Vector Taylor Series based HMM Adaptation for Generalized Cepstrum in Noisy Environment

Soonho Baek [1], Hong-Goo Kang [2]

School of Electrical and Electronic Engineering, Yonsei University
Seoul, Korea
[1] bestboybsh@dsp.yonsei.ac.kr
[2] hgkang@yonsei.ac.kr

***Abstract*—This paper proposes a novel HMM adaptation algorithm for robust automatic speech recognition (ASR) system in noisy environments. The HMM adaptation using vector Taylor series (VTS) significantly improves the ASR performance in noisy environments. Recently, the power normalized cepstral coefficient (PNCC) that replaces a logarithmic mapping function with a power mapping function has been proposed and it is proved that the replacement of the mapping function is robust to additive noise. In this paper, we extend the VTS based approach to the cepstral coefficients obtained by using a power mapping function instead of a logarithmic mapping function. Experimental results indicate that HMM adaptation in the cepstrum obtained by using a power mapping function improves the ASR performance comparing the VTS based conventional approach for mel-frequency cepstral coefficients (MFCCs).**

I. Introduction

Although the performance of automatic speech recognition systems (ASRs) adopting a statistical model based approach has been dramatically improved in clean environment, the presence of acoustic interferences such as background noise still degrades the ASR performance due to the mismatches between the features extracted for the test stage and for the training stage. Therefore, many algorithms try to maintaining the inherent property of speech signals by reducing the acoustic mismatch [1]-[9].

Compensation techniques to minimize the mismatch effect caused by background noise are divided into three groups: signal domain algorithm, feature domain algorithm, and model domain algorithm. Firstly, signal domain algorithm, typically called speech enhancement such as Wiener filter and spectra subtraction, estimate the denoised signal prior to the feature extraction. The well-known speech enhancement algorithm for robust speech recognition is a two stage mel-warped Wiener filter noise reduction method used in the European telecommunications standards institute (ETSI) advanced front end (AFE) [8]. Secondly, feature domain algorithms such as power normalized cepstral coefficient (PNCC) and perceptual linear predictive (PLP) make the features robust to additive noise [3][7]. These type of algorithms aim at extracting robust features from noisy signals. Finally, model domain algorithms such as maximum likelihood linear regression (MLLR), parallel model combination (PMC), and vector Taylor series (VTS) approximation adjust the hidden Markov model (HMM) parameters so that the ASR system becomes better matched to the distorted environment [2][4].

Model adaptation algorithms show more improvement in ASR performance than other types of algorithms. In [2], an approximation based on Taylor series for HMM adaptation is proposed. The static and dynamic means and variances are adjusted by the parameters related to the background noise. The VTS approximation based approach significantly takes advantage of the extra knowledge provided by the model to reduce the data requirements and dramatically improves the ASR performance in noisy environments.

In addition, the PNCC recently proposed by Kim and Stern shows improved performance in noisy conditions compared to conventional MFCC and PLP features. PNCC replaces a natural logarithmic mapping function used for extracting MFCC with a power mapping function. It is indicated that the replacement of spectral mapping function is useful in adverse environments.

In this paper, we propose an VTS based model adaptation, which is an extension of [2] from the approach for MFCCs to the approach for power mapping function based cepstral coefficient. We use the generalized cepstral coefficient obtained by the generalized logarithmic function [10]. It covers much wider range of factors than the fractional power function used for computing PNCCs and PLP. Firstly, we reformulate the generalized cepstral coefficient of the observed speech in noisy environments. Then, the VTS approximation schemes for HMM adaptation is applied to the generalized cepstrum. Experimental results show that the HMM adaptation in the generalized cepstrum based ASR system improves the word accuracy in noisy environments by reducing deletion errors, which is very effective compared to the conventional HMM adaptation approach for MFCC.

The layout of this paper is as follows. Section II introduces the generalized cepstral coefficient. In Section IV, the VTS based conventional approach for MFCC is showed. Its extension to the generalized cepstrum based ASR system is described in Section V. Experimental results on Aurora 2 are given in Section VI. Finally, conclusions are followed in Section VII.

II. Generalized cepstral coefficients

The generalized cepstral coefficient, introduced in [10], is the generic form of cepstral coefficients. Commonly, mel-

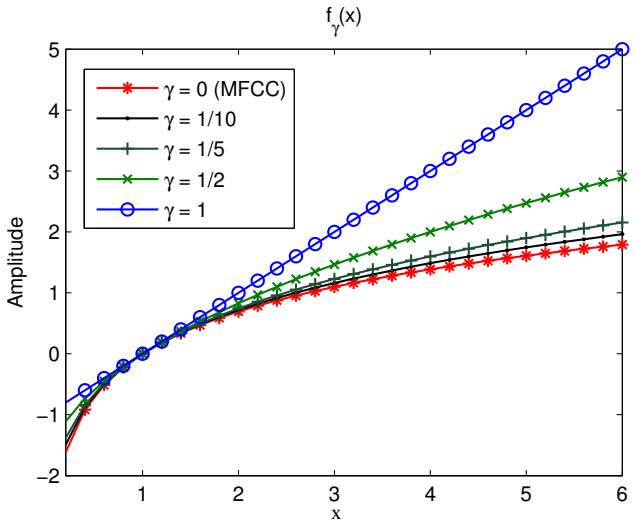

Fig. 1. Curves of generalized logarithmic function.

frequency cepstral coefficients (MFCCs) are used for ASR system. It is obtained by taking a discrete cosine transform to logarithmic spectrum on a nonlinear mel-scale frequency. On the other hand, the generalized cepstral coefficients are obtained from the generalized logarithmic function based spectrum instead of logarithmic spectrum.

