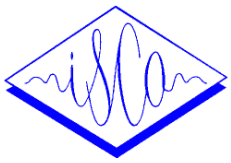

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

ACOUSTICAL PRE-PROCESSING FOR ROBUST SPOKEN LANGUAGE SYSTEMS

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 discuss several issues that concern the development of spoken language systems that are robust to changes in the acoustical environment. We describe the benefit of *joint* compensation for differences in noise level and spectral tilt between close-talking and desk-top microphones, as opposed to independent compensation. For SPHINX, the CMU continuous-speech speaker-independent recognition system, cepstral processing offers the advantages of easier integration, greater computationally efficiency and greater accuracy compared to processing in the spectral domain. We also present algorithms that adapt to new environments by estimating noise level and spectral tilt directly from the input speech, without the need for environment-specific training data.

1. Introduction

There are many sources of acoustical distortion that can degrade the accuracy of speech-recognition systems. For example, obstacles to robustness include additive noise sources such as 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 to compensate for the effects of additive noise *or* distortion through linear filtering have been proposed in the literature. For example, Boll [5] and Berouti *et al.* [4] introduced the spectral subtraction of DFT coefficients to enhance speech corrupted by additive noise, and Porter and Boll [11] used MMSE techniques to estimate the DFT coefficients of the corrupted speech. Spectral equalization to compensate for unknown linear filtering was introduced by Stockham *et al.* [13]. Recent applications of spectral subtraction and spectral equalization for speech recognition systems include the work of Van Compernolle [6] and Stern and Acero [12]. Although relatively successful, the above methods all assume that the various spectral components are statistically independent. Nadas [9], and Erell and Weintraub [7] demonstrated improved performance with an MMSE estimator in which correlation among frequency components is modeled explicitly by modeling the probability density of the speech vector as a mixture of gaussians.

In this paper, we address several issues in building robust spoken language systems that we believe deserve greater attention: interaction between additive noise and spectral equalization, processing in the cepstral domain, and environment adaptation. Although a great deal of work has been directed toward increasing robustness to noise, little work has been reported that optimally combines compensation for additive noise with compensation for linear filtering. Whereas the parameter set of choice by the speech recognition community is largely the cepstrum, most authors perform their processing in the spectral domain. With some exceptions (*e.g.* Nadas [10]), little ground has been broken in the problem of adapting to an acoustical environment directly from input speech, without the need for long-term averages.

Interactions between additive noise and linear filtering, and *joint* compensation for their effects, will be the topic of Section 2. We will show in Section 3 that it is possible to perform enhancement in the cepstral domain in a very efficient fashion that produces recognition accuracy exceeding that obtained in the spectral domain. Section 4 describes several algorithms that are able to track the parameters of the environment, especially spectral tilt, directly from input speech by means of acoustic space normalization.

2. Interaction Between Noise and Equalization

In this section we compare the performance of two algorithms that compensate for additive noise and linear filtering, and we argue that *joint* compensation for these two effects will perform better than cascaded processing which performs separate and independent compensation. We believe that the interaction between the two phenomena will be greater for speech with a low SNR.

Our general goal has been to derive MMSE estimates of the log-spectrum of speech recorded from desk-top microphones that are compensated for the effects of additive noise and linear filtering. We implement noise subtraction using an algorithm referred to as MMSE1 (similar to the approach proposed by Porter and Boll [11]), in which we define $Z_i(\omega_k)$ to be the log-spectrum of the input speech signal at frame i and frequency band k, and $N(\omega_k)$ to be the estimate of the log-spectrum of the noise for the desk-top microphone recordings. It is advantageous to implement compensation for additive noise as an additive correction $f(SNR_{ik})$ of $SNR_{ik} = Z_i(\omega_k) - N(\omega_k)$, the channel SNR, so that the restored log-spectrum has the form

$$\overline{X}_i(\omega_k) = Z_i(\omega_k) + f(Z_i(\omega_k) - N(\omega_k)) \quad (1)$$

The function f is selected to minimize the average squared error between utterances recorded from a desk-top microphone and a standard close-talking microphone, and it is estimated from data recorded stereophonically from the two microphones. At high channel SNRs, the function is zero so that no correction is applied, while at low channel SNRs the correction reflects the average difference in noise levels between speech recorded from the two microphones.

We implement spectral equalization to compensate for linear filtering using an algorithm referred to as EQUAL, which compensates the log-spectra $Z_i(\omega_k)$ of the desk-top microphone by adding a function $Q(\omega_k)$. An approximate ML estimate of Q can be obtained as the difference between the average log-spectra from the two microphones during speech frames that are supposed not to be contaminated by noise. The compensated log-spectrum is

$$\overline{X}_i(\omega_k) = Z_i(\omega_k) + Q(\omega_k) \quad (2)$$

We evaluated these two algorithms (and others) using an alphanumeric database of utterances that were recorded simultaneously in stereo using both the close-talking Sennheiser HMD224 microphone (CLSTK), 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. This database is described in greater detail in [1], and Figures 1 and 2 compare the time-varying frequency response of the word *Yes* recorded from the two microphones with no additional processing.

