ENVIRONMENT NORMALIZATION FOR ROBUST SPEECH RECOGNITION USING DIRECT CEPSTRAL COMPARISON

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

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

ABSTRACT

In this paper we describe and evaluate a series of new algorithms that compensate for the effects of unknown acoustical environments or changes in environment. The algorithms use compensation vectors that are added to the cepstral representations of speech that is input to a speech recognition system. While these vectors are computed from direct frame-by-frame comparisons of cepstra of speech simultaneously recorded in the training environment and various prototype testing environments, the compensation algorithms do not assume that the acoustical characteristics of the actual testing environment are known. The specific compensation vector applied in a given frame depends on either physical attributes such as SNR or presumed phonetic identity. The compensation algorithms are evaluated using the 1992 ARPA 5000-word WSJ/CSR corpus. The best system combines phoneme-based and SNR-based cepstral compensation with cepstral mean normalization, and provides a 66.8% reduction in error rate over baseline processing when tested using a standard suite of unknown microphones.

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]). Many approaches have been considered in the development of robust speech recognition systems including techniques based on autoregressive analysis, the use of special distortion measures, the use of auditory models, and the use of microphone arrays, among many other approaches (as reviewed in [1,2]).

Over the past few years, CMU and other sites have developed a series of algorithms that reduce the effects of environmental variability on speech recognition accuracy [*e.g.* 2,3]. The CMU normalization algorithms are based on three different types of approaches. The first approach is that of *cepstral remapping based on a structural model of the acoustical degradation.* An example of this type of processing is the codeword-dependent cepstral normalization algorithm (CDCN), which assumes that the effects of environmental distortion can be characterized as unknown additive noise combined with unknown linear filtering [4]. The second approach to environmental normalization is that of *high-pass filtering of cepstral coefficients*, as exemplified by the various RASTA algorithms [5] and the practice of cepstral mean removal. The third approach, which is the focus of this paper, is based on *direct cepstral comparisons* of simultaneously-recorded data from different environments on a frame-by-frame basis. We describe some of the more useful cepstral-comparison algorithms in the next section.

2. ENVIRONMENTAL NORMALIZATION USING CEPSTRAL COMPARISON

Environment-normalization algorithms based on cepstral comparison all assume that differences between the training and testing environments can be characterized by an additive correction to the cepstral vectors that represent the speech. The compensation vectors are estimated empirically on the basis of direct frame-by-frame comparisons of the cepstral representations of speech that is simultaneously recorded in the training environment and various testing environments ("stereo data"). The individual algorithms differ in the way the compensation vectors are estimated from training data, and in the way in which the need for stereo data is circumvented when the recognition system analyzes speech from an unknown environment. This general approach has become much more popular with the availability of the ARPA Wall Street Journal corpus, which in its initial phase contained about 31,000 utterances of stereo data recorded in 16 different acoustical environments.

2.1. The SDCN and FCDCN algorithms

SDCN. The simplest compensation algorithm, *SNR-Dependent Cepstral Normalization* (SDCN) [2], applies an additive correction in the cepstral domain that depends exclusively on the instantaneous SNR of the signal. This compensation vector equals the average difference in cepstra between simultaneous stereo recordings of speech samples from both the training and testing environments at each SNR in the testing environment. At high SNRs, this compensation vector primarily compensates for the effects of unknown linear filtering, while at low SNRs the vector provides a form of noise subtraction. The SDCN algorithm is simple and effective, but it requires environment-specific training.

FCDCN. *Fixed codeword-dependent cepstral normalization* (FCDCN) [2] is similar to SDCN, but it provides a greater number of compensation vectors. At each SNR the observed cepstra in the testing environment are also clustered, based on a VQ codebook. The FCDCN algorithm applies an additive correction that depends on both the instantaneous SNR of each frame of input speech, and the VQ codeword location to which the cepstral compensation vector is closest. FCDCN compensation provides

0-7803-1775-0/94 $3.00 © 1994 IEEE

greater recognition accuracy than SDCN, but it also requires environment-specific training.

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.