The generalized logarithmic function is defined by

$$f_\gamma(x) = \frac{1}{\gamma}\left(x^\gamma - 1\right), \quad \gamma \neq 0, \tag{1}$$

where γ is a real value. Fig. 1 shows the curves of the generalized logarithmic function $f_\gamma(x)$ for several values of γ. It is observed that as the value of γ is close to 0, its corresponding curve is close to the logarithmic function, which is mathematically proved in [10]. Consequently, the mapping function $f_\gamma(x)$ is redefined as

$$f_\gamma(x) = \begin{cases} \frac{1}{\gamma}\left(x^\gamma - 1\right) & 0 < \gamma \leq 1 \\ \log x & \gamma = 0 \end{cases}. \tag{2}$$

Note that the derivative of $f_\gamma(x)$ in terms of x is depends on the value of γ. In case $\gamma = 1$, its curve is linear function and the derivative is constant regardless of x. On the other hands, when $\gamma < 1$, the slope of corresponding curve depends on the amplitude of x. The derivative in the small values of x is larger than in the high values of x. It is indicates that if the noise components is even small, the distortion in spectral valleys can be significant. For this reason, MFCC based ASR performance is tremendously degraded in noisy environments.

III. Signal Model

Let the clean speech be corrupted by the channel noise $h(n)$ and the additive noise $d(n)$:

$$y(n) = x(n) * h(n) + d(n). \tag{3}$$

where n is a time index. Assuming that the additive noise and speech signal are independent and the noise has zero mean, the i-th filterbank coefficient of the observed signal is represented by

$$Y_i = X_i H_i + D_i, \tag{4}$$

where X_i represents the filterbank coefficient of the clean speech $x(n)$, H_i the filterbank coefficient of the channel noise $h(n)$, and D_i the filterbank coefficient of the additive noise. In this work, we assume that the length of $h(n)$ is much shorter than the window length.

IV. VTS based conventional Approach for MFCC

In this section, the conventional HMM adaptation algorithm based on VTS approximation for robust speech recognition is described. The conventional algorithm is operated in the mel-frequency cepstral domain which is obtained from the log-spectrum.

A. Cepstral Coefficients in Noisy Environments

Let the length M cepstral coefficient vector of the observed signal be defined by

$$\mathbf{y} = \mathbf{C}[\ \log Y_0 \quad \log Y_1 \quad \cdots \quad \log Y_{N-1}\ ]^{\mathbf{T}} \tag{5}$$

where $\mathbf{C}$ is the $(M \times N)$ DCT matrix and N denotes the number of the filterbank. Combining Eq. (4) with (5) and after some manipulation, the cepstral coefficient vector of the observed signal is represented by

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{C}\log\left(\mathbf{1} + \exp\left(\mathbf{C^{-1}}(\mathbf{d} - \mathbf{x} - \mathbf{h})\right)\right) \tag{6}$$

In practice, the length of cepstral coefficient is smaller than the number of filterbanks, so that $\mathbf{C^{-1}}$ is the pseudo-inverse matrix of $\mathbf{C}$. From Eq. (6), the relationship between the log spectral features representing clean and noisy speech in noisy environments is formulated. It is indicated that when the cepstrum of the additive noise $\mathbf{d}$, the cepstrum of the convolutive noise $\mathbf{h}$, and the cepstrum of the clean speech $\mathbf{x}$ are given, the cepstrum of the corrupted speech can be computed.

B. Model Adaptation

Given Gaussian random variable vectors $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{d}$ with means $\mu_\mathbf{x}$, $\mu_\mathbf{h}$, and $\mu_\mathbf{d}$, and covariance matrices $\sigma_\mathbf{x}$, $\sigma_\mathbf{h}$, and $\sigma_\mathbf{d}$, the cepstrum of the observed speech is approximated by a first order Taylor series expansion at $(\mu_\mathbf{x}, \mu_\mathbf{h}, \mu_\mathbf{d})$ [2]:

$$\begin{aligned}\mathbf{y} \approx\ & \mu_\mathbf{x} + \mu_\mathbf{h} \\ & + \mathbf{C}\log\left(\mathbf{1} + \exp\left(\mathbf{C^{-1}}\left(\mu_\mathbf{d} - \mu_\mathbf{x} - \mu_\mathbf{h}\right)\right)\right) \\ & + \mathbf{F_x}\left(\mathbf{x} - \mu_\mathbf{x}\right) + \mathbf{F_h}\left(\mathbf{h} - \mu_\mathbf{h}\right) + \mathbf{F_d}\left(\mathbf{d} - \mu_\mathbf{d}\right),\end{aligned} \tag{7}$$

where $\mathbf{F_x}$, $\mathbf{F_h}$, and $\mathbf{F_d}$ denote the derivative of $\mathbf{y}$ in terms of $\mathbf{x}$, $\mathbf{h}$, and $\mathbf{d}$ at the point $(\mu_\mathbf{x}, \mu_\mathbf{h}, \mu_\mathbf{d})$, respectively. The $(i,j)^{th}$ entries of matrix $\mathbf{F_x}$, $\mathbf{F_h}$, and $\mathbf{F_d}$ are obtained by

$$\mathbf{F_x}(i,j) = \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_j} - \sum_m \mathbf{C}_{i,m}\left(\bar{\xi}_m + 1\right)^{-1}\mathbf{C}^{-1}_{m,j}, \tag{8}$$

$$\mathbf{F_h}(i,j) = \frac{\partial \mathbf{h}_i}{\partial \mathbf{h}_j} - \sum_m \mathbf{C}_{i,m}\left(\bar{\xi}_m + 1\right)^{-1}\mathbf{C}^{-1}_{m,j}, \tag{9}$$

and

$$\mathbf{F_d}(i,j) = \sum_m \mathbf{C}_{i,m}\left(\bar{\xi}_m + 1\right)^{-1}\mathbf{C}^{-1}_{m,j}, \tag{10}$$

where

$$\bar{\xi}_m = \exp\left\{\sum_i \mathbf{C}^{-1}_{m,i}\left(\mu_{\mathbf{x}_i} + \mu_{\mathbf{h_i}} - \mu_{\mathbf{d}_i}\right)\right\}. \tag{11}$$

Then, the mean and covariance matrix of $\mathbf{y}$ are obtained by

$$\begin{aligned}\mu_{\mathbf{y}} \approx & \mu_{\mathbf{x}} + \mu_{\mathbf{h}} \\ & + \mathbf{C}\log\left(\mathbf{1} + \exp\left(\mathbf{C^{-1}}\left(\mu_{\mathbf{d}} - \mu_{\mathbf{x}} - \mu_{\mathbf{h}}\right)\right)\right),\end{aligned} \tag{12a}$$

$$\sigma_{\mathbf{y}} \approx \mathbf{F_x}\sigma_{\mathbf{x}}\mathbf{F_x}^{\mathbf{T}} + \mathbf{F_h}\sigma_{\mathbf{h}}\mathbf{F_h}^{\mathbf{T}} + \mathbf{F_d}\sigma_{\mathbf{d}}\mathbf{F_d}^{\mathbf{T}}. \tag{12b}$$

The mean and covariance matrix of delta cepstrum are represented by

$$\mu_{\mathbf{\Delta y}} \approx \mathbf{F_x}\mu_{\mathbf{\Delta x}}, \tag{13a}$$

$$\sigma_{\mathbf{\Delta y}} \approx \mathbf{F_x}\sigma_{\mathbf{\Delta x}}\mathbf{F_x}^{\mathbf{T}} + \mathbf{F_d}\sigma_{\mathbf{\Delta d}}\mathbf{F_d}^{\mathbf{T}}. \tag{13b}$$

In addition, for the delta-delta cepstrum, the mean and covariance matrix are given by

$$\mu_{\mathbf{\Delta^2 y}} \approx \mathbf{F_x}\mu_{\mathbf{\Delta^2 x}}, \tag{14a}$$

$$\sigma_{\mathbf{\Delta^2 y}} \approx \mathbf{F_x}\sigma_{\mathbf{\Delta^2 x}}\mathbf{F_x}^{\mathbf{T}} + \mathbf{F_d}\sigma_{\mathbf{\Delta^2 d}}\mathbf{F_d}^{\mathbf{T}}. \tag{14b}$$

V. HMM ADAPTION FOR GENERALIZED CEPSTRAL COEFFICIENTS

At the previous section, the VTS based conventional model adaptation algorithm that operated in the cepstral domain (MFCCs) has been shown. Here, we extend the VTS based model adaptation algorithm to the generalized cepstrum based ASR system, which is introduced at Section II. First, the relationship between the generalized cepstral coefficients obtained from the the observed signal and the clean speech. In addition, the model update equation is reformulated for the acoustic model trained in the generalized cepstral domain.