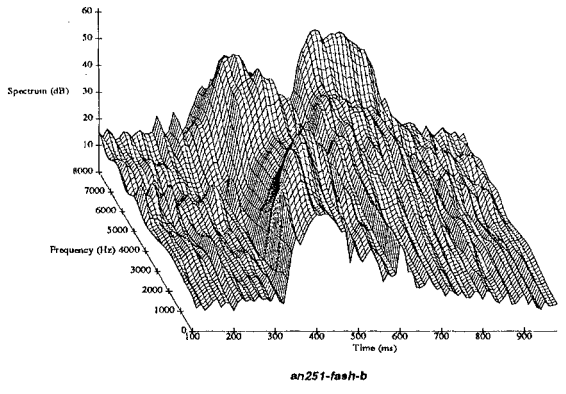

Figure 1: "Yes" with CLSTK and no processing.

Table 1 summarizes the recognition accuracy of CMU's SPHINX [8] system using no preprocessing (BASE), using the EQUAL and MMSE1 algorithms, and using a cascade of the two algorithms (EQ+MM). 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 CLSTK. Use of EQUAL and MMSE1 increase the robustness of the system. The algorithm EQ+MM performs a cascade of spectral equalization and noise suppression:

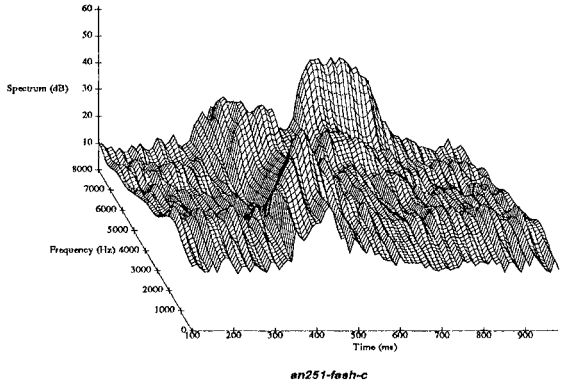

Figure 2: "Yes" with CRPZM and no processing.

$$\overline{X}_i(\omega_k) = Z_i(\omega_k) + f(Z_i(\omega_k) - N(\omega_k)) + Q(\omega_k) \quad (3)$$

The results in Table 1 show that the cascade of compensators for noise and filtering improves recognition accuracy.

One way of performing *joint* compensation for additive noise and filtering is to use a different transformation curve $g_k(SNR_k)$ for every frequency band. An algorithm which we refer to as MMSEN accomplishes this task in the log-frequency domain according to the equation

$$\overline{X}_i(\omega_k) = Z_i(\omega_k) + g_k(Z_i(\omega_k) - N(\omega_k)) \quad (4)$$

The functions $g_k(SNR_k)$ combine noise suppression and spectral equalization:

$$g_k(Z_i(\omega_k) - N(\omega_k)) = f_k(Z_i(\omega_k) - N(\omega_k)) + Q(\omega_k) \quad (5)$$

and are estimated to minimize the average squared error between the two sets of recordings. Since the values of the corrections for high and low channel-SNR are different for every frequency band k, equalization can be achieved, as well as different amounts of noise suppression (thereby handling colored noise unlike the MMSE1 algorithm).

In Table 1 we observe that the MMSEN algorithm produces better recognition than the cascade of EQUAL and MMSE1 for conditions in which the system is trained with the CLSTK microphone and tested using the CRPZM microphone. We note that although we include in the table all the possible cases of training and testing on the CLSTK and CRPZM for the sake of completeness, in practice we are primarily concerned with the performance obtained when the system is trained using the CLSTK microphone.

TRAIN TEST	CLSTK CLSTK	CLSTK CRPZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%
EQUAL	N/A	38.3%	50.9%	76.5%
MMSE1	N/A	48.7%	68.7%	71.4%
EQ+MM	N/A	61.4%	75.8%	74.3%
MMSEN	N/A	66.4%	75.5%	72.3%
SDCN	N/A	67.2%	76.4%	75.5%
FCDCN	N/A	73.1%	79.3%	75.8%
CDCN	85.3%	74.9%	73.7%	77.9%
ISDCN	84.8%	62.1%	71.4%	72.4%

Table 1: Performance of different normalization algorithms. N/A refers to conditions in which compensation reverts to baseline processing.

