ENVIRONMENTAL ROBUSTNESS IN AUTOMATIC SPEECH RECOGNITION

Alejandro Acero and Richard M. Stern

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

Abstract

In this paper we report our initial efforts to make SPHINX, the CMU continuous-speech speaker-independent recognition system, robust to changes in the environment. To deal with differences in noise level and spectral tilt between close-talking and desk-top microphones, we propose two novel methods based on additive corrections in the cepstral domain. In the first algorithm, the additive correction depends on the instantaneous SNR of the signal. In the second technique, EM techniques are used to best match the cepstral vectors of the input utterances to the ensemble of codebook entries representing a standard acoustical ambience. Use of the proposed algorithms dramatically improves recognition accuracy when the system is tested on a microphone other than the one on which it was trained.

1. Introduction

Real applications demand that the performance of speech recognition systems is not affected by changes in the environment. However, it is well known that when a system is trained and tested under different conditions, the recognition rate drops unacceptably. In this study we are concerned with the variability present when different microphones are used in training and testing, and specifically the development of procedures that can signficantly improve the accuracy of speech-recognition systems that use desk-top microphones.

There are many sources of acoustical distortion that can degrade the accuracy of speech-recognition systems. For example, obstacles to robustness include additive noise from machinery, competing talkers, etc., reverberation from surface reflections in a room, and spectral shaping by microphones and the vocal tracts of individual speakers. These sources of distortion cluster into two complementary classes: *additive* noise (as in the first two examples) and distortions resulting from the *convolution* of the speech signal with an unknown linear system (as in the remaining three).

A number of algorithms for speech enhancement have been proposed in the literature. For example, Boll [1] and Berouti *et al.* [2] introduced the spectral subtraction of DFT coefficients, and Porter and Boll [3] used MMSE techniques to estimate the DFT coefficients of corrupted speech. Spectral equalization to compensate for convolved distortions was introduced by Stockham *et al.* [4]. Recent applications of spectral subtraction and spectral equalization include the work of Van Compernolle [5] and Stern and Acero [6]. Although relatively successful, the above methods all depend on the assumption of independence of the spectral estimates across frequencies. Erell and Weintraub [7] demonstrated improved performance with an MMSE estimator in which correlation among frequencies is modeled explicitly.

In this paper we present two algorithms for speech normalization based on additive corrections in the cepstral domain. We have chosen the cepstral domain rather than the frequency domain so that we work directly with the parameters that SPHINX uses, and because speech can be characterized with a smaller number of parameters in the cepstral domain than in the frequency domain. The first algorithm, *SNR-dependent cepstral normalization* (SDCN) is simple and effective, but it cannot be applied to new microphones without microphone-specific training. The second algorithm, *codeword-dependent cepstral normalization* (CDCN) computes an ML estimate for the noise and spectral tilt, and then an MMSE estimate for the speech cepstrum. These algorithms are evaluated using an alphanumeric database in which utterances were recorded simultaneously with two different microphones.

2. A Model of the Environment

We assume that the speech signal $x(t)$ is first passed through a linear filter $h(t)$ whose output is then corrupted by uncorrelated additive noise $n(t)$. We can characterize the power spectral density (PSD) of the processes involved as

$$P_y(f) = P_x(f)\,|H(f)|^2 + P_n(f) \qquad (1)$$

If we let the cepstral vectors $\mathbf{x}$, $\mathbf{n}$, $\mathbf{y}$ and $\mathbf{q}$ represent the Fourier series expansion of $\ln P_x(f)$, $\ln P_n(f)$, $\ln P_y(f)$ and $\ln |H(f)|^2$ respectively, (1) can be rewritten as

$$\mathbf{y} = \mathbf{x} + \mathbf{q} + \mathbf{r}(\mathbf{x},\mathbf{n},\mathbf{q}) \quad \text{or} \quad \mathbf{y} = \mathbf{n} + \mathbf{s}(\mathbf{x},\mathbf{n},\mathbf{q}) \qquad (2)$$

where the correction vectors $\mathbf{r}(\mathbf{x},\mathbf{n},\mathbf{q})$ and $\mathbf{s}(\mathbf{x},\mathbf{n},\mathbf{q})$ are given by

$$\mathbf{r}(\mathbf{x},\mathbf{n},\mathbf{q}) = IDFT\,\{\ln\,(1 + e^{DFT\,[\mathbf{n} - \mathbf{q} - \mathbf{x}]})\} \qquad (3)$$

$$\mathbf{s}(\mathbf{x},\mathbf{n},\mathbf{q}) = IDFT\,\{\ln\,(1 + e^{DFT\,[\mathbf{x} + \mathbf{q} - \mathbf{n}]})\} \qquad (4)$$