A. Generalized Cepstral Coefficients in Noisy Environment

Using Eq. (2) and (4), the generalized logarithmic spectrum of the observed signal is obtained by

$$\begin{aligned}f_\gamma(Y_i) = & f_\gamma(X_i) \\ & + \left(\gamma f_\gamma(X_i) + 1\right) f_\gamma(H_i) \\ & + \left(\gamma f_\gamma(D_i) + 1\right) g_\gamma(\xi_i),\end{aligned} \tag{15}$$

where

$$\xi_i = X_i H_i / D_i, \tag{16}$$

and

$$g_\gamma(x) = f_\gamma(x+1) - f_\gamma(x). \tag{17}$$

It is indicated that the generalized logarithmic spectrum of the observed signal is represented by those obtained from the clean speech and two additive distortion terms. The first distortion term related to the convolutive noise in the frequency domain depends on $f_\gamma(X_i)$ and $f_\gamma(H_i)$. The second related to the additive noise is determined by the spectrum of the additive noise component and SNR depending term.

We now proceed with the representation of the generalized cepstrum of the observed signal. After taking DCT and some manipulation, the generalized cepstral coefficient of the observed signal can be obtained by

$$\mathbf{y}^\gamma = \mathbf{x}^\gamma + \mathbf{M}(\hat{\mathbf{x}}^\gamma)\mathbf{h}^\gamma + \mathbf{M}(\hat{\mathbf{d}}^\gamma)\mathbf{g}^\gamma(\xi), \tag{18}$$

where

$$\mathbf{x}^\gamma = \mathbf{C}\left[\begin{array}{cccc} f_\gamma(X_0) & f_\gamma(X_1) & \cdots & f_\gamma(X_{N-1}) \end{array}\right]^T, \tag{19}$$

$$\mathbf{h}^\gamma = \mathbf{C}\left[\begin{array}{cccc} f_\gamma(H_0) & f_\gamma(H_1) & \cdots & f_\gamma(H_{N-1}) \end{array}\right]^T, \tag{20}$$

$$\mathbf{d}^\gamma = \mathbf{C}\left[\begin{array}{cccc} f_\gamma(D_0) & f_\gamma(D_1) & \cdots & f_\gamma(D_{N-1}) \end{array}\right]^T, \tag{21}$$

$$\mathbf{g}^\gamma(\xi) = \mathbf{C}\left[\begin{array}{cccc} g_\gamma(\xi_0) & g_\gamma(\xi_1) & \cdots & g_\gamma(\xi_{N-1}) \end{array}\right]^T, \tag{22}$$

$$\hat{\mathbf{x}}^\gamma = \left[\begin{array}{cccc} \gamma\mathbf{x}_0^\gamma + 1 & \gamma\mathbf{x}_1^\gamma & \cdots & \gamma\mathbf{x}_{N-1}^\gamma \end{array}\right]^T, \tag{23}$$

$$\hat{\mathbf{d}}^\gamma = \left[\begin{array}{cccc} \gamma\mathbf{d}_0^\gamma + 1 & \gamma\mathbf{d}_1^\gamma & \cdots & \gamma\mathbf{d}_{N-1}^\gamma \end{array}\right]^T, \tag{24}$$

and

$$\mathbf{M}(\mathbf{a})_{i,j} = w_i w_j \left(\frac{a_{j+i}}{2w_{j+i}} + \frac{a_{j-i}}{2w_{j-i}}\right). \tag{25}$$

In Eq. (25), w_i denotes the weighting factor for DCT, which is given by

$$w_i = \left\{\begin{array}{ll} 1/\sqrt{N} & i = 0 \\ \sqrt{2/N} & 1 \le i < N \end{array}\right. \tag{26}$$

In case $\gamma = 0$, since the first entries of the vector $\hat{\mathbf{x}}^\gamma$ and $\hat{\mathbf{d}}^\gamma$ are one and the other entries of those vectors are zero, $\mathbf{M}(\hat{\mathbf{x}}^\gamma)$ and $\mathbf{M}(\hat{\mathbf{d}}^\gamma)$ are identity matrices. Then, it is proved that when $\gamma = 0$, Eq. (18) are equal to Eq. (6). It is indicated that MFCC is the specific case of the generalized cepstral coefficient, as shown at Eq. (1).

B. Model Adaptation

Assuming that $\mathbf{x}^\gamma$, $\mathbf{h}^\gamma$, and $\mathbf{d}^\gamma$ are Gaussian in the generalized cepstral domain with means $\mu_{\mathbf{x}}^\gamma$, $\mu_{\mathbf{h}}^\gamma$, and $\mu_{\mathbf{d}}^\gamma$, and covariance matrices $\sigma_{\mathbf{x}}^\gamma$, $\sigma_{\mathbf{h}}^\gamma$, and $\sigma_{\mathbf{d}}^\gamma$, the generalized cepstrum of the observed signal can be approximated using a first order Taylor series expansion at ($\mu_{\mathbf{x}}^\gamma$, $\mu_{\mathbf{h}}^\gamma$, $\mu_{\mathbf{d}}^\gamma$):

