HMM ADAPTATION USING VECTOR TAYLOR SERIES FOR NOISY SPEECH RECOGNITION

Alex Acero, Li Deng, Trausti Kristjansson, Jerry Zhang

Speech Technology Group
Microsoft Research
One Microsoft Way
Redmond, Washington 98052, USA
http://research.microsoft.com/srg

ABSTRACT

In this paper we address the problem of robustness of speech recognition systems in noisy environments. The goal is to estimate the parameters of a HMM that is matched to a noisy environment, given a HMM trained with clean speech and knowledge of the acoustical environment. We propose a method based on truncated vector Taylor series that approximates the performance of a system trained with that corrupted speech. We also provide insight on the approximations used in the model of the environment and compare them with the lognormal approximation in PMC.

1. Introduction

Speech recognition systems built in the lab with clean speech can often offer very high accuracies. But accuracy degrades significantly when such systems are used in the real world, mostly because of the mismatch between the clean speech in training and the real world speech [1].

If there is a mismatch between acoustical environments, it is sensible to retrain the HMM. This is done in practice for telephone speech where only telephone speech, and no clean high-bandwidth speech, is used in the training phase. Unfortunately, training a large vocabulary speech recognizer requires a very large amount of data, which is often not available for a specific noisy condition. For example, it is difficult to collect a large amount of training data in a car driving at 50 mph, whereas it is much easier to record it at idle speed or in a lab. Often we want to adapt our model given a relatively small sample of speech from the new acoustical environment.

If additive noise is the only degradation, one option is to take a noise waveform from the new acoustical environment, add it to all the utterances in our training database and retrain the system [3]. If the noise characteristics are known ahead of time, this method allows us to adapt the model to the new environment with a relatively small amount of data from the new environment, yet use a large amount of training data. This simple technique can provide good results at no cost during recognition if the noise sample is available offline. If the target acoustical environment also has a different channel, we can also filter all the utterances in the training data prior to retraining [5]. If this noise were not available beforehand, the noise addition and model retraining would need to occur at run-time. This is feasible for speaker-dependent small vocabulary systems where the training data can be kept in memory and where the retraining time can be small, but is probably not feasible for large vocabulary speaker-independent systems because of memory and computational limitations. We can also pool training data from different environments [3].

It is useful to consider whether this retraining can be done only from the HMM itself, instead of requiring all the training waveforms, since this would greatly diminish the memory required. Parallel Model Combination [4] is one such approach that estimates the matched noisy model from the clean HMM using a nonlinear transform, derived from a variant of the model of the environment described in Section 2. PMC makes the approximation that the nonlinear function of Gaussian random vectors is also Gaussian so that the same decoder can be used.

In this paper we propose an approximation based on Taylor series, which is an extension of [8] from the feature space to the model space. Our proposed approach differs from [8] in that we derive expressions for the variance of the noise and delta and delta-delta means and variances. We compare the proposed approximation with the lognormal approximation in PMC. The proposed method approaches the error rate of a noisy-matched trained system.

In Section 2 we describe the model of the environment. Section 3 analyzes the empirical distribution of the corrupted speech. Section 4 introduces the new approach. The approximations and a comparison with PMC are analyzed in section 5. Section 6 presents some experimental results.

2. A Model of the Environment

Figure 1 shows a commonly used model for the acoustical environment [1], which assumes the speech signal $x[m]$ is corrupted by additive noise $n[m]$ and channel distortion $h[m]$:

$$y[m] = x[m] * h[m] + n[m] \tag{1}$$

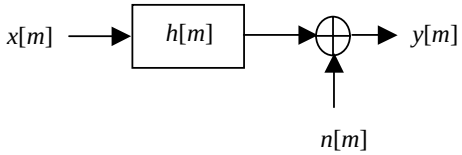

Figure 1. A model of the acoustical environment.

It is convenient to express Eq. (1) in the frequency domain. To do that, we window the signal, take a $2K$-point DFT and then take the square of the magnitude:

$$\begin{aligned} |Y(f_k)|^2 &= |X(f_k)|^2 |H(f_k)|^2 + |N(f_k)|^2 \\ &+ 2\,\mathrm{Re}\{X(f_k)H(f_k)N^*(f_k)\} \end{aligned} \tag{2}$$

where $k = 0,1,\cdots,K$.

