NOISE SUPPRESSION WITH UNSUPERVISED JOINT SPEAKER ADAPTATION AND NOISE MIXTURE MODEL ESTIMATION

Masakiyo Fujimoto *Shinji Watanabe*[†] *Tomohiro Nakatani*

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0237 Japan

ABSTRACT

The estimation of an accurate noise model is a crucial problem for model-based noise suppression including a vector Taylor series (VTS)-based approach. The variation of the speaker characteristics is also a crucial factor as regards the model-based noise suppression. As a result, a speaker adaptation technique plays an important role in the model-based noise suppression. To deal with former problem, we have already proposed an unsupervised estimation method for a noise mixture model. Therefore, this paper proposes a joint processing method that simultaneously achieves speaker adaptation and noise mixture model estimation. This joint processing is realized by using minimum mean squared error (MMSE) estimates of clean speech and noise. Although VTS-based approach involves non-linear transformation, the MMSE estimates make it possible to flexibly estimate accurate parameters for the joint processing without the influences of non-linear VTS transformation. In the evaluation, the proposed method provided an improvement compared with results obtained using only noise mixture model estimation.

Index Terms— noise suppression, noise mixture model, speaker adaptation, MMSE estimation

1. INTRODUCTION

Noise robustness is a crucial problem as regards the practical use of automatic speech recognition (ASR). To ensure noise robustness for ASR, various noise robust techniques have been proposed. As the front-end processing of ASR, robust feature extraction [1] and noise suppression [2]-[6] reduce the influence of interfering noise from observed noisy speech signals. On the other hand, back-end processing including model compensation [7, 8] and model adaptation [9, 10, 11] provide acoustic models that accurately represent the acoustic features of the corrupted speech signals.

In recent progress in the research field, statistical model-based front-end processing has been widely used and its effectiveness is well known. Representative techniques include the minimum mean squared error (MMSE)-based approach [3] and the vector Taylor series (VTS)-based approach [4]. Especially, the VTS-based approach has attracted attention as a powerful tool for noise robust ASR.

The VTS-based approach compensates models of observed noisy speech with models of clean speech and noise based on a non-linear mismatch function [4]. The variation of the speaker characteristics and the noise conditions are adjusted by using an EM algorithm-based parameter update with a linear approximation based on the Taylor series expansion. An accurate update of the parameter, especially an estimation of the noise model parameters, is a crucial factor in the VTS-based approach.

As mentioned in our previous work [6], although the typical VTS-based approach employs a single Gaussian distribution, a noise model with a single Gaussian distribution is unsuitable for a non-stationary noise that has a multi-peak distribution. If the noise has a multi-peak distribution, a mixture model, e.g., a Gaussian mixture model (GMM), should be used for the noise model. However, the estimation of hidden variables, namely noise mixture components, is computationally intractable in the conventional VTS-based approaches. For this problem, we have proposed an unsupervised estimation method for the noise mixture model by utilizing the MMSE estimates of noise signals [6], and showed a significant improvement in ASR accuracy in highly non-stationary noise environments.

With model-based noise suppression, although a speaker independent (SI) model is typically used for the clean speech model, the use of a speaker dependent (SD) model is a reasonable way to improve the noise suppression. The accuracy of a clean speech model seriously affects the estimation accuracy of the noise model and the other parameters. Thus, speaker adaptation of the SI clean speech model becomes an important factor for the VTS-based approach. However, speaker adaptation is also an intractable problem in the VTS-based approach, because the observed data is restricted to noisy speech signals when we utilize the VTS-based approach. Namely, we cannot use sufficient data for the speaker adaptation due to the unobservability of the clean speech signal. As a solution to this problem, Chin *et al.*, used pre-trained speaker adaptation parameter sets, and found an appropriate parameter based on the EM algorithm [8]. However, this method cannot fully represent the characteristics of the target speaker in the adapted model, because this method merely selects the adaptation parameters of a similar speaker. To overcome this problem, we utilize the MMSE estimates of clean speech signal for the speaker adaptation. The MMSE estimates of clean speech signals are obtained by using the approach described in [6].

