Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 451–464
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Variational noise model composition through model perturbation for robust speech recognition with time-varying background noise

Wooil Kim, John H.L. Hansen *

Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA

Received 9 August 2009; received in revised form 29 August 2010; accepted 1 December 2010
Available online 28 December 2010

Abstract

This study proposes a novel model composition method to improve speech recognition performance in time-varying background noise conditions. It is suggested that each element of the cepstral coefficients represents the frequency degree of the changing components in the envelope of the log-spectrum. With this motivation, in the proposed method, variational noise models are formulated by selectively applying perturbation factors to the mean parameters of a basis model, resulting in a collection of noise models that more accurately reflect the natural range of spectral patterns seen in the log-spectral domain. The basis noise model is obtained from the silence segments of the input speech. The perturbation factors are designed separately for changes in the energy level and spectral envelope. The proposed variational model composition (VMC) method is employed to generate multiple environmental models for our previously proposed parallel combined gaussian mixture model (PCGMM) based feature compensation algorithm. The mixture sharing technique is integrated to reduce computational expenses, caused by employing the variational models. Experimental results prove that the proposed method is considerably more effective at increasing speech recognition performance in time-varying background noise conditions, with +31.31%, +10.65%, and +20.54% average relative improvements in word error rate for speech babble, background music, and real-life in-vehicle noise conditions respectively, compared to the original basic PCGMM method.

Keywords: Variational model composition (VMC); Time-varying noise; Feature compensation; Multiple environmental models; Robust speech recognition

1. Introduction

Acoustic mismatch between training and operating conditions of an actual speech recognition system is one of the primary factors severely degrading recognition performance. To minimize this mismatch, extensive research has been conducted in recent decades, which includes many types of speech/feature enhancement methods such as spectral subtraction (Boll, 1979; Martin, 1994), cepstral mean normalization, and a variety of feature compensation schemes (Ephraim and Malah, 1984; Hansen and Clements, 1991; Hansen, 1994; Moreno et al., 1998; Kim, 2002; Sasou et al., 2004; Stouten et al., 2004; Kim and Hansen, 2009a). Various model adaptation techniques have also been successfully employed such as maximum a posteriori (MAP) (Gauvain and Lee, 1994), maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995) and parallel model combination (PMC) (Varga and Moore, 1990; Gales and Young, 1996). Recently, missing-feature methods have shown promising results (Cook et al., 2001; Raj et al., 2004; Kim and Hansen, 2009b), and some advanced schemes utilize no prior knowledge of the background noise (Kim and Stern, 2006).

Existing conventional methods have achieved successful results for improving speech recognition performance in noisy environments. These methods however, generally assume the target background noise to be stationary or slowly changing over the input speech duration, or that

* Corresponding author. Tel.: +1 972 883 2910; fax: +1 972 883 2710.
E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).
URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.12.001

even an a prior amount of knowledge of the noise signal is available. To estimate the background noise, many of these existing methods estimate the noise signal from the leading silence duration employing a voice activity detector. Various approaches for adaptive noise estimation have also been employed, however, they obtain only representative characteristics of the target noise signal (Martin, 1994, 2001; Hirsch and Ehrlicher, 1995; Gales and Young, 1996; Kim et al., 1997; Moreno et al., 1998; Frey et al., 2001; ETSI, 2002; Kim, 2002; Kim and Hansen, 2009a), and fail to reflect the true dynamic changes of background noise over both time and frequency domains. Therefore, they continue to suffer from ineffectiveness in time-varying background noise conditions, where the noise characteristics need to be effectively estimated as time evolves. In particular, a Monte Carlo based method was proposed utilizing the states of the speech model (i.e., Hidden Markov Model (HMM)) which are obtained by Viterbi decoding, where the updated noise model is used for updating the noise-corrupted HMM (Yao and Nakamura, 2001). Such an approach is out of the scope of our paper, where the feature compensation method is considered as a front-end procedure independent from back-end speech recognizer.

Speech signals are severely corrupted by time-varying background noise in real-life scenarios, and many of these reported advancements have yet to be employed in actual noisy speech recognition scenarios (i.e., speech obtained in the actual environment including noise, Lombard effect, and potential task induced stress). Such actual examples can be easily found in the corpora of in-vehicle scenarios such as UTDrive (Angkititrakul et al., 2007, 2009) and CU-Move (Hansen et al., 2004), or spoken document retrieval of diverse audio data such as the National Gallery of Spoken Word (NGSW) (Hansen et al., 2005) and the Collaborative Digitization Program (CDP) (Kim and Hansen, 2007), and others, which make the transition of speech recognition technology to real environments most challenging in everyday real-life.

In this study, a novel model composition method is proposed to address time-varying background noise for improved speech recognition. Our motivation is that each order of the cepstral coefficients represents a frequency degree of the changing components in the log-spectrum envelope (Deller et al., 2000). In the proposed method, variational noise models are generated by selectively applying perturbation factors to the mean parameters of a basis noise model aimed at achieving a range of spectral patterns to reflect the background noise signal included during the speech interval. The proposed variational model composition method is employed to generate multiple environmental models for our previously proposed gaussian mixture model (GMM) based feature compensation algorithm (Kim and Hansen, 2009a). In order to reduce the computational expense caused by employing the variational models, a mixture sharing method (Kim and Hansen, 2009a) is integrated into the proposed feature compensation scheme employing the variational model composition method. The proposed method will be evaluated on various types of background noise including speech babble and background music within the Aurora 2.0 evaluation framework (Hirsch and Pearce, 2000). The CU-Move corpus (Hansen et al., 2004) is also used for performance evaluation to prove the effectiveness of the proposed scheme in a real-life in-vehicle scenario.

The paper is organized as follows. First, the motivation of the proposed variational model composition method is presented and the detailed procedure of the proposed method is described in Section 2. A multiple-model based feature compensation method, as an application of the proposed study, is presented in Section 3 which has been developed in our previous study. The mixture sharing technique is presented in Section 4. The representative experimental procedures and results are presented and discussed in Section 5. Finally, Section 6 presents our conclusions and discussion for future work.

2. Variational model composition

In this section, a novel method is proposed to effectively estimate time-varying background noise corrupting the speech utterance by using information contained in the neighboring silent segments. As initial knowledge for our discussion, first, the effect on log-spectral coefficients caused by adding a gain to the cepstral coefficients is presented. From the fundamentals of the cepstrum, which is obtained by a discrete cosine transformation (DCT) of the log-spectrum, each order of the obtained cepstral coefficients represents the frequency of the log-spectrum envelope changes (i.e., *quefrency* Deller et al., 2000). For example, the lower order cepstral coefficients indicate a measure of the slowly changing components in the envelope of the log-spectrum, having the 0th cepstral coefficient represent the DC component (i.e., energy) of the log-spectrum at a frame. Therefore, applying a weight to each order of the cepstral coefficients could generate a variation of the original cepstrum in terms of the frequency of the envelope change along the log-spectral axis.

Assume that a vector of cepstral coefficients $\mathbf{x}$ consists of 0th to $(N-1)$th components. A variation of the cepstrum vector can be obtained by adding a gain vector $\mathbf{g}$ as follows:

$$\widetilde{\mathbf{x}} = \mathbf{x} + \mathbf{g}. \tag{1}$$

If the gain is applied only to the 0th coefficient such as $\mathbf{g} = [\pm g\, 0\, 0 \ldots 0]$, the log-spectral coefficients, which can be calculated by an inverse DCT of the obtained variation $\widetilde{\mathbf{x}}$, will have a different energy level from the original log-spectrum. In Fig. 1, the plots in (a) show log-spectra of the variations which are generated by adding gains of $\pm g$ to the 0th cepstral coefficient. The plain solid line indicates the original log-spectral coefficients and the lines with solid or open circles indicate the resulting log-spectrum by applying $+g$ and $-g$ to the 0th cepstral component respectively. We can see the two variations have different energy levels, while maintaining an identical spectral envelope