Statistically, the expected value of the last term in Eq. (2) is zero since $x[m]$ and $n[m]$ are statistically independent. In practice, this term is not zero (see Figure 7), though it is small if we average over a range of frequencies, as is often done

when computing the popular mel-cepstrum [2]. When using a filterbank, we can obtain a relationship for the energies at each of the *M* filters:

$$|Y(f_i)|^2 = |X(f_i)|^2 |H(f_i)|^2 + |N(f_i)|^2 \quad (3)$$

An implicit assumption of Eq. (2) and (3) is that the length of $h[n]$, the filter's impulse response, is much shorter than the window length $2N$. That means that for filters with long reverberation times, Eq. (3) is inaccurate. For example, for $|N(f)|^2 = 0$, a window shift of T and a filter's impulse response $h[n] = \delta[n-T]$, we have $Y_t[f_m] = X_{t-1}[f_m]$, i.e. the output spectrum at frame t does not depend on the input spectrum at that frame. This is a more serious assumption, which is why speech recognition systems tend to fail under long reverberation times [9].

Taking logarithms in Eq. (3) and after some algebraic manipulation we obtain

$$\begin{aligned} \ln|Y(f_i)|^2 &= \ln|X(f_i)|^2 + \ln|H(f_i)|^2 \\ &+ \ln\left(1+\exp\left(\ln|N(f_i)|^2 - \ln|X(f_i)|^2 - \ln|H(f_i)|^2\right)\right) \end{aligned} \quad (4)$$

Let's define the following length (M+1) cepstrum vectors

$$\begin{aligned} \mathbf{x} &= \mathbf{C}\left(\ln|X(f_0)|^2 \quad \ln|X(f_1)|^2 \quad \cdots \quad \ln|X(f_M)|^2\right) \\ \mathbf{h} &= \mathbf{C}\left(\ln|H(f_0)|^2 \quad \ln|H(f_1)|^2 \quad \cdots \quad \ln|H(f_M)|^2\right) \\ \mathbf{n} &= \mathbf{C}\left(\ln|N(f_0)|^2 \quad \ln|N(f_1)|^2 \quad \cdots \quad \ln|N(f_M)|^2\right) \\ \mathbf{y} &= \mathbf{C}\left(\ln|Y(f_0)|^2 \quad \ln|Y(f_1)|^2 \quad \cdots \quad \ln|Y(f_M)|^2\right) \end{aligned} \quad (5)$$

with **C** being the DCT matrix. Combining Eq. (4) with (5) results in

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{n} - \mathbf{x} - \mathbf{h}) \quad (6)$$

where the nonlinear function $\mathbf{g}(\mathbf{z})$ is given by

$$\mathbf{g}(\mathbf{z}) = \mathbf{C}\ln\left(1 + e^{\mathbf{C}^{-1}\mathbf{z}}\right) \quad (7)$$

Eq. (6) and (7) say that we can compute the cepstrum of the corrupted speech if we know the cepstrum of the clean speech **x**, the cepstrum of the noise **n** and the cepstrum of the filter **h**. In practice, the DCT matrix **C** is not square, so that dimension of the cepstrum vector is smaller than the number of filters. This means that we are losing resolution when going back to the frequency domain, and thus Eq. (6) and (7) represent only an approximation.

3. Empirical Distribution of Corrupted Speech

In continuous-density HMM-based speech recognition systems, the output pdf for the cepstrum of the clean speech **x** is typically a mixture of Gaussians. In this section, we examine the effect to that distribution under the model of the environment of Figure 1. Even if we assume that the noise **n** follows a Gaussian distribution, the cepstrum of the corrupted speech **y** in Eq. (6) is no longer a mixture of Gaussians because of the non-linearity in Eq. (7). Nonetheless it is convenient to assume it still follows a mixture of Gaussians because that way we can use the same decoder we use for clean speech. Furthermore, it is generally assumed that *each* mixture component is still Gaussian after undergoing the transform in Eq. (7) because of expediency [4].

