Nonlinear Noise Compensation in Feature Domain with Numerical Methods

Hui Jiang

Technical Report CS-2003-02

Feburary 6, 2003

Department of Computer Science

4700 Keele Street North York, Ontario M3J 1P3 Canada

Non-linear Noise Compensation in Feature Domain with Numerical Methods

Hui Jiang
Department of Computer Science, York University
4700 Keele Street, Toronto, Ontario, M3J 1P3, CANADA
Jan. 29, 2003

1 Introduction

In the past decade, the performance of automatic speech recognition (ASR) has been significantly improved. More and more ASR systems are being deployed in many real-field applications. In many situations, these speech recognition systems must be operated in some adverse environments, where ambient noise becomes the major hurdle to achieve high-accuracy recognition performance. How to improve environmental robustness of ASR has been intensively studied in the speech community. In the literature, a variety of noisy speech recognition techniques usually fall into two main categories. In the first one, we try to remove or compensate the effect of noise in speech signals prior to the actual recognition procedure. The noise compensation methods can be performed in the time domain (such as many early speech enhancement methods), the spectral domain, or the real feature domain used by most speech recognizers, such as the log-cepstrum, LPC-cepstrum, MFCC, or etc. It has been shown that the methods applied to the ASR feature domains usually yield the better performance in terms of improving ASR noise robustness. The most popular techniques in this category include spectral subtraction[13, 5], Wiener filtering[4] transformation based on stereo data[1, 6, 7], linear noise compensation based on Taylor series approximation[14, 10, 2], feature domain stochastic matching[17], and so on. In second category, the effect of noise is compensated within speech recognition procedure. It usually involves adapting or modifying acoustic models (usually HMM's) of the ASR systems to match the noisy speech feature in a new testing environment. The methods applied in the HMM model domain always are more computationally expensive than others. The representative methods in the category include parallel model combination (PMC)[8], model adaptation using MLLR (maximum likelihood linear regression)[12] or MAP (maximum *a posteriori*) [9], Jacobian environment adaptation[16], speech and noise decomposition[20], model space stochastic matching[17]. It is well known that the distortion caused by additive ambient noises is highly non-linear in the log-spectral or cepstral domain. However, due to computational complexity issue, most noise compensation methods for ASR are approximated by some linear functions, such as in simple bias removal, an affine transformation, linear regression, first order Taylor series expansion,

and so on. In the literature, there are only some very limited efforts to compensate noise with any non-linear ways, such as higher order Taylor series expansion, neural networks under the framework of stochastic matching[19].

In this study, we propose to compensate additive noise in the log-spectral domain based on its original non-linear distortion function. We assume the clean speech follows a Gaussian mixture model in the log-spectral domain and noise signal is a single Gaussian distribution. Given any noisy speech observation, we estimate the clean speech by using the original nonlinear distortion function among noise, clean and noisy speech based on the MMSE (minimum mean square error) criterion. The MMSE estimation of clean speech ends up with a complex integral. In this work, we propose an efficient algorithm to use some numerical methods to solve the integral. At last, the estimated clean speech will be mapped from the log-spectral domain into the MFCC domain, and sent to a speech recognizer for the recognition results.

2 Environmental Model for Speech in Additive noise

Assume we have clean speech $x(t)$ in the time domain and $x(t)$ is corrupted by an independent ambient noise $n(t)$ (also in the time domain). The resultant noisy speech can be expressed in the time domain as:

$$y(t) = x(t) + n(t) \tag{1}$$

Usually we can assume $x(t)$ and $n(t)$ are statistically independent.

If we convert the signals into the log-spectrum domain (either linear or Mel-scale), the above simple relation becomes a complex nonlinear function (see [1]). For d-th filter bank (or d-th frequency bin), we have

$$\mathbf{y}_d = \mathbf{x}_d + \ln\left(1 + e^{\mathbf{n}_d - \mathbf{x}_d}\right) \tag{2}$$

