NONLINEAR NOISE COMPENSATION IN FEATURE DOMAIN FOR SPEECH RECOGNITION WITH NUMERICAL METHODS

Hui Jiang *Qi Wang*

Department of Computer Science, York University
4700 Keele Street, Toronto, Ontario, M3J 1P3, CANADA
Email: {hj,qwang}@cs.yorku.ca

ABSTRACT

In this paper, we propose to compensate noise in the log-spectral domain for robust speech recognition based on a nonlinear environmental model. In our approach, starting from the original nonlinear speech distortion model in the feature domain, we derive the minimum mean square error (MMSE) estimation of clean speech signal given a noisy observation, which turns out to be an integral of a complex nonlinear function. In this work, we propose to use a numerical method to solve the above nonlinear integral. It requires higher computational complexity than the normal linear approximation methods but it is usually affordable since calculation is performed entirely in the pre-processing feature domain without involving any change in speech decoders. Experimental results show that the proposed nonlinear method outperforms the conventional Vector Taylor Series (VTS) method in terms of ASR performance when dealing with artificial white Gaussian noises as well as true hands-free noisy speech, especially in low SNR levels.

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, Wiener filtering, transformation based on stereo data, linear noise compensation based on Taylor series approximation, feature domain stochastic matching, 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), model adaptation using MLLR (maximum likelihood linear regression) or MAP (maximum *a posteriori*), Jacobian environment adaptation, speech and noise decomposition, model space stochastic matching. 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 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[8]. In a recent work [5], a nonlinear method called "optimal filtering" is proposed to compensate noisy speech.

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 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. The proposal method has been examined in many robust speech recognition experiments. The results show that the proposed nonlinear method outperforms the conventional Vector Taylor Series (VTS) method in terms of ASR performance when dealing with artificial white Gaussian noises as well as true hands-free noisy speech, especially in low SNR levels.

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 $\mathbf{n}$:

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

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

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

3. MMSE ESTIMATION OF CLEAN SPEECH

Based on the above non-linear environmental model, for any given noisy speech feature vector $\mathbf{y}$, 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 vector $\mathbf{x} = \{x_1, x_2, \cdots, x_D\}$ in the log-spectral domain follows a multivariate Gaussian mixture model (GMM) as:

$$\begin{aligned} p(\mathbf{x}) &= \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{x} \mid \mu_{xk}, \sigma_{xk}^2) \\ &= \sum_{k=1}^{K} w_k \cdot \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{xkd}^2}} \cdot e^{-\frac{(x_d - \mu_{xkd})^2}{2\sigma_{xkd}^2}} \end{aligned} \tag{6}$$

where $\mu_{xk} = \{\mu_{xk1}, \mu_{xk2}, \cdots, \mu_{xkD}\}$ and $\sigma_{xk}^2 = \{\sigma_{xk1}, \sigma_{xk2}, \cdots, \sigma_{xkD}\}$ are mean and variance vectors of k-th Gaussian mixture, 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:

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n} \mid \mu_n, \sigma_n^2) = \prod_{d=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{nd}^2}} \cdot e^{-\frac{(n_d - \mu_{nd})^2}{2\sigma_{nd}^2}} \tag{7}$$

where $\mu_n = \{\mu_{n1}, \mu_{n2}, \cdots, \mu_{nD}\}$ and $\sigma_n^2 = \{\sigma_{n1}, \sigma_{n2}, \cdots, \sigma_{nD}\}$ are mean and variance vectors 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, if we assume independence among all vector dimensions, $p(\mathbf{y}|\mathbf{x})$ can be derived as:

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

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

3.2. MMSE Estimation of Clean Speech

Given a noisy speech vector $\mathbf{y}_0 = \{y_{01}, y_{02}, \cdots, y_{0D}\}$, it is well known that the MMSE estimation $\hat{\mathbf{x}} = \{\hat{x}_0, \hat{x}_1, \cdots, \hat{x}_D\}$ of clean speech is calculated as $\hat{\mathbf{x}} = \mathrm{E}_{\mathbf{x}}[\mathbf{x} \mid \mathbf{y}_0]$. Therefore, we have