With the proposed method, since only given utterance is available for the speaker adaptation, the amount of adaptation data becomes small. Thus, we adapted the parameters of the SI clean speech model based on the bias-based adaptation [10], i.e., $\boldsymbol{\mu}_{SD} = \boldsymbol{\mu}_{SI} + \boldsymbol{b}$, where $\boldsymbol{\mu}_{SD}$, $\boldsymbol{\mu}_{SI}$, and $\boldsymbol{b}$ denote the mean vectors of the SD and the SI clean speech model, and the bias vector, respectively[1]. With this method, we assume that the bias vector $\boldsymbol{b}$ is a common parameter for all the Gaussian components included in the SI clean speech model.

Based on the above considerations, we propose a joint processing method that simultaneously achieves speaker adaptation and the previously proposed noise mixture model estimation by utilizing MMSE estimates of the clean speech and noise. The evaluation results prove that the proposed method provides further improvements in ASR accuracy in highly non-stationary noise environments.

2. MODEL COMPENSATION BASED ON VTS

The VTS-based approach compensates the SI clean speech model for differences in speaker characteristics and noise conditions by us-

[†]Shinji Watanabe is now with Mitsubishi Electric Research Laboratories.

[1]This processing involves the adaptation of channel difference.

ing a non-linear mismatch function [4]. In this paper, we utilize an SI clean speech model with two internal states, i.e., states of silence (speech absent) and speech (speech activity). Each state is modeled in advance by a GMM with K Gaussians in the M-dimensional logarithm output energy of the Mel-filter bank (LMFB) domain.

2.1. Formulation of VTS

In the LMFB domain, the mismatch function between the mean vectors of observed signals $\boldsymbol{\mu}_O$ and $\boldsymbol{\mu}_{SI}$ is derived as follows:

$$\begin{aligned}\boldsymbol{\mu}_{O,j,k} &= \boldsymbol{\mu}_{SI,j,k} + \boldsymbol{b} + \log\left(\mathbf{1} + \exp\left(\boldsymbol{\mu}_N - \boldsymbol{\mu}_{SI,j,k} - \boldsymbol{b}\right)\right) \\ &= h\left(\boldsymbol{\mu}_{SI,j,k}, \boldsymbol{b}, \boldsymbol{\mu}_N\right) ,\end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_N$, j, and k denote mean vector of the noise and the indices of the state and the Gaussian component, respectively. The operations $\log(\cdot)$ and $\exp(\cdot)$ are independently applied to each vector element, and $\mathbf{1} = \{1, \cdots, 1\}^{\mathrm{T}}$.

Based on Eq. (1), typical VTS compensates for the difference between the initial parameters and the target parameters by using the first order Taylor series-based linear approximation as follows:

$$w_{O,j,k} = w_{SI,j,k} \tag{2}$$

$$\begin{aligned}\boldsymbol{\mu}_{O,j,k} &\simeq h\left(\boldsymbol{\mu}_{SI,j,k}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{\mu}}_N\right) \\ &\quad + \left(\boldsymbol{I} - \boldsymbol{H}_{j,k}\right)\left(\boldsymbol{b} - \bar{\boldsymbol{b}}\right) + \boldsymbol{H}_{j,k}\left(\boldsymbol{\mu}_N - \bar{\boldsymbol{\mu}}_N\right)\end{aligned} \tag{3}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{O,j,k} &\simeq \left(\boldsymbol{I} - \boldsymbol{H}_{j,k}\right)\boldsymbol{\Sigma}_{SI,j,k}\left(\boldsymbol{I} - \boldsymbol{H}_{j,k}\right)^{\mathrm{T}} + \boldsymbol{H}_{j,k}\boldsymbol{\Sigma}_N\boldsymbol{H}_{j,k}^{\mathrm{T}} \\ &= g\left(\boldsymbol{\Sigma}_{SI,j,k}, \boldsymbol{\Sigma}_N, \boldsymbol{H}_{j,k}\right) ,\end{aligned} \tag{4}$$

