Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 32546, 7 pages
doi:10.1155/2007/32546

Research Article

Model Compensation Approach Based on Nonuniform Spectral Compression Features for Noisy Speech Recognition

Geng-Xin Ning, Gang Wei, and Kam-Keung Chu

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China

Received 8 October 2005; Revised 20 December 2006; Accepted 20 December 2006

Recommended by Douglas O'Shaughnessy

This paper presents a novel model compensation (MC) method for the features of mel-frequency cepstral coefficients (MFCCs) with signal-to-noise-ratio- (SNR-) dependent nonuniform spectral compression (SNSC). Though these new MFCCs derived from a SNSC scheme have been shown to be robust features under matched case, they suffer from serious mismatch when the reference models are trained at different SNRs and in different environments. To solve this drawback, a compressed mismatch function is defined for the static observations with nonuniform spectral compression. The means and variances of the static features with spectral compression are derived according to this mismatch function. Experimental results show that the proposed method is able to provide recognition accuracy better than conventional MC methods when using uncompressed features especially at very low SNR under different noises. Moreover, the new compensation method has a computational complexity slightly above that of conventional MC methods.

1. INTRODUCTION

The problem of achieving robust speech recognition in noisy environments has aroused much interest in the past decades. However, drastic degradation of performance may still occur when a recognizer operates under noisy circumstances. Resolutions to this problem can be generally divided into three categories: inherently robust feature representation [1], speech enhancement schemes [2], and model-based compensation [3–6]. More details are reviewed in [7]. Recently, different speech analyses based on psychoacoustics have been reported in the literature [8]. The well-known perceptual linear prediction (PLP) [9] uses critical band filtering followed by equal-loudness pre-emphasis to simulate, respectively, the frequency resolution and frequency sensitivity of the auditory system. Cubic-root spectral magnitude compression with a fixed compression root is subsequently used to approximate the intensity-to-loudness conversion. However, it is suboptimal to use a constant root for compressing all the filter bank outputs, because employing a constant compression root would over-compress some outputs and under-compress other outputs at the same time.

A new kind of noise-resistant feature by employing a SNR-dependent nonuniform spectral compression scheme was presented in [1], which compress the corrupted speech spectrum by a SNR-dependent root value. [1] has shown that the SNSC derived mel-frequency cepstral coefficients (SNSC-MFCC) features are able to provide recognition accuracy better than the conventional MFCC features and cubic-root compressed features. In a SNSC scheme, the compressed speech spectra in the linear-spectral domain, $\mathbf{Y}_k$, is expressed as

$$\mathbf{Y}_k = (Y_k)^{\alpha_k} \quad \text{for } 0 \le \alpha_k \le 1,\ Y_k > 1, \tag{1}$$

where Y_k is the kth mel-scale filter bank output of a corrupted speech segment and α_k is the compression root for the kth filter band, which is SNR-dependent. However, since α_k is SNR-dependent, estimation of noise is required in the training session for finding α_k under a particular noise type and global SNR. Thus models estimated by training in this way should only be used for a recognizing task under the same global SNR and noise environment.

So as not to reestimate the model when adopting a SNSC scheme, we need to compensate the models for the mismatch

caused by the compression root. This paper presents a compensation scheme to compensate the recognition models trained with clean and uncompressed training data for mel-frequency cepstral coefficients SNSC-MFCC features in various noisy environments. In this scheme, we start with using conventional MC methods such as the PMC [3, 4] method or the VTS [6] approach, to produce compensated models for features of no compression. The means and variances of the compressed mismatch function are derived in the paper. With the use of Gaussian-Hermite numerical integrals [10], a model compensation procedure is developed. Most importantly, the new compensation scheme is applicable to any conventional model compensation method. The experimental results of the paper show that the new compensated models provide very good accuracy in recognizing SNSC-MFCC features at different SNRs in different noisy environments. The computational complexity of the proposed MC-SNSC method is comparable with conventional MC methods. We call our new scheme the model compensation approach based on SNR nonuniform spectral compression (MC-SNSC).