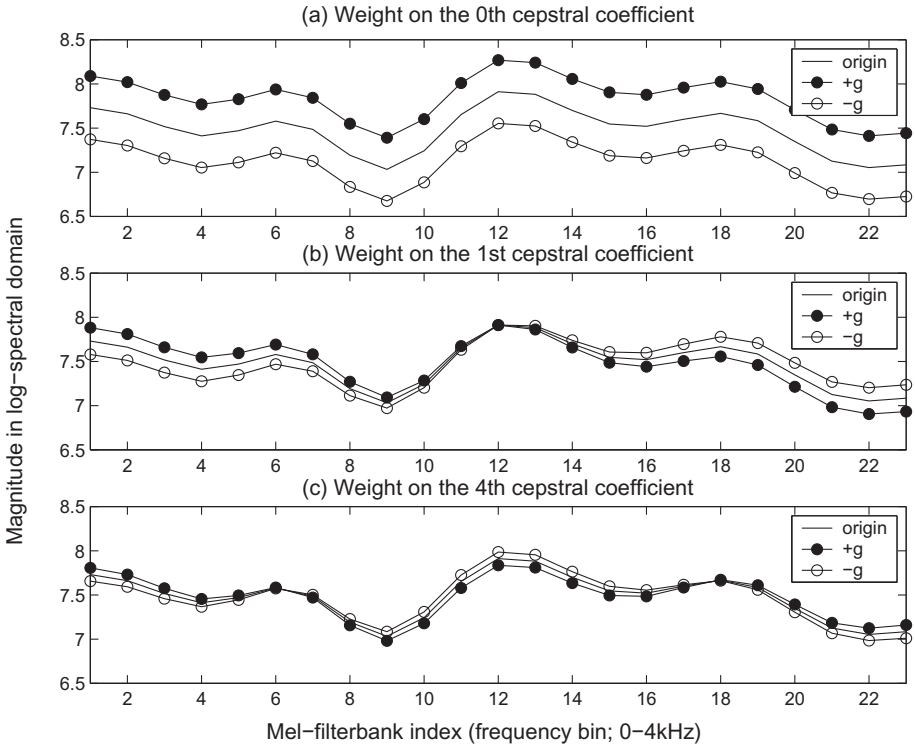

Fig. 1. Examples of variational log-spectral coefficients generated by applying a weight to the (a) 0th, (b) 1st, and (c) 4th cepstral coefficients.

shape in the original coefficients. Plots (b) and (c) present the log-spectra of the variations generated by applying weights only on the 1st and 4th cepstral components respectively. The variations in (b) show a smooth change of the envelope in the log-spectrum, with larger changes in low and high frequency bins, and the plots of the variations in (c) are varying relatively faster with several cross-over points over frequency bins from 0 to 4 kHz.

With this motivation, we believe that a range of models could be generated by *parameter perturbation*, which can be accomplished by applying a combination of weights to the mean parameter of an original model in the cepstral domain. In our proposed method, it is assumed that: (i) a basis noise model can be obtained from periods of silence within the input speech, and (ii) the variations (in terms of energy level or change of spectral envelope) of the estimated basis model might represent the target time-varying noise included in the speech duration. The *variational models* are generated by selectively applying weights on each component of the mean vector of the basis model in the cepstral domain (i.e., mean parameter perturbation). Here, we propose a novel algorithm to generate a collection of variational noise models by model perturbation in the following sections.

2.1. Step 1: basis model estimation

A basis noise model is obtained from silent duration segments within the input speech, which generally exists at beginning and end parts of an utterance. The basis model is estimated as a single Gaussian pdf $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ in the cepstral domain. The parameter vectors $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ consist of a total of N individual components as follows:

$$\boldsymbol{\mu} = [\mu_1\, \mu_2\, \ldots\, \mu_N], \quad \boldsymbol{\sigma} = [\sigma_1\, \sigma_2\, \ldots\, \sigma_N]. \tag{2}$$

While $\boldsymbol{\sigma}$ represents the standard deviation vector, the variance vector $\boldsymbol{\sigma}^2$ is actually used.

2.2. Step 2: Variational component determination

The V largest components $\{v_1, v_2, \ldots, v_V\}$ from the variance vector $\boldsymbol{\sigma}^2$ of the basis model obtained from Step 1 are selected. These terms are named the *variational components*, which are considered to have a statistically large range of variations from the original components (i.e., mean parameters of the basis model). In particular, the component v_1 is determined as the index of the 0th cepstral coefficient for the purpose of assigning a separate perturbation factor from the other components. In this paper, we use 13 cepstral coefficients including c0 are used for the feature vector and locate c0 as the first component (i.e., c0–c12). Therefore, the v_1 is forcedly set to be the first component in the proposed method, even though σ_{v_1} (i.e., variance of c0) will in general have the largest value. The remaining components $\{v_2, v_3, \ldots, v_V\}$ are determined in a size-ordered rank[1] as follows:

$$\sigma_{v_2} \geqslant \sigma_{v_3} \geqslant \cdots \geqslant \sigma_{v_V}. \tag{3}$$

[1] $\sigma^2_{v_2} \geqslant \sigma^2_{v_3} \geqslant \cdots \geqslant \sigma^2_{v_V}$ has the same order as $\sigma_{v_2} \geqslant \sigma_{v_3} \geqslant \cdots \geqslant \sigma_{v_V}$, since all standard deviations are positive values.

2.3. Step 3: model composition by mean perturbation

A variation of the mean vector is generated by selectively applying gains to the mean parameters of the basis model. The elements of the mean vector for applying the gains correspond to the determined variational components $\{v_1, v_2, \ldots, v_V\}$. From the motivation presented, it can be seen that applying a gain to the 0th component of the mean in the cepstral domain (i.e., v_1) will change the energy level of a basis noise model, while perturbing the other components will affect the frequency distribution of the envelope in the log-spectral domain (e.g., as illustrated in the example log-spectral feature plots in Fig. 1).

In our experiment, it was found that applying different styles of *perturbation factors* for v_1 and $\{v_2, v_3, \ldots, v_V\}$ separately is more effective at increasing the ability of the noise model to characterize the unseen time-varying background noise structure. It is noted that the noise during speech is "unseen" because we are only able to capture a snapshot of the noise spectral structure during periods of leading/trailing silence. According to our finding, a variation of the mean vector $\tilde{\boldsymbol{\mu}} = [\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_N]$ is obtained by applying two different styles of perturbation factors p_E and p_S to the variational components via Eq. (4):

$$\tilde{\mu}_i = \begin{cases} \mu_i + p_E, & \text{if } i = v_1, \\ \mu_i + p_S, & \text{else if } i \in \{v_2, v_3, \ldots, v_V\}, \\ \mu_i, & \text{otherwise.} \end{cases} \tag{4}$$

Here, the proposed perturbation factors p_E and p_S have three different types respectively as follows:

$$p_E = \{0, -\alpha\mu_i, \quad \text{or} \quad \alpha\mu_i\}, \tag{5}$$

$$p_S = \{0, -\beta\sigma_i, \quad \text{or} \quad \beta\sigma_i\}. \tag{6}$$

The combinations of these three types of perturbation factors for the V variational components generate a collection of variational models $\{\tilde{\boldsymbol{\lambda}} = (\tilde{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2)\}$ consisting of a total 3^V members. All variation models share the same original variance vector $\boldsymbol{\sigma}^2$ of the basis noise model. Determination of the coefficients for the perturbation factors α and β, and the number of the variational components V, will be discussed in Section 5 where we present the experimental results.

Table 1 presents an example of the variational model composition method proposed in this section. Here, the feature vector consists of 13 distinct cepstral coefficients (i.e., c0–c12), a set of 3 variational components are employed (i.e., $V = 3$). The first, fourth, and second indexes are selected for the variational components in this example (i.e., $\{v_1, v_2, v_3\} = \{1, 4, 2\}$). The three columns under the "Perturbation factor" show the combinations of the perturbation factors for the determined variational components $\{v_1, v_2, v_3\}$. Here, 27 (=3^3) represents the number of variational models that can be generated by employing the variations of the mean parameter of the basis model $\tilde{\boldsymbol{\mu}}_1$ to $\tilde{\boldsymbol{\mu}}_{27}$. Fig. 2 illustrates examples of the variational noise models (i.e., mean parameters in the log-spectral domain) generated by the example presented in Table 1. The responses show various spectral patterns generated by combinations of the weights at the selected variational cepstral components using a basis model which is presented as the dashed line in each figure. In this paper, we name our proposed method as the variational model composition (VMC) method. In the next section, we integrate this scheme into our feature compensation method.

Table 1
An example of the proposed variational model composition with $V = 3$.