If $\mathbf{x}(0) + \mathbf{q}(0) \gg \mathbf{n}(0)$ (*i.e.* high SNR), $\mathbf{r} \approx \mathbf{0}$, and $\mathbf{y} \approx \mathbf{x} + \mathbf{q}$. On the other hand, when $\mathbf{x}(0) + \mathbf{q}(0) \ll \mathbf{n}(0)$ (*i.e.* low SNR), $\mathbf{s} \approx \mathbf{0}$, and $\mathbf{y} \approx \mathbf{n}$. We can obtain an estimate $\hat{P}_y(f)$ of the PSD $P_y(f)$ from a sample function of the process $y(t)$ (*i.e.* a frame of speech that is assumed to be locally stationary). If $\mathbf{z}$ represents the

Fourier expansion of $\ln \hat{P}_y(f)$, our goal is to estimate the uncorrupted vectors $\mathbf{X} = \mathbf{x}_0,...\mathbf{x}_{N-1}$ of an utterance given the observations $\mathbf{Z} = \mathbf{z}_0,...\mathbf{z}_{N-1}$.

3. SNR-Dependent Cepstral Normalization

If we assume that the estimation error is negligible (*i.e.* $p(\mathbf{z}/\mathbf{y}) = \delta(\mathbf{z}-\mathbf{y})$), and that the correction vector $\mathbf{r}$ in (3) depends only on $\mathbf{x}(0) - \mathbf{n}(0)$ (*i.e.* that we can apply an average correction to all spectral shapes with the same SNR), then we can estimate $\hat{\mathbf{x}}$ by the expression

$$\hat{\mathbf{x}} = \mathbf{z} - \mathbf{w}(SNR) \tag{5}$$

which subtracts from the observed vector a correction $\mathbf{w}$ that depends only on the instantaneous SNR of the observed signal, $\mathbf{z}(0) - \mathbf{n}(0)$. We have estimated these compensation vectors $\mathbf{w}(SNR)$ by computing the average difference between cepstral vectors for the test condition versus a standard acoustical environment from simultaneous stereo recordings. Although this technique performs acceptably, it has the disadvantage that new microphones must be "calibrated" by collecting long-term statistics from a new stereo database. Since only long-term averages are used, the SDCN is clearly not able to model a non-stationary environment.

4. Codeword-Dependent Cepstral Normalization

A robust speech recognizer should be immune to the transformation described by (1). To reverse the effects of $H(f)$ and $P_n(f)$ we have to solve two problems:

1. Estimate $\mathbf{q}$ and $\mathbf{n}$, the equalization and noise vectors, given the observations $\mathbf{Z}$ for an utterance. An ML estimator of the parameter vectors will be used.
2. Estimate the uncorrupted vector $\mathbf{x}$ given the observation for that frame $\mathbf{z}$ and the equalization and noise vectors $\mathbf{q}$ and $\mathbf{n}$. For this task we will use an MMSE estimator.

