Figure 1 illustrates some typical compensation vectors obtained with the FCDCN algorithm, computed using the ARPA standard close-talking Sennheiser HMD-414 microphone and the unidirectional desktop PCC-160 microphone used as the testing environment. The vectors are computed at the extreme SNRs of 0 and 29 dB, as well as at 5 dB. The horizontal axis represents frequency, warped nonlinearly according to the mel scale, with a maximum frequency of 8000 Hz. We note that the spectral profile of the compensation vector varies with SNR, and that especially for the intermediate SNRs the various VQ clusters require compensation vectors of different spectral shapes.

2.2. Phone-Dependent Cepstral Normalization (PDCN)

It is also possible to select additive cepstral compensation vectors on the basis of the current phoneme hypothesis in the search process, rather than according to physical parameters such as SNR or VQ codeword identity as in MFCDCN. This approach to

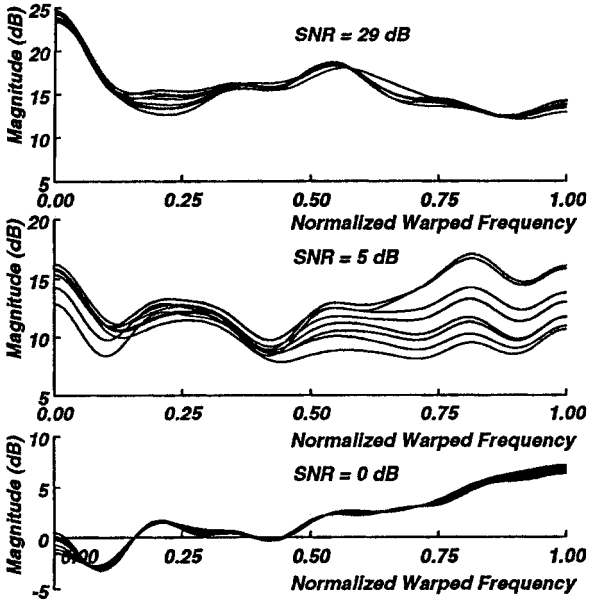

Figure 1. Comparison of compensation vectors using the FCDCN method with the PCC-160 unidirectional desktop microphone, at three different signal-to-noise ratios. The maximum SNR used by the FCDCN algorithm is 29 dB.

environmental compensation is referred to as *phone-dependent cepstral normalization* (PDCN), and is described in this section.

Estimation of PDCN compensation vectors. In the current implementation of PDCN, we develop compensation vectors that are specific to individual phonetical events, using a base phone set of 51 phonemes, including silence but excluding other types of non-lexical events. This is accomplished by running the decoder in supervised mode using CLSTLK data and correction transcriptions. All CLSTLK utterances are divided into phonetic segments. For every phonetic label, a difference vector is computed by accumulating the difference between the cepstral vector in a given frame in the CLSTLK training data, and its counterpart in the secondary environment.

Figure 2 illustrates some typical compensation vectors obtained with the PDCN algorithm, again computed using the CLSTLK microphone, with the unidirectional desktop PCC-160 microphone used as the testing environment. Typical compensation vectors are shown for three front vowels, three nasals, and three fricatives. The compensation vectors for the various types of phonemes show systematic differences in shape that are related to phoneme type.

Application of PDCN compensation in recognition. The SPHINX-II system uses the senone [3], a generalized state-based probability density function, as the basic unit to compute the likelihood from acoustical models. Multiple compensated cepstral vectors are formed in PDCN by adding various compensation vectors such as those depicted in Fig. 2, to the incoming cepstra. The compensation vectors are selected frame by frame, based on the presumed phoneme identity. The amount of computation needed for this procedure is reduced because in SPHINX-II, each senone corresponds to only one distinctive base phoneme. Each cepstral vector is normalized with a PDCN compensation vector that corresponds to its base phonetic identity.