It is difficult to visualize the effect on the distribution given the non-linearities involved. To provide some insight, let's consider Eq. (4) for a given frequency when no filtering is done, *i.e.* $H(f) = 1$, with $n = \ln|N(f)|^2$, $x = \ln|X(f)|^2$:

$$y = x + \ln\left(1 + \exp(n - x)\right) \quad (8)$$

Now let's assume that both x and n are Gaussian random variables. We can use Monte Carlo simulation to draw a large number of points from those two Gaussian distributions, and obtain the corresponding noisy values y using Eq. (8). Figure 2 shows the resulting distribution for several values of σ_x. We fixed $\mu_n = 0dB$, since it is only a relative level, and set $\sigma_n = 2dB$, a typical value. We also set $\mu_x = 25dB$ and see that the resulting distribution can be bimodal when σ_x is very large. Similar graphs are shown in [4][8]. Fortunately, for modern speech recognition systems that have many Gaussian components, σ_x is never that large and the resulting distribution is unimodal.

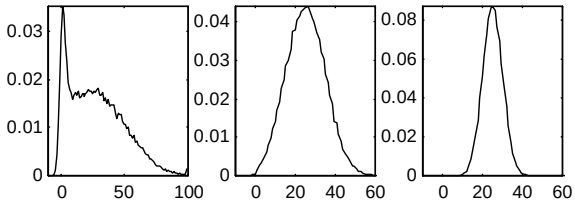

Figure 2 Distributions of y in Eq. (8) for $\mu_n = 0dB$, $\sigma_n = 2dB$, $\mu_x = 25dB$ and σ_x of 25, 10 and 5 dB respectively.

Figure 3 shows the distribution of y for two values of μ_x given the same values for the noise distribution, $\mu_n = 0dB$ and $\sigma_n = 2dB$, and a more realistic value for $\sigma_x = 5dB$. We see that the distribution is unimodal, though not necessarily symmetric, particularly for low SNR ($\mu_x - \mu_n$).

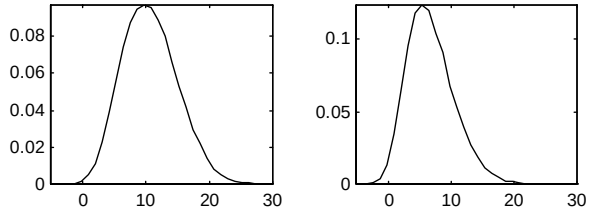

Figure 3 Distributions of y in Eq. (8) for $\mu_n = 0dB$, $\sigma_n = 2dB$, $\sigma_x = 5dB$ and μ_x of 10 and 5 dB respectively.

4. Vector Taylor Series (VTS)

Moreno [8] suggests the use of Taylor series to approximate the non-linearity in Eq. (7) as a feature preprocessor with a Gaussian mixture in the spectral domain. Here we extend that work to the model space, whose Gaussians are in the cepstral domain, include the covariance of the noise and extend it to the delta and delta-delta features.

Assume that **x**, **h** and **n** are Gaussian with means $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_h$ and $\boldsymbol{\mu}_n$ and covariance matrices $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\Sigma}_n$ respectively, and furthermore that **x**, **h** and **n** are independent. After algebraic manipulation it can be shown that the Jacobian of Eq. (6) with

respect to **x**, **h**, and **n** evaluated at $\boldsymbol{\mu} = \boldsymbol{\mu}_{\mathbf{n}} - \boldsymbol{\mu}_{\mathbf{x}} - \boldsymbol{\mu}_{\mathbf{h}}$, can be expressed as

$$\left.\frac{\partial \mathbf{y}}{\partial \mathbf{x}}\right|_{(\boldsymbol{\mu}_n,\boldsymbol{\mu}_x,\boldsymbol{\mu}_h)} = \left.\frac{\partial \mathbf{y}}{\partial \mathbf{h}}\right|_{(\boldsymbol{\mu}_n,\boldsymbol{\mu}_x,\boldsymbol{\mu}_h)} = \mathbf{A}$$
$$\left.\frac{\partial \mathbf{y}}{\partial \mathbf{n}}\right|_{(\boldsymbol{\mu}_n,\boldsymbol{\mu}_x,\boldsymbol{\mu}_h)} = \mathbf{I} - \mathbf{A} \tag{9}$$

where the matrix **A** being given by

$$\mathbf{A} = \mathbf{C}\mathbf{F}\mathbf{C}^{-1} \tag{10}$$

and **F** is a diagonal matrix whose elements are given by vector $\mathbf{f}(\boldsymbol{\mu})$, which in turn is given by

$$\mathbf{f}(\boldsymbol{\mu}) = \frac{1}{1+e^{\mathbf{C}^{-1}\boldsymbol{\mu}}} \tag{11}$$