In the absence of exact statistics for the AR spectral estimator, we modeled the distribution $p(\mathbf{z}/\mathbf{y})$ as a multivariate gaussian $N_z(\mathbf{y}, \Gamma)$. We have confirmed the validity of this assumption empirically for the signal processing in SPHINX. The probability density function of $\mathbf{x}$ will be assumed to be a mixture of K gaussian densities with means $\mathbf{c}_k$, covariance matrices $\mathbf{C}_k$, and weights p_k:

$$p(\mathbf{x}) = \sum_{k=0}^{K-1} p_k \, p(\mathbf{x}/k) = \sum_{k=0}^{K-1} p_k \, N_x(\mathbf{c}_k, \mathbf{C}_k) \tag{6}$$

4.1. MMSE Estimator of the Cepstral Vector

The MMSE estimate for $\mathbf{x}$ has the form

$$\hat{\mathbf{x}}_{MMSE} = \frac{\sum_{k=0}^{K-1} p_k \int \mathbf{x} \, p(\mathbf{z}/\mathbf{x}, \mathbf{n}, \mathbf{q}, k) \, p(\mathbf{x}/k) \, d\mathbf{x}}{\sum_{k=0}^{K-1} p_k \int p(\mathbf{z}/\mathbf{x}, \mathbf{n}, \mathbf{q}, k) \, p(\mathbf{x}/k) \, d\mathbf{x}} \tag{7}$$

Since the true MMSE estimate for $\mathbf{x}$ cannot be obtained directly due to nonlinearities introduced by the correction vectors, we obtained approximate estimates with the following procedure:

1. We assign the mixture component 0 to the noise event, and assume that the elements of the covariance matrix $\mathbf{C}_0$ are much smaller than the corresponding elements from Γ. This implies that $p(\mathbf{x}/k=0) \approx \delta(\mathbf{x}-\mathbf{c}_0)$.
2. All other components are assumed to belong to some class of speech event. We assume that the elements of their covariance matrices $\mathbf{C}_k$ are much larger than the corresponding elements of Γ, which implies that $p(\mathbf{z}/\mathbf{x}, \mathbf{n}, \mathbf{q}, k) \approx \delta(\mathbf{z}-\mathbf{x}-\mathbf{q}-\mathbf{r})$.

With these approximations, the estimate has the form

$$\hat{\mathbf{x}}_{MMSE} = f_0 \mathbf{c}_0 + \sum_{k=1}^{K-1} f_k \hat{\mathbf{x}}_k \tag{8}$$

where $\hat{\mathbf{x}}_k = \mathbf{z} - \mathbf{q} - \mathbf{r}_k$ and (9)

$$f_k = \frac{\dfrac{p_k}{|\mathbf{C}_k|^{1/2}} \exp(-d_k/2)}{\dfrac{p_0}{|\Gamma|^{1/2}} \exp(-d_0/2) + \sum_{l=1}^{K-1} \dfrac{p_l}{|\mathbf{C}_l|^{1/2}} \exp(-d_l/2)} \tag{10}$$

$$d_0 = (\hat{\mathbf{x}}_0 - \mathbf{c}_0)\, \Gamma^{-1} (\hat{\mathbf{x}}_0 - \mathbf{c}_0); \quad d_k = (\hat{\mathbf{x}}_k - \mathbf{c}_k)\, \mathbf{C}_k^{-1} (\hat{\mathbf{x}}_k - \mathbf{c}_k)$$

In this procedure the correction vectors $\mathbf{r}_k = \mathbf{r}(\mathbf{c}_k, \mathbf{n}, \mathbf{q})$ and $\mathbf{s}_k = \mathbf{s}(\mathbf{c}_k, \mathbf{n}, \mathbf{q})$ are no longer a function of $\mathbf{x}$, so the cepstral normalization is *codeword-dependent*.

4.2. ML Estimation of Noise and Spectral Tilt

If no *a priori* information is given about the noise and equalization vectors $\mathbf{n}$ and $\mathbf{q}$, the optimum estimation method is maximum likelihood:

$$(\hat{\mathbf{n}}_{ML}, \hat{\mathbf{q}}_{ML}) = \operatorname{argmax} \; p(Z/\mathbf{q}, \mathbf{n}) \tag{11}$$

By assuming that different frames are independent from each other, we can use the expression

$$\ln p(Z/\mathbf{q}, \mathbf{n}) = \sum_{i=0}^{N-1} \ln p(\mathbf{z}_i/\mathbf{q}, \mathbf{n}) \tag{12}$$

whose maximization leads to

$$\nabla_n \ln p(Z/\mathbf{q}, \mathbf{n}) = \sum_{i=0}^{N-1} \frac{\nabla_n p(\mathbf{z}_i/\mathbf{q}, \mathbf{n})}{p(\mathbf{z}_i/\mathbf{q}, \mathbf{n})} = \mathbf{0} \tag{13}$$

A similar expression can be derived for $\mathbf{q}$.