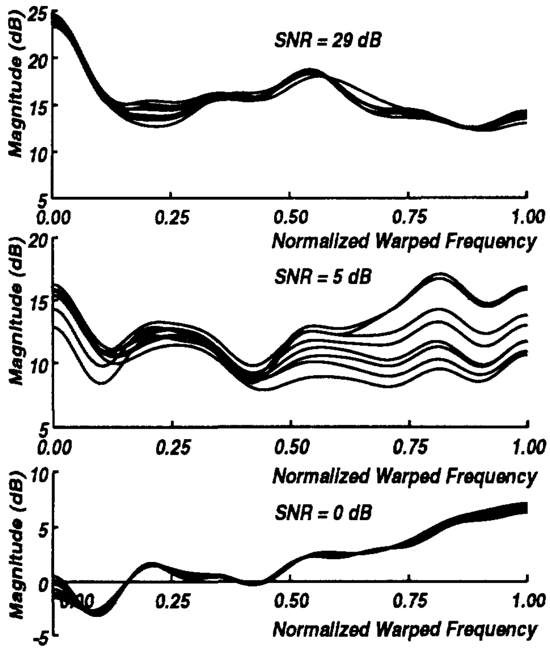

Figure 1: Comparison of compensation vectors using the FCDCN method with the PCC-160 unidirectional desktop microphone, at three different signal-to-noise ratios. The maximum SNR used by the FCDCN algorithm is 29 dB.

2.2. MFCDCN and related algorithms

MFCDCN. *Multiple fixed codeword-dependent cepstral normalization* (MFCDCN) is a simple extension of the FCDCN algorithm [2] that has the advantage of not requiring that the identity of the testing environment be known *a priori.* In MFCDCN, compensation vectors are precomputed in parallel for each of a set of testing environments using the FCDCN training procedure. When an utterance from some unknown environment is input to the recognition system, the system first determines which of the testing environments in the training data is most similar to the current testing environment. Compensation vectors for the chosen testing environment are applied to normalize the utterance according to the expression $\hat{x}_i = z_i + r[k_i, l_i, e]$ where k_i, l_i, i, and e are the VQ codeword index, instantaneous frame SNR, frame index and the index of the chosen environment, respectively, and $\hat{x}$, z, and r are the compensated (transformed) data, original data and compensation vectors, respectively.

Environment selection. We have made use of two schemes for environment selection. In the first procedure, referred to as *selection by compensation,* compensation vectors computed using each of the possible testing environments are applied successively to the incoming test utterance. The environment e is chosen that minimizes the average residual VQ distortion over the entire utterance.

In the second approach, referred to as *environment-specific VQ,* codebooks that are specific to each environment are generated from the original uncompensated speech. Environment selection is accomplished by vector quantizing the incoming test utterance using each environment-specific codebook in turn and choosing the (uncompensated) testing environment that is closest to the incoming speech in terms of VQ distortion.

Using data from the 11/92 ARPA Wall Street Journal corpus, the selection-by-compensation method produces environment-selection errors 28.8% of the time for data from one of the 15 "secondary" environments and no selection errors for data obtained using the close-talking Sennheiser microphone used in the training data. The environment-specific VQ approach produces a 14.2% misjudgment rate for data using secondary microphones and 0.3% for Sennheiser mic data. Both methods produce similar speech recognition accuracy. The latter method is similar in spirit to an approach by BBN [4], in which each incoming utterance is classified into one of seven groups of acoustical environments.

Interpolated FCDCN. The MFCDCN algorithm described above applies compensation from the single environment in the training set that is believed to have acoustical characteristics that most closely resemble those of the testing environment. In some cases, however, the testing environment does not closely resemble any single environment in the training set. In that case, interpolating the compensation vectors of several environments may be more helpful than using compensation vectors from a single (incorrect) environment.

For these reasons, the Interpolated Fixed Codeword Dependent Cepstral Normalization algorithm (IFCDCN) estimates compensation vectors for new environments by linear interpolation of several of the compensation vectors that had been precomputed for environments in the training database:

$$\hat{r}[k, l] = \sum_{e=1}^{E} f_e \cdot r[k, l, e]$$

where $\hat{r}[k, l]$, $r[k,l,e]$, and f_e are the estimated compensation vectors, the environment-specific compensation vector for the e^{th} environment, and the weighting factor for the e^{th} environment, respectively.

The weighting factors for each environment are also based on residual VQ distortion:

$$f_e = \frac{p(e|\bar{Z})}{\sum_{i=1}^{E} p(i|\bar{Z})} = \frac{exp\{D_e/(2\sigma^2)\}}{\sum_{i=1}^{E} exp\{D_i/(2\sigma^2)\}}$$

where σ is the codebook standard deviation using clean speech and $\bar{Z}$ is the testing utterance. With the present training and testing data we have generally used a value of 3 for E.

2.3. Phone-Dependent Cepstral Normalization (PDCN) and related algorithms

In this section, we describe a new family of algorithms, referred to as phone-dependent normalization procedures, which compensate for environmental variation based on the presumed phoneme identity of individual acoustical segments during the search process. This approach has the advantage that information from the acoustic-phonetic and language models as well as the constraints arising from the search process can be used to determine the most effective form of environmental compensation.