The structure of this paper is as follows. The SNSC method is briefly reviewed in Section 2. In Section 3, we will introduce the MC-SNSC approach. Series of experimental results along with discussion and analyses are then presented in Section 4. Our conclusions on this study will be given in the final section.

2. SNR-DEPENDENT NONUNIFORM SPECTRAL COMPRESSION

The functional diagram of the generation of SNSC-MFCC features is depicted in Figure 1. The testing utterance is segmented into frames using a Hamming window. The frequency spectra of the speech segments are computed via discrete Fourier transform (DFT). Their squared magnitude spectra are passed to the mel-scaled filter bank. After the mel-scaled bandpass filtering, the spectral compression is applied to the outputs as in (1). Taking the log of the compressed outputs and then the discrete cosine transform, we obtain the SNSC-MFCC features.

Simulated by the spectrally partial masking effect, the compression function α_k is defined as

$$\alpha_k = (1 - A_0)\left(1 - e^{-[\log(Y_k/\tilde{N}_k)-\beta]/\gamma}\right) \cdot u\left(\log\left(\frac{Y_k}{N_k}\right) - \beta\right) + A_0, \tag{2}$$

where A_0 is the floor compression root, β is the cutoff parameter to function as the just-audible threshold, γ is the parameter to control the steepness of the compression function, and $u(\cdot)$ is the unit step function. For SNR less than the cutoff, (2) yields the floor compression value. The compression function produces small α_k at a steep rate of change for small band SNR above the cutoff and large α_k asymptotically close to one at a gradual rate for large band SNR. This SNSC scheme renders the filter bank outputs of low SNR less contributed to the resulting speech features while the outputs of high SNR are largely emphasized.

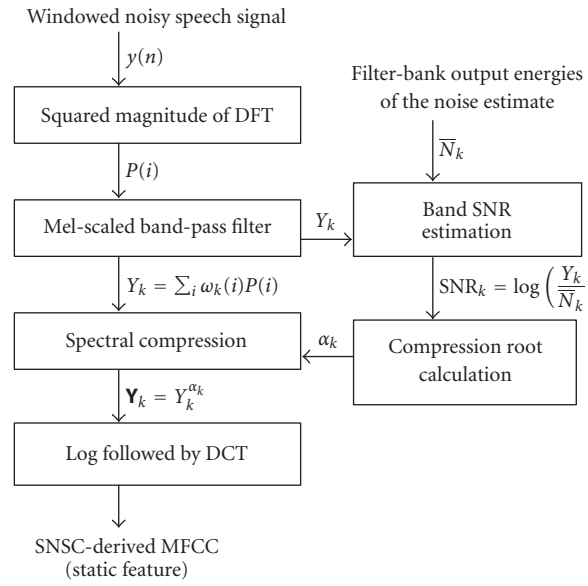

Figure 1: Procedure of the SNSC scheme.

The mismatch function Y_k of the kth mel-filter bank output, which is modeled as the sum of the noise energy N_k and the clean speech energy X_k in the linear-spectral domain, is expressed as

$$Y_k = X_k + N_k. \tag{3}$$

We define the clean speech and noise segment in the Log-spectral domain as $X_k^{(l)}$ and $N_k^{(l)}$, respectively, then the mismatch function in the log-spectral domain is expressed as

$$Y_k^{(l)} = \log\left(e^{X_k^{(l)}} + e^{N_k^{(l)}}\right). \tag{4}$$

Thus the compressed mismatch function for the SNSC in the log-spectral domain is expressed as

$$\mathbf{Y}_k^{(l)} = \alpha_k Y_k^{(l)}, \tag{5}$$

where

$$\alpha_k = (1 - A_0)\left(1 - e^{-(Y_k^{(l)} - N_k^{(l)} - \beta)/\gamma}\right) \cdot u\left(Y_k^{(l)} - N_k^{(l)} - \beta\right) + A_0. \tag{6}$$

In this paper, we make the following assumptions in order to facilitate the derivations of the MC procedures. (1) The recognition model is a standard HMM with mixture Gaussian output probability distributions. The transition probabilities and mixture component weights of the models are assumed to be unaffected by the additive noise. (2) The background noise is additive, stationary, and independent of the speech.