By using the approximations described above we can express the distribution of $\mathbf{z}_i$ as

$$p(\mathbf{z}_i/\mathbf{q}, \mathbf{n}) = \frac{p_0}{|\Gamma|^{1/2}} \exp(-\frac{1}{2}(\mathbf{z}_i - \mathbf{n} - \mathbf{s}_0)\, \Gamma^{-1} (\mathbf{z}_i - \mathbf{n} - \mathbf{s}_0)) +$$

$$\sum_{k=1}^{K-1} \frac{p_{i_k}}{|\mathbf{C}_k|^{1/2}} \exp(-\frac{1}{2}(\mathbf{z}_i - \mathbf{q} - \mathbf{r}_k - \mathbf{c}_k)\, \mathbf{C}_k^{-1} (\mathbf{z}_i - \mathbf{q} - \mathbf{r}_k - \mathbf{c}_k))$$

except for a constant factor. The first term is expressed as a function of the noise $\mathbf{n}$ explicitly to reflect the fact that the noise codeword ($k = 0$) is largely insensitive to $\mathbf{q}$, and depends mostly on $\mathbf{n}$. Similarly, the other codewords are largely insensitive to $\mathbf{n}$ and depend mostly on $\mathbf{q}$.

Since (13) leads to a highly nonlinear equation, we will use a variant of the well-known EM algorithm [8] that has been extensively used in the literature to obtain ML solutions with incomplete data:

1. Assume initial values of $\hat{\mathbf{n}}^{(0)}$ and $\hat{\mathbf{q}}^{(0)}$ for $j = 1$.
2. **Estimate** the correction vectors $\mathbf{r}_k$ and $\mathbf{s}_k$ given $\hat{\mathbf{n}}^{(j-1)}$, $\hat{\mathbf{q}}^{(j-1)}$, and $\mathbf{x} = \mathbf{c}_k$ according to (3) and (4).
3. **Maximize** the log-likelihood (12). The new estimates for $\hat{\mathbf{n}}^{(j)}$ and $\hat{\mathbf{q}}^{(j)}$ are

$$\hat{\mathbf{n}}^{(j)} = \hat{\mathbf{n}}^{(j-1)} + \frac{\sum_{i=0}^{N-1} f_{i_0} \hat{\mathbf{q}}_{i_0}}{\sum_{i=0}^{N-1} f_{i_0}} - \hat{\mathbf{q}}^{(j-1)} \tag{14}$$

$$\hat{\mathbf{q}}^{(j)} = \frac{\sum_{i=0}^{N-1} \sum_{k=1}^{K-1} f_{i_k} \hat{\mathbf{q}}_{i_k}}{\sum_{i=0}^{N-1} \sum_{k=1}^{K-1} f_{i_k}} \tag{15}$$

$$\text{where} \quad \hat{\mathbf{q}}_{i_k} = \mathbf{z}_i - \mathbf{c}_k - \mathbf{r}_k \tag{16}$$

4. Stop if convergence has been reached, otherwise go to Step 2.

4.3. Implementation and Discussion

For simplicity, all the covariance matrices $\mathbf{C}_k$ are assumed to be equal to $\sigma^2 \mathbf{I}$. We also assumed that Γ equals $\gamma^2 \mathbf{I}$, which is actually not the case when frequency warping is performed. The codebook elements $\{\mathbf{c}_k\}$ are estimated with a standard Lloyd algorithm. Furthermore, all p_k are considered identical, except for $p_0$which is somewhat greater.

Unlike in previous studies where estimates of the power normalization factor, spectral equalization function, and noise are obtained independently, these quantities are jointly estimated in CDCN using a common maximum likelihood framework that is based on *a priori* knowledge of the speech signal. In CDCN, power normalization is acomplished by the appropriate $\mathbf{q}(0)$. The criterion used in CDCN is the minimization of the distortion between the cepstral vectors of the input utterance and the ensemble of codebook entries of the normalized speech, rendering the need for long-term averages unnecessary. Since CDCN only requires a single utterance in order to estimate noise and spectral tilt, it can better capture the non-stationarity of the environment. Moreover, in a real application, long-term averages may not be available for every speaker and new microphone.