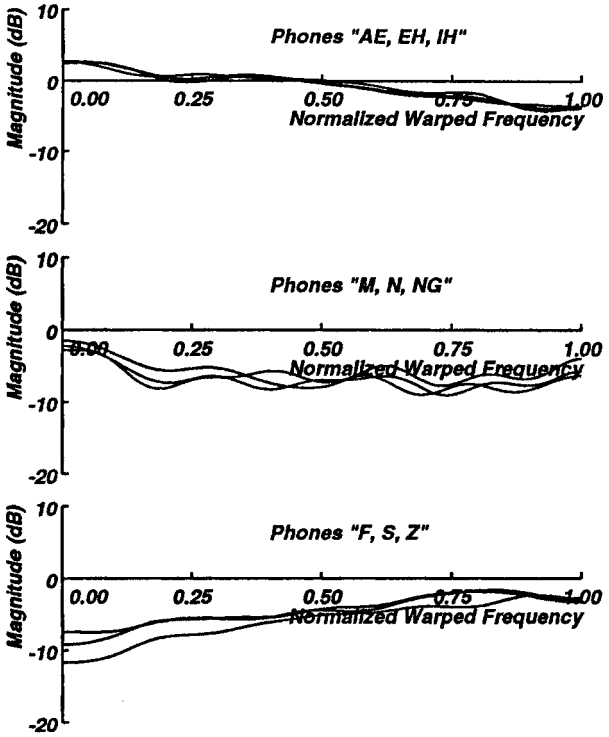

Figure 2. Comparison of compensation vectors used in the PDCN method with the PCC-160 unidirectional desktop microphone, for (from top to bottom) three front vowels, three nasals, and three fricatives.

2.3. Blind Environment Selection

A number of ensembles of compensation vectors in MFCDCN and PDCN are compiled for a number of environments, of which one must be selected for the compensation process. We considered two procedures for environment selection.

The first procedure, referred to as *selection by compensation*, applies compensation vectors from each possible environment successively to the incoming test utterance. The environment is chosen that minimizes the average residual VQ distortion over the entire utterance. The second procedure, referred to as the *Gaussian environment classifier*, models each environment with mixtures of Gaussian densities. The chosen environment is the one that maximizes the probability of observing the incoming cepstra. This approach is similar to one proposed previously by BBN [5]. The two methods produce similar speech recognition accuracy for most domains of interest.

2.4. Environmental Interpolation

In cases where the testing environment does not closely resemble any particular environment used to develop compensation parameters for MFCDCN or PDCN, interpolating the compensation vectors of several environments can be more helpful than using compensation vectors from a single (incorrect) environment. We refer to interpolated versions of the MFCDCN and PDCN algorithms as IMFCDCN and IPDCN, respectively. In both cases, compensation vectors for new environments are obtained by linear interpolation of several of the MFCDCN compensation vectors, respectively. The weighting factors for each environment are set to equal the probability that that environment is present given the observed incoming cepstral vectors. In the work described in this paper, interpolation was generally carried out over the best three environments. In the case of IPDCN, the interpolation made use of the closest four Gaussian mixtures.

2.5. Codebook Adaptation (DCCA and BWCA)

Compensation procedures like MFCDCN and PDCN apply additive corrections to incoming vectors. An alternative approach is to use information about environmental differences to modify the internal templates to which these incoming feature vectors are compared, the means and variances of the vector quantization (VQ) codebook. In this section we discuss two approaches to *codebook adaptation*, in which we modify the mean vectors and/or covariance matrices of the VQ codebooks in order to compensate for acoustical differences between training and testing environments.

Dual-Channel Codebook Adaptation (DCCA). *Dual-Channel Codebook Adaptation (DCCA)* exploits the existence of speech that is simultaneously recorded using the CLSTLK microphone and a number of secondary microphones. This information is used to modify the means and variances of the mixture densities that comprise the probability density functions for the senones used in SPHINX-II. Specifically, VQ encoding is performed on speech from the CLSTLK microphone processed with CMN. The output VQ labels are shared by the CLSTLK data and the corresponding data in the secondary (or target) environment. For each subspace in the CLSTLK training environment, we generate the corresponding means and variances for the target environment. Thus, a one-to-one mapping between the means and variances of the cepstral space of the CLSTLK training condition and that of the target condition is established.

Baum-Welch Codebook Adaptation (BWCA). There are many applications in which stereo data simultaneously recorded in the CLSTLK and target environments are unavailable. In these circumstances, transformations can be developed between environments using the adaptation utterances in conjunction with the Baum-Welch algorithm.