If we assume the independence between all different filter banks, then we can drop the subscript d for clarity. We just repeat the same operation for all different filter banks (or feature dimensions). Hereafter, we use letters in bond to represent the corresponding signals in the log-cepstrum domain, i.e., $\mathbf{y}$ denotes noisy speech in the log-cepstrum domain, $\mathbf{x}$ for clean speech and $\mathbf{n}$ for noise. Then, we can have the following three equivalent functions for $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{n}$:

$$\mathbf{y} = \mathbf{x} + \ln\left(1 + e^{\mathbf{n}-\mathbf{x}}\right) \tag{3}$$

$$\mathbf{x} = \ln\left(e^{\mathbf{y}} - e^{n}\right) \tag{4}$$

$$\mathbf{n} = \mathbf{x} + \ln\left(e^{\mathbf{y}-\mathbf{x}} - 1\right) \tag{5}$$

3 MMSE Estimation of Clean Speech

Based on the above non-linear environmental model, for any given noisy speech feature $\mathbf{y}$, we will try to estimate a clean speech $\hat{\mathbf{x}}$ in the MMSE (minimum mean square error) sense. Without losing generality, we assume the clean speech feature $\mathbf{x}$ in the log-spectral domain follows a Gaussian mixture model (GMM) as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_{xk}, \sigma_{xk}^2) \tag{6}$$

where μ_{xk} and σ_{xk}^2 are mean and variance of k-th Gaussian mixand, and w_k is the weight of k-th mixand with the constraint $\sum_{k=1}^{K} w_k = 1$. The GMM model of speech signals may be constant for all frames in an utterances, or may change from one frame to another. In the former case, we can train a generic GMM from clean speech data. In the latter one, for any a particular feature vector, we can use a proper HMM state from the whole HMM sets for clean speech.

Besides, we assume noise signals in the log-spectral domain follows a single Gaussian distribution as:

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n} \mid \mu_{\mathbf{n}}, \sigma_{\mathbf{n}}^{\mathbf{2}}) \tag{7}$$

where μ_n and σ_n^2 are mean and variance of noise signals. They can be estimated from some initial noise frames in an utterance. Alternatively, if the clean speech distribution $p(\mathbf{x})$ is known, μ_n and σ_n^2 can also be refined based on the EM algorithm.

3.1 Deriving the distribution for noisy speech y

Given the pdf's of clean speech $\mathbf{x}$ and noise $\mathbf{n}$ in eqs.(6) and (7), as well as the environmental model for noisy speech $\mathbf{y}$ in eq.(3), here, we are interested in deriving a conditional distribution of $\mathbf{y}$ given clean speech $\mathbf{x}$, i.e., $p(\mathbf{y}|\mathbf{x})$. If $\mathbf{x}$ is given, $\mathbf{y}$ can be viewed as a transformation from the Gaussian random variable $\mathbf{n}$ (with its distribution in eq.(7)) according to eq.(3). If $\mathbf{x}$ is fixed, from eq.(3) we know the transformation from $\mathbf{n}$ to $\mathbf{y}$ is a one-to-one monotonic mapping. Therefore, $p(\mathbf{y}|\mathbf{x})$ can be derived as: (see [15] for the theorem on transformation of random variables)

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &\equiv \left|\frac{\mathrm{d}\mathbf{n}}{\mathrm{d}\mathbf{y}}\right| \cdot p(\mathbf{n})\Big|_{\mathbf{n}=\mathbf{x}+\ln(e^{\mathbf{y}-\mathbf{x}}-1)} \\ &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \cdot \frac{e^{\mathbf{y}-\mathbf{x}}}{e^{\mathbf{y}-\mathbf{x}}-1} \cdot e^{-\frac{1}{2\sigma_n^2}\left[\mathbf{x}+\ln(e^{\mathbf{y}-\mathbf{x}}-1)-\mu_n\right]^2} \\ &= \frac{1}{\sqrt{2\pi\sigma_n^2}} \cdot \frac{\psi(\mathbf{x},\mathbf{y})}{\psi(\mathbf{x},\mathbf{y})-1} \cdot e^{-\frac{[\mathbf{x}-\mu_n+\ln(\psi(\mathbf{x},\mathbf{y})-1)]^2}{2\sigma_n^2}} \end{aligned} \tag{8}$$

where we denote $\psi(\mathbf{x},\mathbf{y}) = e^{\mathbf{y}-\mathbf{x}}$.