PDCN. In the current implementation of *phone-dependent cepstral normalization* (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. Labelled phonetic segments for training PDCN compensation are produced by running the decoder in supervised mode using the correct transcription of the incoming speech. For each phoneme, compensation vectors are derived by averaging the difference between cepstral coefficients obtained from the training environment and a given testing environment, using the same stereo pairs of training sentences that were used for MFCDCN. This approach is similar to SDCN except that the different compensation vectors are calculated according to phonetic identity rather than according to instantaneous frame SNR values. The compensation vectors in PDCN are described as follows,

$$\mathbf{c}[p] = \frac{\sum_{u=1}^{A} \sum_{i=1}^{T_u} (\mathbf{x}_i - \mathbf{z}_i)\, \delta(f_i - p)}{\sum_{u=1}^{A} \sum_{i=1}^{T_u} \delta(f_i - p)}$$

where f_i is the phoneme for frame i, and T_u is length of the uth utterance out of A sentences from each of training environments in stereo databases.

The SPHINX-II system uses the senone [6,8], a generalized state-based probability density function, as the basic unit to compute the likelihood from acoustical models. The probability density function for senone s for frame i for the cepstral vector $\mathbf{z}_i$ of incoming speech can be expressed as

$$p_{s_{z_i}} = \sum_{m_z=1}^{B} w_{m_z} N(\mathbf{z}_i; \mu_{m_z}, \sigma_{m_z})$$

where m_z stands for the index of the best B Gaussian mixtures of senone s for frame i, and μ_{m_z}, σ_{m_z}, and w_{m_z} are the corresponding mean, standard deviation, and weight for the m_z^{th} mixture of senone s.

Similar to before, multiple compensated cepstral vectors are formed by adding various compensation vectors to incoming cepstra, $\hat{\mathbf{x}}_{i,p} = \mathbf{z}_i + \mathbf{c}[p]$, on a frame-by-frame basis. This is a simple process in the present implementation because each senone corresponds to only one distinctive base phoneme. As a result, senone probabilities can be calculated directly in terms of *compensated* incoming speech vectors, by assuming the phonetic identity that corresponds to a given senone. Using this approach, the senone probability with PDCN is re-written as

$$p_{s_{\hat{x}_i}} = \sum_{m_{\hat{x}}=1}^{B} w_{m_{\hat{x}}} N(\hat{\mathbf{x}}_{i,p_s}; \mu_{m_{\hat{x}}}, \sigma_{m_{\hat{x}}})$$

where $m_{\hat{x}}$ is the index of the best B Gaussian mixtures for senone s at frame i with respect to the PDCN-normalized cepstral vector $\hat{x}_{i,p_s}$, for the corresponding phonetic label for senone s.

Compensation vectors are exploited by the decoder during the process of searching for the optimal sequence of states in the HMM, and scores used to evaluate hypotheses are calculated using the compensated cepstral vectors. The increase in computation incurred by PDCN is very minor and arises primarily from an increase in the number of vector quantization operations performed on the 51 alternatives for each cepstal vector.

SNR-Dependent PDCN (SPDCN). The performance of PDCN can be further improved by further partitioning the compensation vectors in terms of SNR (as is done with SDCN and FCDCN). The estimation of compensation vectors for SPDCN can then be expressed as

$$\mathbf{c}[p, l] = \frac{\sum_{u=1}^{A} \sum_{i=1}^{T_u} (\mathbf{x}_i - \mathbf{z}_i)\, \delta(f_i - p)\, \delta(s_i - l)}{\sum_{u=1}^{A} \sum_{i=1}^{T_u} \delta(f_i - p)\, \delta(s_i - l)}$$

where s_i is the instantaneous frame SNR of z_i. We chose a range of 30 dB of SNR in our current implementation.

Interpolated PDCN (IPDCN). PDCN, like SDCN and FCDCN assumes the existence of a database of utterances recorded in stereo in the training and testing environments. As in the case of MFCDCN, the PDCN algorithm can be extended to cases where the testing environment is unknown by developing ensembles of PDCN compensation vectors for a variety of testing environments, and applying to incoming utterance either the set of compensation vectors from the "closest" environment used to train the algorithm (MPDCN), or an interpolation of compensation vectors from several of the closest environments (IPDCN). In the current implementation of IPDCN, we use environment-specific VQ means and variances for environment selection to obtain the 3 closest environments with the best 4 Gaussian mixtures contributing to the interpolation weights.