Variational model	Perturbation factor			Parameter computation ($\tilde{\sigma}_{k,i} = \sigma_i$ for all i)
	v_1	v_2	v_3	
$\tilde{\boldsymbol{\mu}}_1$	0	0	0	$\tilde{\mu}_{1,i} = \mu_i$ for all i
$\tilde{\boldsymbol{\mu}}_2$	$-\alpha$	0	0	$\tilde{\mu}_{2,1} = \mu_1(1-\alpha)$, $\tilde{\mu}_{2,i} = \mu_i$ for all other i
$\tilde{\boldsymbol{\mu}}_3$	α	0	0	$\tilde{\mu}_{3,1} = \mu_1(1+\alpha)$, $\tilde{\mu}_{3,i} = \mu_i$ for all other i
⋮	⋮	⋮	⋮	⋮
$\tilde{\boldsymbol{\mu}}_7$	0	β	0	$\tilde{\mu}_{7,4} = \mu_4 + \beta\sigma_4$, $\tilde{\mu}_{7,i} = \mu_i$ for all other i
$\tilde{\boldsymbol{\mu}}_8$	$-\alpha$	β	0	$\tilde{\mu}_{8,1} = \mu_1(1-\alpha)$, $\tilde{\mu}_{8,4} = \mu_4 + \beta\sigma_4$, $\tilde{\mu}_{8,i} = \mu_i$ for all other i
$\tilde{\boldsymbol{\mu}}_9$	α	β	0	$\tilde{\mu}_{9,1} = \mu_1(1+\alpha)$, $\tilde{\mu}_{9,4} = \mu_4 + \beta\sigma_4$, $\tilde{\mu}_{9,i} = \mu_i$ for all other i
$\tilde{\boldsymbol{\mu}}_{10}$	0	0	$-\beta$	$\tilde{\mu}_{10,2} = \mu_2 - \beta\sigma_2$, $\tilde{\mu}_{10,i} = \mu_i$ for all other i
⋮	⋮	⋮	⋮	⋮
$\tilde{\boldsymbol{\mu}}_{26}$	$-\alpha$	β	β	$\tilde{\mu}_{26,1} = \mu_1(1-\alpha)$, $\tilde{\mu}_{26,2} = \mu_2 + \beta\sigma_2$, $\tilde{\mu}_{26,4} = \mu_4 + \beta\sigma_4$, $\tilde{\mu}_{26,i} = \mu_i$ for all other i
$\tilde{\boldsymbol{\mu}}_{27}$	α	β	β	$\tilde{\mu}_{27,1} = \mu_1(1+\alpha)$, $\tilde{\mu}_{27,2} = \mu_2 + \beta\sigma_2$, $\tilde{\mu}_{27,4} = \mu_4 + \beta\sigma_4$, $\tilde{\mu}_{27,i} = \mu_i$ for all other i

Basis model: $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_{13}]$, $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_{13}]$.
Variational Components: $\{v_1, v_2, v_3\} = \{1, 4, 2\}$.

3. PCGMM-based feature compensation employing variational model composition

In this section, as an application of the proposed variational model composition method, the parallel combined gaussian mixture model (PCGMM) based feature compensation algorithm is presented, which has been proposed in our previous study (Kim and Hansen, 2009a) to address time-varying background noise for speech recognition. In the PCGMM method, parameters of the noise-corrupted speech GMM are obtained through a model combination procedure using clean speech and noise GMMs.

The clean speech model $\{\omega_k, \boldsymbol{\mu}_{\mathbf{x},k}, \boldsymbol{\Sigma}_{\mathbf{x},k}\}$ consists of K Gaussian components and the noise model is estimated with a single Gaussian pdf $\{\boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\Sigma}_{\mathbf{n}}\}$ both in the cepstral

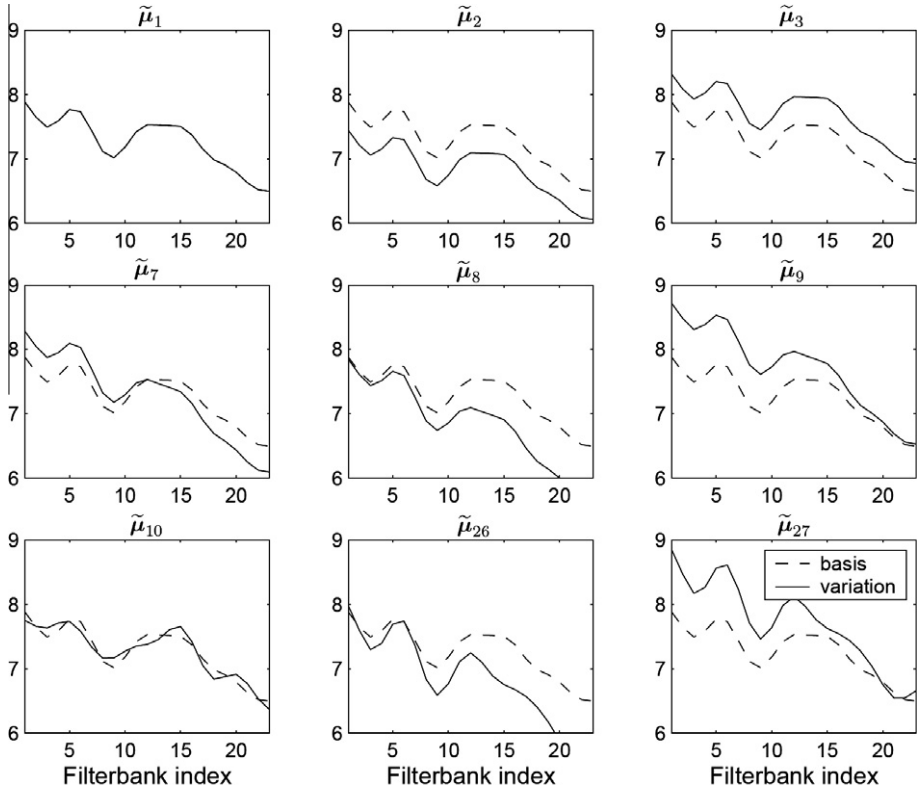

Fig. 2. Mean parameters of variational models in log-spectral domain generated by the example variational model from Table 1.

domain. Before combining the models, first, the model parameters need to be converted into the log-spectral domain using an inverse discrete cosine transform (DCT) as follows:

$$\boldsymbol{\mu}^{\{ls\}} = \mathbf{C}^{-1}\boldsymbol{\mu},$$
$$\boldsymbol{\Sigma}^{\{ls\}} = \mathbf{C}^{-1}\boldsymbol{\Sigma}(\mathbf{C}^{-1})^{T}. \tag{7}$$

Next, they need to be converted to the linear-spectral domain by Eq. (8), resulting in the parameters for a log-normal distribution:

$$\mu_i^{\{lin\}} = \exp\left(\mu_i^{\{ls\}} + \Sigma_{ii}^{\{ls\}}/2\right),$$
$$\Sigma_{ij}^{\{lin\}} = \mu_i^{\{lin\}}\mu_j^{\{lin\}}\left[\exp\left(\Sigma_{ij}^{\{ls\}}\right) - 1\right], \tag{8}$$

where i and j indicate the element index of the mean vector and covariance matrix. For the model combination of the PCGMM method, we employ "log-normal approximation" method, where it is assumed that the addition of two log-normal distributions also results in a log-normal formulation (Gales and Young, 1996; Kim and Hansen, 2009a). The mean and covariance of the noise-corrupted speech in the linear-spectral domain are obtained by

$$\boldsymbol{\mu}_{\mathbf{y},k}^{\{lin\}} = \boldsymbol{\mu}_{\mathbf{x},k}^{\{lin\}} + g\boldsymbol{\mu}_{\mathbf{n}}^{\{lin\}},$$
$$\boldsymbol{\Sigma}_{\mathbf{y},k}^{\{lin\}} = \boldsymbol{\Sigma}_{\mathbf{x},k}^{\{lin\}} + g^2\boldsymbol{\Sigma}_{\mathbf{n}}^{\{lin\}}, \tag{9}$$

where g denotes a factor for addition gain and it is set to 0.5 in this study. The obtained parameters of the noise-corrupted speech model in Eq. (9) need to be converted back to the log-spectral domain using an approximation equation (Gales and Young, 1996) as follows:

$$\mu_i^{\{ls\}} \approx \log\left(\mu_i^{\{lin\}}\right) - \frac{1}{2}\log\left(\frac{\Sigma_{ii}^{\{lin\}}}{\left(\mu_i^{\{lin\}}\right)^2} + 1\right),$$
$$\Sigma_{ij}^{\{ls\}} \approx \log\left(\frac{\Sigma_{ij}^{\{lin\}}}{\mu_i^{\{lin\}}\mu_j^{\{lin\}}} + 1\right). \tag{10}$$

Finally, the mean and covariance obtained by Eq. (10) must be returned to the cepstral domain via the DCT transform, which is the inverse process of Eq. (7). The resulting GMM of the noise-corrupted speech $\{\omega_k, \boldsymbol{\mu}_{\mathbf{y},k}, \boldsymbol{\Sigma}_{\mathbf{y},k}\}$ also consists of same K number of Gaussian components and the same weight parameter ω_k is just used as the clean speech model.

A constant bias transformation of the mean parameters of the clean speech model is assumed in the cepstral domain under an additive noisy environment, which is the assumption generally taken by other data-driven methods (Moreno, 1996; Moreno et al., 1998) as follows:

$$\boldsymbol{\mu}_{\mathbf{y},k} = \boldsymbol{\mu}_{\mathbf{x},k} + \mathbf{r}_k, \tag{11}$$

where $\boldsymbol{\mu}_{\mathbf{y},k}$ and $\boldsymbol{\mu}_{\mathbf{x},k}$ denote mean vectors of the kth component of GMMs for noise corrupted speech $\mathbf{y}$ and clean speech $\mathbf{x}$ respectively. The bias term $\mathbf{r}_k$ is estimated by Eq. (11), once the mean parameters of the clean speech model and corresponding noise-corrupted speech model

are obtained by the model combination procedure as presented through Eqs. (7)–(10).