3.2 MMSE Estimation of Clean Speech

Given a noisy speech observation $\mathbf{y}_0$, it is well known that the MMSE estimation $\hat{\mathbf{x}}$ of clean speech is calculated as $\hat{\mathbf{x}} = \mathrm{E}_{\mathbf{x}}[\mathbf{x} \mid \mathbf{y}_0]$. From eq.(4), we can see given $\mathbf{y}_0$ the valid range for $\mathbf{x}$ is $(-\infty, \mathbf{y}_0]$. Therefore, we have

$$\begin{aligned}
\hat{\mathbf{x}} &= \mathrm{E}_{\mathbf{x}}[\mathbf{x} \mid \mathbf{y}_0] = \int_{-\infty}^{\mathbf{y}_0} \mathbf{x} \cdot p(\mathbf{x} \mid \mathbf{y}_0)\, \mathrm{d}\mathbf{x} = \int_{-\infty}^{\mathbf{y}_0} \frac{\mathbf{x} \cdot p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x})}{p(\mathbf{y}_0)}\, \mathrm{d}\mathbf{x} \\
&= \frac{\int_{-\infty}^{\mathbf{y}_0} \mathbf{x} \cdot p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x})\, \mathrm{d}\mathbf{x}}{\int_{-\infty}^{\mathbf{y}_0} p(\mathbf{x}) \cdot p(\mathbf{y}_0|\mathbf{x})\, \mathrm{d}\mathbf{x}} \\
&= \frac{\sum_{k=1}^{K} w_k \cdot \int_{-\infty}^{\mathbf{y}_0} \mathbf{x} \cdot \mathcal{N}(\mathbf{x}|\mu_{xk}, \sigma_{xk}^2) \cdot p(\mathbf{y}_0|\mathbf{x})\, \mathrm{d}\mathbf{x}}{\sum_{k=1}^{K} w_k \cdot \int_{-\infty}^{\mathbf{y}_0} \mathcal{N}(\mathbf{x}|\mu_{xk}, \sigma_{xk}^2) \cdot p(\mathbf{y}_0|\mathbf{x})\, \mathrm{d}\mathbf{x}}
\end{aligned} \tag{9}$$

Then we replace $p(\mathbf{y}_0|\mathbf{x})$ with the right hand of eq.(8), we finally can derive $\hat{\mathbf{x}}$ as:

$$\hat{\mathbf{x}} = \frac{\sum_{k=1}^{K} w_k \int_{-\infty}^{\mathbf{y}_0} \mathbf{x} \cdot \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0)\, \mathrm{d}\mathbf{x}}{\sum_{k=1}^{K} w_k \int_{-\infty}^{\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0)\, \mathrm{d}\mathbf{x}} \tag{10}$$

with

$$\mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = \frac{1}{2\pi\sigma_{xk}\sigma_n} \cdot \frac{\psi(\mathbf{x}, \mathbf{y}_0)}{\psi(\mathbf{x}, \mathbf{y}_0) - 1} \cdot e^{-\frac{(\mathbf{x}-\mu_{xk})^2}{2\sigma_{xk}^2}} \cdot e^{-\frac{[\mathbf{x}-\mu_n+\ln(\psi(\mathbf{x},\mathbf{y}_0)-1)]^2}{2\sigma_n^2}} \tag{11}$$

3.3 A Numerical Solution

Obviously, we need solve the integral calculation in eq.(10) with some numerical methods. Since we have $\lim_{\mathbf{x}\to-\infty} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = \lim_{\mathbf{x}\to-\infty} \mathbf{x} \cdot \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = 0$ and $\lim_{\mathbf{x}\to\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = 0$ (see Appendix A for derivations), we can define the lower bound l and the upper bound u for the numerical integral as follows:

$$l_k = -\varepsilon_1 \cdot \sigma_{xk} \quad (\varepsilon_1 > 3) \tag{12}$$

$$u_k = \min(\mathbf{y}_0,\ \varepsilon_2 \cdot \sigma_{xk}) \quad (\varepsilon_2 > 3) \tag{13}$$