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

From eq.(4), we can see if given y_{0d} the valid range for x_d is $(-\infty, y_{0d}]$. If we still assume the independence among all vector dimensions, we can calculate each dimension $\hat{x}_e$ $(e = 1, 2, \cdots, D)$ of $\hat{\mathbf{x}}$ independently. We replace $p(\mathbf{y}_0|\mathbf{x})$ with the right hand of eq.(8), we finally derive $\hat{x}_e$ as:

$$\hat{x}_e = \frac{\sum_{k=1}^{K} w_k \cdot A_{ke} \cdot \prod_{d=1, d\neq e}^{D} B_{kd}}{\sum_{k=1}^{K} w_k \cdot \prod_{d=1}^{D} B_{kd}} \tag{10}$$

with

$$A_{ke} = \int_{-\infty}^{y_{0e}} x_e \cdot \mathcal{U}_k(x_e|y_{0e})\, \mathrm{d}x_e \tag{11}$$

$$B_{kd} = \int_{-\infty}^{y_{0d}} \mathcal{U}_k(x_d|y_{0d})\, \mathrm{d}x_d \tag{12}$$

$$\begin{aligned} \mathcal{U}_k(x_d|y_{0d}) &= \frac{1}{2\pi\sigma_{xkd}\sigma_{nd}} \frac{\psi(x_d, y_{0d})}{\psi(x_d, y_{0d}) - 1} e^{-\frac{(x_d - \mu_{xkd})^2}{2\sigma_{xkd}^2}} \\ &\quad e^{-\frac{\left[x_d - \mu_{nd} + \ln(\psi(x_d, y_{d0}) - 1)\right]^2}{2\sigma_{nd}^2}} \end{aligned} \tag{13}$$

3.3. A Numerical Solution

Obviously, we need solve the integral calculation in eq.(10) with some numerical methods. Since we have $\lim_{x\to-\infty} \mathcal{U}_k(x|y_0) = \lim_{x\to-\infty} x \cdot \mathcal{U}_k(x|y_0) = 0$ and $\lim_{x\to y_0} \mathcal{U}_k(x|y_0) = 0$ (see [3] for derivation), we can define the lower bound l and the upper bound u for the numerical integral as follows:

$$u_{kd} = y_{0d} \tag{14}$$

$$l_{kd} = \min(y_{0d}, u_{kd}) - \varepsilon \cdot \sigma_{xkd} \quad (\varepsilon > 3) \tag{15}$$

Then we uniformly partition the interval $[l_{kd}, u_{kd}]$ into J equal-length segments as:

$$l_{kd} = x_{kd0} < x_{kd1} < x_{kd2} < \cdots < x_{kdJ-1} < x_{kdJ} = u_{kd} \tag{16}$$

where we have $x_{kdj+1} = x_{kdj} + \Delta_{kd}$ with $\Delta_{kd} = \frac{u_{kd}-l_{kd}}{J}$. We use a linear approximation in each of these segments $[x_{kdj}, x_{kdj+1}]$, so the equation (10) can be approximated as:

$$\hat{x}_e = \frac{\sum_{k=1}^{K} w_k \cdot \mathcal{M}_{ke} \cdot \prod_{d=1, d\neq e}^{D} \mathcal{N}_{kd}}{\sum_{k=1}^{K} w_k \cdot \prod_{d=1}^{D} \mathcal{N}_{kd}} \tag{17}$$

where

$$\begin{aligned} \mathcal{M}_{ke} &= \Delta_{ke}\Big[x_{ke0}\mathcal{U}_k(x_{ke0}|y_{0e}) + x_{keJ}\mathcal{U}_k(x_{keJ}|y_{0e}) \\ &+ 2\sum_{j=2}^{J-1} x_{kej}\mathcal{U}_k(x_{kej}|y_{0e})\Big] \end{aligned} \tag{18}$$

$$\mathcal{N}_{kd} = \Delta_{kd}\Big[\mathcal{U}_k(x_{kd0}|y_{0d}) + \mathcal{U}_k(x_{kdJ}|y_{0d}) + 2\sum_{j=2}^{J-1} \mathcal{U}_k(x_{kdj}|y_{0d})\Big] \tag{19}$$

