Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080621

A VTS-based Feature Compensation Method using Noisy Speech HMMs

Yongjoo Chung[*]

Department of Electronics, Keimyung University, 704701 Daegu, South Korea

Received: 22 Oct. 2013, Revised: 20 Jan. 2014, Accepted: 21 Jan. 2014
Published online: 1 Nov. 2014

Abstract: In conventional Vector Taylor Series (VTS) based noisy speech recognition methods, Hidden Markov Models (HMMs) are trained using clean speech, and the parameters of the clean speech HMM are adapted to test noisy speech, or the original clean speech is estimated from the test noisy speech. However, these approaches have a drawback in that acoustic models trained using noisy speech cannot be used in recognition. In noisy speech recognition, improved performance is generally expected by employing noisy acoustic models produced by methods such as Multi-condition Training (MTR) and Multi-Model-based Speech Recognition framework (MMSR). Motivated by this idea, a method has been developed that can make use of the noisy acoustic models in the VTS algorithm where additive noise was adapted for the speech feature compensation. In this paper, we modified the previous method to adapt channel noise as well as additive noise. A mathematical relation was derived in the log-spectrum domain between the test and training noisy speech considering both channel and additive noise. After approximating the relation using VTS, Minimum Mean Square Error (MMSE) estimation of the training noisy speech is obtained from the test noisy speech based on the relation. The proposed method was applied to noisy speech HMMs trained by MTR and MMSR and could reduce the relative word error rate by 7% and 8%, respectively, in the noisy speech recognition experiments on the Aurora 2 database.

Keywords: noisy speech recognition, MTR, MMSR, VTS, MMSE.

1 Introduction

Despite many technical advances, accurate speech recognition in noisy environments still remains a difficult problem. The techniques cannot fully overcome the performance degradation caused by channel and additive noise. Broadly categorized, there are two different approaches used to improve the performance in noisy speech recognition. In one of the approaches, test noisy speech or trained acoustic models are compensated to reduce the mismatch between them [1,2,3,4,5]. In particular, compensation based on VTS has been known to perform quite well in noisy conditions [6,7,8,9].

In another approach, noisy speech was directly used to produce noisy speech HMMs during training [10,11,12]. MTR [13] and MMSR [14,15] are representatives of this approach. The environment-dependent HMMs make it possible to cope with test noisy speech without any compensation algorithm. In the MTR method, noisy speech signals under various noise conditions are collected and used for training the HMM. MMSR was recently proposed to improve the sharpness of probability density functions in acoustic models of MTR, and successful results using MMSR were demonstrated [14,15,16]. In contrast to MTR, where a single HMM set is constructed, multiple HMM sets corresponding to various noise types and signal-to-noise ratio (SNR) values are produced during training, and a single HMM set which is closest to test noisy speech among multiple HMM sets is selected for recognition.

Although the noisy speech HMM performs rather well by itself, its performance would be improved further by applying compensation. In a previous study, a novel mathematical relation between test and training noisy speech was derived in the log-spectrum domain [16]. After approximating the relation using VTS, the performance of the noisy speech HMM could be improved by compensating the feature vectors of the test noisy speech. The MMSE estimation of training noisy speech (not clean speech) conditioned on the test noisy speech was used for recognition, which could further reduce the mismatch between the test noisy speech and the acoustic models of the noisy speech HMM. However,

[*] Corresponding author e-mail: yjjung@kmu.ac.kr

in the previous study, the channel noise was not considered in the compensation, which probably had a negative effect on improving the performance on Set C of the Aurora 2 database. In this study, the previous algorithm was modified to compensate the test noisy speech considering both the channel and additive noise. The detailed mathematical formulation is derived, and MTR as well as MMSR are used for producing the noisy speech HMM.

This paper is organized as follows. A review on MTR and MMSR is presented in Section 2, and compensation of the test noisy speech based on the noisy speech HMM is described in Section 3. The experimental procedure and results are presented and discussed in Section 4. Finally, conclusions are given in Section 5.

2 Training of Noisy Speech HMMs

In this study, both MTR and MMSR are used to produce the noisy speech HMM. Although MMSR is known to have some advantages over the MTR method [14,15], it is rather controversial regarding which method is better in performance for noisy speech recognition. In this regard, both methods will be used to find the more appropriate one in the proposed feature-compensation method.

In Figure 1, a schematic diagram of MTR and MMSR for training noisy speech HMMs is shown. In MTR, a collection of noisy speech signals with various noise types (Subway, Babble, Car, Exhibition) and SNR values (0, 5, 10, 15, 20 and dB) is used to construct a single set of noisy speech HMM. In MMSR, multiple HMM sets are constructed, and each of them corresponds to a different noise type (Subway, Babble, Car, Exhibition) and SNR value (0-30 dB in 2 dB intervals). A single HMM set which is closest to the test noisy speech is selected for recognition based on the estimated SNR value and noise type of the test speech. Since MTR method combines a number of noise conditions to train a single HMM set, it tends to reduce the phonetic sharpness of the acoustic models in their probability density functions of the HMM. MMSR method can overcome the weakness of MTR by choosing a specific single HMM set which is most appropriate to the test noisy speech. However, the errors in selecting the closest HMM set will incur misrecognition, causing performance degradation in the MMSR.