In Baum-Welch codebook adaptation, mean vectors and covariance matrices, along with senones, are re-estimated and updated using the Baum-Welch algorithm [6] during each iteration of training process. To compensate for the effect of changes in acoustical environments, the Baum-Welch approach is used to transform the means and covariances toward the cepstral space of the target testing environments. This is exactly like baseline training, except that only a few adaptation utterances are available, and that number of free parameters to be estimated (*i.e.* the means and variances of the VQ codewords) is very small.

3. EXPERIMENTAL RESULTS

In this and the following section we describe the results of a series of experiments that compare the recognition accuracy of the various algorithms described in Sec. 2 using the ARPA CSR Wall Street Journal task. The 7000 WSJ0 utterances recorded using the CLSTLK microphone were used for the training corpus, and the system was evaluated using the 330 utterances from secondary microphones in the ARPA WSJ 1992 evaluation test set, which has a closed vocabulary of 5000 words.

To expedite processing, we used a smaller and faster version of SPHINX-II than the implementation used for the official ATIS and CSR Hub evaluations. The faster system differs from the hub evaluation system in four ways: it uses a bigram grammar (rather than a trigram grammar), it uses only one codebook (rather than 27 phone-dependent codebooks), it performs sex classification on the basis of VQ distortion of the incoming cepstral vectors, and it uses a forward pass only, rather than three passes. While error rates obtained using the faster system is 50 percent greater than the corresponding rates obtained using the official evaluation system, the relative errors observed using the different compensation procedures are the same.

The conventional SPHINX-II system uses signal processing that extracts Mel-frequency cepstral coefficients (MFCC) over an analysis range of 130 to 6800 Hz. When speech is determined to be of telephone bandwidth using the Gaussian environment classifier described in Sec. 2.3, the analysis bandwidth is reduced. This is accomplished by performing the normal DFT analysis with the normal 16,000-Hz sampling rate, but only retaining DFT coefficients after the triangular frequency smoothing from center frequencies of 200 to 3700 Hz. Reduced-bandwidth MFCC coefficients are obtained by performing the discrete-cosine transform only on these frequency-weighted DFT coefficients. Two sets of VQ codebooks and senones are used, one for telephone speech using a wideband front-end analysis and another for non-telephone speech.

3.1. Comparison of MFCDCN, IFCDCN, PDCN, and IPDCN

Figure 3 and Table 1 compare word error rates obtained using various processing schemes along with the corresponding reduction of word error rates with respect to the baseline with CMN. Compensation vectors used for these comparisons were developed from training data that include the testing environments. Table 2 summarizes similar results that were obtained when the actual testing environment was excluded from the set of data used to develop the compensation vectors.

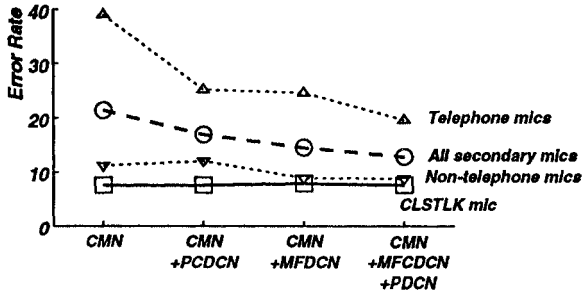

Figure 3. Comparison of recognition error rates using CMN, PDCN, and MFCDCN. Results are tabulated separately for the non-telephone and telephone microphones of the 1992 WSJ secondary-microphone evaluation data.

The results of Figure 3 and Table 1 indicate that PDCN when applied in isolation provides a recognition error rate that is not as good as that obtained using MFCDCN. Nevertheless, the effects of PDCN and MFCDCN are complementary in that the use of the two algorithms in combination provides a lower error rate

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	14.5
CMN+IMFCDCN	7.8	15.1
CMN+PDCN	7.9	16.9
CMN+IPDCN	7.7	16.5
CMN+MFCDCN+PDCN	7.6	12.9

Table 1: Word error rates obtained on the secondary-mic data from the 1992 WSJ evaluation test using CMN, MFCDCN, and PDCN with and without environment interpolation.