The notations for the description of variables in the paper are defined as follows. The superscripts (l) mean the

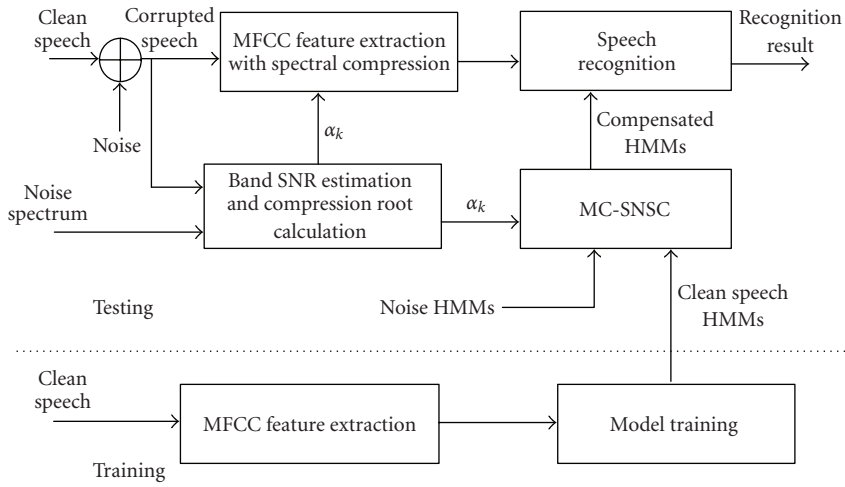

Figure 2: Processing stages for MC-SNSC approach.

log-spectral domains. When the variables have no superscript, they are the variables in the linear-spectral domain. The model parameters of the background noise model and the noise-corrupted speech model are capped with $\tilde{}$ and $\hat{}$, respectively.

3. MODEL COMPENSATION APPROACH BASED ON THE SNSC SCHEME

Figure 2 shows the functional diagram of the recognition system using model compensation for SNSC-MFCC features. In the training phase, clean speech HMMs are trained from standard MFCC features of which no compression is applied or the compression root is just equal to one. During the feature extraction in the testing phase, the SNSC scheme as described in (1) is used to compress each filter bank output. The clean HMMs are combined with the noise model to construct the corrupted speech models to recognize the SNSC-MFCC features using MC-SNSC approach.

There are no closed-form solutions for the moments of the mismatch function in (5) and (6). The expectations are multidimensional integrals for which we need to use computationally expensive numerical integrations to calculate the model parameters. With the use of assumption (2) and an additional assumption that the two random variables $Y_k^{(l)}$ and $N_k^{(l)}$ are uncorrelated, we can reduce the dimensionality of the integration. Using the Gauss-Hermite numerical integral method, we derive the procedures for computing the means and variances of the static features in the log-spectral domain in the next subsections.

3.1. Mean compensation

Using the compressed mismatch function described in (5), the mean of the static SNSC-MFCC feature in the log-spectral domain is given by

$$\begin{aligned}\hat{\mu}_{\mathbf{Y}_k}^{(l)} &= (1-A_0)\\ &\quad\cdot\Big(\mathrm{E}\{Y_k^{(l)}\cdot \mathrm{u}(Y_k^{(l)}-N_k^{(l)}-\beta)\}\\ &\qquad-\mathrm{E}\Big\{\mathrm{e}^{-(Y_k^{(l)}-N_k^{(l)}-\beta)/\gamma}\cdot Y_k^{(l)}\cdot \mathrm{u}(Y_k^{(l)}-N_k^{(l)}-\beta)\Big\}\Big)\\ &\quad+A_0\cdot\mathrm{E}\{Y_k^{(l)}\}.\end{aligned} \tag{7}$$

For the sake of simplifying the expression, we define

$$g(\gamma)=\mathrm{E}\Big\{\mathrm{e}^{-(Y_k^{(l)}-N_k^{(l)}-\beta)/\gamma}Y_k^{(l)}\mathrm{u}(Y_k^{(l)}-N_k^{(l)}-\beta)\Big\}. \tag{8}$$