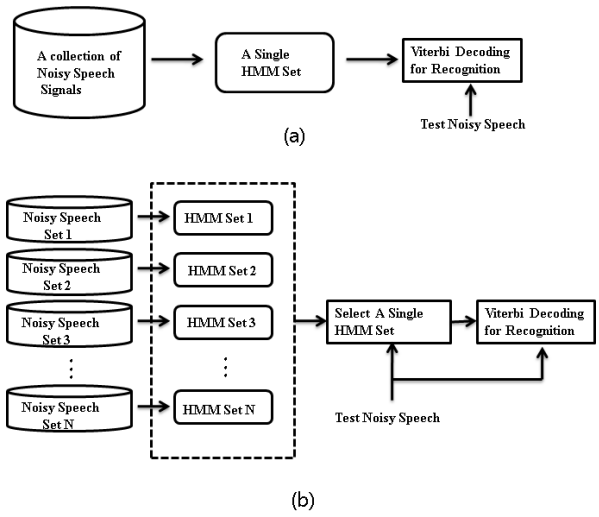

Fig. 1: A Schematic diagram of training noisy speech HMMs (a) MTR (b) MMSR

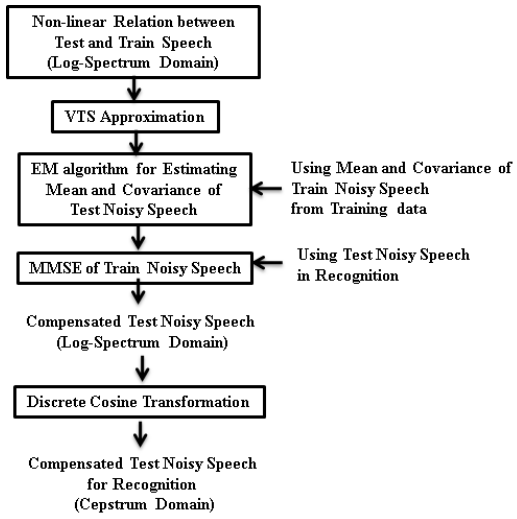

Fig. 2: Block diagram of the proposed feature compensation method

3 Feature Compensation in the Presence of Channel Noise

Unlike the pervious study [16] where only additive noise is assumed to exist, we derive, in this paper, a speech feature compensation method assuming that there is also channel noise in the test speech. The relation between training and test noisy speech is first derived in log-spectrum domain. Since the relation is non-linear, it is approximated using the VTS to obtain the mean vectors and covariance matrices of the test noisy speech given the statistics of training noisy speech obtained during the training. The statistics of the test noisy speech are used to obtain MMSE estimation of the training noisy speech, which is used as a feature vector for recognition after Discrete Cosine Transformation (DCT). The block diagram of the whole process is shown in Figure 2. A more detailed explanation of this process is given in the next subsections.

3.1 Relation Between Test and Training Noisy Speech in Log-Spectrum Domain

Assuming that channel noise is present in the test noisy speech, log-spectrum vector $\mathbf{x}$ of the clean speech and y of the noisy speech are related as follows:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})) \tag{1}$$

where $\mathbf{n}$ and $\mathbf{h}$ are the log-spectrum vector of additive and convolution noise, respectively, and $\mathbf{i}$ is a unity vector. Based on Equation (1), the log-spectrum vector $\mathbf{y}$ of the test noisy speech and $\mathbf{y}_{Tr}$ of the training noisy speech can be expressed as follows, assuming that there is no channel noise in the training noisy speech for the convenience of analysis:

$$\mathbf{y}_{Tr} = \mathbf{x} + \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr}) \tag{2}$$

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) \tag{3}$$

$$\mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr}) = \log(\mathbf{i} + \exp(\mathbf{n}_{Tr} - \mathbf{x})) \tag{4}$$

$$\mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) = \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})) \tag{5}$$

where $\mathbf{n}$ and $\mathbf{n}_{Tr}$ represent the additive noise contained in the test and training noisy speech, respectively. $\mathbf{n}_{Tr}$ is determined during training, and $\mathbf{n}$ is estimated using test noisy speech in recognition.

By combining Equations (2) and (3), the test noisy speech y can be expressed in terms of the training noisy speech $\mathbf{y}_{Tr}$ as follows:

$$\mathbf{y} = \mathbf{y}_{Tr} + \mathbf{h} + \mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr}) \tag{6}$$

$$\begin{aligned}[\mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr})]_i &= \log\left(\frac{[\mathbf{i} + \exp(\mathbf{n} - \mathbf{x} - \mathbf{h})]_i}{[\mathbf{i} + \exp(\mathbf{n}_{Tr} - \mathbf{x})]_i}\right) \\ &= \log\left(\frac{[\exp(\mathbf{x}) + \exp(\mathbf{n} - \mathbf{h})]_i}{[\exp(\mathbf{x}) + \exp(\mathbf{n}_{Tr})]_i}\right)\end{aligned} \tag{7}$$

Here, $[\cdot]_i$ represents the i-th element of a vector.