The utilization of multiple environmental models is considered to be effective for compensating input features adaptively under time-varying noisy conditions (Kim and Hansen, 2009). In the multiple model method, a sequential posterior probability of each possible environment is estimated over the incoming noisy speech. Given the input noisy speech feature vectors $\mathbf{Y}_t = [\mathbf{y}_{t-d+1}, \mathbf{y}_{t-d+2}, \ldots, \mathbf{y}_t]^T$ over a d interval, the sequential posterior probability of a specific environment GMM G_i among all models can be written as:

$$p(G_i|\mathbf{Y}_t) = \frac{P(G_i)p(\mathbf{Y}_t|G_i)}{\sum_{e=1}^{E} P(G_e)p(\mathbf{Y}_t|G_e)}, \tag{12}$$

where $p(\mathbf{Y}_t|G_i) = \prod_{\tau=t-d}^{t} p(\mathbf{y}_\tau|G_i)$ and $P(G_i)$ is a prior probability of each environment i represented as a GMM. Based on Eq. (12), the clean feature at frame t is reconstructed by the weighted combination of the compensation terms obtained from a set of E multiple environments as follows:

$$\tilde{\mathbf{x}}_{t,MMSE} \cong \mathbf{y}_t - \sum_{e=1}^{E} p(G_e|\mathbf{Y}_t) \sum_{k=1}^{K} \mathbf{r}_{e,k} p(k|G_e, \mathbf{y}_t), \tag{13}$$

where $\mathbf{r}_{e,k}$ is a constant bias term from the kth Gaussian component of the eth environment model and $p(k|G_e, \mathbf{y}_t)$ is the posterior probability for environment G_e. Here, we use 3 frames for the interval d in our experiment.

The variational noise models obtained by the proposed variational model composition method in this study are used to generate the environmental models $\{G_e\}$, which are estimated through the model combination procedure using the clean speech GMM and the obtained variational noise models. With V number of variational components, $3^V(=E)$ environmental models are generated, and then the corresponding bias terms $\{\mathbf{r}_{e,k}\}$ are also obtained by Eq. (11). A uniform prior probability is set on all the obtained environmental models in this study, which could be modified in future scenarios based on known acoustic noise conditions (e.g., for in-vehicle applications, wind/road noise is more common than wiper-blade or horn noises). Fig. 3 shows the resulting integrated block diagram of PCGMM-based feature compensation employing the proposed variational model composition method.

4. Computational reduction by mixture sharing

As expected, as the number of variational components V increases, the number of the generated variational models exponentially increase as 3^V. The amount of computation for the PCGMM method depends primarily on the number of Gaussian components to be computed. Consequently, the computational expense increases in proportion to the number of multiple models generated by the variational model composition method. In this section, in an effort to reduce the computational complexity caused by employing the variational model composition, our previously proposed technique is introduced, where the statistically similar components among the multiple environment models are effectively shared (Kim and Hansen, 2009a).

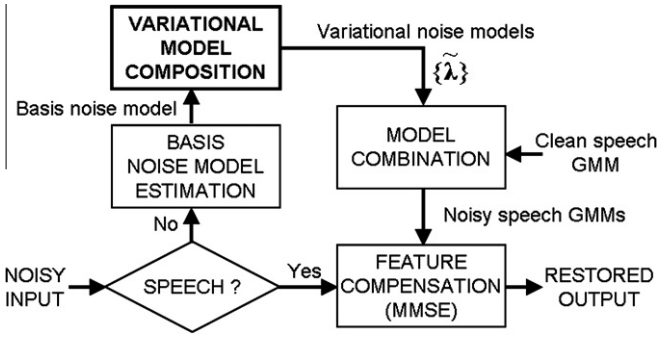

Fig. 3. Block diagram of the PCGMM method employing the proposed variational model composition method.

In the mixture sharing method, the Gaussian components which are statistically similar to each other across the different environmental models are selected and the common components for sharing are generated through a smoothing step of the similar components. Suppose there are a total of E environmental models (i.e., noise corrupted speech GMMs) $\{G_1, G_2, \ldots, G_E\}$ obtained by combining the $E(=3^V)$ variational noise models with the clean speech GMM. The procedure of selecting the similar components is presented in the following Steps 0–3, where $\mathbf{D}$ is the set of distances between the Gaussian components, and $\mathbf{C}_S$ is the set of shared Gaussian components:

- **Step 0:** $\mathbf{D} = \{d_1, d_2, \ldots, d_K\}$, $\mathbf{C}_S = \emptyset$

$$d_k = \sum_{e=2}^{E} KL\_dist(g_{1,k}, g_{e,k}), \quad 1 \leqslant k \leqslant K. \tag{14}$$

- **Step 1:** $\widehat{k} = \arg\min d_k \in \mathbf{D}$.
- **Step 2:** $\mathbf{C}_S = \mathbf{C}_S^k \cup \{\widehat{k}\}$, $\mathbf{D} = \mathbf{D} - \left\{d_{\widehat{k}}\right\}$.
- **Step 3:** if $N(\mathbf{C}_S) = K_S$, then stop, else go back to **Step 1**.

In the steps, d_k is the sum of Kullback–Leibler distances (e.g., $KL\_dist(\cdot)$) between the kth Gaussian component of each environmental model $g_{e,k}$ and the kth Gaussian component of the first environmental model $g_{1,k}$, and $N(\cdot)$ denotes the number of resulting shared elements. The first environmental model plays the pivot role in computing the distance to the Gaussians in the models. The order of the environments from 1st to Eth can be arbitrarily determined. Finally, the Gaussian search process is halted when the combined Gaussian set $\mathbf{C}_S$ reaches the desired K_S number of Gaussian components, which are now tagged as similar pdfs across the noisy speech models. The parameters of the merged Gaussian components which are shared are computed as follows:

$$\boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}} = \frac{1}{E} \sum_{e=1}^{E} \boldsymbol{\mu}_{\mathbf{y},e,k}, \quad k \in \mathbf{C}_S, \tag{15}$$

$$\boldsymbol{\Sigma}_{\mathbf{y},k}^{\{S\}} = \frac{1}{E}\sum_{e=1}^{E}\left(\boldsymbol{\Sigma}_{\mathbf{y},e,k} + \left(\boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}}\right)\left(\boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}}\right)^{T}\right), \quad k \in \mathbf{C}_S. \tag{16}$$

The likelihood functions which contain the unique Gaussian components included in set $\mathbf{C}_S$ are replaced by the merged Gaussian components:

$$p(\mathbf{y}|e,k) = \begin{cases} p\left(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}}, \boldsymbol{\Sigma}_{\mathbf{y},k}^{\{S\}}\right), & \text{if } k \in \mathbf{C}_S, \\ p\left(\mathbf{y}; \boldsymbol{\mu}_{\mathbf{y},e,k}, \boldsymbol{\Sigma}_{\mathbf{y},e,k}\right), & \text{otherwise.} \end{cases} \tag{17}$$

The constant bias terms used for feature reconstruction in Eq. (13) are also shared if their indices are included in set $\mathbf{C}_S$:

$$\mathbf{r}_{e,k} = \begin{cases} \boldsymbol{\mu}_{\mathbf{y},k}^{\{S\}} - \boldsymbol{\mu}_{\mathbf{x},k}, & \text{if } k \in \mathbf{C}_S \\ \boldsymbol{\mu}_{\mathbf{y},e,k} - \boldsymbol{\mu}_{\mathbf{x},k}, & \text{otherwise.} \end{cases} \tag{18}$$

The computations over the $E \times K$ number of Gaussian likelihood functions can be reduced to $K_S + E(K - K_S)$, leading to a computational reduction by as much as $(E-1)K_S$ via sharing the components.

5. Experimental results

5.1. Experimental setup and baseline performance

Our evaluations of the proposed method were performed within the Aurora 2.0 evaluation framework as developed by the European Language Resources Association (ELRA) (Hirsch and Pearce, 2000). The task is connected English-language digits consisting of eleven words, with each whole word represented by a continuous-density Hidden Markov Model (HMM) with 16 states and 3 mixtures per state. The feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed for all experiments (ETSI, 2000). An analysis window of 25 ms duration is used with a 10 ms skip rate for 8-kHz speech data. The computed 23 Mel-filterbank outputs are transformed to 13 cepstrum coefficients including c0 (i.e., c0–c12). The first and second order time derivatives are also included, resulting in a final feature vector of 39 dimensions.