3. Processing in the Cepstral Domain

A large number of speech-recognition systems, including the ones developed by AT&T Bell Laboratories, BBN, CMU, MIT Lincoln Laboratory, and SRI, use features derived from the cepstrum of speech as the basis for classification. This is partially a consequence of the smaller number of parameters needed to represent speech compared to the frequency domain. Also, since cepstral components are approximately uncorrelated with each other, the common assumption that their probability density can be modeled as a mixture of gaussians sharing a diagonal covariance matrix is more valid than in the case of log-spectrum components.

For these reasons, and for the sake of computational efficiency and better system integration, it is desirable to perform compensation for noise and filtering in the same cepstral domain that is used for speech recognition. Recently, Acero and Stern [1] proposed the *SNR-Dependent Cepstral Normalization* (SDCN) algorithm that performs joint compensation for noise and filtering while operating directly in the cepstral domain. If we define $\mathbf{z}_i$ as the cepstrum vector of the degraded speech, and $\mathbf{n}$ the cepstrum of the noise, the compensated cepstral vector has the form:

$$\hat{\mathbf{x}}_i[k] = \mathbf{z}_i[k] + h_k(\mathbf{z}_i[0] - \mathbf{n}[0]) = \mathbf{z}_i[k] + \mathbf{w}(SNR_i) \qquad (6)$$

where the correction vector $\mathbf{w}$ is a function of the *frame-SNR* $SNR_i = \mathbf{z}_i[0] - \mathbf{n}[0]$. Equations (4) and (6) are quite similar, with the former expressed in the spectral domain and the latter in the cepstral domain, except that the argument in (4) is different for every component (the channel SNR), while the argument in (6) is the same for all components (the frame-SNR). As before, the correction vector $\mathbf{w}$ compensates primarily for noise at low SNR, and for filtering at high SNR. These correction vectors are again estimated from a stereo training database to minimize the squared error between the two sets of recordings.

Results in Table 1 show that SDCN performs slightly better than MMSEN, indicating that it is advantageous to perform environmental normalization in the same domain of the parameter space of the speech recognizer. The SDCN algorithm is extremely computationally efficient, and we found that effective compensation can be achieved by adjusting only the first two components of the cepstral vector [3].

Although the simple SDCN algorithm performs remarkably well, it assigns the same compensation to all the vectors with the same SNR in the Vector Quantization (VQ) codebook used to represent speech in the classification process. The more recent *Fixed Codeword-Dependent Cepstral Normalization* (FCDCN) [3] operates in similar fashion, but develops correction vectors that are *codeword-dependent*. In the FCDCN algorithm, the environment normalization and VQ are integrated into one single step. The correction vectors are estimated by an EM algorithm described in [3], since they cannot be computed using a simple average as was the case with the previously-described algorithms. Recognition results using the FCDCN algorithm are also shown in Table 1, and they are better than those observed in all previous algorithms. In fact, a study conducted with additional testing sets demonstrated that the recognition accuracy using the SPHINX system and the FCDCN algorithm when trained on speech from a standard microphone and tested on speech from a desk-top microphone is no worse on the average than the accuracy obtained when the system is trained and tested using the desk-top microphone. The FCDCN algorithm is also very computationally efficient, and we found that it is only necessary to compensate the first four cepstral components to achieve optimal recognition accuracy.

4. Acoustic Space Normalization

Although the FCDCN algorithm is very accurate and computationally efficient, offering a great degree of integration within the SPHINX system, it also requires a stereophonically-recorded database of training data to estimate the correction vectors for each new environment considered. Since such a resource may not always be available in practice, we now consider two algorithms that can compensate for noise and filtering without environment-specific training.

In [1] we first described the *Codeword-Dependent Cepstral Normalization* (CDCN) algorithm. This algorithm finds ML estimates (by an EM algorithm) of $\mathbf{n}$ and $\mathbf{q}$ that best match the cepstral vectors of a universal acoustic space with an ensemble of cepstral vectors from the target environment. The universal acoustic space is defined as a mixture of gaussian densities with a fixed correction vector for each mixture component. An MMSE estimate is obtained for every cepstral vector that weights the contribution of all mixture components. As can be seen from Table 1, the recognition accuracy obtained using the CDCN algorithm exceeds that of almost all other algorithms considered, and CDCN has the additional advantage that it does not need to be trained using stereophonically-recorded data. Unfortunately, it is also more computationally burdensome than the other algorithms considered. Figure 3 provides an example of the word *Yes* recorded using the CRPZM microphone, with CDCN processing.