From Equations (2) and (4),

$$\mathbf{y}_{Tr} = \mathbf{x} + \log(\mathbf{i} + \exp(\mathbf{n}_{Tr} - \mathbf{x})) \tag{8}$$

Taking the exponential of both sides of Equation (8),

$$\exp(\mathbf{x}) = \exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) \tag{9}$$

Substituting (9) into (7),

$$\begin{aligned}&[\mathbf{g}(\mathbf{x}, \mathbf{n}, \mathbf{h}) - \mathbf{g}_0(\mathbf{x}, \mathbf{n}_{Tr})]_i \\ &= \log\left(\frac{[\exp(\mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr}) + \exp(\mathbf{n} - \mathbf{h})]_i}{[\exp(\mathbf{y}_{Tr})]_i}\right) \\ &= [\log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr} - \mathbf{y}_{Tr}))]_i \\ &= [G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr})]_i\end{aligned} \tag{10}$$

If Equation (10) is substituted back into Equation (6), the relation between log-spectrum vectors of the test and training noisy speech can be obtained as follows:

$$\begin{aligned}\mathbf{y} &= \mathbf{y}_{Tr} + G(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}) = \mathbf{y}_{Tr} + \mathbf{h} \\ &+ \log(\mathbf{i} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{y}_{Tr}) - \exp(\mathbf{n}_{Tr} - \mathbf{y}_{Tr}))\end{aligned} \tag{11}$$

3.2 Estimating Statistics of Test Noisy Speech

From Equation (11), the mean and covariance of the test noisy speech $\mathbf{y}$ can be estimated. Equation (11) is expanded using a first-order VTS around the initial value $\mathbf{n}_0$ $\mathbf{h}_0$ of $\mathbf{n}$,$\mathbf{h}$ and the mean of the training noisy speech $\mu_{y_{Tr}} = E\{\mathbf{y}_{Tr}\}$ to obtain the following equation. $\mathbf{n}_{Tr}$ is assumed to be fixed.

$$\begin{aligned}\mathbf{y} = \mathbf{y}_{Tr} + \mathbf{h} &+ G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr}) \\ &+ \nabla_{\mathbf{y}_{Tr}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{y}_{Tr} - \mu_{\mathbf{y}_{Tr}}) \\ &+ \nabla_{\mathbf{n}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{n} - \mathbf{n}_0) \\ &+ \nabla_{\mathbf{h}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{h} - \mathbf{h}_0)\end{aligned} \tag{12}$$

$$\begin{aligned}&\left[\nabla_{\mathbf{y}_{Tr}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})\right]_{ii} = \\ &\frac{[\exp(\mathbf{n}_{Tr}) - \exp(\mathbf{n}_0 - \mathbf{h}_0)]_i}{[\exp(\mu_{\mathbf{y}_{Tr}}) + \exp(\mathbf{n}_0 - \mathbf{h}_0) - \exp(\mathbf{n}_{Tr})]_i}\end{aligned} \tag{13}$$

$$\begin{aligned}&[\nabla_{\mathbf{n}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})]_{ii} = \\ &\frac{[\exp(\mathbf{n}_0 - \mathbf{h}_0)]_i}{[\exp(\mu_{\mathbf{y}_{Tr}}) + \exp(\mathbf{n}_0 - \mathbf{h}_0) - \exp(\mathbf{n}_{Tr})]_i}\end{aligned} \tag{14}$$

$$\begin{aligned}[\nabla_{\mathbf{h}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})]_{ii} \; &= \\ &- [\nabla_{\mathbf{n}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})]_{ii}\end{aligned} \tag{15}$$

Here, $[\cdot]_{ii}$ represents the i-th diagonal element of a matrix. Using Equation (12), the mean μ_y and covariance σ_y of the test noisy speech $\mathbf{y}$ can be expressed from the mean $\mu_{\mathbf{y}_{Tr}}$ and covariance $\sigma_{\mathbf{y}_{Tr}}$ of the training noisy speech $\mathbf{y}_{Tr}$ as follows:

$$\begin{aligned}\mu_y = \mu_{y_{Tr}} + \mathbf{h} &+ G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr}) \\ &+ \nabla_{\mathbf{n}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{n} - \mathbf{n}_0) \\ &+ \nabla_{\mathbf{h}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})(\mathbf{h} - \mathbf{h}_0)\end{aligned} \tag{16}$$

$$\begin{aligned}\sigma_y = &\left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})\right) \sigma_{y_{Tr}} \\ &\cdot \left(\mathbf{I} + \nabla_{\mathbf{y}_{Tr}} G(\mu_{\mathbf{y}_{Tr}}, \mathbf{n}_0, \mathbf{h}_0, \mathbf{n}_{Tr})\right)^T\end{aligned} \tag{17}$$