As pointed out in [5], when $(y_d - \mu_{nd})/\sigma_{nd}$ grows, the function $\mathcal{U}_k$ in eq.(13) may converge to a Dirac δ-function centered somewhere near y_d. In this paper, we adopt a simple solution to deal with this problem. If $(y_d - \mu_{nd})/\sigma_{nd} > \tau$ (τ is a preset threshold; In our experiments, we fix $\tau = 10.0$), the above function will be approximated with a δ-function centered at y_d, i.e., $\mathcal{N}(y_d|\mu_{xk}, \sigma^2_{xk}) \cdot \delta(x_d - y_d)$. In this case, the above integral can be calculated in closed-form without using the above numerical method.

4. NONLINEAR 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 stage.

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 $\vec{\mathbf{Y}}$, as in [2]. For simplicity, we don't refine the noise variance σ_n^2. We simply re-scale the noise variances for all dimensions with a constant ρ. From experiments, we find an acceptable range for ρ is [2.0,4.0]. In the following experiments, we fix $\rho = 3.0$ unless stated explicitly.
3. Based on the refined noise model $p(\mathbf{n})$ and clean model $p(\mathbf{x})$, we compensate $\vec{\mathbf{Y}}$ frame by 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 eq. (17) 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. EXPERIMENTS

5.1. Database and Experimental Setup

Our noise compensation algorithm is evaluated on an hands-free database (CARVUI database [2]) recorded inside a moving car. The data was collected in Murray Hill, NJ area, under various driving conditions (highway/city roads) and noise environments (with or without radio/music in the background). About 2/3rd of the recordings contain music and babble noise in the background. Simultaneous recordings were made using a close-talking microphone and a 16-channel microphone array of first order hypercardioid microphones mounted on the visor. A total of 56 speakers participated in the data collection, including many non-native speakers of American English. The recorded text is made of various materials, including phonetically balanced TIMIT sentences, some digits strings with 1 to 7 digitsm and about 85 short command words, like "window up", "turn radio off", etc. The speech material from 50 speakers is used for training, and the 6 remaining speakers is used for test, leading to a total of 4417 utterances available for training and 993 utterances for test. The data is recorded at 24kHz sampling rate and is down-sampled to 8kHz and followed by a MFCC feature extraction step for our speech recognition experiments. The recognition task consists of command words and digits strings of unspecified length, modeled by a finite state grammar. In our experiments, data from 2 channels only are used. The first one is the close-talking microphone (CT), the second one is a single channel from the microphone array, referred to as hands-free data (HF) hereafter. The average SNR is about 21 dB for the CT channel and 8dB for the HF channel. In our experiments, we used 39-dimension feature vector, consisting of 12 MFCC's and C_0 energy, their delta and delta-delta. A set of tri-phone models is built on the CT training data using a decision tree tying algorithm with aggressive tying given the limited amount of training data. The standard Viterbi decoder is tuned on the CT test data, leading to a string error rate (SER) of 3.7%. When recognizing the HF test data, the performance degrades significantly, down to 26.6% in SER. For the noise compensation purpose, a 128-Gaussian mixture model is trained in the MFCC domain on the same CT training data and then mapped to the log-spectral domain.

Based on the above experimental setup, in this study, we will examine the proposed nonlinear noise compensation method in

[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.

[2] We acknowledge Bell Labs, Lucent Technology to allow us to evaluate our algorithm on the CARVUI database.