Then we uniformly (better not) partition the interval $[l_k, u_k]$ into J equal-length segments as:

$$l_k = \mathbf{x}_{k0} < \mathbf{x}_{k1} < \mathbf{x}_{k2} < \cdots < \mathbf{x}_{kJ-1} < \mathbf{x}_{kJ} = u_k \tag{14}$$

where we have $\mathbf{x}_{kj+1} = \mathbf{x}_{kj} + \Delta_k$ with $\Delta_k = \frac{u_k - l_k}{J}$. We use a linear approximation in each of these segments $[\mathbf{x}_{kj}, \mathbf{x}_{kj+1}]$, so the equation (10) can be approximated as:

$$\hat{\mathbf{x}} = \frac{\sum_{k=1}^{K} w_k \Delta_k \left[\mathbf{x}_{k0}\mathcal{U}_k(\mathbf{x}_{k0}|\mathbf{y}_0) + \mathbf{x}_{kJ}\mathcal{U}_k(\mathbf{x}_{kJ}|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathbf{x}_{kj}\mathcal{U}_k(\mathbf{x}_{kj}|\mathbf{y}_0)\right]}{\sum_{k=1}^{K} w_k \Delta_k \left[\mathcal{U}_k(\mathbf{x}_{k0}|\mathbf{y}_0) + \mathcal{U}_k(\mathbf{x}_{kJ}|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathcal{U}_k(\mathbf{x}_{kj}|\mathbf{y}_0)\right]} \tag{15}$$

4 Non-linear Noise Compensation for Robust Speech Recognition

It is well known that mismatches caused by additive noise corruption can seriously degrade performance of speech recognition. In this study, we assume that we have a set of HMM models trained from clean speech data. These HMM models will be used to recognize some noisy speech utterances. We know, most speech recognition systems use speech feature in the cepstral domain, e.g., MFCC's. But the above non-linear noise compensation method must be performed in the log-cepstral domain. First of all, we train a GMM model for clean speech in the log-cepstral domain, i.e. $p(\mathbf{x})$, based on clean speech data in training set. Then model parameters for $p(\mathbf{x})$ will be fixed during noise compensation procedure in this study.

For each test noisy speech utterance, we compute the feature vectors in the log-spectral domain as $\vec{\mathbf{Y}} = \{\vec{\mathbf{y}}_1, \vec{\mathbf{y}}_2, \cdots, \vec{\mathbf{y}}_T\}$, then we do

1. Initialize the mean μ_n and variance σ_n of the noise distribution $p(\mathbf{x})$ using the first N frames of the utterance. We typically use $N = 10$.[1]

2. Given clean speech model $p(\mathbf{x})$, refine the noise mean μ_n according to the EM algorithm based on the whole utterance $\vec{\mathbf{Y}}$. It is also possible to refine the noise variance in this step. (see Appendix B for details.)

3. Based on the refined noise model $p(\mathbf{n})$ and clean model $p(\mathbf{x})$, we compensate $\vec{\mathbf{Y}}$ a frame by a frame. More specifically, for each vector dimension $\mathbf{y}_{td}$ in each frame $\{\vec{\mathbf{y}}_t \mid 1 \leq t \leq T\}$, we use equation (15) to obtain its MMSE estimation.

4. The compensated vectors are mapped from the log-spectral domain into the MFCC domain by using the DCT transformation. Then the resultant feature vectors can be sent to the recognizer for recognition results.

5 Computational Complexity Issue

In the above approach, given any noisy speech utterance, we have to repeat the numerical integral in eq.(15) for each dimension in every frame. For each of these integrals, we have to calculate the value of function $\mathcal{U}_k(\mathbf{x}|\mathbf{y})$ for J different sampling points of $\mathbf{x}$ and for every Gaussian component in clean speech model $p(\mathbf{x})$. Furthermore, the calculation of function $\mathcal{U}_k(\mathbf{x}|\mathbf{y})$ involves to call functions $\exp(\cdot)$ and $\ln(\cdot)$ several times. Obviously, the overall computational complexity is very high. In this section, we will consider several possible ways to reduce computational complexity of the algorithm.

[1] We assume the first 10 frames, i.e. 100 msec in usual frame rate, of each utterance are non-speech segment, which is reasonable in most situations.