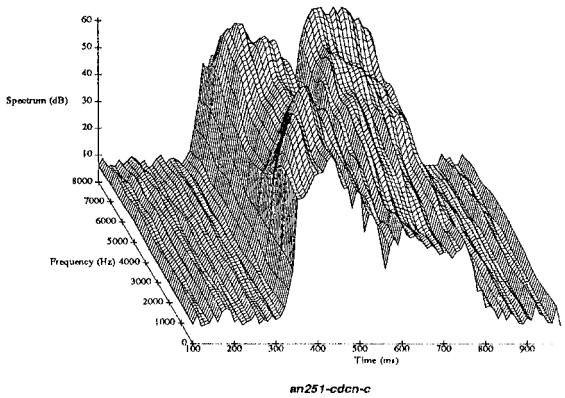

Figure 3: "Yes" with CRPZM and CDCN.

The *Interpolated SNR-Dependent Cepstral Normalization* (ISDCN) algorithm [2] was developed as a means to combine the computational simplicity of the SDCN algorithm with the environmental adaptation capabilities of the CDCN algorithm. This algorithm can be thought of as a modification of the SDCN algorithm, with a correction vector that is interpolated between the noise vector $\mathbf{n}$ at low SNR and the equalization vector $\mathbf{q}$ at high SNR:

$$\mathbf{w}(\mathbf{n}, \mathbf{q}, SNR) = \mathbf{n} + (\mathbf{q} - \mathbf{n}) f(SNR) \quad (7)$$

An approximate ML estimate of $\mathbf{q}$ was obtained by minimization of the accumulated VQ distortion. The interpolating function in (7) was arbitrarily chosen to be the sigmoid function

$$f_i(x) = 1 / [1 + \exp(-\alpha x + \beta)] \quad \alpha > 0 \quad (8)$$

with parameters α and β chosen empirically.

In principle, the estimates of $\mathbf{n}$ and $\mathbf{q}$ for both the CDCN and ISDCN algorithms can be performed on the basis of previous samples of speech, thereby decoupling environmental adaptation from the speech recognition itself. An implementation of the ISDCN algorithm in this fashion produced no additional degradation in recognition accuracy [3].

5. Summary

In this paper we addressed three issues that are important in building robust spoken language systems. We have shown that a *joint* compensation for noise and filtering yields a higher recognition accuracy than an independent compensation because it deals better with colored noise and does not rely on high-SNR speech to estimate the equalization vector.

We have shown that for systems like SPHINX whose parameter vector is the cepstrum, processing in the cepstral domain produces a greater degree of integration, greater efficiency, and higher accuracy than approaches that operate on the spectrum.

Finally, we have described algorithms that will adapt rapidly to a new environment by finding the noise and equalization necessary to normalize the target acoustic space to a universal standard space. The environmental parameters can be slowly updated from past speech frames.

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.

References

1. A. Acero and R. M. Stern, "Environmental Robustness in Automatic Speech Recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 849-852.

2. A. Acero and R. Stern, "Towards Environment-Independent Spoken Language Systems", *Proc. Speech and Natural Language Workshop, Hidden Valley, PA*, Morgan Kaufmann, Jun. 1990.

3. A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, PhD dissertation, Carnegie Mellon University, Sep. 1990.

4. M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", in *Speech Enhancement*, J. S. Lim, ed., Prentice Hall, Englewood Cliffs, NJ, Signal Processing, Vol. 1, 1983, pp. 69-73.

5. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. 27, No. 2, April 1979, pp. 113-120.

6. D. Van Compernolle, "Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Glasgow, UK*, May 1989, pp. 258-261.

7. A. Erell and M. Weintraub, "Estimation Using Log-Spectral-Distance Criterion for Noise-Robust Speech Recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Albuquerque, NM*, April 1990, pp. 853-856.

8. K. F. Lee et al., "The SPHINX Speech Recognition System", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Glasgow, UK*, May 1989, pp. 445-448.

9. A. J. Nadas, D. Nahamoo and M. A. Picheny, "Speech Recognition Using Noise-Adaptive Prototypes", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New York, NY*, 1988, pp. 517-520.

10. A. J. Nadas, D. Nahamoo and M. A. Picheny, "Adaptive labeling: Normalization of Speech by Adaptive Transformations Based on Vector Quantization", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New York, NY*, 1988, pp. 521-524.

11. J. E. Porter and S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, San Diego, CA*, May 1984, pp. 18A.2.1.

12. R. Stern and A. Acero, "Acoustical Pre-processing for Robust Speech Recognition", *Proc. Speech and Natural Language Workshop, Cape Cod, MA*, Morgan Kaufmann, Oct. 1989, pp. 311-318.

13. T. G. Stockham, T. M. Cannon and R. B. Ingebretsen, "Blind Deconvolution Through Digital Signal Processing", *Proc. of the IEEE*, Vol. 63, No. 4, Apr. 1975, pp. 678-692.