Towards Environment-Independent Spoken Language Systems

Alejandro Acero and Richard M. Stern

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

Abstract

In this paper we discuss recent results from our 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 describe two novel methods based on additive corrections in the cepstral domain. In the first algorithm, an additive correction is imposed that 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 these algorithms dramatically improves recognition accuracy when the system is tested on a microphone other than the one on which it was trained.

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 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 [3] and Berouti *et al.* [2] introduced the spectral subtraction of DFT coefficients, and Porter and Boll [11] used MMSE techniques to estimate the DFT coefficients of corrupted speech. Spectral equalization to compensate for convolved distortions 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 [5] and Stern and Acero [12]. Although relatively successful, the above methods all depend on the assumption of independence of the spectral estimates across frequencies. Erell and Weintraub [6] demonstrated improved performance with an MMSE estimator in which correlation among frequencies is modeled explicitly. Acero and Stern [1] proposed an approach to environment normalization in the cepstral domain, going beyond the noise stripping problem.

In this paper we present two algorithms for speech normalization based on additive corrections in the cepstral domain and compare them to techniques that operate in the frequency domain. We have chosen the cepstral domain rather than the frequency domain so that can 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) uses the speech knowledge represented in a codebook to estimate the noise and spectral equalization necessary for the environmental normalization. We also describe an interpolated SDCN algorithm (ISDCN) which combines the simplicity of SDCN and the normalization capabilities of CDCN. These algorithms are evaluated with a number of microphones using an alphanumeric database in which utterances were recorded simultaneously with two different microphones.

Experimental Procedures

The alphanumeric database and system used for these experiments has been described previously [12] [1]. Briefly, the database contains 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 task has a vocabulary of 104 words that are highly confusable. A simplified version of SPHINX with no grammar was used.

Baseline recognition results obtained by training and testing SPHINX using this database are shown in the first two columns of Table 1. 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. 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 [12]. It can also be seen that the "cross conditions" (training on one microphone and testing using the other) produce a very large degradation in recognition accuracy.

Independent Compensation for Noise and Filtering

In this section we examine the performance of SPHINX under some of the techniques that have been used in the literature to combat noise and spectral tilt: multi-style training, short-time liftering, spectral subtraction, and spectral equalization.

Multi-Style Training

Multi-style training is a technique in which the training set includes data representing different conditions so that the resulting HMM models are more robust to this variability. This simple approach has been used successfully in the field of speech styles [10] and speaker independence [9]. The price one must pay for the robustness is a degradation in performance for cases in which the training and testing are done with the same condition.

An experiment was carried out in which all the speech recorded from the CLSTK and the CRPZM microphones were used in training (Table 1). As expected, robustness is gained by using multi-style training, but at the expense of sacrificing performance with respect to the case of train and test on the same conditions.

TRAIN	CLSTK	CRPZM	MULTI
Test **CLSTK**	85.3%	36.9%	78.3%
Test **CRPZM**	18.6%	76.5%	69.7%

Table 1: Comparison of recognition accuracy of SPHINX under different training and testing conditions. CLSTK is the Sennheiser HMD224, CRPZM is the Crown PZM6sf and MULTI means that the data from both microphones were used in training

Liftering

Many studies have examined several potential distortion measures for speech recognition in noise. Most of these measures involve unequal weightings of the mean-square distance between cepstral coefficients of the reference and test utterances. The motivation for weighting distances between cepstral vectors is twofold: it provides some variance normalization for the coefficients and it makes the system more robust to noise and spectral tilt by giving less weight to the low-order cepstral coefficients. We tried in our system several weighting measures that have been proposed in the literature including the inverse of the intra-cluster variance as defined by Tokhura [14], the exponential lifter with s=1.0 which Junqua [8] found to be optimum, and the bandpass liftering method defined by Juang [7].

Unfortunately, we found that application of these techniques produced essentially no improvement for clean speech and only a very small improvement for corrupted speech. Since the frequency-warping transformation in SPHINX alters the variances of the coefficients, some other set of weights may prove more effective.

Spectral Subtraction and Equalization

In spectral subtraction and equalization it is assumed that the speech signal $x(t)$ is degraded by linear filtering and/or uncorrelated additive noise, as depicted in Fig. 1. The goal of the compensation is to reverse the effects of these degradations.