$$\begin{aligned}\mathbf{y}^\gamma \approx & \mu_{\mathbf{x}}^\gamma + \mathbf{M}(\mu_{\mathbf{x}}^\gamma)\mu_{\mathbf{h}}^\gamma + \mathbf{M}(\mu_{\mathbf{n}}^\gamma)\mathbf{g}^\gamma(\bar{\xi}) \\ & + \mathbf{F}_{\mathbf{x}}^\gamma(\mathbf{x}^\gamma - \mu_{\mathbf{x}}^\gamma) + \mathbf{F}_{\mathbf{h}}^\gamma(\mathbf{h}^\gamma - \mu_{\mathbf{h}}^\gamma) + \mathbf{F}_{\mathbf{d}}^\gamma(\mathbf{d}^\gamma - \mu_{\mathbf{d}}^\gamma)\end{aligned} \tag{27}$$

where $\mathbf{F_x}^\gamma$, $\mathbf{F_h}^\gamma$, and $\mathbf{F_d}^\gamma$ denote the derivative of $\mathbf{y}^\gamma$ in terms of $\mathbf{x}^\gamma$, $\mathbf{h}^\gamma$, and $\mathbf{d}^\gamma$ at the point ($\mu_{\mathbf{x}}$, $\mu_{\mathbf{h}}$, $\mu_{\mathbf{d}}$), respectively, and

$$\bar{\xi} = \frac{\left(\gamma\mathbf{C}^{-1}\mu_{\mathbf{x}}^\gamma + 1\right)^{1/\gamma}\left(\gamma\mathbf{C}^{-1}\mu_{\mathbf{h}}^\gamma + 1\right)^{1/\gamma}}{\left(\gamma\mathbf{C}^{-1}\mu_{\mathbf{d}}^\gamma + 1\right)^{1/\gamma}}. \tag{28}$$

The $(i,j)^{th}$ entry of $\mathbf{F_x}$, $\mathbf{F_h}$, and $\mathbf{F_d}$ are represented by

$$\begin{aligned}\mathbf{F}_{\mathbf{x}}^\gamma(i,j) = & \frac{\partial\mathbf{x}_i^\gamma}{\partial\mathbf{x}_j^\gamma} + w_i\sum_{m'=0}\frac{w_{m'}\mathbf{h}_{m'}^\gamma}{2w_{m'+i}}\frac{\partial\hat{\mathbf{x}}_{m'+i}^\gamma}{\partial\mathbf{x}_j^\gamma} \\ & + w_i\sum_{m'=0}\frac{w_{m'}\mathbf{h}_{m'}^\gamma}{2w_{m'-i}}\frac{\partial\hat{\mathbf{x}}_{m'-i}^\gamma}{\partial\mathbf{x}_j^\gamma} \\ & + w_i\sum_{m'=0}\left(\frac{w_{m'}\hat{\mathbf{d}}_{m'+i}^\gamma}{2w_{m'+i}} + \frac{w_{m'}\hat{\mathbf{d}}_{m'-i}^\gamma}{2w_{m'-i}}\right)\frac{\partial\mathbf{g}_{m'}^\gamma}{\partial\mathbf{x}_j^\gamma},\end{aligned} \tag{29}$$

$$\mathbf{F}_{\mathbf{h}}^{\gamma}(i,j) = w_i \sum_{m'=0} \left(\frac{w_{m'} \hat{\mathbf{x}}_{m'+i}^{\gamma}}{2w_{m'+i}} + \frac{w_{m'} \hat{\mathbf{x}}_{m'-i}^{\gamma}}{2w_{m'-i}} \right) \frac{\partial \mathbf{h}_{m'}^{\gamma}}{\partial \mathbf{h}_j^{\gamma}} + w_i \sum_{m'=0} \left(\frac{w_{m'} \hat{\mathbf{d}}_{m'+i}^{\gamma}}{2w_{m'+i}} + \frac{w_{m'} \hat{\mathbf{d}}_{m'-i}^{\gamma}}{2w_{m'-i}} \right) \frac{\partial \mathbf{g}_{m'}^{\gamma}}{\partial \mathbf{h}_j^{\gamma}}, \quad (30)$$

$$\mathbf{F}_{\mathbf{d}}^{\gamma}(i,j) = w_i \sum_{m'=0} \left(\frac{w_{m'} \hat{\mathbf{d}}_{m'+i}^{\gamma}}{2w_{m'+i}} + \frac{w_{m'} \hat{\mathbf{d}}_{m'-i}^{\gamma}}{2w_{m'-i}} \right) \frac{\partial \mathbf{g}_{m'}^{\gamma}}{\partial \mathbf{d}_j^{\gamma}} + w_i \sum_{m'=0} \frac{w_{m'} \mathbf{g}_{m'}^{\gamma}}{2w_{m'+i}} \frac{\partial \hat{\mathbf{d}}_{m'+i}^{\gamma}}{\partial \mathbf{d}_j^{\gamma}} + w_i \sum_{m'=0} \frac{w_{m'} \mathbf{g}_{m'}^{\gamma}}{2w_{m'-i}} \frac{\partial \hat{\mathbf{d}}_{m'-i}^{\gamma}}{\partial \mathbf{d}_j^{\gamma}}, \quad (31)$$