Using Eq. (9) we can then approximate Eq. (6) by a first order Taylor series expansion around $(\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{h}})$ as

$$\begin{aligned}\mathbf{y} &\approx \boldsymbol{\mu}_x + \boldsymbol{\mu}_h + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x - \boldsymbol{\mu}_h) \\ &+\mathbf{A}(\mathbf{x} - \boldsymbol{\mu}_x) + \mathbf{A}(\mathbf{h} - \boldsymbol{\mu}_h) + (\mathbf{I} - \mathbf{A})(\mathbf{n} - \boldsymbol{\mu}_n)\end{aligned} \tag{12}$$

The mean of $\mathbf{y}$, $\boldsymbol{\mu}_y$, can be obtained from Eq. (12) as

$$\boldsymbol{\mu}_y \approx \boldsymbol{\mu}_x + \boldsymbol{\mu}_h + \mathbf{g}(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x - \boldsymbol{\mu}_h) \tag{13}$$

and its covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}}$ by

$$\boldsymbol{\Sigma}_{\mathbf{y}} \approx \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{A}^T + \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{h}}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}_{\mathbf{n}}(\mathbf{I} - \mathbf{A})^T \tag{14}$$

so that even if $\boldsymbol{\Sigma}_x$, $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\Sigma}_n$ are diagonal, $\boldsymbol{\Sigma}_{\mathbf{y}}$ is no longer diagonal. Nonetheless, it can be assumed it is diagonal because this way, we can transform a clean HMM to a corrupted HMM that has the same functional form and use a decoder that has been optimized for diagonal covariance matrices.

To compute the means and covariance matrices of the delta and delta-delta parameters, let's take the derivative of the approximation of **y** in Eq. (12) with respect to time:

$$\frac{\partial \mathbf{y}}{\partial t} \approx \mathbf{A}\frac{\partial \mathbf{x}}{\partial t} \tag{15}$$

Since delta-cepstrum is computed through $\Delta\mathbf{x}_t = \mathbf{x}_{t+2} - \mathbf{x}_{t-2}$, Gopinath *et al* [6] showed that it is related to the derivative by

$$\Delta\mathbf{x} \approx 4\frac{\partial \mathbf{x}_t}{\partial t} \tag{16}$$

so that

$$\boldsymbol{\mu}_{\Delta\mathbf{y}} \approx \mathbf{A}\boldsymbol{\mu}_{\Delta\mathbf{x}} \tag{17}$$

and similarly

$$\boldsymbol{\Sigma}_{\Delta\mathbf{y}} \approx \mathbf{A}\boldsymbol{\Sigma}_{\Delta\mathbf{x}}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}_{\Delta\mathbf{n}}(\mathbf{I} - \mathbf{A})^T \tag{18}$$

where we assumed that **h** is constant within an utterance so that $\Delta\mathbf{h} = 0$.

Similarly, for the delta-delta cepstrum, the mean is given by

$$\boldsymbol{\mu}_{\Delta^2\mathbf{y}} \approx \mathbf{A}\boldsymbol{\mu}_{\Delta^2\mathbf{x}} \tag{19}$$

and the covariance matrix is given by

$$\boldsymbol{\Sigma}_{\Delta^2\mathbf{y}} \approx \mathbf{A}\boldsymbol{\Sigma}_{\Delta^2\mathbf{x}}\mathbf{A}^T + (\mathbf{I} - \mathbf{A})\boldsymbol{\Sigma}_{\Delta^2\mathbf{n}}(\mathbf{I} - \mathbf{A})^T \tag{20}$$

where we again assumed that **h** is constant within an utterance so that $\Delta^2\mathbf{h} = 0$.

Eq. (13), (17) and (19) resemble the MLLR adaptation formulae [7] for the means, though in this case the matrix is different for each Gaussian and is heavily constrained.