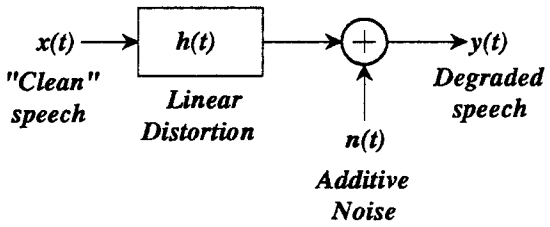

Figure 1: Model of the degradation.

Using the notation of Fig. 1, we can characterize the power spectral density (PSD) of the processes involved as

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

Spectral equalization techniques attempt to compensate for the filter $h(t)$, while spectral subtraction techniques attempt to remove the effects of the noise from the signal. We compare the performance of the following different implementations of spectral subtraction and equalization techniques in Table 2.

- A spectral equalization algorithm (EQUAL) that is similar to the approach of [13]. It compensates for the effects of the linear filtering, but not the additive noise, as described in [12].
- A direct implementation of the original power spectral subtraction rule (PSUB) on 32 frequency bands obtained via a real DFT of the cepstrum vector. The restored cepstrum is obtained with an inverse DFT.
- An implementation of Boll's algorithm (MMSE1) [4], in which a transformation is applied to all the frequency bands of the CRPZM speech that minimizes the mean squared error relative to the CLSTK speech. The log-power correction in each frequency band depended only on the instantaneous SNR in that band.
- An implementation of magnitude spectral subtraction (MSUB) described in [12] that incorporates over- and under-subtraction depending on the SNR as suggested by [2]. In [12] it was noted that a cascade of the EQUAL and MSUB algorithms did not yield any further improvement in recognition accuracy because they interact nonlinearly.

The different criteria used in PSUB, MSUB, produce different curves that relate the effective SNR of the input and output. Some of these curves are shown in Figure 2.

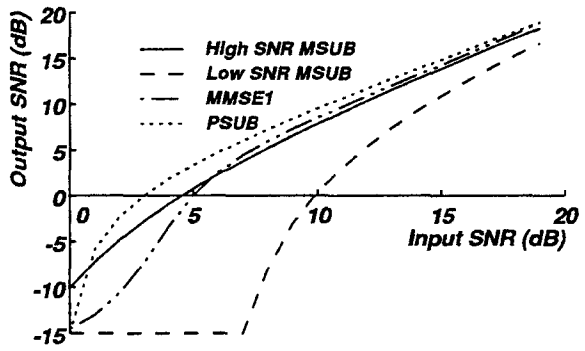

Figure 2: Input-Output transformation curves for PSUB, MSUB and MMSE1. The SNR is defined as the log-power of the signal in a frequency band minus the log-power of the noise in that band. The transformation for MSUB is not a single curve but a family of curves that depend on the total SNR for a given frame.

TRAIN TEST	CLSTK CLSTK	CLSTK CRPZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%
EQUAL	85.3%	38.3%	50.9%	76.5%
PSUB	82.2%	37.2%	62.0%	64.7%
MMSE1	85.3%	48.7%	68.7%	71.4%
MSUB	82.7%	64.8%	75.1%	72.8%

Table 2: Performance of different equalization and spectral subtraction algorithms. EQUAL and MMSE1 were applied only to the CRPZM speech while PSUB and MSUB were applied to both the CLSTK and the CRPZM speech.

For the most part these algorithms provide increasing degrees of compensation, but their recognition accuracy under the "cross" conditions is still much worse than that obtained even with the system is trained and tested on the CRPZM. We have found that the above techniques produce many output frames that do not constitute legitimate speech vectors, especially at low SNR, because they do not take into account correlations across frequency. That problem, along with the nonlinear interaction of the subtraction and normalization processes motivated us to consider new algorithms which jointly compensate for noise and filtering, and with some attention paid to the spectral profile of the compensated speech.

Joint Compensation for Noise and Filtering

In this section we discuss two algorithms that perform noise suppression and spectral-tilt compensation *jointly* in the cepstrum by means of additive corrections.

If we let the cepstral vectors **x**, **n**, **y** and **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

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

where the correction vector $\mathbf{r}(\mathbf{x}, \mathbf{n}, \mathbf{q})$ is given by

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

Let **z** be an estimate of **y** obtained through our spectral estimation algorithm. Our goal is to recover the uncorrupted vectors $\mathbf{X} = \mathbf{x}_0, \ldots \mathbf{x}_{N-1}$ of an utterance given the observations $\mathbf{Z} = \mathbf{z}_0, \ldots \mathbf{z}_{N-1}$ and our knowledge of the environment **n** and **q**.