where

$$\frac{\partial \mathbf{g}_i^{\gamma}}{\partial \mathbf{x}_j^{\gamma}} = \sum_m \mathbf{C}_{i,m} \left\{ \bar{\xi}_m \left(\bar{\xi}_m + 1 \right)^{\gamma-1} + \left(\bar{\xi}_m \right)^{\gamma} \right\} \cdot \left(\gamma \sum_k \mathbf{C}_{m,k}^{-1} \mu_{\mathbf{x}_k} + 1 \right)^{-1} \mathbf{C}_{m,j}^{-1}, \quad (32)$$

$$\frac{\partial \mathbf{g}_i^{\gamma}}{\partial \mathbf{h}_j^{\gamma}} = \sum_m \mathbf{C}_{i,m} \left\{ \bar{\xi}_m \left(\bar{\xi}_m + 1 \right)^{\gamma-1} + \left(\bar{\xi}_m \right)^{\gamma} \right\} \cdot \left(\gamma \sum_k \mathbf{C}_{m,k}^{-1} \mu_{\mathbf{h}_k} + 1 \right)^{-1} \mathbf{C}_{m,j}^{-1}, \quad (33)$$

$$\frac{\partial \mathbf{g}_i^{\gamma}}{\partial \mathbf{d}_j^{\gamma}} = \sum_m \mathbf{C}_{i,m} \left\{ \bar{\xi}_m \left(\bar{\xi}_m + 1 \right)^{\gamma-1} + \left(\bar{\xi}_m \right)^{\gamma} \right\} \cdot \left(\gamma \sum_k \mathbf{C}_{m,k}^{-1} \mu_{\mathbf{d}_k} + 1 \right)^{-1} \mathbf{C}_{m,j}^{-1}, \quad (34)$$

$$\frac{\partial \hat{\mathbf{x}}_i^{\gamma}}{\partial \mathbf{x}_j^{\gamma}} = \gamma \sum_{m=0} \mathbf{C}_{i,m} \mathbf{C}_{m,j}^{-1}, \quad (35)$$

and

$$\frac{\partial \hat{\mathbf{d}}_i^{\gamma}}{\partial \mathbf{d}_j^{\gamma}} = \gamma \sum_{m=0} \mathbf{C}_{i,m} \mathbf{C}_{m,j}^{-1}. \quad (36)$$

Then, the mean and covariance matrices of $\mathbf{y}^{\gamma}$, its delta cepstrum, and its delta-delta cepstrum can be obtained by

$$\mu_{\mathbf{y}}^{\gamma} \approx \mu_{\mathbf{x}}^{\gamma} + \mathbf{M}\left(\mu_{\mathbf{x}}^{\gamma}\right) \mu_{\mathbf{h}}^{\gamma} + \mathbf{M}\left(\mu_{\mathbf{d}}^{\gamma}\right) \mathbf{g}^{\gamma}\left(\bar{\xi}\right) \quad (37a)$$

$$\sigma_{\mathbf{y}}^{\gamma} \approx \mathbf{F}_{\mathbf{x}}^{\gamma} \sigma_{\mathbf{x}} (\mathbf{F}_{\mathbf{x}}^{\gamma})^{\mathbf{T}} + \mathbf{F}_{\mathbf{h}}^{\gamma} \sigma_{\mathbf{h}} (\mathbf{F}_{\mathbf{h}}^{\gamma})^{\mathbf{T}} + \mathbf{F}_{\mathbf{d}}^{\gamma} \sigma_{\mathbf{d}} (\mathbf{F}_{\mathbf{d}}^{\gamma})^{\mathbf{T}} \quad (37b)$$

$$\mu_{\Delta \mathbf{y}}^{\gamma} \approx \mathbf{F}_{\mathbf{x}}^{\gamma} \mu_{\Delta \mathbf{x}}^{\gamma} \quad (38a)$$

$$\sigma_{\Delta \mathbf{y}}^{\gamma} \approx \mathbf{F}_{\mathbf{x}}^{\gamma} \sigma_{\Delta \mathbf{x}} (\mathbf{F}_{\mathbf{x}}^{\gamma})^{\mathbf{T}} + \mathbf{F}_{\mathbf{d}}^{\gamma} \sigma_{\Delta \mathbf{d}} (\mathbf{F}_{\mathbf{d}}^{\gamma})^{\mathbf{T}} \quad (38b)$$

$$\mu_{\Delta^2 \mathbf{y}}^{\gamma} \approx \mathbf{F}_{\mathbf{x}}^{\gamma} \mu_{\Delta^2 \mathbf{x}}^{\gamma} \quad (39a)$$

$$\sigma_{\Delta^2 \mathbf{y}}^{\gamma} \approx \mathbf{F}_{\mathbf{x}}^{\gamma} \sigma_{\Delta^2 \mathbf{x}} (\mathbf{F}_{\mathbf{x}}^{\gamma})^{\mathbf{T}} + \mathbf{F}_{\mathbf{d}}^{\gamma} \sigma_{\Delta^2 \mathbf{d}} (\mathbf{F}_{\mathbf{d}}^{\gamma})^{\mathbf{T}} \quad (39b)$$