3. EXPERIMENTAL RESULTS

The MFCDCN, IFCDCN, PDCN and related algorithms were evaluated using the SPHINX-II recognition system [6] in the context of the ARPA 5000-word closed-vocabulary task consisting of dictated sentences from the Wall Street Journal. The system was trained using the WSJ0 training corpus and has 7000 senones. The testing corpus consists of utterances from a set of "secondary" microphones including desktop microphones, stand-mounted microphones and telephone handsets and speakerphones. We also compared recognition accuracy for the same system using two types of cepstral high-pass filtering: the RASTA filter [5] as implemented in the SRI ARPA system [7], and cepstral mean normalization (CMN).

PROCESSING METHOD	CLSTLK mic	% Dec.	OTHER mics	% Dec.
Baseline	8.1	–	38.5	–
RASTA	9.0	-11.1	28.0	27.3
CMN	7.6	6.2	21.4	44.4
MFCDCN	8.1	0	16.7	56.6
IFCDCN	8.4	-3.7	16.7	56.6
CMN+MFCDCN	8.1	0	14.5	62.3
CMN+IFCDCN	8.4	-3.7	14.8	61.6
CMN+PDCN	–	–	15.7	59.2
CMN+MFCDCN +PDCN	–	–	12.8	66.8

Table 1: Percentage of word errors and corresponding error rate reduction for different processing schemes on the test corpus for the ARPA 11/92 5000-word, closed-vocabulary task using sentences from the Wall Street Journal.

Table 1 compares word error rates obtained using the various processing schemes along with the corresponding reduction of word error rate with respect to the baseline (no processing). The system was trained on the standard Sennheiser closetalking HMD-414 microphone (CLSTLK), and tested using either the CLSTLK mic or one of several secondary microphones (OTHER). Table 2 summarizes similar results obtained with various combinations of MFCDCN IFCDCN, PDCN, and IPDCN when the actual testing environment was excluded from the set of data used to develop the compensation vectors..

METHOD	CLSTLK mic	% Dec.	OTHER mics	% Dec.
Baseline	8.1	–	38.5	–
CMN+MFCDCN	8.1	0	16.1	58.2
CMN+MFCDCN +PDCN	8.1	0	14.8	61.6
CMN+IFCDCN	8.4	-3.7	14.8	61.6
CMN+IFCDCN +IPDCN	8.4	-3.7	13.5	64.9

Table 2: Recognition accuracy obtained for the same task as in Table 1, but with the testing environments excluded from the corpus used to develop compensation vectors.

The high baseline word error rate obtained testing with alternate microphones demonstrates the impact of mismatches of training and testing environments. Among other results we note the following: (1) Cepstral highpass filtering (RASTA and CMN) is simple and quite effective, but the use of direct cepstral comparison (MFCDCN or IFCDCN) provides substantial further decreases in error rate. (2) The combination of highpass filtering and direct cepstral comparison provides a modest additional decrease in error rate. (3) PDCN provides a benefit comparable to that of MFCDCN, but best results are obtained when these two algorithms are used in consort. (4) If the testing environment is unknown or unavailable when compensation vectors are computed, better results are obtained by use of interpolated compensation (as with IFCDCN and IPDCN).

4. SUMMARY

We describe a family of environmental normalization algorithms that apply additive corrections to incoming cepstral vectors based on either SNR (as in the MFCDCN algorithm) or presumed phoneme identity (as in PDCN). When evaluated in the context of experimental results using the 1992 ARPA WSJ/CSR task, best results were obtained using a combination of MFCDCN, PDCN, and cepstral mean normalization, which collectively reduce the error rate observed with secondary microphones by 66.8%. Interpolated implementations of these algorithms are also described for application in which the acoustics of the testing environment are unknown. While the MFCDCN and PDCN are described in the context of the SPHINX-II system with semi-continuous HMMs, they can easily be implemented in recognition systems using discrete-density or continuous-density HMMs.

ACKNOWLEDGMENTS

This research was sponsored by the Department of the Navy, Naval Research Laboratory, under Grant No. N00014-93-2005. 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. The authors would like thank Mei-Yuh Hwang, Xuedong Huang, Yoshiaki Ohshima, Raj Reddy and the rest of the CMU speech group for their contributions to this work.

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. 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*, pages 865-868, March 1992.
4. 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.
5. Hermansky, H., Morgan, N., and Hirsch, H., "Recognition of Speech In Additive and Convolutional Noise Based on RASTA Processing", *ICASSP-93*, pages II-83-86, April, 1993.
6. 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.
7. Murveit, H., Butzberger, M, and Weitraub, M., "Reduced Channel Dependence for Speech Recognition", *Proc. DARPA Spoken and Natural Language* Workshop, Feb., 1992.
8. Hwang, M.-Y. and Huang, X.D. "Shared-Distribution Hidden Markov Models for Speech Recognition", IEEE Trans. on Speech and Audio Processing, 1:414-420, 1993.