SDCN Algorithm

SNR-Dependent Cepstral Normalization (SDCN) is a simple algorithm that applies a fixed additive correction vector **w** to the cepstral coefficients that depends exclusively on the instantaneous SNR of the input frame.

$$\hat{\mathbf{x}} = \mathbf{z} - \mathbf{w}(SNR) \quad (4)$$

At high SNR, inspection of equations (1), (2) and (3) indicates that $\mathbf{x}(0) + \mathbf{q}(0) \gg \mathbf{n}(0)$, $\mathbf{r} \approx \mathbf{0}$, and $\mathbf{y} \approx \mathbf{x} + \mathbf{q}$. On the other hand at low SNR, $\mathbf{x}(0) + \mathbf{q}(0) \ll \mathbf{n}(0)$ and $\mathbf{y} \approx \mathbf{n}$. Hence, the SDCN algorithm performs spectral equalization at high SNR and noise suppression at low SNR.

SNR is estimated in the SDCN algorithm as $\mathbf{z}(0) - \mathbf{n}(0)$. This is not the true signal-to-noise ratio but it is related to it and easier to compute. The compensation vectors $\mathbf{w}(SNR)$ were estimated with an MMSE criterion by computing the average difference between cepstral vectors for the test condition versus a standard acoustical environment from simultaneous stereo recordings. We have observed that applying a correction to just c_0 and c_1 yields basically the same results than if all the cepstrum coefficients are normalized.

For the sake of comparison between algorithms operating in the spectral domain and the cepstral domain, we developed an algorithm called MMSEN that accomplishes noise suppression and spectral equalization jointly using different transformations for every frequency band. MMSEN is similar in concept to SDCN except that it operates in the spectral (rather than cepstral) domain. As is seen in Table 4, SDCN performs slightly better than MMSEN, and it is more computationally efficient as well.

TRAIN TEST	CLSTK CLSTK	CLSTK CRPZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%
MMSEN	85.3%	66.4%	75.5%	72.3%
SDCN	85.3%	67.2%	76.4%	75.5%

Table 3: Performance of the MMSEN and SDCN algorithms when compared with the baseline.

Although liftering provided very little improvement for our baseline system, this technique is actually complementary to SDCN: liftering techniques can be viewed as a variance normalization while SDCN is a bias-compensation algorithm. Using SDCN and the algorithm of Juang [7] with $p = 12$ and values of the parameter L ranging from 0 to 6, we observed a modest improvement over pure SDCN (from 67.2% to 72.3%) when training using the CLSTK and testing with the CRPZM microphone.

CDCN Algorithm

Although the SDCN 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, SDCN is clearly not able to model a non-stationary environment. The second new algorithm, *Codeword-Dependent Cepstral Normalization* (CDCN), was proposed to circumvent these problems.

The CDCN algorithm attempts to determine the fixed equalization and noise vectors $\mathbf{q}$ and $\mathbf{n}$ that provide an ensemble of compensated cepstral vectors $\hat{\mathbf{x}}$ that are collectively closest to the set of locations of legitimate VQ codewords. The correction vector will be different for every codebook vector.

The $\mathbf{q}$ and $\mathbf{n}$ are estimated using ML techniques via the EM algorithm since no close-form expression can be obtained. The compensated vectors $\hat{\mathbf{x}}$ are estimated using MMSE techniques. The reader is referred to [1] for the details of this algorithm.

Results and Discussion

Table 4 describes the recognition accuracy of the original SPHINX system with no preprocessing, and with the SDCN and CDCN algorithms. Use of the CDCN algorithm brings the performance obtained when training on the CLSTK 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.

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. 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.

In Figures 3, 4, 5 and 6 we show 3-D representations of an utterance with the CLSTK and no processing, the CRPZM with no processing, SDCN, and CDCN respectively. While it can be seen that noise suppression is achieved with both SDCN and CDCN, the CDCN algorithm provides greater compensation for spectral tilt.

TRAIN TEST	CLSTK CLSTK	CLSTK CRPZM	CRPZM CLSTK	CRPZM CRPZM
BASE	85.3%	18.6%	36.9%	76.5%
SDCN	85.3%	67.2%	76.4%	75.5%
CDCN	85.3%	74.9%	73.7%	77.9%

Table 4: Comparison of recognition accuracy of SPHINX with no processing, SDCN and CDCN algorithms. The system was trained and tested using all combinations of the CLSTK and CRPZM microphones.