First of all, from eq.(15), for each $\mathbf{y}_0$, we must repeat the numerical integral for every Gaussian components in clean speech model $p(\mathbf{x})$. If we use a GMM model with 128 or more Gaussian mixands, the computation in this step is very expensive. One simple solution is that for any given $\mathbf{y}_0$ we only calculate the integral for the most significant Gaussian component corresponding to $\mathbf{y}_0$ or the N-best Gaussian components ($N < 5$). In this way, the total computational complexity can be largely reduced. In order to select the most significant Gaussian component for any $\mathbf{y}_0$, we can first use the conventional linear noise compensation method (e.g., the one in [2]) to get a rough estimation $\bar{\mathbf{x}}_0$ for $\mathbf{y}_0$, then the most signification mixture component for $\mathbf{y}_0$ is selected as:

$$k^* = \arg\max_{k=1}^{K} w_k \cdot \mathcal{N}(\bar{\mathbf{x}}_0 \mid \mu_{xk}, \sigma_{xk}^2) \tag{16}$$

Similarly, the N-best mixture components for for $\mathbf{y}_0$ can be selected as:

$$\Sigma^{(N)} = \arg\max_{k=1}^{(N)} w_k \cdot \mathcal{N}(\bar{\mathbf{x}}_0 \mid \mu_{xk}, \sigma_{xk}^2) \tag{17}$$

where $\arg\max^{(N)}$ denotes the operation to select the N-best ones, and the N-best set is denoted as $\Sigma^{(N)}$, and usually we have $N << K$.

Accordingly, the estimation in eq.(15) can be simplified as:

$$\hat{\mathbf{x}} = \frac{\mathbf{x}_0 \mathcal{U}_{k^*}(\mathbf{x}_0|\mathbf{y}_0) + \mathbf{x}_J \mathcal{U}_{k^*}(\mathbf{x}_J|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathbf{x}_j \mathcal{U}_{k^*}(\mathbf{x}_j|\mathbf{y}_0)}{\mathcal{U}_{k^*}(\mathbf{x}_0|\mathbf{y}_0) + \mathcal{U}_{k^*}(\mathbf{x}_J|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathcal{U}_{k^*}(\mathbf{x}_j|\mathbf{y}_0)} \tag{18}$$

or

$$\hat{\mathbf{x}} = \frac{\sum_{k\in\Sigma^{(N)}} w_k \Delta_k \left[\mathbf{x}_{k0} \mathcal{U}_k(\mathbf{x}_{k0}|\mathbf{y}_0) + \mathbf{x}_{kJ} \mathcal{U}_k(\mathbf{x}_{kJ}|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathbf{x}_{kj} \mathcal{U}_k(\mathbf{x}_{kj}|\mathbf{y}_0)\right]}{\sum_{k\in\Sigma^{(N)}} w_k \Delta_k \left[\mathcal{U}_k(\mathbf{x}_{k0}|\mathbf{y}_0) + \mathcal{U}_k(\mathbf{x}_{kJ}|\mathbf{y}_0) + 2\sum_{j=2}^{J-1} \mathcal{U}_k(\mathbf{x}_{kj}|\mathbf{y}_0)\right]} \tag{19}$$

Secondly, since the functions $\exp(\cdot)$ and $\ln(\cdot)$ will be called repeatedly when calculating $\mathcal{U}_k(\mathbf{x}|\mathbf{y}_0)$, they must be tabulated in memory for quick look-up.

Thirdly, if we can arrange to cache $\mathcal{U}_k(\mathbf{x}|\mathbf{y}_0)$ during the calculation, then the nonlinear noise compensation can be further accelerated greatly.

6 Discussion

The proposed method can be evaluated in robust speech recognition tasks. In the first set of experiments, we are going to test with some artificial data, where clean speech data is artificially corrupted by computer-generated white Gaussian noises in the time domain. In the second set of experiments, the method is used to recognition some hands-free speech data, which is recorded from some distant microphones in a running car environment.