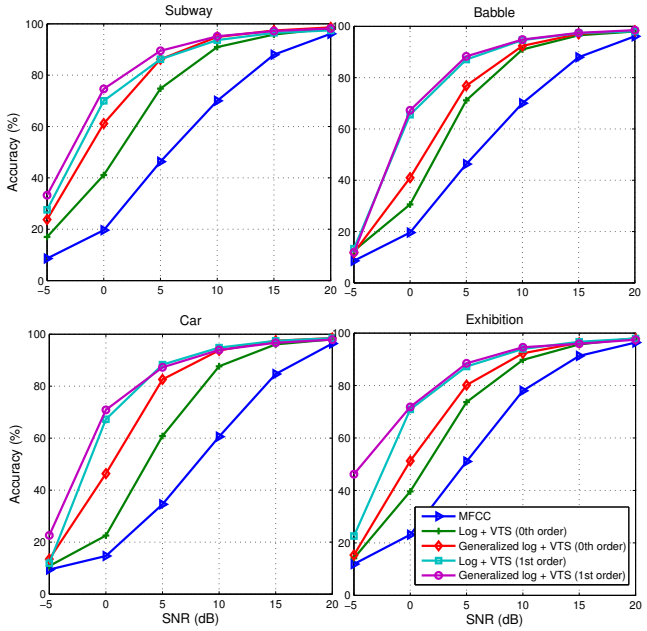

Fig. 2. Recognition performance for test Set A of Aurora 2.

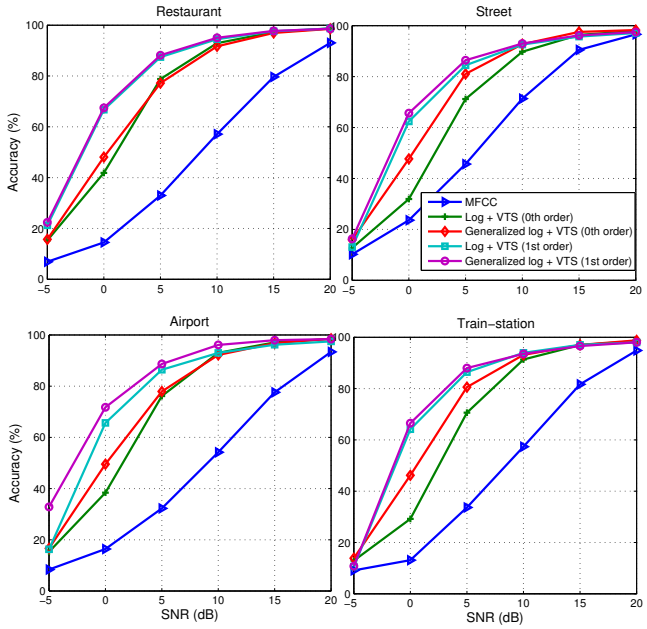

Fig. 3. Recognition performance for test Set B of Aurora 2.

When $\gamma = 0$, Eq. (37), (38), and (39) become equivalent to Eq. (12), Eq. (13), and Eq. (14), respectively. It is indicated that the proposed HMM update equation is the generic form of the conventional approach.

VI. Experiments

In this section, we compare the ASR performance of VTS approximations of order zero and one in the cepstral domain

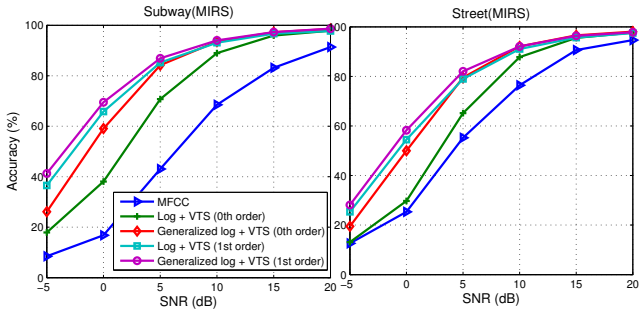

Fig. 4. Recognition performance for test Set C of Aurora 2.

and in the generalized cepstral domain. The VTS based approach of order zero models the effect of the environment on clean speech distributions only as a shift of the mean by Eq. (37). The word accuracy for Aurora 2 database and its error pattern are showed.

A. Experimental Setups

The recognition experiment is conducted with Aurora 2 database, which is a noisy speech database distributed by European telecommunications standards institute (ETSI). The source speech is the downsampling version of TIDIGITS consisting of English connected digit strings. The different types of the noise are added to clean speech with various SNRs. Test data of Aurora 2 database are composed of three different sets such as Test set A, Test set B and Test set C. In Test set C, speech samples are filtered with an MIRS characteristic to show the influence of recognition performance in the telephone channel distortion environment.