Then the mean parameters of the static corrupted and compressed features are expressed as

$$\hat{\mu}_{\mathbf{Y}_k}^{(l)}=(1-A_0)[g(\infty)-g(\gamma)]+A_0\cdot\hat{\mu}_{Y_k}^{(l)}. \tag{9}$$

Using the Gauss-Hermite integral, $g(\gamma)$ is calculated as

$$g(\gamma)=\left[\frac{\hat{\Sigma}_{Y_{kk}}^{(l)}}{\sqrt{2\pi\Psi_k}}\mathrm{e}^{-[\Phi_k+\Psi_k/(2\gamma)]^2/2\Psi_k}+\Omega_k S(\gamma)\right]\mathrm{e}^{(\Phi_k+\Psi_k/(2\gamma))/\gamma} \tag{10}$$

with

$$S(\gamma)\cong\frac{1}{2}-\frac{1}{2\sqrt{\pi}}\sum_{i=1}^{n}\omega_i\,\mathrm{erf}\left(\frac{\sqrt{\tilde{\Sigma}_{N_{kk}}^{(l)}}}{\sqrt{\hat{\Sigma}_{Y_{kk}}^{(l)}}}t_i+\frac{\Phi_k+\Psi_k/\gamma}{\sqrt{2\hat{\Sigma}_{Y_{kk}}^{(l)}}}\right), \tag{11}$$

where $\Phi_k=\tilde{\mu}_{N_k}^{(l)}-\hat{\mu}_{Y_k}^{(l)}+\beta$, $\Psi_k=\tilde{\Sigma}_{N_{kk}}^{(l)}+\hat{\Sigma}_{Y_{kk}}^{(l)}$, $\Omega_k=\hat{\mu}_{Y_k}^{(l)}-(1/\gamma)\hat{\Sigma}_{Y_{kk}}^{(l)}$, and $\mathrm{erf}(\cdot)$ is the error function. The parameters t_i and ω_i for $i=1$ to n are, respectively, the abscissas and the weights of the nth-order Hermite polynomial $H_n(t)$ [10].

3.2. Variance compensation

The diagonal elements of the covariance matrix of the SNSC-MFCC static features are given by

$$\hat{\Sigma}_{\mathbf{Y}_{kk}}^{(l)} = \mathrm{E}\left\{\left(\mathbf{Y}_k^{(l)}\right)^2\right\} - \left(\hat{\mu}_{\mathbf{Y}_k}^{(l)}\right)^2 = (1-{A_0}^2) f(\infty) - 2(1-A_0) f(\gamma) + (1-A_0)^2 f\left(\frac{\gamma}{2}\right) + {A_0}^2 \cdot \left[\left(\hat{\mu}_{Y_k}^{(l)}\right)^2 + \hat{\Sigma}_{Y_{kk}}^{(l)}\right] - \left(\hat{\mu}_{\mathbf{Y}_k}^{(l)}\right)^2, \tag{12}$$

where

$$\begin{aligned} f(\gamma) &= \mathrm{E}\left\{\left(Y_k^{(l)}\right)^2 \cdot \mathrm{e}^{-\left(Y_k^{(l)} - N_k^{(l)} - \beta\right)/\gamma} \cdot \mathrm{u}\left(Y_k^{(l)} - N_k^{(l)} - \beta\right)\right\} \\ &= \mathrm{e}^{(\Phi_k + \Psi_k/(2\gamma))/\gamma} \cdot \left[\frac{\hat{\Sigma}_{Y_{kk}}^{(l)}}{\sqrt{2\pi\Psi_k}} \cdot \mathrm{e}^{-(\Phi_k + \Psi_k/\gamma)^2/2\Psi_k} \cdot \left(\frac{\hat{\Sigma}_{Y_{kk}}^{(l)}\Phi_k}{\Psi_k} + 2\hat{\mu}_{Y_k}^{(l)} - \frac{\hat{\Sigma}_{Y_{kk}}^{(l)}}{\gamma}\right) + \left(\hat{\Sigma}_{Y_{kk}}^{(l)} + {\Omega_k}^2\right) \cdot S(\gamma)\right]. \end{aligned} \tag{13}$$