COMPENSATION ALGORITHM	CLSTLK mic	OTHER mics
CMN (baseline)	7.6	21.4
CMN+MFCDCN	7.6	16.1
CMN+IMFCDCN	7.6	14.8
CMN+PDCN	7.6	16.9
CMN+IPDCN	7.6	16.8
CMN+MFCDCN+PDCN	7.6	14.8
CMN+IMFCDCN+IPDCN	7.6	13.5

Table 2: Word error rates obtained using CMN, MFCDCN, and PDCN as in Table 1, but with the testing environments excluded from the corpus used to develop compensation vectors.

than was observed with either algorithm applied by itself, resulting in 39.7 percent fewer errors than with CMN alone. In fact, for the non-telephone microphones, the error rate obtained for the compensated secondary microphones is only 15.7 percent worse than that obtained training and testing using the CLSTLK microphone. The results in Table 2 demonstrate that the use of environment interpolation is helpful when the testing environment is not included in the set used to develop compensation vectors. As seen in Table 1, environmental interpolation degrades performance slightly when the actual testing environment is included in the development of the compensation vectors.

3.2. Performance of Codebook Adaptation

Table 3 compares word error rates obtained with the DCCA and BWCA as described in Sec. 2.5 with error rates obtained with CMN and MFCDCN. The Baum-Welch codebook adaptation was implemented with four iterations of re-estimation of the means of the codebook. (Means and variances were re-estimated in a pilot experiment, but with no improvement in performance.) These results indicate that the effectiveness of codebook adaptation used in isolation to reduce error rate is about equal to that of MFCDCN, but that the application of DCCA in conjunction with MFCDCN can provide further improvements in recognition accuracy.

4. SUMMARY AND CONCLUSIONS

In this paper we describe a number of procedures that improve the recognition accuracy of the SPHINX-II system in unknown acoustical environments. We found that the use of MFCDFN and

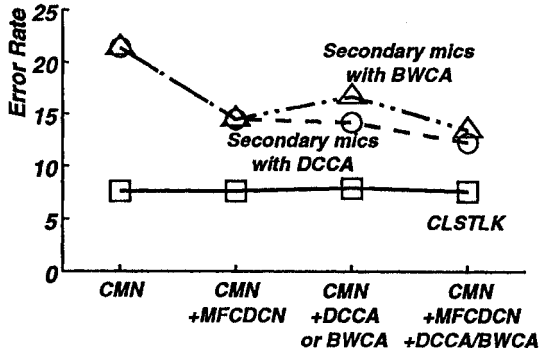

Figure 4. Comparison of recognition error rates using CMN, MFCDCN, DCCA, and BWCA, using the 1992 WSJ secondary-microphone evaluation data

phone-dependent cepstral normalization reduces the error rate by 40 percent compared to that obtained with CMN alone. The use of Baum-Welch codebook adaptation with MFCDCN reduces the error rate by 37 percent compared to that obtained with CMN alone.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We thank Raj Reddy, Mei-Yuh Hwang, and the rest of the speech group for their contributions to this work, and Yoshiaki Ohshima in particular for seminal discussions on reduced-bandwidth frequency analysis.

REFERENCES

1. Juang, B.-H., "Speech Recognition in Adverse Environments", *Computer Speech and Language*, 5:275-294, 1991.
2. Acero, A., *Acoustical and Environmental Robustness in Automatic SPeech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.
3. Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., and Rosenfeld, R., "The SPHINX-II Speech Recognition System: An Overview", *Computer Speech and Language*, 2:137-148, 1993.
4. Liu, F.H., Stern, R.M., Huang, X.D., and Acero A., "Efficient Cepstral Normalization for Robust Speech Recognition," *Proceedings of ARPA Speech and Natural Language Workshop*, pp. 69-74, Princeton, March 1993.
5. Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., and Zavaliagkos, G., "Comparative Experiments on Large Vocabulary Speech Recognition", *Proc. ARPA Human Language Technology Workshop*, March, 1993.
6. Huang, X.D., Ariki, Y., and Jack, M., *Hidden Markov Models for Speech Recognition*, Edinburgh University Press, Edinburgh, U.K., 1990.