3.3 Estimation of Noise Parameter

The training noisy speech vector $\sigma_{\mathbf{y}_{Tr}}$ is assumed to be distributed as a mixture of Gaussians with mean vectors and covariance matrices obtained through a vector quantization (VQ) process. The mixture Gaussian distribution is separately estimated for each noisy type and the SNR value using the same noisy training data to produce the multiple HMM sets in MMSR. Assuming also that the log-spectrum vector $\mathbf{y}$ of the test noisy speech is a mixture of Gaussians, the distribution of $\mathbf{y}$ as a function of unknown noise vector $\mathbf{n}, \mathbf{h}$ can be defined as follows using Equation (16) and (17),

$$p(\mathbf{y}|\mathbf{n}, \mathbf{h}) = \sum_{m=1}^{M} p_m N(\mu_{\mathbf{y},m}, \sigma_{\mathbf{y},m}) \tag{18}$$

where $N(\mu_{\mathbf{y},m}, \sigma_{\mathbf{y},m})$ is the m-th Gaussian distribution with a mean vector $\mu_{\mathbf{y},m}$ and covariance matrix $\sigma_{\mathbf{y},m}$. p_m is the mixture weight of the m-th component. Note that the mean vector $\mu_{\mathbf{y},m}$ and covariance matrix $\sigma_{\mathbf{y},m}$ are themselves fully parameterized by the noise vectors $\mathbf{n}$ and $\mathbf{h}$, which are treated just as parameters, not random variables; only the noisy speech vectors were treated as random variables.

Given a sequence of log-spectrum vectors for the test noisy speech $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots \mathbf{y}_T\}$, the log-likelihood for the sequence is defined as follows using Equation (18):

$$L(\mathbf{Y}|\mathbf{n},\mathbf{h}) = \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{n},\mathbf{h}) \tag{19}$$

An iterative Expectation Maximization (EM) algorithm is used to re-estimate the noise vector $\psi = \{\mathbf{n},\mathbf{h}\}$ by maximizing Equation (19). In the EM algorithm, an auxiliary function $Q(\psi,\bar{\psi})$ is written as follows:

$$\begin{aligned} Q(\psi,\bar{\psi}) &= E\{L(\mathbf{Y}|\bar{\psi})|\mathbf{Y},\psi\} \\ &= \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h}) \log p(\mathbf{y}_t,m|\bar{\mathbf{n}},\bar{\mathbf{h}}) \end{aligned} \tag{20}$$

The symbol ψ represents the noise vector $\mathbf{n},\mathbf{h}$, which is already known, and $\bar{\psi}$ is the unknown noise vector $\bar{\mathbf{n}},\bar{\mathbf{h}}$, which should be estimated. $Q(\psi,\bar{\psi})$ can be expanded as:

$$\begin{aligned} Q(\psi,\bar{\psi}) &= \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})[\log p_m + \frac{D}{2}\log 2\pi \\ &-\frac{D}{2}\log|\sigma_{\mathbf{y},m}| - \frac{1}{2}(\mathbf{y}_t-\bar{\mu}_{y,m})^T\sigma_{y,m}^{-1}(\mathbf{y}_t-\bar{\mu}_{y,m})] \end{aligned}$$

$$\begin{aligned} \bar{\mu}_{y,m} &= \mu_{y_{Tr},m} + \bar{\mathbf{h}} + G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr}) \\ &+\nabla_{\mathbf{n}}G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr})(\bar{\mathbf{n}}-\mathbf{n}_0) \\ &+\nabla_{\mathbf{h}}G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr})(\bar{\mathbf{h}}-\mathbf{h}_0) \end{aligned} \tag{21}$$

Next, to re-estimate $\mathbf{n},\mathbf{h}$ in Equation (21), the derivative of the auxiliary function with respect to $\bar{\mathbf{n}},\bar{\mathbf{h}}$ must be taken and set equal to 0.

$$\begin{aligned} &\nabla_{\bar{\mathbf{n}}}Q(\psi,\bar{\psi}) = \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})\cdot \\ &[\nabla_{\mathbf{n}}G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr})^T\sigma_{y,m}^{-1}(\mathbf{y}_t-\bar{\mu}_{y,m})] = 0 \end{aligned}$$

$$\begin{aligned} &\bar{\mathbf{n}} = [\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})\nabla_{\mathbf{n}}G(\cdot)^T\sigma_{y,m}^{-1}\nabla_{\mathbf{n}}G(\cdot)]^{-1}\cdot \\ &[\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})\nabla_{\mathbf{n}}G(\cdot)^T\sigma_{y,m}^{-1}(\mathbf{y}_t-(\mu_{y_{Tr},m} \\ &+\bar{\mathbf{h}}+G(\cdot)-\nabla_{\mathbf{n}}G(\cdot)\mathbf{n}_0+\nabla_{\mathbf{h}}G(\cdot)(\bar{\mathbf{h}}-\mathbf{h}_0))] \end{aligned} \tag{22}$$