SNR	baseline	VTS	new
∞ dB	3.7	4.0	4.1
15 dB	30.8	20.2	19.6
10 dB	54.2	32.7	31.2
5 dB	77.3	57.3	50.9
0 dB	87.6	74.2	69.4

Table 1: The ASR performance comparison in string error rate (in %) by using different noise compensation methods. The *baseline* means the corrupted CT data is directly sent to the decoder without any process, the *VTS* means the data is pre-processed by the VTS method, the *new* means the data is pre-processed by our nonlinear noise compensation method.

terms of ASR performance improvements and compare with other conventional methods with a linear approximation, such as the VTS method in [4, 2]. In our new method, we set the threshold $\tau = 10.0$, the noise variance scale $\rho = 3.0$, the lower bound parameter in eq.(15) $\varepsilon = 3.0$. And we use the interval segment number $J = 100$.

5.2. Experiments on artificial white Gaussian noises

In the first set of experiments, we evaluate our method on artificial white Gaussian noises. The computer-generated white Gassuan noises are added into the CT test data in the time domain at various SNR levels. The corrupted CT test data are processed in the log-spectral domain with our nonlinear noise compensation method or the VTS method. Then the compensated speech is converted into MFCC's and sent to the decoder for recognition which uses the triphone HMM's trained in the original CT training data set. The recognition results are shown in Table 1. From the results, it is shown that the baseline performance drops quickly as the noise level increases, from 3.7% in SER in clean CT data down to 87.6% in SER when white noise is added at SNR 0dB level. The VTS method largely improves recognition performance over the baseline across all examined SNR levels in mismatched conditions. When the VTS method is used to process the clean CT data, it degrades the recognition performance slightly, down from 3.7% in SER to 4.0%. The results also show that our new nonlinear noise compensation method clearly outperforms the conventional linear approximation VTS method in all mismatched conditions. Especially when the noise level is high, the improvement over VTS is quite significant, e.g., in SNR 5dB level the new method achieves 6.4% absolute SER reduction over the VTS, from 57.3% down to 50.9%. When processing the clean CT data, the new method obtains a similar performance, 4.1% in SER, as the VTS.

5.3. Experiments on hands-free (HF) data

Our new approach is also evaluated in some true hands-free noisy speech data. In this set of experiments, the HF test data set is pre-processed by using VTS or our new nonlinear method before sending to the decoder for recognition based on the pre-trained triphone HMMs. The ASR performance comparison is shown in Table 2. From the results, we can see the baseline performance (without any noise compensation) drops to 26.6% in SER and the VTS achieves 19.9% while our new method yields 19.0% in SER in this case, which is a moderate improvement over the VTS.

	baseline	VTS	new
HF	26.6	19.9	19.0

Table 2: The ASR performance comparison in string error rate (in %) by using different noise compensation methods to pre-process true hands-free (HF) noisy speech data.

6. DISCUSSIONS

In some cases, the new approach achieves some encouraging performance gain over the VTS method but it requires much more computation than the VTS in the pre-processing stage. It is still possible to speedup the numerical integral calculation in the new method further, e.g., in many case if we can predict the total integral values in eqs.(18) and (19) are small, we may be able to use some simple approximation to avoid the numerical method. Moreover, when we derive the original distortion model in eq.(2), we simply ignore the phase difference between speech and noise signals. The experiments show that such a simplification significantly affects the accuracy of the MMSE estimation when we strictly follow the distortion function. At last, unlike in 1st order VTS, noise variance σ_n also plays a role in the new method. It becomes a critical issue how to estimate noise variance precisely. All these issues are still under investigation and we will be able to report more results in the conference.

7. 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] H. Jiang, " Nonlinear Noise Compensation in Feature Domain with Numerical Methods," *Technical Report CS-2003-02*, Department of Computer Science, York University, February 2003. (http://www.cs.yorku.ca/techreports/2003/CS-2003-02.html)

[4] 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.

[5] T.-A. Myrvoll and S. Nakamura, "Optimal Filtering of Noisy Cepstral Coefficients for Robust ASR," *submitted to IEEE 2003 Automatic Speech Recognition and Understanding (ASRU) Workshop*, December 2003.

[6] 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.

[7] 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.

[8] 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.