The HMMs of the speech recognizer were trained using a database that contains 8,440 utterances of clean speech from the Aurora 2.0 database. In order to evaluate performance under time-varying background noise conditions, speech babble condition was selected from the Aurora 2.0 test database, and a new test data set was generated by combining clean speech samples with background music which consists of prelude parts of ten Korean popular songs with varying degrees of beat and tempo. Each test set consists of 1,001 samples at five different SNRs: 0, 5, 10, 15, and 20 dB. Figs. 4 and 5 present example time waveforms and spectrograms of speech babble and background music which are used as additive noise signals in the experiments.

The performance of the baseline system with no compensation was examined with comparison to several existing preprocessing algorithms in terms of speech recognition performance. The framework throughout this study is a clean condition trained HMM, so we focus only on speech/feature enhancement methods for the performance comparison, and do not consider acoustic model (i.e., HMM) adaptation methods (e.g., MAP, MLLR, PMC, etc.). spectral subtraction (SS) (Boll, 1979; Martin, 1994) combined with cepstral mean normalization (CMN) was selected as one of the conventional algorithms. This represents one of the most commonly used techniques for additive noise suppression and removal of channel distortion respectively. We also evaluated a feature compensation method, Vector Taylor Series (VTS) for performance comparison where the noisy speech GMM is adaptively estimated using the expectation–maximization (EM) algorithm over each test utterance (Moreno et al., 1998). The advanced front-end (AFE) algorithm developed by ETSI was also evaluated as one state-of-the-art method, which contains an iterative Wiener filter and blind equalization (ETSI, 2002). Table 2 demonstrates speech recognition performance (i.e., Word Error Rate, WER) of the baseline

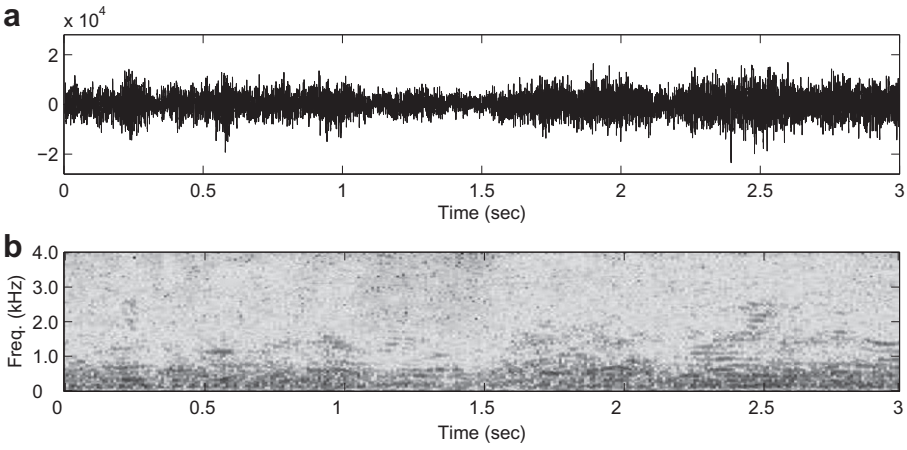

Fig. 4. A sample of speech babble: (a) time domain, and (b) spectrogram.

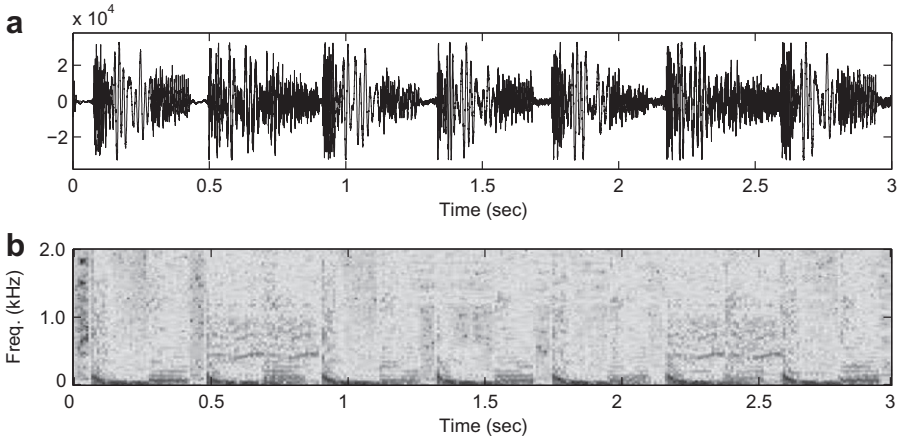

Fig. 5. A sample of background music: (a) time domain, and (b) spectrogram.

Table 2
Recognition performance of baseline system and conventional methods in speech babble and background music conditions (WER, %). Noise-free performance is 1.18% WER.

	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Speech babble						
Baseline	88.88	71.13	44.38	21.13	7.47	46.60
SS + CMN	54.90	26.24	10.97	4.63	2.48	19.84
VTS	55.83	25.51	9.49	4.05	2.57	19.49
AFE	42.17	19.41	8.13	3.99	1.87	15.11
Background music						
Baseline	74.27	51.34	28.11	12.19	4.84	34.15
SS + CMN	54.52	29.93	15.24	7.28	3.39	22.07
VTS	54.67	31.60	16.08	8.92	4.69	23.19
AFE	44.43	25.55	11.72	6.76	2.99	18.29

system and the conventional algorithms on speech babble and background music conditions.

5.2. Determination of the perturbation factor

In this section, we discuss selection of the perturbation factor for the proposed variational model composition by assessing performance across a range of perturbation factors. The performance was evaluated using the speech recognition ability of the reconstructed speech employing the PCGMM method and the variational model composition method. First, we will observe the performance dependency on the perturbation factor p_E as shown in Fig. 6. To see the impact of only the p_E on recognition performance, a single variational component (i.e., $V = 1$) was used and the WER performance was plotted as a change of α from 0 to 0.1 for p_E over four kinds of background noise conditions[2]. Here, the WER is an average value of all SNR conditions (i.e., 0, 5, 10, 15, and 20 dB) for each background noise and the plot with the black-filled circles reflects average performance of the four kinds of noise conditions. The performance of the case with $\alpha = 0$ indicates the basic PCGMM method employing only a basis model without the variational model composition method, which is a target system for performance comparison of the proposed VMC–PCGMM. It is interesting to note that each plot shows a concave shape, providing a local minimum WER in the range of 0.05 to 0.07 for α values. These results suggest that a suitable value for α needs to be determined to achieve effective performance in the proposed variational noise model composition method. We believe that a properly determined α will be effective in generating a noise model with an energy level matched to the actual background noise corrupting the input speech, since α is applied to the first variational component v_1 which corresponds to the 0th indexed cepstral coefficient. Based on the average performance plot, 0.06 was selected for α of the perturbation factor p_E in all following experiments.

Next, we discuss the determination of the perturbation factor p_S which is applied to the remaining variational components $\{v_2, v_3, \ldots, v_V\}$. Here, α is fixed as 0.06 and β for p_S is varied from 0 to 1.0 as shown in Fig. 7. The performance dependency of β is clearly seen as a change in the number of variational components $V \in (2, 3, 4, 5)$. The case where $\beta = 0$ indicates the identical case of the VMC with $V = 1$ and $\alpha = 0.06$. Also, the overall WER value in Fig. 7 reflects the average of the cases of speech babble and background music for all SNR cases, which shows how effective the scheme is when employing the perturbation factor p_S. It can be seen that the plot for each V forms a roughly concave shape, providing a local minimum WER at a certain point. This indicates that the suitable determination of β for higher order variational components (i.e., $v_2, v_3, \ldots$) is effective at increasing performance compared to VMC alone with $V = 1$. It also can be seen that the plot with the higher value of V produces a lower WER response, which suggests that larger number of variational components is useful for increasing recognition performance when employing VMC. We obtained a 0.34% improvement in WER for $V = 5$ with $\beta = 0.7$ compared to the VMC with $V = 1$ (i.e., from 13.04% to 12.70%). From these results, it is

[2] The test data for car and subway noise was obtained from Aurora2.0 (Hirsch and Pearce, 2000).

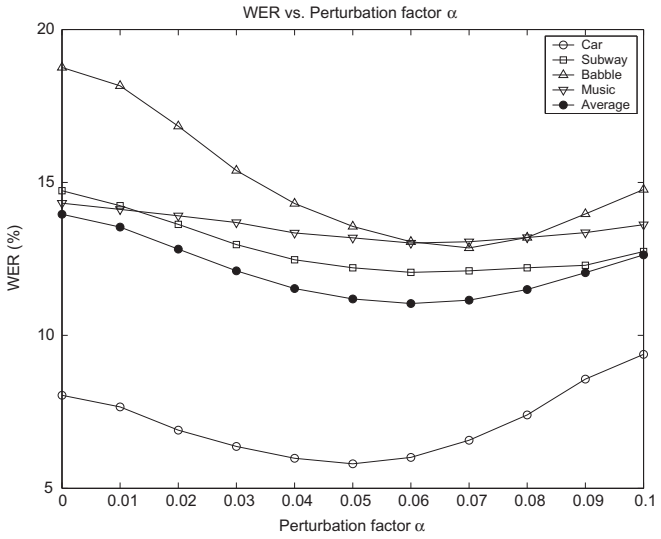

Fig. 6. Recognition performance as change of α for perturbation factor p_E (WER, %).