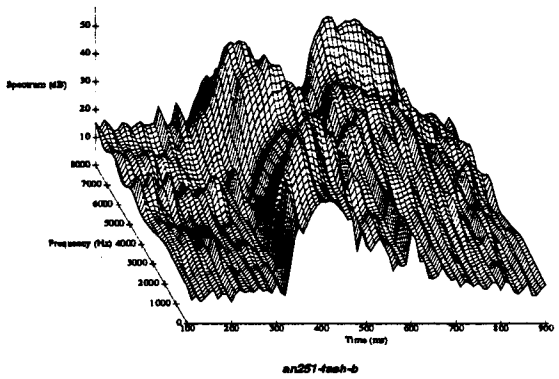

Figure 3: "Yes" with CLSTK and no processing.

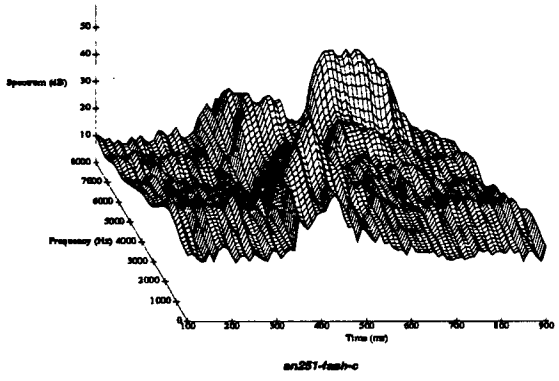

Figure 4: "Yes" with CRPZM and no processing.

Results with other microphones

To confirm the ability of the CDCN algorithm to adapt to new environmental conditions, a series of tests was performed with 5 new stereo speech databases. The test data were all collected after development of the CDCN algorithm was completed. In all cases the system was trained using the Sennheiser HMD224. The "second" microphones (with which the system was *not* trained) were:

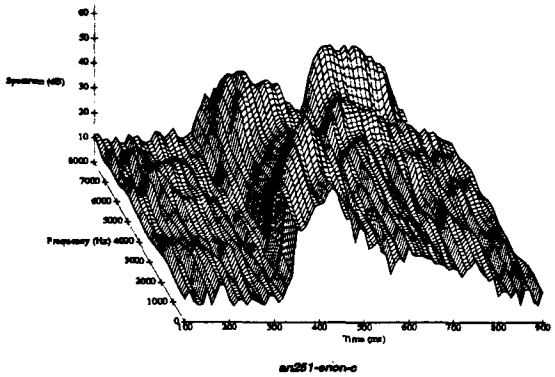

Figure 5: "Yes" with CRPZM and SDCN.

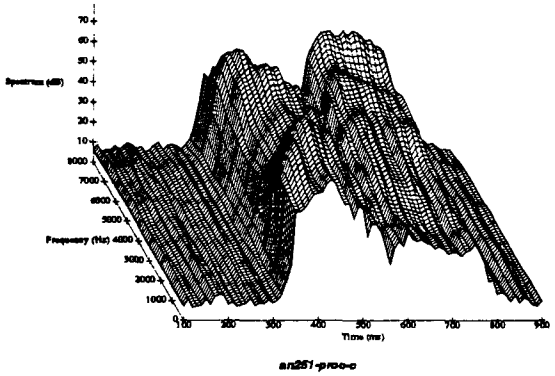

Figure 6: "Yes" with CRPZM and CDCN.

- The Crown PCC160 desk-top phase-coherent cardioid microphone (CRPCC160). (This is the new DARPA "standard" desk-top microphone.)
- An independent test set using the Crown PZM6fs.
- The Sennheiser 518 dynamic cardioid, hand-held microphone (SENN518).
- The Sennheiser ME80 electret supercardioid stand-mounted microphone (SENNME80).
- An HME lavalier microphone that also used an FM receiver (HME).

TEST	CLSTK	CRPCC160
BASE	82.4%	70.2%
CDCN	81.0%	78.5%

TEST	CLSTK	CRPZM6FS
BASE	84.8%	41.8%
CDCN	83.3%	73.9%

In Table 5 we compare results using the CDCN algorithm to baseline performance. With this algorithm great robustness is obtained across microphones. However, there is a slight drop in performance when training and testing on the Sennheiser HMD224. We believe that one cause for this is that estimates of **q** and **n** are not very good for short utterances.