5. Evaluation

5.1. Database

An alphanumeric database has been collected that consists of 1018 training utterances (74 different speakers) and 140 (10 different speakers) testing utterances. These utterances were recorded simultaneously in stereo using both the close-talking Sennheiser HMD224 microphone (CLSTLK), a standard in previous DARPA evaluations, and a desk-top Crown PZM6fs microphone (CRPZM). The recordings with the CRPZM exhibit not only background noise but also key clicks from workstations, interference from other talkers, and reverberation. The database consists of strings of letters, numbers and a few control words, that were naturally elicited in the context of a task in which speakers spelled their names, addresses and other personal information, and entered some random letter and digit strings. A total of 106 vocabulary items appear in the vocabulary, of which about 40 were rarely uttered.

Figure 1 compares averaged spectra from the database for frames believed to contain speech and background noise from each of the two microphones. By comparing these curves, it can be seen that the average SNR using the CLSTLK is about 25 dB. The signals from the CRPZM, on the other hand, exhibit an SNR of less than 10 dB for frequencies below 1500 Hz and about 15 dB for frequencies above 2000 Hz. Furthermore, the response of the CRPZM exhibits a greater spectral tilt than that of the CLSTLK.

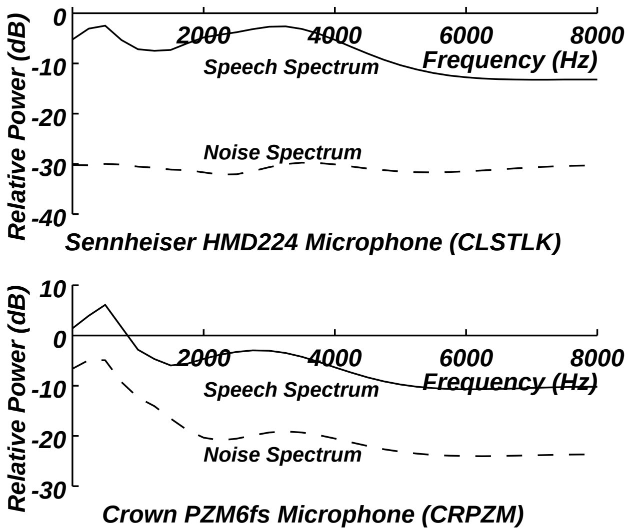

Figure 1: Average speech and noise spectra from the stereo alphanumeric database obtained using the CLSTLK and CRPZM microphones. The separation of the two curves in each panel provides an indication of SNR for each microphone. It can also be seen that the CRPZM produces greater spectral tilt.

5.2. The Recognition System

The first stages of signal processing in the evaluation system are virtually identical to those that have been reported for the SPHINX speech recognition system previously [9], except that the number of cepstral coefficients before frequency warping was increased from 12 to 32 to provide better frequency resolution after frequency warping. This led to a relative improvement of 5 percent in the baseline performance of SPHINX. All algorithms operate on the cepstral vectors computed by the SPHINX front end. The normalized cepstra, differenced cepstra, and combined power and differenced power parameters are vector quantized into three different codebooks. A simplified version of SPHINX with no grammar was used for the experiments.

5.3. Results

Table 1 describes the recognition accuracy of the original SPHINX system with no preprocessing, with conventional spectral equalization and spectral subtraction as described in [6], and with the SDCN and CDCN algorithms. These results were tabulated using current standard DARPA evaluation protocols. With no processing, training and testing using the CRPZM degrades recognition accuracy by about 60 percent relative to that obtained by training and testing on the CLSTLK. Although most of the "new" errors introduced by the CRPZM were confusions of silence or noise segments with weak phonetic events, a significant percentage was also due to crosstalk [6]. Use of the CDCN algorithm brings the performance obtained when training on the CLSTLK and testing on the CRPZM to the level observed when the system is trained and tested on the CRPZM. Moreover, use of CDCN improves performance obtained when training and testing on the CRPZM to a level greater than the baseline performance. The much simpler SDCN algorithm also provides considerable improvement in performance when the system is trained and tested on two different microphones.