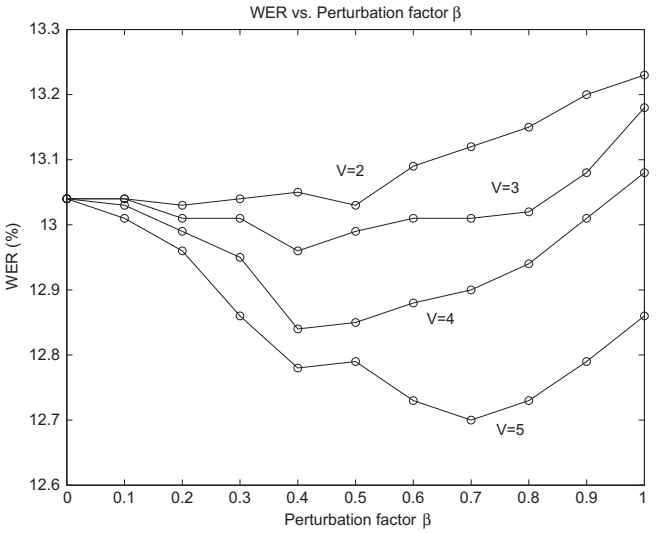

Fig. 7. Recognition performance as change of β for perturbation factor p_S (WER, %).

suggested that a suitable value for β and increasing the number of V are both effective at estimating a noise model with a more precise variation of the spectral envelope that reflects the background noise signal.

Next, the effectiveness of the proposed scheme for determining the variational components is presented in Fig. 8. The plot with the solid circles is obtained with the proposed method, where the variational components v_2 to v_6 are determined by the order of the variance size as shown in Eq. (4). For the plot with the empty circles, the v_2 to v_6 were selected randomly. We conducted 10 independent trials of the experiments for random selection. Each circle indicates the average WER of the 10-time trials with the standard deviation of the obtained WERs using the small bars. Both plots were obtained with $\alpha = 0.06$ and $\beta = 0.4$. It seems that there is no significant difference in performance between the proposed determination method for the variational components and the random selections in the plots of Fig. 8. However, it needs to be considered that the performance of the random selection fluctuates as presented by the standard deviation depending on selection of the perturbation factors. It should be noted that the performance of the proposed method always appears within a range of the standard deviations from the averaged performance of the random selection (except $V = 2$ for the speech babble case). These results suggest that the proposed selection scheme for the perturbation factor based on the size-ordered rank of the variance is an effective way to provide effective performance as a change of the number of variational components.

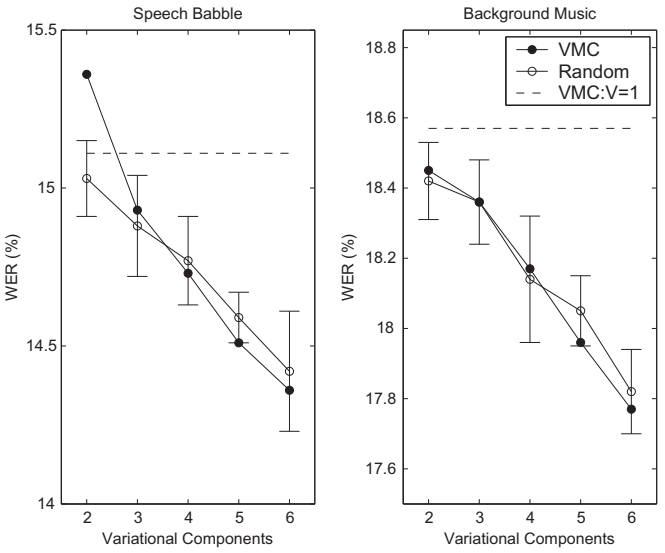

Fig. 8. Performance comparison to random selections for variational components (WER, %).

5.3. Performance evaluation of the PCGMM employing the variational model composition

Table 3 shows performance of the PCGMM method employing the proposed Variational Model Composition (VMC) for speech babble and background music conditions. Performance is compared to the basic PCGMM method (i.e., a single noise model based approach) in terms of relative improvement in WER. Here, we estimated the noise model as a Gaussian pdf from the silence (i.e., non-speech) duration at the beginning and end parts of each utterance which consists of a total of 24 frames. The estimated noise model is used as a target noise model for the basic PCGMM, and as the basis noise model for the proposed VMC–PCGMM method. Based on the performance analysis from Section 5.2, $\alpha = 0.06$ and $\beta = 0.4$ are used for the perturbation factors p_E and p_S, respectively. By considering performance and computational expenses, the number of variational components was set to 4 (i.e., $V = 4$), resulting in a collection of $81 (= 3^4)$ unique variational noise models.

Table 3
Recognition performance of the proposed VMC–PCGMM method in speech babble and background music conditions (WER, %).

	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Speech babble						
PCGMM	58.43	22.70	7.26	3.36	2.06	18.76
VMC–PCGMM	**39.87**	**14.63**	**5.20**	**2.90**	**1.84**	**12.89**
(Relative improvement)	**(+31.76)**	**(+35.55)**	**(+28.37)**	**(+13.69)**	**(+10.68)**	**(+31.31)**
Background music						
PCGMM	37.03	19.96	8.67	4.01	1.94	14.32
VMC–PCGMM	**32.27**	**18.17**	**7.84**	**3.76**	**1.94**	**12.80**
(Relative improvement)	**(+12.85)**	**(+8.97)**	**(+9.57)**	**(+6.23)**	**(+0.00)**	**(+10.65)**

From the results in Table 3, there are considerable relative improvements in WER by employing the proposed variational model composition method. We obtained +31.31% and +10.65% average relative improvements in WER compared to the basic PCGMM method for babble and music noise conditions. This suggests that the proposed VMC method is significantly more effective in generating candidate noise models that reflect the actual background noise signal, which represents the unseen noise within the speech utterance that are not effectively estimated with the conventional single noise model method. The performance comparison to the baseline and other conventional methods at different SNR conditions is also included in Fig. 9. We can see that the proposed VMC–PCGMM method consistently outperforms all existing methods including the AFE and VTS methods for all SNR conditions with speech babble and background music interfering signals.

To prove the effectiveness of the proposed method in relatively slowly changing noise conditions (compared to speech babble and music), we also evaluated the performance on car and subway noise conditions, which were obtained from Aurora 2.0 (Hirsch and Pearce, 2000). The performance comparison is presented in Fig. 10. We can see that the proposed VMC–PCGMM method again shows significantly improved performance for both car and subway conditions compared to the basic PCGMM and other conventional methods. Our analysis of the results suggests that the precise estimate of the energy level of the background noise (i.e., impact of the perturbation factor p_E) has a strong impact to the performance improvement for car and subway background noise cases, where the spectral patterns are not significantly different from the silence and speech duration segments, compared with the time-varying babble and music conditions. The performance evaluation for the four background noise types is summarized in Tables 4 and 5, and Fig. 11 with averaged WER across all. It is noted that the proposed VMC–PCGMM method outperforms all other conventional methods for all types and all SNRs of background noise conditions. The proposed VMC was particularly effective at increasing recognition performance in adverse SNR conditions as low as 0 and 5 dB, with a +24.60% and +20.40% relative improvement respectively compared to the previous basic PCGMM method.

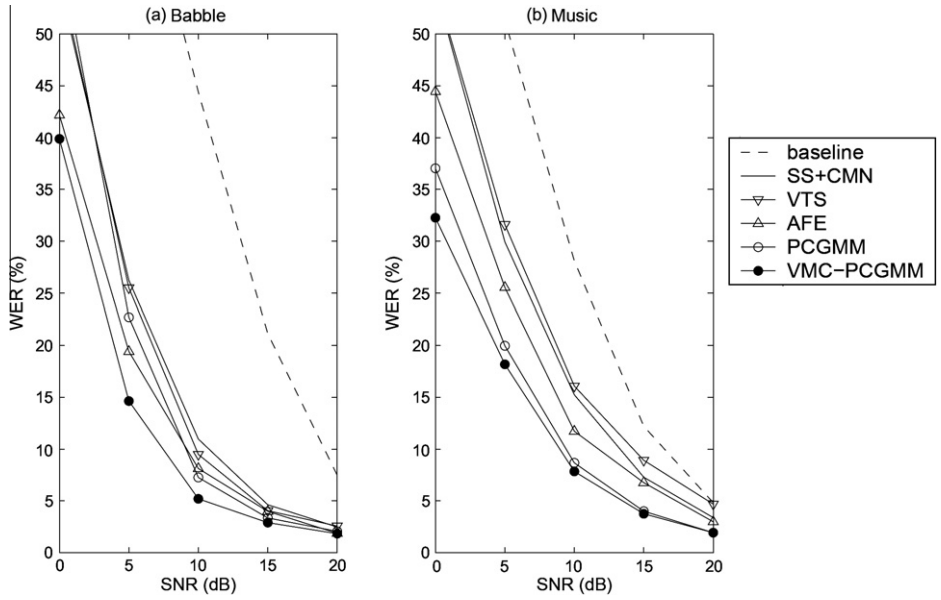

Fig. 9. Performance comparison at different SNR conditions: (a) speech babble and (b) background music conditions (WER, %).