We are interested in estimating the environmental parameters $\boldsymbol{\mu}_{\mathbf{n}}$, $\boldsymbol{\mu}_{\mathbf{h}}$ and $\boldsymbol{\Sigma}_{\mathbf{n}}$ given a set of *T* observation frames $\mathbf{y}_t$. This estimation can be done iteratively using the EM algorithm on Eq. (12). If the noise process is stationary, $\boldsymbol{\Sigma}_{\Delta\mathbf{n}}$ could be approximated, assuming independence between $\mathbf{n}_{t+2}$ and $\mathbf{n}_{t-2}$, by $\boldsymbol{\Sigma}_{\Delta\mathbf{n}} = 2\boldsymbol{\Sigma}_{\mathbf{n}}$. Similarly, $\boldsymbol{\Sigma}_{\Delta^2\mathbf{n}}$ could be approximated, assuming independence between $\Delta\mathbf{n}_{t+1}$ and $\Delta\mathbf{n}_{t-1}$, by $\boldsymbol{\Sigma}_{\Delta^2\mathbf{n}} = 4\boldsymbol{\Sigma}_{\mathbf{n}}$. If the noise process is not stationary, it is best to estimate $\boldsymbol{\Sigma}_{\Delta\mathbf{n}}$ and $\boldsymbol{\Sigma}_{\Delta^2\mathbf{n}}$ from input data directly.

5. Analysis of the Approximations

There are three main variants of the PMC method [1] depending on how the means and covariance matrices are computed. *Numerical integration* is the most accurate way of estimating the mean and covariance matrix of each transformed Gaussian component, but it also is very computationally expensive. *Data-driven PMC* (DMPC) is another variant that obtains the mean and covariance matrix through Monte Carlo simulation, and that requires a sample of at least 100 vectors per Gaussian to obtain similarly accurate results. Finally, the popular *lognormal approximation* is yet another variant of PMC that approximates the sum of two lognormal distributions as lognormal, but is not as accurate as the two above and also cannot be used for the delta and delta-delta parameters.

Here we compare the Monte Carlo approximation, the lognormal approximation, and the VTS approximation of Section 4. For simplicity the simulations were done in the spectral domain and not the cepstral domain, as it is simpler to interpret the results. In Figure 4 we show the mean and standard deviation of *y* in dB from Eq. (8) as a function of the μ_x where $\sigma_x = 10\text{dB}$, $\mu_n = 0\text{dB}$ and $\sigma_n = 2\text{dB}$. We see that the VTS approximation is more accurate than the lognormal approximation for the mean and especially for the standard deviation.

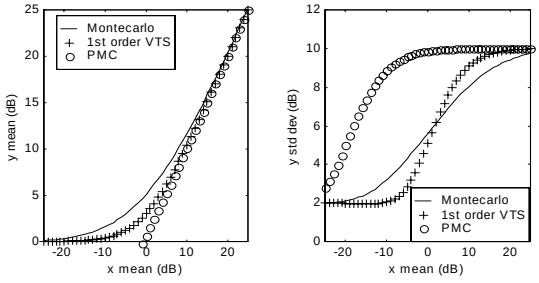

Figure 4 Means and standard deviation of *y* in Eq. (8) for the MonteCarlo method, lognormal and VTS approximations for $\mu_n = 0\text{dB}$, $\sigma_n = 2\text{dB}$, $\sigma_x = 10\text{dB}$ and μ_x varying from –25dB to 25dB.

Figure 5 is similar to Figure 4 only that $\sigma_x = 5\text{dB}$, a more realistic number in speech recognition systems. In this case, both the lognormal approximation and the first order VTS approximation are good estimates of the mean, though the standard deviation estimated through the lognormal approximation in PMC is not as good as that obtained through first order VTS.

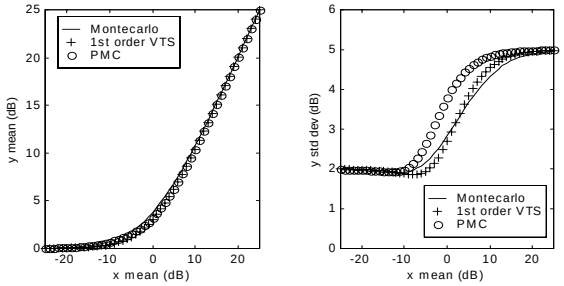

Figure 5 Means and standard deviation of *y* in Eq. (8) for the MonteCarlo method, lognormal and VTS approximations for $\mu_n = 0\text{dB}$, $\sigma_n = 2\text{dB}$, $\sigma_x = 5\text{dB}$ and μ_x varying from –25dB to 25dB.

Figure 6 shows that in practice, VTS is doing a reasonable job of approximating the noisy distribution, though the cross terms are not negligible as shown in Figure 7.