Appendix

A Proof of $\lim_{\mathbf{x}\to\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = 0$

First of all, we have

$$\lim_{\mathbf{x}\to\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = \lim_{\mathbf{x}\to\mathbf{y}_0} \frac{1}{2\pi\sigma_{xk}\sigma_n} \cdot \frac{\psi(\mathbf{x},\mathbf{y}_0)}{\psi(\mathbf{x},\mathbf{y}_0)-1} \cdot e^{-\frac{(\mathbf{x}-\mu_{xk})^2}{2\sigma_{xk}^2}} \cdot e^{-\frac{[\mathbf{x}-\mu_n+\ln(\psi(\mathbf{x},\mathbf{y}_0)-1)]^2}{2\sigma_n^2}} \tag{20}$$

We define $\psi = e^{\mathbf{y}_0-\mathbf{x}} - 1$, then we have

$$\lim_{\mathbf{x}\to\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = C \cdot \lim_{\psi\to 0} \frac{1}{\psi} \cdot e^{-\frac{(K+\ln\psi)^2}{2\sigma_n^2}} \tag{21}$$

where C and K are constants. Moreover, we define $z = -\ln\psi$, we have

$$\lim_{\mathbf{x}\to\mathbf{y}_0} \mathcal{U}_k(\mathbf{x}|\mathbf{y}_0) = C \cdot \lim_{z\to\infty} e^z \cdot e^{-\frac{(z-K)^2}{2\sigma_n^2}} = 0 \tag{22}$$

B The EM algorithm to refine noise model $p(\mathbf{n})$

Assume we know clean speech model $p(\mathbf{x})$ as shown in eq.(6), we assume the noise model is a single Gaussian distribution as shown in eq.(7). The noise model parameters μ_n and σ_n are unknown and initialized as $\mu_n^{(0)}$ and $\sigma_n^{(0)}$ in the first step. Now we are interested in refining the noise model parameters from a noisy speech utterance $\vec{\mathbf{Y}}$ based on the EM algorithm. Since we assume independence among all feature vector dimensions, we can perform the EM-based re-estimation for every vector dimension separately as follows.

In order to obtain a close-form solution, in this step, we use a linear approximation for the environmental model in eq.(3) based on the first order Taylor series expansion. If we expand around the point $(\mathbf{x}_0, \mathbf{n}_0)$, we have

$$\begin{aligned} \mathbf{y} &= f(\mathbf{x},\mathbf{n}) = \mathbf{x} + \ln\left(1+e^{\mathbf{n}-\mathbf{x}}\right) \\ &\approx f(\mathbf{x}_0,\mathbf{n}_0) + f'_{\mathbf{x}}(\mathbf{x}_0,\mathbf{n}_0)\cdot(\mathbf{x}-\mathbf{x}_0) + f'_{\mathbf{n}}(\mathbf{x}_0,\mathbf{n}_0)\cdot(\mathbf{n}-\mathbf{n}_0) \\ &= A_0\cdot\mathbf{x} + B_0\cdot\mathbf{n} + C_0 \end{aligned} \tag{23}$$

where

$$A_0 = f'_{\mathbf{x}}(\mathbf{x}_0,\mathbf{n}_0) = \frac{1}{1+e^{\mathbf{n}_0-\mathbf{x}_0}} \tag{24}$$

$$B_0 = f'_{\mathbf{n}}(\mathbf{x}_0,\mathbf{n}_0) = \frac{e^{\mathbf{n}_0-\mathbf{x}_0}}{1+e^{\mathbf{n}_0-\mathbf{x}_0}} \tag{25}$$

$$C_0 = f(\mathbf{x}_0,\mathbf{n}_0) - f'_{\mathbf{x}}(\mathbf{x}_0,\mathbf{n}_0)\cdot\mathbf{x}_0 - f'_{\mathbf{n}}(\mathbf{x}_0,\mathbf{n}_0)\cdot\mathbf{n}_0 \tag{26}$$

If we use the above linear distortion function, the p.d.f. of noisy speech $\mathbf{y}$ is also a GMM model as:

$$p(\mathbf{y}) = \sum_{k=1}^{K} \mathcal{N}(\mathbf{y} \mid \mu_{yk}, \sigma_{yk}^2) \tag{27}$$

with

$$\mu_{yk} = A_{0k} \cdot \mu_{xk} + B_{0k} \cdot \mu_n + C_{0k} \tag{28}$$

$$\sigma_{yk}^2 = A_{0k}^2 \cdot \sigma_{xk}^2 + B_{0k}^2 \cdot \sigma_n^2 \tag{29}$$

where $\{A_{0k}, B_{0k}, C_{0k}\}$ are coefficients when expanding first order Taylor series around the means of Gaussian components (μ_{xk}, μ_n).