$$\begin{aligned} &\nabla_{\bar{\mathbf{h}}}Q(\psi,\bar{\psi}) = \sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})\cdot \\ &[\mathbf{I}+\nabla_{\mathbf{h}}G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr})^T\sigma_{y,m}^{-1}(\mathbf{y}_t-\bar{\mu}_{y,m})] = 0 \end{aligned}$$

$$\begin{aligned} &\bar{\mathbf{h}} = [\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})\cdot \\ &(\mathbf{I}+\nabla_{\mathbf{h}}G(\cdot))^T\sigma_{y,m}^{-1}(\mathbf{I}+\nabla_{\mathbf{h}}G(\cdot))]^{-1} \\ &[\sum_{t=1}^{T}\sum_{m=1}^{M} p(m|\mathbf{y}_t,\mathbf{n},\mathbf{h})(\mathbf{I}+\nabla_{\mathbf{h}}G(\cdot)^T\sigma_{y,m}^{-1}(\mathbf{y}_t-(\mu_{y_{Tr},m} \\ &+G(\cdot)+\nabla_{\mathbf{n}}G(\cdot)(\bar{\mathbf{n}}-\mathbf{n}_0)-\nabla_{\mathbf{h}}G(\cdot)\mathbf{h}_0))] \end{aligned} \tag{23}$$

$$G(\cdot) \equiv G(\mu_{\mathbf{y}_{Tr},m},\mathbf{n}_0,\mathbf{h}_0,\mathbf{n}_{Tr}) \tag{24}$$

The noise vector $\bar{\mathbf{n}},\bar{\mathbf{h}}$ derived from Equations (22) and (23) is substituted into $\mathbf{n},\mathbf{h}$ in Equations (16) and (17) to adapt the mean and covariance of the test noisy speech. The likelihood function from Equation (19) and the auxiliary function from Equation (20) are consequently updated. This process is iterated until the log-likelihood function from Equation (19) converges. After the convergence, an MMSE estimation of the training noisy speech is performed and used for recognition.

3.4 MMSE of Training Noisy Speech

The MMSE of training speech $\mathbf{y}_{Tr}$ given the test speech $\mathbf{y}$ is expressed as follows:

$$\hat{\mathbf{y}}_{Tr,MMSE} = E(\mathbf{y}_{Tr}\ |\mathbf{y}) = \int \mathbf{y}_{Tr}p(\mathbf{y}_{Tr}|\mathbf{y})d\mathbf{y}_{Tr} \tag{25}$$

From Equation (11),

$$\mathbf{y}_{Tr} = \mathbf{y} - \mathbf{h} - G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr}) \tag{26}$$

Substituting Equation (26) into (25) and approximating $G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr})$ by a VTS of order zero around $\mu_{y_{Tr},m}$, the following relationship is obtained:

$$\begin{aligned} \hat{\mathbf{y}}_{Tr,MMSE} &= \mathbf{y}-\mathbf{h}-\int G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr})p(\mathbf{y}_{Tr}|\mathbf{y})d\mathbf{y}_{Tr} \\ &= \mathbf{y}-\mathbf{h}-\int \sum_{m=1}^{M} G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr})p(\mathbf{y}_{Tr},m|\mathbf{y})d\mathbf{y}_{Tr} \\ &= \mathbf{y}-\mathbf{h}-\sum_{m=1}^{M} p(m|\mathbf{y})\int G(\mathbf{y}_{Tr},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr})p(\mathbf{y}_{Tr}|m,\mathbf{y})d\mathbf{y}_{Tr} \\ &\cong \mathbf{y}-\mathbf{h}-\sum_{m=1}^{M} p(m|\mathbf{y})G(\mu_{Tr,m},\mathbf{n},\mathbf{h},\mathbf{n}_{Tr}) \end{aligned} \tag{27}$$

The DCT of the log-spectrum vector $\hat{\mathbf{y}}_{Tr,MMSE}$ is taken to find a 13th-order cepstrum vector. The c0 component in the cepstrum vector is replaced with log-energy. The delta and acceleration (delta-delta) coefficients of the cepstrum vector are also calculated to obtain a 39-dimentional feature vector which is used for the speech recognition experiments described in the next section.

4 Experiments and Results

To verify the effectiveness of the proposed feature compensation method, experiments were conducted on the Aurora 2 database. The MTR training data consists of clean and noisy speech signals contaminated by various kinds of noise signals (Subway, Babble, Car, Exhibition) with SNR ranges from 0 to 20 dB in 5 dB intervals. New noisy speech data was additionally generated for MMSR training, which consists of noisy speech signals with the 4 noise types (Subway, Babble, Car, Exhibition) and SNR values from 0 to 30 dB in 2 dB intervals.

Three test sets (Set A, Set B, Set C) are used for recognition experiments. They are corrupted by a range of noise types with SNRs of -5, 0, 5, 10, 15, and 20 dB. Set A is corrupted by additive noise whose types (Subway, Babble, Car, Exhibition) are known during training, while Set B is corrupted with unknown types (Restaurant, Street, Airport, Train Station) of additive noises, and Set C is corrupted by a combination of convolution and additive noise (Subway, Street).