TEST	CLSTK	SENN518
BASE	87.2%	84.5%
CDCN	82.2%	83.3%

TEST	CLSTK	SENNME80
BASE	83.7%	71.4%
CDCN	81.5%	80.7%

TEST	HME	CRPCC160
BASE	55.9%	56.3%
CDCN	81.7%	72.2%

Table 5: Analysis of performance of SPHINX for the baseline and the CDCN algorithm. Two microphones were recorded in stereo in each case. The microphones compared are the Sennheiser HMD224, 518, ME80, the Crown PZM6FS and PCC160, and the HME microphone. Training was done with the Sennheiser HMD224 in all cases.

Interpolated SDCN

One of the deficiencies of the SDCN algorithm is the inability to adapt to new environments since the correction vectors are derived from a stereo database of our "standard" Sennheiser HMD224 and the new microphone. By using an MMSE criterion that included some *a priori* information about the distribution of speech (a codebook), the SDCN can estimate the parameters of the environment **q** and **n** just as CDCN does.

As we have noted above, the correction vector in SDCN, **w**, has the asymptotic value of the noise vector **n** at low SNR and of the equalization vector **q** at high SNR. In interpolated SDCN (ISDCN) the dependence on SNR is modelled as follows:

$$\mathbf{w}_i(SNR) = \mathbf{n}_i + (\mathbf{q}_i - \mathbf{n}_i) f_i(SNR) \quad (5)$$

where $f_i(SNR)$ is chosen to be the sigmoid function

$$f_i(x) = 1/[1 + \exp(-\alpha_i x + \beta_i)] \quad (6)$$

In this evaluation α_i and β_i were set empirically to 3.0 for $i > 0$ and 6.0 for $i = 0$. The vectors **n** and **q** were determined by an EM algorithm whose objective function is the minimization of the total VQ distortion.

In evaluating the ISDCN algorithm we also varied the amount of speech used for estimation of **q** and **n**. Since these parameters are normally estimated over the course of only a single utterance, the estimates of **q** and **n** will exhibit a large variance for short utterances. We believe this is one of the causes for the slight degradation in performance in Table 5 observed when the system was trained and tested using the CLSTK microphone.

We compared the recognition accuracy with the ISDCN algorithm using estimates of the model parameters obtained by considering only one utterance at a time, and with estimates obtained using all 14 utterances spoken by a given speaker. Estimating the model parameters from all utterances for a speaker produced an accuracy of 85.9%, which is slightly higher than the baseline 85.3%. (The corresponding recognition accuracy working an utterance at a time was 84.8%.) These results lead us to believe that CDCN could also benefit from a longer estimation time, and will be analyzed in future work.

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.

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. 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. M. Berouti, R. Schwartz and J. Makhoul. *Signal Processing*. Volume 1: Enhancement of Speech Corrupted by Acoustic Noise. In *Speech Enhancement*, J. S. Lim, Ed., Prentice Hall, Englewood Cliffs, NJ, 1983, pp. 69-73.

3. S. F. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Trans. Acoustics, Speech and Signal Processing 27*, 2 (April 1979), 113-120.

4. S. Boll, J. Porter and L. G. Bahler. Robust Syntax Free Speech Recognition. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, New York, NY, 1988, pp. 179-182.

5. 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.

6. A. Erell and M. Weintraub. Spectral Estimation for Noise Robust Speech Recognition. Proc. Speech and Natural Language Workshop, Cape Cod, MA, Oct., 1989.

7. B. H. Juang, L. R. Rabiner and J. G. Wilpon. "On the Use of Bandpass Liftering in Speech Recognition". *IEEE Trans. Acoustics, Speech and Signal Processing ASSP-35* (Jul. 1987), 947-954.

8. J. C. Junqua and H. Wakita. A Comparative Study of Cepstral Lifters and Distance Measures for All-Pole Models of Speech in Noise. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Glasgow, UK, May, 1989, pp. 476-479.

9. 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.

10. R. P. Lippmann, E. A. Martin and D.B. Paul. Multi-Style Training for Robust Isolated-Word Speech Recognition. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Dallas, TX, April, 1987, pp. 705-708.

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, 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 63*, 4 (Apr. 1975), 678-692.

14. Y. Tokhura. "A Weighted Cepstral Distance Measure for Speech Recognition". *IEEE Trans. Acoustics, Speech and Signal Processing ASSP-35* (Oct. 1987), 1414-1422.