The computations of the off-diagonal elements of the covariance matrix of static models involve two dimensional Gaussian-Hermite numerical integrals. To reduce the computational complexity, the off-diagonal elements are approximated as

$$\hat{\Sigma}_{\mathbf{Y}_{lk}}^{(l)} = \hat{\Sigma}_{(\alpha Y)_{lk}}^{(l)} \approx \lambda_{lk} \mathrm{E}\{\alpha_l\} \mathrm{E}\{\alpha_k\} \hat{\Sigma}_{Y_{lk}}^{(l)}, \tag{14}$$

where λ_{lk} is a scaling factor defined as

$$\lambda_{lk} = \lambda_{kl} = \sqrt{\rho_{kk}\rho_{ll}}, \qquad \rho_{kk} = \frac{\hat{\Sigma}_{\mathbf{Y}_{kk}}^{(l)}}{\hat{\Sigma}_{Y_{kk}}^{(l)}} \tag{15}$$

in order to ensure that the off-diagonal elements are smaller than the corresponding diagonal elements.

3.3. Corrupted models of noncompressed features

The above MC-SNSC procedures need the compensated static models of noncompressed corrupted speech in the log-spectral domain, $\{\hat{\mu}_{Y_k}^{(l)}, \hat{\Sigma}_{Y_{kl}}^{(l)}\}$. They can be obtained from any conventional model-based compensation methods such as the PMC method [3, 4] or the VTS (Vector Taylor series) [6].

In the log-normal PMC method, the kth elements of the mean vectors and the (k, l)th elements of the covariance matrices of the clean speech models in the linear-spectral domain are related to the log-spectral domain as

$$\mu_{X_k} = \mathrm{e}^{\mu_{X_k}^{(l)} + (1/2)\Sigma_{X_{kk}}^{(l)}}, \qquad \Sigma_{X_{kl}} = \mu_{X_k}\mu_{X_l}\left(\mathrm{e}^{\Sigma_{X_{kl}}^{(l)}} - 1\right). \tag{16}$$

In the linear-spectral domain, the noise is assumed to be additive and independent of the speech. The corrupted speech model parameters in this domain are obtained by combining the clean speech models and the noise model as

$$\hat{\boldsymbol{\mu}}_Y = \boldsymbol{\mu}_X + \tilde{\boldsymbol{\mu}}_N, \qquad \hat{\boldsymbol{\Sigma}}_Y = \boldsymbol{\Sigma}_X + \tilde{\boldsymbol{\Sigma}}_N. \tag{17}$$

Table 1: Index table for the ten compensation methods.

Index	Method
⟨1⟩	Mismatched case on MFCC
⟨2⟩	Mismatched case on SNSC-MFCC
⟨3⟩	Matched case on MFCC
⟨4⟩	Matched case on SNSC-MFCC
⟨5⟩	Log-add PMC on MFCC
⟨6⟩	MC-SNSC + log-add PMC on SNSC-MFCC
⟨7⟩	Log-normal PMC on MFCC
⟨8⟩	MC-SNSC + log-normal PMC on SNSC-MFCC
⟨9⟩	VTS-1 on MFCC
⟨10⟩	MC-SNSC + VTS-1 on SNSC-MFCC

After model combination, the model parameters are mapped back to the log-spectral domain as

$$\begin{aligned} \hat{\mu}_{Y_k}^{(l)} &= \log\left(\hat{\mu}_{Y_k}\right) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{Y_{kk}}}{\left(\hat{\mu}_{Y_k}\right)^2} + 1\right), \\ \hat{\Sigma}_{Y_{kl}}^{(l)} &= \log\left(\frac{\hat{\Sigma}_{Y_{kk}}}{\hat{\mu}_{Y_k}\hat{\mu}_{Y_l}} + 1\right). \end{aligned} \tag{18}$$

For the log-add PMC, the mean compensation is described as

$$\hat{\mu}_{Y_k}^{(l)} = \log\left(\mathrm{e}^{\mu_{X_k}^{(l)}} + \mathrm{e}^{\tilde{\mu}_{N_k}^{(l)}}\right). \tag{19}$$

This method only compensates for the mean but not the variance. It thus has low computational complexity. However, its performance becomes unsatisfactory at low SNR. This scheme can be viewed as the zeroth-order VTS (denoted as VTS-0).