For the feature vector, a noise-robust version of Mel-Frequency Cepstral Coefficients (MFCCs) called AFE (Advanced Front-End) was used. AFE is known to significantly reduce the word error rates in noisy speech recognition [17]. The 12th-order MFCCs with the 0th-order cepstral coefficient set aside are appended with the log-energy to form a 13th order basic feature vector along with their delta and acceleration coefficients to construct a 39th-order feature vector for each frame.

The acoustic models were trained using both the Complex Back End (CBE) and Simple Back End (SBE) scripts, which are each separately defined for the Aurora 2 database. For the SBE model, the HMM for each digit consists of 16 states with 3 Gaussian mixtures in each state. In addition, a three-state silence model with 6 Gaussian mixtures per state and a one-state pause model tied with the center state of the silence model are used. For the CBE, the number of mixtures in each state is increased to 20 and 36 for the digit and silence models, respectively. The hidden Markov model toolkit (HTK) was employed to train and test the HMM used in this study [18].

Table 1 shows the word error rates (WERs) of the proposed feature compensation method in comparison with the conventional methods for the Aurora 2 database. MTR-MMSE/MMSR-MMSE are methods in our previous study [16], where only additive noise is adapted for the speech feature compensation while channel noise is additionally compensated in the proposed MTR-MMSE+H/MMSR-MMSE+H. MTR-MMSE and MMSR-MMSE differ in the type of noisy speech HMM used for compensation.

Conventional MTR and MMSR method improve the performance of the baseline system which was trained using clean speech data. The baseline system scores 12.97% WER on average, whereas MTR and MMSR achieve WERs of 8.22 % and 8.17%, respectively.

Table 1: WERs (%) of the proposed feature compensation methods using SBE models compared to conventional methods for Aurora 2 database

Method	Set A	Set B	Set C	Ave.
Baseline	12.25	12.90	14.56	12.97
MTR	7.70	8.23	9.26	8.22
MMSR	6.78	9.56	8.17	8.17
MTR-MMSE (additive noise)	7.54	7.75	9.18	7.95
MTR-MMSE+H (additive+ channel noise)	7.61	7.52	8.45	7.74
MMSR-MMSE (additive noise)	6.71	8.98	7.92	7.86
MMSR-MMSE+H (additive+ channel noise)	6.61	8.67	7.60	7.63

Although MMSR performs slightly better than MTR, their difference is minor.

By using the feature compensation, the performance of MTR and MMSR could be improved further. As shown in Table 1, MTR-MMSE and MMSR-MMSE achieve 7.95% and 7.86% average WERs by adapting the additive noise, providing 3.3% and 3.8% relative improvement over MTR and MMSR, respectively. By additionally compensating the channel noise, we could further improve the recognition performance of the MTR-MMSE/MMSR-MMSE. As shown in the table, MTR-MMSE+H achieve 7.74% average WER further reducing the WER of MTR-MMSE. This is mainly due to the significant performance improvement in Set C. Similar results have been also found in MMSR-MMSE+H as shown in Table 1

Figure 3 shows the relative improvement (%) achieved by proposed methods over MTR and MMSR. The figure shows that the improvement is more prominent for Set B and Set C than for Set A. This is expected because the acoustic mismatch between the test noisy speech and the noisy speech HMM is greater for Set B and Set C than for Set A. For example, since the acoustic mismatch is very small for Set A in the MMSR, the smallest relative improvement of just 1% is achieved by the MMSR-MMSE.

When only additive noise is adapted, the relative improvement is more prominent in Set B than Set A and Set C. Since Set A has similar acoustic characteristic to the noisy speech HMMs, the noise adaptation could not have much impact on improving the recognition performance. Also, the channel mismatch in Set C cannot be overcome by just compensating the additive noise.

When the channel noise is additionally adapted in the MTR-MMSE+H, we can observe significant performance improvement in Set C as expected. While MTR-MMSE could achieve the relative improvement of about 1% over MTR for Set C, it increases nearly to 9% in the MTR-MMSE+H. Similar result was also observed when applying the channel adaptation to MMSR. Meanwhile, we could also see some performance improvement in Set

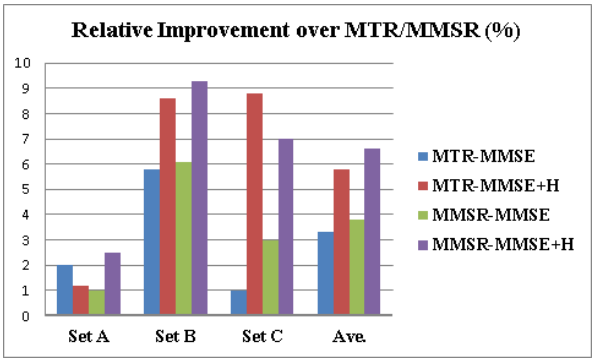

Fig. 3: Relative Improvement (%) in WERs achieved by MTR-MMSE/MTR-MMSE+H and MMSR-MMSE/MMSR-MMSE+H over MTR and MMSE, respectively, when using SBE models.