Given the utterance $\{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T\}$, the iterative maximum likelihood (ML) estimation for μ_n and σ_n can be derived based on the EM algorithm as:

$$\mu_n^{(i+1)} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{x}_t)(\mathbf{y}_t - A_{0k}\mu_{xk} - C_{0k})/B_{0k}}{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{x}_t)} \tag{30}$$

$$\sigma_n^{(i+1)} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{x}_t)(\mathbf{y}_t - A_{0k}\mu_{xk} - B_{0k}\mu_n^{(i+1)} - C_{0k})^2}{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{x}_t)}} \tag{31}$$

where

$$p(k|\mathbf{x}_t) = \frac{\mathcal{N}(\mathbf{x}_t \mid \mu_{yk}, \sigma_{yk}^2)}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{x}_t \mid \mu_{yk}, \sigma_{yk}^2)} \tag{32}$$

In which $\{\mu_{yk}, \sigma_{yk}^2\}$ are derived from eqs. (28) and (29) based on the current noise model parameters $\{\mu_n^{(i)}, \sigma_n^{(i)}\}$.

The above iterative estimation in eqs.(30) and (31) continues until some convergency conditions are met.

References

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

[2] M. Afify and O. Siohan, "Sequential noise Estimation with Optimal Forgetting for Robust Speech Recognition," *Proc. of ICASSP '2001*, Salt Lake City, May 2001.

[3] M. Afify and O. Siohan, "Sequential estimation with optimal forgetting for robust speech recognition", *submitted to IEEE Trans. on Speech and Audio Processing*, 2002.

[4] A.D. Berstein and I.D. Shallom, "A Hypothesized Wiener filtering approach to noisy speech recognition," *Proc. of ICASSP*, pp.913-916, 1991.

[5] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on Acoustic, Speech and Signal Processing*, Vol. ASSP-27, pp.113-120, Apr. 1979.

[6] L. Deng, A. Acero, M. Plumpe and X.-D. Huang, "Large Vocabulary Speech Recognition under Adverse Acoustic Environments," *Proc. of ICSLP-2000*, Beijing, China, October 2000.

[7] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the AURORA2 Database," *Proc. of Eurospeech 2001*, Aalborg, Denmark, September 2001.

[8] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, 1996.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on Speech and Audio Processing*, Vol. 2, pp.291-298, Apr. 1994.

[10] N.S. Kim, "Statistical Linear Approximation for Environment Compensation," *IEEE Signal Processing Letters*, Vol.5, no. 1, pp8-10, January 1998.

[11] N.S. Kim, "nonstationary Environment Compensation based on Sequential Estimation," *IEEE Signal Processing Letters*, Vol. 5, no. 3, pp.57-59, March 1998.

[12] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer and Speech Language*, Vol. 9, pp.171-185, 1995.

[13] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of IEEE*, Vol. 67, pp.1586-1604, 1979.

[14] P.J. Moreno, B. Raj and R.M. Stern, "A Vector Taylor Series Approach for Environment-Independent Speech Recognition," *Proc. of ICASSP'96*, pp.733-736, Atlanta, GA, May 1996.

[15] P.Z. Peebles, *Probability, Random Variables, and Random Signal Principles*, 3rd edition, McGraw-Hill Inc..

[16] S. Sagayama, Y. Yamaguchi, S. Takahashi and J. Takahashi, " Jacobian approach to fast acoustic model adaptation," *Proc. of ICASSP*, pp.835-838, 1997.

[17] A. Sankar and C.-H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 3, pp.190-202, 1996.

[18] O. Siohan and C.-H. Lee, "Iterative noise and channel estimation under the stochastic matching algorithm framework," *IEEE Signal Processing Letters* , Vol. 4, No. 11, pp.304-306, Nov 1997.

[19] A. Surendran, M. Rahim and C.-H. Lee. "Non-linear Compensation for Stochastic Matching", *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 7, pp. 643-655, 1999.

[20] A.P. Varga and R.K. Moore, "Hidden Markov Model Decomposition of speech and noise," *Proceedings of ICASSP*, pp.845-848, 1990.