The simulation setup is designed by the method introduced in [11] with HTK toolkit v3.4. A 39 component feature vector including 13 MFCCs and its delta and acceleration coefficients is used for the recognizer. The acoustic model that has eleven whole word HMMs with 16 states and 3 Gaussian mixtures is trained by using clean training set comprised of 8440 utterances.

In the VTS based model adaptation approach, it is assumed that the mean and covariance matrices of the generalized cepstrum of the noise signal and the channel distortion are known. In practical situations, these environment parameters can be estimated using a traditional iterative EM approach from noisy signals. In addition, the value of γ is set to 0.2.

B. Recognition Results

Fig. 2, 3, and 4 show the word accuracy of the VTS based model adaptation approach of order zero and one. When 0^{th} order VTS approach is used, the generalized cepstrum based system has a significant performance improvement comparing MFCC based system. In addition, in the case of 1^{st} order VTS approach, the word accuracy is increased by model adaptation in the generalized cepstral domain.

Fig. 5 depicts the error patterns that count the recognition errors by dividing them into tree types: deletion error, substitution error, and insertion error. In addition, the averaged

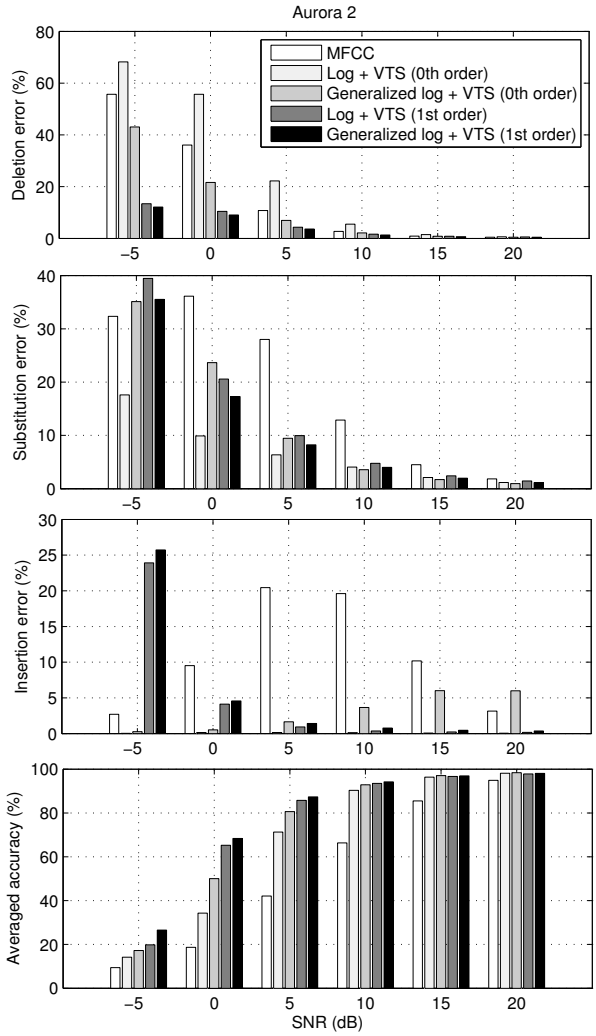

Fig. 5. Error distributions of the ASR system.

accuracy is also showed. Note that the model adaptation in the generalized cepstral domain reduces deletion errors and substitution errors. While insertion errors are increased, the total ASR performance is improved. The average relative WER improvement are 5.26% and 2.11% in the case of 0^{th} and 1^{st} order VTS approaches, respectively.

VII. Conclusion

In this paper, the method for estimating HMM parameters composed of the generalized cepstrum under noisy environments has been proposed. We extended the VTS based approach to the generalized cepstrum based ASR system. It was proved that the expression of the generalized cepstral coefficient is the generic form of MFCCs and the VTS approximation in the generalized cepstrum is useful for ASR in noisy environments. Notably, the number of deletion errors are reduced compared to the conventional approach for MFCCs.

REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech communication*, vol. 34, no. 3, pp. 267–285, 2001.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *The Proceedings of the 6th International Conference on Spoken Language Processing (Volume)*, 2000.

[3] H. Hermansky, "Perceptual linear predictive analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, 1990.

[4] M. J. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[5] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "A unified framework of hmm adaptation with joint compensation of additive and convolutive distortions," *Computer Speech & Language*, vol. 23, no. 3, pp. 389–405, 2009.

[6] C. Kim and R. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," in *INTERSPEECH*, 2009, pp. 28–31.

[7] ——, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. , 2012, pp. 4101–4104.

[8] E. Standard, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," *ETSI ES*, vol. 202, no. 050, p. v1, 2007.

[9] J. Lee, S. Baek, and H.-G. Kang, "Signal and feature domain enhancement approaches for robust speech recognition," in *8th International Conference on Information, Communications and Signal Processing (ICICS)*, dec. 2011, pp. 1 –4.

[10] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 5, pp. 1087–1089, 1984.

[11] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.