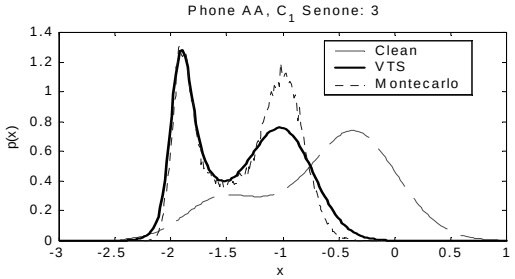

Figure 6. Distribution of **y**[1] of senone 3 of /AA/ for 10dB white noise: clean model (dashed line), VTS (solid line) and Monte-Carlo simulation (dotted line).

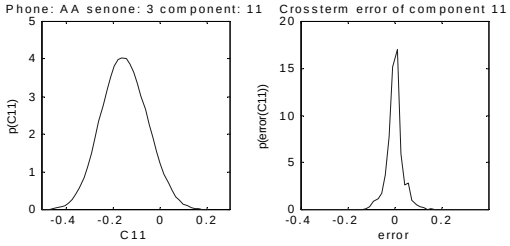

Figure 7. Distribution of **y**[11] of senone 3 of /AA/ for 10dB white noise (left) and error histogram in this component due to the omission of the cross-terms (right).

6. Experimental Results

For evaluation we used the standard 5000-word continuous speech speaker-independent Wall Street Journal task with a bigram language model, a training set of about 16,000 sentences and a test set of 167 sentences. The baseline system is a tied-state continuous-density HMM with 6000 states and 20 Gaussians per state and a 33-dimensional feature vector composed of static, delta and delta-delta MFCC. The baseline word error rate under the clean acoustic environment is 4.87%. Office noise was added to the test set at SNRs ranging from 20 to -10 dB. The error rate of the clean uncompensated model for office noise increases to 55% at -10dB. For the matched condition, noise was also added to the training data. The proposed algorithm approached the matched noisy conditions.

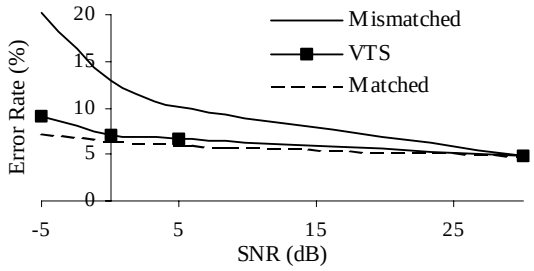

Figure 8 Word error rate as a function of the testing data SNR (dB) for both a system trained on clean data (solid line), a system trained on noisy data at the same SNR as the testing set (dotted line) and the VTS algorithm. Office noise at different SNRs is added.

7. Conclusions and Future Work

A method for estimating the HMM model parameters under noise and channel distortions has been presented, based on truncated Taylor series. The performance of this method is close to that of a system trained with that corrupted speech. The Taylor series approximation appears to be more accurate than the lognormal approximation in PMC.

Future work includes estimation of the environment parameters from data using maximum likelihood, in a EM fashion, or MAP if prior knowledge about the environment is available.

References

[1] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Kluwer Academic Publishers, 1993.

[2] S. Davis and P. Mermelstein. "Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences". *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol 28, No 4, pp. 28, 1980.

[3] L. Deng, A. Acero, M Plumpe and X. Huang. "Large Vocabulary Speech Recognition Under Adverse Acoustic Environments". *Proc. ICSLP*, Beijing, China, 2000.

[4] M. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 1995.

[5] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer. "Training of HMM with Filtered Speech Material for Hands-Free Speech Recognition". In *Proc. Of ICASSP*, Mar. 1999, pp. 449-452.

[6] R.A. Gopinath *et al*. "Robust Speech Recognition in Noise: Performance of the IBM Continuous Speech Recognizer on the ARPA Noise Spoke Task". *Proc. ARPA Workshop on Spoken Language Systems Technology*. 1995.

[7] C.J. Leggetter, P.C. Woodland. "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression". *Proc. of the Int. Conf. on Spoken Language Processing*, Yokohama, Japan, 1994.

[8] P. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University, 1996.

[9] T. M. Sullivan R. M. Stern. "Multi-Microphone Correlation-Based Processing for Robust Speech Recognition". *Proc. Int. Conf. On Acoustics, Speech and Signal Processing*, Minneapolis, Minnesota, 1993.