The VTS method is to approximate the mismatch function by a finite length Taylor series, and the expectation of this Taylor series is taken to find the corrupted speech model parameters. A higher-order Taylor series can yield a better solution but its computational complexity is very expensive. Thus VTS-0 and first-order VTS (VTS-1) [6] are employed commonly. Using the VTS-1 method, the compensation of the mean is the same as the log-add PMC, and the covariance matrix $\hat{\boldsymbol{\Sigma}}_Y^{(l)}$ is compensated as

$$\hat{\boldsymbol{\Sigma}}_Y^{(l)} = \mathbf{M}\boldsymbol{\Sigma}_Y^{(l)}\mathbf{M}^T + (\mathbf{I} - \mathbf{M})\tilde{\boldsymbol{\Sigma}}_N^{(l)}(\mathbf{I} - \mathbf{M})^T, \tag{20}$$

where $\mathbf{M}$ is the diagonal matrix whose elements are expressed as

$$M_k = \frac{1}{1 + \mathrm{e}^{\left(\tilde{\mu}_{N_k}^{(l)} - \mu_{X_k}^{(l)}\right)}}. \tag{21}$$

As a brief summary, the MC-SNSC method uses the background noise model and the uncompressed corrupted-speech models to compute the compressed corrupted speech models. The band SNR-dependent SNSC is employed in this scheme to compress the features so as to emphasize the signal components of high SNR and de-emphasize the highly

Table 2: Word recognition rate (WRR) (%) from ten methods in different noise environments.

Noise	SNR/dB	⟨1⟩	⟨2⟩	⟨3⟩	⟨4⟩	⟨5⟩	⟨6⟩[(1)]	⟨7⟩	⟨8⟩[(2)]	⟨9⟩	⟨10⟩[(3)]
White	Clean	97.72	97.72	97.72	97.72	97.72	**97.72**	97.72	**97.72**	97.72	**97.72**
	30	94.21	96.43	97.42	97.00	96.90	**97.00**	96.78	**96.72**	96.17	**97.19**
	10	29.63	72.46	94.36	94.53	89.78	**92.05**	90.10	**92.52**	89.88	**93.26**
	5	11.48	53.64	90.60	91.27	81.43	**86.42**	83.80	**88.18**	85.67	**90.39**
	0	6.65	31.93	80.83	84.75	63.63	**72.52**	71.94	**80.09**	78.22	**84.65**
	−5	5.00	12.83	61.07	69.34	37.62	**48.18**	50.28	**61.29**	58.20	**68.62**
	Avg.*	7.71	32.80	77.50	81.79	60.89	**69.04**	68.67	**76.52**	74.03	**81.22**
Pink	Clean	97.72	97.72	97.72	97.72	97.72	**97.72**	97.72	**97.72**	97.72	**97.72**
	30	96.72	96.84	97.65	97.07	97.21	**97.15**	97.19	**97.41**	96.41	**97.10**
	10	40.77	81.91	94.66	95.43	90.78	**93.68**	92.16	**94.10**	92.31	**94.28**
	5	16.80	63.96	90.72	92.35	82.11	**88.45**	86.83	**90.04**	88.95	**91.92**
	0	7.92	34.28	83.09	86.02	61.52	**73.26**	75.70	**81.05**	82.44	**86.13**
	−5	5.22	11.07	64.21	70.26	29.57	**44.16**	48.54	**58.79**	63.21	**68.72**
	Avg.*	9.98	36.44	79.34	82.88	57.73	**68.62**	70.36	**76.63**	78.20	**82.26**
Factory	Clean	97.72	97.72	97.72	97.72	97.72	**97.72**	97.72	**97.72**	97.72	**97.72**
	30	97.13	96.38	97.43	97.14	97.11	**97.04**	97.43	**97.59**	96.92	**97.29**
	10	45.99	75.23	93.41	94.89	91.90	**93.43**	92.43	**92.74**	91.96	**93.23**
	5	20.84	55.41	89.17	91.79	83.63	**87.94**	86.31	**88.37**	87.42	**90.45**
	0	9.42	30.50	78.53	83.57	63.31	**71.34**	74.45	**78.40**	77.47	**81.19**
	−5	6.67	12.11	59.46	65.05	35.19	**41.60**	50.96	**54.81**	58.21	**61.32**
	Avg.*	12.31	32.67	75.72	80.13	60.89	**66.96**	70.91	**73.86**	74.37	**77.66**