Table 2: WERs (%) of the proposed feature compensation methods using CBE models compared to conventional methods for Aurora 2 database.

Method	Set A	Set B	Set C	Ave.
Baseline	11.58	12.10	13.68	12.20
MTR	6.04	6.82	7.22	6.59
MMSR	6.17	9.0	7.97	7.66
MTR-MMSE (additive noise)	5.9	6.33	7.11	6.31
MTR-MMSE+H (additive+ channel noise)	5.92	6.23	6.37	6.13
MMSR-MMSE (additive noise)	5.86	8.17	7.72	7.16
MMSR-MMSE+H (additive+ channel noise)	5.84	8.03	7.51	7.05

B when applying the channel compensation, although it is not as significant as in Set C. The use of channel parameter **h** in the compensation algorithm seems to contribute to reduce the noise type mismatch in some degree.

The proposed methods were also applied to the noisy speech HMM trained with the CBE script to verify whether the proposed method could work well when the acoustic modeling becomes more complex. In Table 2, we can observe significant performance improvement when using the CBE script compared with the SBE script and it is more prominent in MTR than MMSR. The increased number of mixtures in each state of the HMM may have greatly contributed to sharpening the acoustic modeling in MTR. Although MMSR had comparable performance with MTR in the SBE script, MTR significantly outperforms MMSR in the CBE script. Figure 4 shows the relative improvement of the Baseline, MTR, and MMSR in the CBE script over SBE script. MTR has a large reduction in WER (19.8%) compared with the other methods (Baseline: 5.9%, MMSR: 6.2%).

As in the SBE script, the performance of MTR and MMSR with the CBE script could be further improved by

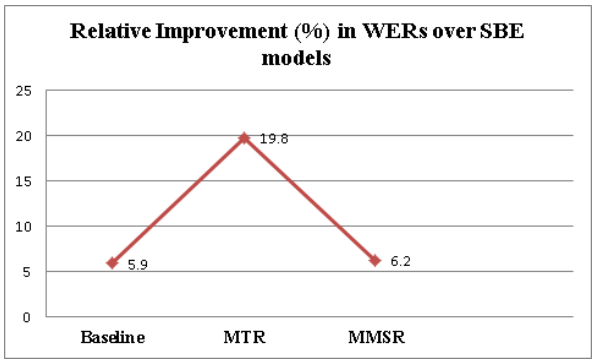

Fig. 4: Relative improvement in WERs (%) of CBE models over SBE models.

compensating the additive noise. As shown in Table 2, MTR-MMSE achieves 6.31% WER, providing 4.24% relative improvement over MTR. Similarly, MMSR-MMSE also show improved performance over MMSR. Since MTR outperforms MMSR in the CBE script, the WER of MTR-MMSE (6.31%) is shown to be much smaller than MMSR-MMSE (7.16%).

We can also obtain additional performance gain by applying the channel compensation in the CBE script. MTR-MMSE+H achieve 6.37% WER in Set C which compares to 7.11% WER by the MTR-MMSE. Similarly, MMSR-MMSE+H achieve 7.51% WER for Set C compared with 7.72% WER by the MMSR-MMSE. These improvements in Set C lead to the overall performance gain in MTR-MMSE+H (6.13% WER) and MMSR-MMSE+H (7.05% WER) compared with MTR-MMSE (6.31% WER) and MMSR-MMSE (7.16%), respectively.

Figure 5 shows the relative improvement achieved by the feature compensation methods over MTR and MMSR in the CBE script. The overall performance trend is similar to the SBE script in Figure 3. The performance improvement is more significant for Set B and Set C than for Set A. MTR-MMSE and MMSR-MMSE are shown to be very effective in improving the performance in Set B. MTR-MMSE+H and MMSR-MMSE+H could achieve significant performance improvement in Set C by adapting the channel noise in addition to the additive noise. Although the performance of MMSR-MMSE and MMSR-MMSE+H is inferior to that of MTR-MMSE and MTR-MMSE+H, we can see from the recognition results that the effect of the feature compensation is manifest irrespective of the type of the noisy speech HMM.

5 Conclusions

In this study, a VTS-based feature compensation method has been applied to noisy speech HMMs. In particular, we propose to adapt the channel noise for the compensation to improve the performance the previous method which takes into consideration only additive noise. The channel and additive noise were adapted to reduce the mismatch between the test noisy speech and the noisy speech

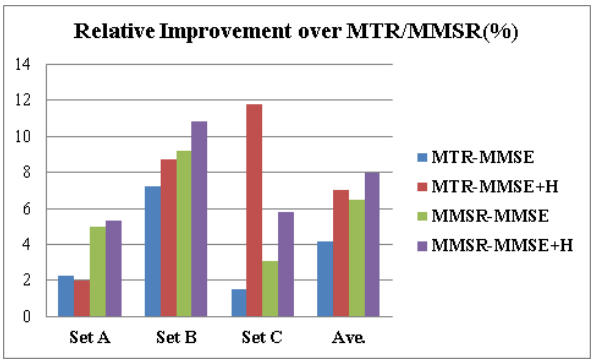

Fig. 5: Relative improvements in WERs (%) achieved by MTR-MMSE/MTR-MMSE+H and MMSR-MMSE/MMSR-MMSE+H over MTR and MMSE, respectively, when using CBE models.