TRAIN TEST	**CLSTLK CLSTLK**	**CLSTLK CRPZM**	**CRPZM CLSTLK**	**CRPZM CRPZM**
BASE	85.3%	18.6%	36.9%	76.5%
EQUAL	85.3%	38.3%	50.9%	76.5%
SUB	82.7%	64.8%	75.1%	72.8%
SDCN	85.3%	67.2%	76.4%	75.5%
CDCN	85.3%	74.9%	73.7%	77.9%

Table 1: Comparison of recognition accuracy of SPHINX with no processing, spectral equalization, spectral subtraction, and the SDCN and CDCN algorithms. The system was trained and tested using all combinations of the CLSTLK and CRPZM microphones.

6. Conclusions

We described and evaluated two algorithms to make SPHINX more robust with respect to changes of microphone and acoustical environment. With the first algorithm, *SNR-dependent cepstral normalization*, a correction vector is added that depends exclusively on the instantaneous SNR of the input. While SDCN is very simple, it provides a considerable improvement in performance when the system is trained and tested on different microphones, while maintaining the same performance for the case of training and testing on the same microphone. Two drawbacks of the method are that the system must be retrained using a stereo database for each new microphone considered, and that the normalization is based on long-term statistical models.

The second algorithm, *codeword-dependent cepstral normalization*, uses a maximum likelihood technique to estimate noise and spectral tilt in the context of an iterative algorithm similar to the EM algorithm. With CDCN, the system can adapt to new speakers, microphones, and environments without the need for collecting statistics about them *a priori*. By not relying on long-term *a priori* information, the CDCN algorithm can dynamically adapt to changes in the acoustical environment as well.

Both algorithms provided dramatic improvement in performance when SPHINX is trained on one microphone and tested on another, without degrading recognition accuracy obtained when the same microhone was used for training and testing.

7. Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163. 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 Defense Advanced Research Projects Agency or the US Government. We thank Joel Douglas, Kai-Fu Lee, Robert Weide, Raj Reddy, and the rest of the speech group for their contributions to this work.

References

1. S. F. Boll, ‘‘Suppression of Acoustic Noise in Speech Using Spectral Subtraction’’, *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 27, 1979, pp. 113-120.

2. M. Berouti, R. Schwartz and J. Makhoul, ‘‘Enhancement of Speech Corrupted by Acoustic Noise’’, in *Speech Enhancement*, J. S. Lim, ed., Prentice Hall, 1983, Englewood Cliffs, NJ, 1979, pp. 69-73.

3. J. E. Porter and S. F. Boll, ‘‘Optimal Estimators for Spectral Restoration of Noisy Speech’’, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New York, NY*, 1984, pp. 18A.2.1-18A.2.1-4.

4. T. G. Stockham, T. M. Cannon and R. B. Ingebretsen, ‘‘Blind Deconvolution Through Digital Signal Processing’’, *Proc. IEEE*, Vol. 63, 1975, pp. 678-692.

5. D. Van Compernolle, ‘‘Noise Adaptation in a Hidden Markov Model Speech Recognition System’’, *Computer, Speech and Language*, Vol. 3, 1989, pp. 151-167.

6. R. Stern and A. Acero, ‘‘Acoustical Preprocessing for Automatic Speech Recognition’’, *Proc. DARPA Speech and Natural Language Workshop*, Oct. 1989.

7. A. Erell and M. Weintraub, ‘‘Spectral Estimation for Noise Robust Speech Recognition’’, *Proc. DARPA Speech and Natural Language Workshop*, Oct. 1989.

8. N.M. Laird, A.P. Dempster and D.B. Rubin, ‘‘Maximum Likelihood from Incomplete Data via the EM algorithm’’, *Ann. Roy. Stat. Soc.*, Dec 1987, pp. 1-38.

9. K.F. Lee and H.W. Hon, ‘‘The SPHINX Speech Recognition System’’, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New York, NY*, Apr. 1989.