[(1,2,3)] For the Gauss-Hermite integral, $n = 4$ is employed. *Average WRR (%) between −5 and 5 dB.

noisy ones. The compressed corrupted speech models are then used for recognizing the SNSC-compressed testing features.

4. EVALUATION

In this section, three noise types from the NOISEX-92 database are used in the evaluation experiments including white, pink, and factor noises. The speech database used for the evaluation of the MC-SNSC techniques is TI-20 database from Ti-Digits which contains 20 isolated words, including digits "0" to "9" plus ten extra commands like "help" and "repeat." The speech database was spoken by 16 speakers (8 males and 8 females), and we select 2 and 16 utterances for training and testing, respectively, from each speaker and each word (641 utterances for training and 5081 utterances for testing). The length of the analysis frame (Hamming windowed) is 32 milliseconds, and the frame rate is 9.6 milliseconds. The feature vector is composed of 13 static cepstral coefficients.

A word-based HMM with six states and four mixture Gaussian densities per state is used as the reference model. In the training mode, we train the system with the clean speech utterances to produce clean models and corrupted speech for the matched case. In the testing, the ten speech recognition methods as listed in Table 1 are used for the performance evaluation. These nine methods are two mismatched and two matched cases; three conventional model-based compensation methods: the log-normal, the log-add PMC, and the first order VTS (denoted as VTS-1); and these three conventional methods plus the MC-SNSC method.

For our MC-SNSC approach, an average background noise power spectrum is needed to estimate the background noise model, and to estimate the band SNR for calculating the SNSC-derived features in the testing phase. The average noise power spectrum is calculated by using 200 nonoverlapping frames of noise data and is scaled according to a specified global SNR. The global SNR for an utterance is defined as

$$\mathrm{SNR}_{\mathrm{global}} = 10\log_{10}\frac{\sum_{m=1}^{O}\sum_{k=0}^{Q/2}P_m(k)}{O\sum_{k=0}^{Q/2}g^2\overline{N}(k)}, \tag{22}$$

where $\{P_m(k)\}$ is the clean speech power spectrum of the mth frame, $\{\overline{N}(k)\}$ is the nonscaled average noise power spectrum, O is the total number of frame for the utterance, Q is the FFT size, and g is the scaling factor to scale the ratio according to a specified $\mathrm{SNR}_{\mathrm{global}}$. Thus, the corrupted speech is produced by

$$y(i) = x(i) + g \cdot n(i), \tag{23}$$

where $y(i)$ is the corrupted speech, $x(i)$ and $n(i)$ are the clean speech and the nonscaled noise signal, respectively.

Table 3: Computational complexity of each MC method.

Method	Number of operations	Total
Log-add PMC	$2M(N+1)+M$	725
Log-add PMC	$2M(N+1)+M$	
+ MC-SNSC	$+2M^2+(3n+41)M$	3300
Log-normal PMC	$MN(2M+N+3)+2M(3M+2)$	25300
MC-SNSC +	$MN(2M+N+3)+2M(3M+2)$	27875
log-normal PMC	$+2M^2+(3n+41)M$	
VTS-1	$MN(2M+N+3)+6M^2+8M$	25400
MC-SNSC +	$MN(2M+N+3)$	27975
VTS-1	$+8M^2+(3n+49)M$	

Experimental results for three different additive noises are shown in Table 2. For the MC-SNSC method, the parameters (A_0, β, γ) are set according to lots of testing experiments. The method can obtain good performance when the parameters are set in the area of $A_0 \in [0.7, 0.9]$, $\beta \in [-0.6, 0.6]$, and $\gamma \in [1, 2]$. In this work, we fix the parameter set as $A_0 = 0.75$, $\beta = -0.4$, and $\gamma = 1$.