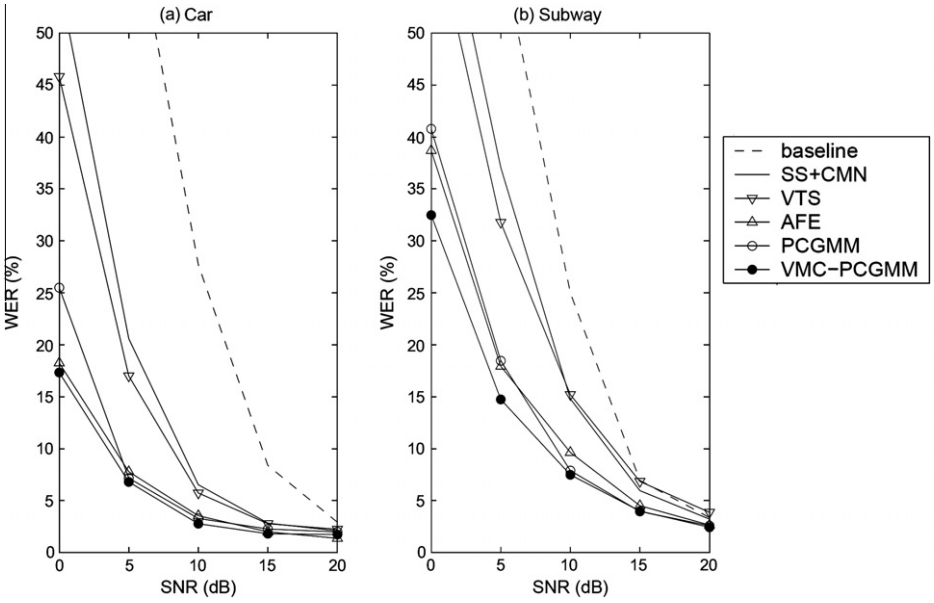

Fig. 10. Performance comparison at different SNR conditions: (a) car and (b) subway noise conditions (WER, %).

Table 4
Performance comparison in four types of background noise conditions as average over all SNRs; 0, 5, 10, 15, and 20 dB (WER, %).

	Car	Subway	Babble	Music	Average
Baseline	38.20	35.26	46.60	34.15	38.55
SS + CMN	17.41	25.93	19.84	22.07	21.31
VTS	14.71	23.89	19.49	23.19	20.32
AFE	6.59	14.68	15.11	18.29	13.67
PCGMM	8.04	14.73	18.76	14.32	13.96
VMC–PCGMM	**6.09**	**12.23**	**12.89**	**12.80**	**11.00**
(Relative improvement)	**(+24.27)**	**(+17.00)**	**(+31.31)**	**(+10.65)**	**(+21.23)**

5.4. VMC–PCGMM method with the mixture sharing technique

Table 6 presents performance of the VMC–PCGMM method employing the mixture sharing technique previously described in Section 4. The number XX associated with VMC–PCGMM–SXX indicates the number of shared Gaussian components K_S from Section 4. All WERs in Table 6 are averaged values across the 4 types of background noise and across all 5 SNR cases (i.e., average WER over 20 kinds of background noise conditions).

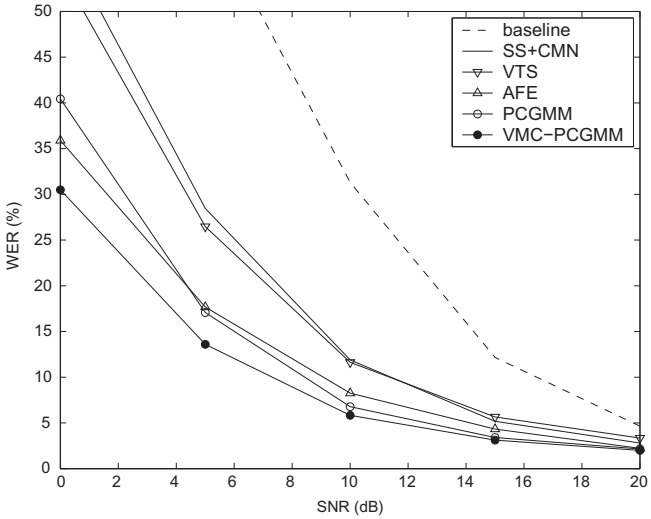

Fig. 11. WER (%) performance comparison at different SNR conditions averaged over four types of background noise conditions.

The results show that mixture sharing is useful for reducing the computational complexity, while maintaining original performance at reasonable levels. In order to investigate the relationship between performance and computational

Table 5
Performance comparison in all SNRs conditions as average over four types of background noise conditions (WER, %).

	0 dB	5 dB	10 dB	15 dB	20 dB	Average
Baseline	83.49	61.16	31.30	12.16	4.65	38.55
SS + CMN	58.25	28.45	11.89	5.18	2.80	21.31
VTS	54.51	26.47	11.63	5.65	3.34	20.32
AFE	35.89	17.67	8.26	4.32	2.22	13.67
PCGMM	40.43	17.07	6.78	3.40	2.15	13.96
VMC–PCGMM	**30.49**	**13.19**	**5.83**	**3.12**	**1.99**	**11.00**
(Relative improvement)	**(+24.60)**	**(+20.40)**	**(+14.02)**	**(+8.31)**	**(+7.46)**	**(+21.23)**

Table 6
Performance of the VMC–PCGMM employing mixture sharing method, with averaged performance over all noise conditions.

	Recognition performance (%)			Computational complexity	
	WER	Relative improvement	Differ.	# of Gaussians	Reduct. (%)
PCGMM	13.96	–	–	128	–
VMC–PCGMM	11.00	+21.23	–	10,368	–
VMC–PCGMM-S16	11.01	+21.17	−0.06	9088	12.35
VMC–PCGMM-S32	11.07	+20.75	−0.48	7808	24.69
VMC–PCGMM-S64	11.60	+16.91	−4.32	5248	49.38
VMC–PCGMM-S96	12.25	+12.30	−8.93	2688	74.07
VMC–PCGMM-S128	13.62	+2.47	−18.75	128	98.77

expense brought by mixture sharing, the relative WER versus the number of Gaussian components to be computed are also presented in Table 6. The column "Differ." under "Recognition performance" indicates the performance difference in terms of relative WER compared to the non-sharing case which is the original VMC–PCGMM. The values in the column "# of Gaussians" are the number of Gaussian components to be computed for VMC–PCGMM processing. In the non-sharing case (VMC–PCGMM), the calculation of the Gaussian probability terms requires 10,368 ($=128 \times 3^4$) components which are obtained by employing a 128-component GMM for the clean speech model and $V = 4$ for the variational components. The values in the column "Reduct." reflects the reduction percent in the number of Gaussian components for computation compared to the total number of components. In the case of VMC–PCGMM-S16 and VMC–PCGMM-S32, a 12.35% and 24.69% reduction in the computation complexity is obtained with only a 0.06% and 0.48% loss in relative WER compared to the non-sharing case. When 64 components were shared for VMC–PCGMM-S64, a 49.38% computational reduction was achieved only with a 4.32% decrease in relative improvement of WER. As discussed in Section 4, performance should degrade when the number of shared components increases. However, the experimental results here demonstrate that a reasonable selection of the number of shared components will result in a significant reduction in computational complexity with only a slight change in overall WER. The mixture sharing technique will be useful in applying the proposed VMC–PCGMM method to small footprint size mobile devices with limited storage and computational resources.

5.5. Real-life condition: CU-Move corpus

The proposed VMC–PCGMM method was also evaluated on a real-life in-vehicle speech condition obtained from the CU-Move corpus (Hansen et al., 2004). The CU-Move project was designed to develop reliable hands-free car navigation systems employing a mixed-initiative dialog. This requires robust speech recognition across changing acoustic conditions. The CU-Move database consists of five parts: (i) route navigation commands or requests, (ii) digit strings of telephone and credit card numbers, (iii) street names and addresses including spelling, (iv) phonetically-balanced sentences, and (v) Wizard of Oz interactive navigation conversations. A total of 500 speakers, balanced across gender and age, produced over 600 GB of data during a 6-month collection effort across the United States. The database and noise conditions are discussed in detail in (Hansen et al., 2004). For the evaluation in this study, we selected 949 utterances (length of 1 h and 40 min) spoken by 20 different speakers (9 males and 11 females), which were collected in Minneapolis, MN. The test samples represent an average 8.48 dB[3] SNR calculated by the NIST STNR Speech Quality Assurance software (http://www.nist.gov/speech).