HMM. The experimental results confirmed that the proposed feature compensation method is very effective in reducing the mismatch occurring in noisy speech recognition using MTR and MMSR based noisy speech HMMs. The feature compensation algorithm was applied to HMMs trained with the CBE script as well as the SBE script to test the robustness of the proposed method against varying HMM complexities and improved performance was found in both of them. The best result (6.13% average WER) was obtained in the Aurora 2 database when the feature compensation was applied to the MTR with the CBE script producing 7.0% relative improvement in average WER over conventional MTR method. When only additive noise is adapted as in the previous study, the improvement was mainly observed in Set B of the Aurora 2 database. However, by using the channel adaptation algorithm proposed in this paper, we could also observe significant performance improvement for Set C in the Aurora 2 database. The approach is distinguished from conventional VTS-based methods in the sense that noisy speech HMM instead of the clean HMM is used for the speech feature compensation and both additive and channel noise is adapted for noise-robust speech recognition.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) fubded by the Ministry of Education, Science and Technology (2012-0002615).
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] S.F Ball, Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust., Speech, Signal Process, **27**, 113-120 (1979).
[2] M.J.F Gales, Model based techniques for noise-robust speech recognition, Ph.D. Dissertation, University of Cambridge, (1996).
[3] W. Kim, J.H.L Hansen, Feature compensation in the cepstral domain employing model combination. Speech Communication, **51**, 83-96 (2009).
[4] Y. Gong, Speech recognition in noisy environments: A survey. Speech Communication, **16**, 261-291 (1995).
[5] S. Sagayama, Y Yamaguchi, S Takahashi, Jacobian adaptation of noisy speech models, Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding. Santa Barbara, California, USA, 396-403 (1997).
[6] P.J Moreno, Speech Recognition in noisy environments, Ph.D. Dissertation, Carnegie Mellon University, (1996).
[7] D.Y Kim, C.K Un, N.S Kim, Speech recognition in noisy environments using first-order vector Taylor series, Speech Communication, **24**, 39-49 (1998).
[8] A. Acero, L. Deng, T. Kristjansson and J. Zhang, HMM adaptation using vector Taylor series for noisy speech recognition, Proceedings of the International Conference on Spoken Language Processing, Beijing, China, 869-872 (2000).
[9] J. Li, L. Deng, D. Yu, Y. Gong and A. Acero, High-performance HMM adaptation with joint compensation of additive and convolutive distortion via vector Taylor series, Proceedings of ASRU, Kyoto, Japan, 65-70 (2007).
[10] M. Akbacak, J.H.L Hansen, Environmental sniffing: noise knowledge estimation for robust speech systems, Proceedings of the International Conference of Acoustics, Speech and Signal Processing, Hongkong, China, 113-116 (2003).
[11] M. Akbacak, J.H.L. Hansen, Environmental sniffing: noise knowledge estimation for robust speech systems, IEEE Trans. Audio, Speech and Language Process, **15**, 465-477 (2007).
[12] R.P. Lippmann, E.A. Martin, D.B. Paul, Multi-style training for robust isolated-word speech recognition, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Dallas, Texas, USA, 705-708 (1987).
[13] H.G Hirsch, D. Pearce, The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, Proceedings of the International Conference on Spoken Language Processing, Bejing, China, 18-20 (2000).
[14] H. Xu, Z.H Tan, P. Dalsgaard, B. Lindberg, Robust speech recognition on noise and SNR classification a multiple-model framework, Proceedings of INTERSPEECH, Lisboa, Portugal, 977-980 (2005).
[15] H. Xu, X.H Tan, P. Dalsgaard, B. Lindberg, Noise condition dependent training based on noise classification and SNR estimation, IEEE Trans. Audio, Speech, Language Process, **15**, 2431-2443 (2007).
[16] Y. Chung and J.H.L. Hansen, Compensation of SNR and noise type mismatch using an environmental sniffing based speech recognition solution, EURASIP Journal on Audio, Speech, and Music Processing, **2013**, 1-14 (2013).
[17] ETSI draft standard doc., Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm. ETSI Standard ES 202 050, (2002).

[18] S. Young, HTK: Hidden Markov Model Toolkit V3.4.1. Cambridge Univ. Eng. Dept. Speech Group, (1993).

Yongjoo Chung received the PhD degree in electrical engineering from Korea Advanced Science and Technology, He is currently a Professor with the Department of Electronics Engineering at Keimyung University, Daegu, S. Korea. He has published over 30 papers in international peer reviewed journals. His research interests are in the areas of speech recognition, multimedia signal processing and pattern recognition.