The results show that all MC methods can achieve good performance for the three additive noises at low SNR. For the sake of comparison, we define an average performance gain G_{ave} of a MC method as the average of the difference of the recognition rates in absolute percentage of the MC method using MC-SNSC and its original counterpart over the four noises. For the -5 dB case, the G_{ave} of the MC-SNSC plus the log-add PMC, the MC-SNSC plus the log-normal PMC, the MC-SNSC plus the VTS-1 are 11%, 10.5%, and 5%, respectively. For 0 dB case, the G_{ave} of the three methods are 9.5%, 7%, and 4.3%, respectively. The experimental results also show that the MC-SNSC scheme can enhance the performance of the original method under the four noises for all SNR cases. It is worth noting that at low SNR as 0, -5 dB, even MC-SNSC gives a better performance than the matched case based on MFCC features.

These experimental results reveal that the new MC-SNSC scheme can deal with different types of additive noise and yield remarkable recognition performance, which is attributed to the noise-resistant feature extraction (SNSC scheme) [1] and pertinent model compensation.

Table 3 lists the number of multiplication, division, logarithm, and exponential operations for each technique to update the parameters of a single mixture density for static parameters, where N and M are the dimensions of features in the cepstral domain and the log-spectral domain, respectively. It can be seen that the computational complexity of the MC-SNSC plus the conventional MC methods is comparable to that of the conventional MC methods. However, the MC-SNSC is more effective than the conventional model compensation methods.

5. CONCLUSION

A novel model compensation approach for robust SNSC-MFCC features is presented in this paper. Meanwhile a compressed mismatch function is defined for the static observations with nonuniform spectral compression. The model-based compensation method for compressed feature has been derived, which employs a Gauss-Hermite integral and the conventional MC approach. The experimental outcome demonstrates that the MC-SNSC approach can cope with different kinds of noises automatically with enhanced recognition accuracy substantially, especially in low SNR in comparison with the conventional MC approaches. In addition, the complexity of the MC approach plus the MC-SNSC method is not very expensive and it is comparable with a correspondent MC approach.

ACKNOWLEDGMENTS

This work was supported by the Nature Science Fund of China (no. 60502041), the Doctoral Program Fund of Guangdong Natural Science Foundation (no. 05300146), and the Natural Science Youth Fund of South China University of Technology.

REFERENCES

[1] K. K. Chu and S. H. Leung, "SNR-dependent non-uniform spectral compression for noisy speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 1, pp. 973–976, Montreal, Quebec, Canada, May 2004.

[2] T. Lotter, C. Benien, and P. Vary, "Multichannel direction-independent speech enhancement using spectral amplitude estimation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1147–1156, 2003.

[3] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, vol. 12, no. 3, pp. 231–239, 1993.

[4] M. J. F. 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.-W. Hung, J.-L. Shen, and L.-S. Lee, "New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 842–855, 2001.

[6] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 2, pp. 733–736, Atlanta, Ga, USA, May 1996.

[7] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[8] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*, Springer, New York, NY, USA, 2nd edition, 1999.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[10] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, NY, USA, 1972.

Geng-Xin Ning was born in January 1981. He received the B.S. degree from Jilin University, Changchun, China, and the Ph.D. degree from South China University of Technology, Guangzhou, China, in 2001 and 2006, respectively. He is currently a lecturer in the School of Electronic and Information Engineering, South China University of Technology. His research interests are speech coding and speech recognition.

Gang Wei was born in January 1963. He received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree from South China University of Technology, Guangzhou, China, in 1984, 1987, and 1990, respectively. He is currently a Professor in the School of Electronic and Information Engineering, South China University of Technology. His research interests are signal processing and personal communications.

Kam-Keung Chu received the B.S. degree from City University of Hong Kong, Hong Kong, in 2005. His research interest is speech recognition. He received the B.S. degree honors in applied physics from City University of Hong Kong in 2000. He further pursued his study in the Department of Electronic Engineering in the same university and got his M.Phil. degree for research in speech recogniton. His research interests include speech recognition in noisy environment and sensation of sound by human in noisy environment.