Table 7
Recognition performance comparison for the CU-Move corpus (WER, %).

Baseline	70.02
SS + CMN	39.90
VTS	48.31
AFE	31.45
PCGMM	30.53
VMC–PCGMM	**24.26**
(Relative improvement)	**(+20.54)**

Tables 7 and 8 show the performance evaluation of the proposed VMC–PCGMM method on the CU-Move corpus. Table 7 demonstrates the VMC–PCGMM method has a significant improvement compared to the basic PCGMM and other conventional methods for the real-life in-vehicle condition as well. The results confirm that WER improvement is also realized on real data as well as artificially generated background noise conditions discussed previously. Table 8 shows the performance of VMC–PCGMM employing the mixture sharing method for the CU-Move corpus. The evaluation results are very similar to the results shown in Table 6, showing a 49.38% computational reduction with only a 2.85% loss in relative improvement for the case of the VMC–PCGMM-S64. The results here prove that the proposed VMC–PCGMM

[3] 0 dB and 5 dB SNR test samples of the car noise condition of Aurora2.0 show 7.15 dB and 11.66 dB average SNRs, respectively using the NIST tool.

Table 8
Performance of the VMC–PCGMM employing the mixture sharing method for the CU-Move corpus.

	Recognition performance (%)			Computational complexity	
	WER	Relative improvement	Differ.	# of Gaussians	Reduct. (%)
PCGMM	30.53	–	–	128	–
VMC–PCGMM	24.26	+20.54	–	10,368	–
VMC–PCGMM-S16	24.29	+20.44	−0.10	9088	12.35
VMC–PCGMM-S32	24.30	+20.41	−0.13	7808	24.69
VMC–PCGMM-S64	25.13	+17.69	−2.85	5248	49.38
VMC–PCGMM-S96	26.62	+12.81	−7.73	2688	74.07
VMC–PCGMM-S128	29.85	+2.23	−18.31	128	98.77

method is highly applicable to real-life in-vehicle conditions for increasing speech recognition performance.

6. Conclusion

In this study, a novel model composition method was proposed to improve speech recognition performance in time-varying background noise conditions. In the proposed method, a basis noise model was estimated from silent duration segments, followed by the creation of variational noise models which were generated by selectively applying perturbation factors to the mean parameters of the basis model. The proposed VMC method was applied to the multiple-model based PCGMM algorithm and a mixture sharing technique was also integrated to reduce overall computational expenses. Suitable values for the perturbation factors were determined through a series of pilot experiments to maximize performance of speech recognition in various types of changing background noise conditions. The procedure for determining the variational components was conducted in a size-ordered rank of the variance values and shown to be effective versus a comparison to a randomly selected set of trials.

The performance evaluation was demonstrated within the Aurora 2.0 framework using four types of background noise as well as the CU-Move in-vehicle corpus. Experimental results demonstrated that the proposed method is considerably effective at increasing speech recognition performance in unknown time-varying background noise conditions. By employing the mixture sharing method, considerable computational reduction was also achieved with only a slight loss in recognition performance. We obtained +31.31%, +10.65% and +20.54% average relative improvements in WER for speech babble, background music, and real-life in-vehicle conditions respectively, compared to the previous basic PCGMM method. This proves that the variational noise model composition generates a noise space that can effectively address the time-varying nature of the background noise.

The proposed method can be employed for a range of applications in the speech processing area, where reliable noise estimation is required in time-varying background noise environments. It is possible that the amount of noise samples are not sufficient to obtain a multiple number of model parameters through training over the samples. This is expected to have a complex spectral configuration with many variations in both time and frequency domains. It is possible however, to obtain a simple noise model from the available samples, and then a number of variational noise models can be generated by the proposed algorithm to produce candidates for the expected spectral patterns of the noise signals.

Acknowledgements

This work was supported by the USAF under a subcontract to RADC, Inc., Contract FA8750-09-C-0067 (Approved for public release. Distribution unlimited). A preliminary study of this work was presented at the Interspeech-2009, Brighton, UK, September 2009 (Kim and Hansen, 2009c).

References

Angkititrakul, P., Petracca, M., Sathyanarayana, A., Hansen, J.H.L., 2007. UTDrive: driver behavior and speech interactive systems for in-vehicle environments. In: IEEE Intelligent Vehicle Symposium, pp. 566–569.

Angkititrakul, P., Hansen, J.H.L., Choi, S., Creek, T., Hayes, J., Kim, J., Kwak, D., Noecker, L.T., Phan, A., 2009. UTDrive: the smart vehicle project. In: In-Vehicle Corpus and Signal Processing for Driver Behavior. Springer (Chapter 5).

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Cook, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34 (3), 267–285.

Deller Jr., J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals. IEEE Press.

Ephraim, Y., Malah, D., 1984. Speech enhancement using minimum mean square error short time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

ETSI Standard Document, 2000. ETSI ES 201 108 v1.1.2 (2000-04).

ETSI Standard Document, 2002. ETSI ES 202 050 v1.1.1 (2002-10).

Frey, J., Deng, L., Acero, A., Kristjansson, T.T., 2001. ALGONQUIN: iterating Laplace's method to remove multiple types of acoustic distortions for robust speech recognition. In: Eurospeech-2001.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4 (5), 352–359.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Hansen, J.H.L., Clements, M., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39 (4), 795–805.

Hansen, J.H.L., 1994. Morphological constrained enhancement with adaptive cepstral compensation (MCE–ACC) for speech recognition in noise and lombard effect. IEEE Trans. Speech Audio Process. 2 (4), 598–614.

Hansen, J.H.L., Zhang, X., Akbacak, M., Yapanel, U., Pellom, B., Ward, W., Angkititrakul, P., 2004. CU-Move: advances for in-vehicle speech systems for route navigation. In: DSP for In-Vehicle and Mobile Systems. Springer (Chapter 2).

Hansen, J.H.L., Huang, R., Chou, B., Beadle, M., Deller Jr., J.R., Gurijala, A.R., Kurimo, M., Angkititrakul, P., 2005. SpeechFind: advances in spoken document retrieval for a national gallery of the spoken word. IEEE Trans. Speech Audio Process. 13 (5), 712–730.

Hirsch, H.G., Ehrlicher, C., 1995. Noise estimation technique for robust speech recognition. In: ICASSP-95, pp. 153–156.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: ISCA ITRW ASR2000.

Kim, N., Kim, D., Kim, S., 1997. Application of sequential estimation to time-varying environment compensation. In: IEEE ASRU-1997, pp. 389–395.

Kim, N.S., 2002. Feature domain compensation of nonstationary noise for robust speech recognition. Speech Commun. 37, 231–248.

Kim, W., Stern, R.M., 2006. Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise. In: ICASSP-2006, pp. 305–308.

Kim, W., Hansen, J.H.L., 2007. SpeechFind for CDP: advances in spoken document retrieval for the US collaborative digitization program. In: IEEE ASRU-2007, pp. 687–692.

Kim, W., Hansen, J.H.L., 2009a. Feature compensation in the cepstral domain employing model combination. Speech Commun. 51 (2), 83–96.

Kim, W., Hansen, J.H.L., 2009b. Time-frequency correlation based missing-feature reconstruction for robust speech recognition in band-restricted conditions. IEEE Trans. Audio Speech Lang. Process. 17 (7), 1292–1304.

Kim, W., Hansen, J.H.L., 2009c. Variational model composition for robust speech recognition with time-varying background noise. In: Interspeech-2009, pp. 2399–2402.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput. Speech Lang. 9, 171–185.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: EUSIPCO-94, pp. 1182–1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. Thesis. Carnegie Mellon University.

Moreno, P.J., Raj, B., Stern, R.M., 1998. Data-driven environmental compensation for speech recognition: a unified approach. Speech Commun. 24 (4), 267–285.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Commun. 43 (4), 275–296.

Sasou, A., Tanaka, T., Nakamura, S., Asano, F., 2004. HMM-based feature compensation methods: an evaluation using the Aurora2. In: ICSLP-2004, pp. 121–124.

Stouten, V., Van hamme, H., Wambacq, P., 2004. Joint removal of additive and convolutional noise with model-based feature enhancement. In: ICASSP-2004, pp. 949–952.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: ICASSP-90, pp. 845–848.

Yao, K., Nakamyra, S., 2001. Sequential noise compensation by sequential Monte Carlo method. Adv. Neural Info. Process. Syst., 14.