with the Jacobian matrix $\boldsymbol{H}_{j,k} = diag\left\{\partial h\left(\boldsymbol{\mu}_{SI,j,k}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{\mu}}_N\right)/\partial\bar{\boldsymbol{\mu}}_N\right\}$, where $w_{O,j,k}$, $\boldsymbol{\Sigma}_{O,j,k}$, $w_{SI,j,k}$, and $\boldsymbol{\Sigma}_{SI,j,k}$ denote the Gaussian weights and the diagonal variance matrices of the compensated and SI clean speech models, respectively. $\boldsymbol{\mu}_N$, $\boldsymbol{\Sigma}_N$, and $\boldsymbol{I}$ denote the mean vector and variance matrix of the target noise model and the unit matrix, respectively. The initial mean vector of the noise model is derived as $\bar{\boldsymbol{\mu}}_N = \frac{1}{U}\sum_{t=0}^{U-1}\boldsymbol{O}_t$, where $\boldsymbol{O}_t$ denotes the vector of the observed noisy speech signal in the LMFB domain at the t-th frame. The initial bias vector is set to $\bar{\boldsymbol{b}} = \mathbf{0}$.

Based on the EM-algorithm, the bias vector $\boldsymbol{b}$ and the parameter set of the noise model $\boldsymbol{\lambda}_N = \{\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\}$ are optimized as the parameters that maximize the following cost function $Q_O(\cdot)$:

$$\begin{aligned}\left\{\hat{\boldsymbol{b}}, \hat{\boldsymbol{\lambda}}_N\right\} &= \arg\max_{\boldsymbol{b},\boldsymbol{\lambda}_N} Q_O\left(\boldsymbol{O}_{0:T-1}, \boldsymbol{\lambda}_O | \boldsymbol{b}^{(i)}, \boldsymbol{\lambda}_N^{(i)}\right) \\ &= \arg\max_{\boldsymbol{b},\boldsymbol{\lambda}_N} Q_O\left(\boldsymbol{O}_{0:T-1}, \mathrm{VTS}\left(\boldsymbol{\lambda}_{SI}, \boldsymbol{b}, \boldsymbol{\lambda}_N\right) | \boldsymbol{b}^{(i)}, \boldsymbol{\lambda}_N^{(i)}\right) ,\end{aligned} \tag{5}$$

where the subscript $0 : T-1 = 0, \cdots, T-1$, where T denotes the amount of frame. i denotes the iteration index of the EM-algorithm. $\boldsymbol{\lambda}_{SI} = \left\{w_{SI,j,k}, \boldsymbol{\mu}_{SI,j,k}, \boldsymbol{\Sigma}_{SI,j,k}\right\}$ and $\boldsymbol{\lambda}_O = \left\{w_{O,j,k}, \boldsymbol{\mu}_{O,j,k}, \boldsymbol{\Sigma}_{O,j,k}\right\}$, respectively. The operation $\mathrm{VTS}(\cdot)$ is the VTS transformation given by Eqs. (2) to (4).

2.2. Problem with VTS-based approach

The typical VTS-based approach employs a single Gaussian distribution for the noise model. As mentioned in our previous work [6], a noise model with a single Gaussian distribution is unsuitable for a non-stationary noise that has a multi-peak distribution. In the VTS-based approach, the parameter set $\boldsymbol{\lambda}_N$ is estimated with the criterion of Eq. (5). However, this processing is an indirect approach for estimating $\boldsymbol{\lambda}_N$. Therefore, we should estimate $\boldsymbol{\lambda}_N$ by using only observed signal $\boldsymbol{O}_t$ and the non-linear function of the VTS transformation, because the noise signal cannot be observed directly. In this case, the estimation of the hidden variables, namely the noise mixture components, is computationally intractable.

In addition, we cannot obtain a closed form solution of noise variance matrix $\boldsymbol{\Sigma}_N$ in the conventional VTS scheme due to nonlinearity of the mismatch function. Usually, a gradient-based approach such as Newton's method is used to estimate $\boldsymbol{\Sigma}_N$. In other cases, $\boldsymbol{\Sigma}_N$ is not updated. In this case, we cannot obtain accurate parameter estimates even for a noise model with a single Gaussian distribution. Here, the estimation of bias vector $\boldsymbol{b}$ is also computationally intractable due to the same problem.

3. ESTIMATION OF BIAS VECTOR AND NOISE MODEL

To overcome the problem with the VTS-based approach, the proposed method employs an unsupervised technique for estimating the bias vector and noise mixture model by using the LMFB vectors of clean speech $\boldsymbol{S}_t$ and noise $\boldsymbol{N}_t$ given by the MMSE estimators.

The MMSE estimates of $\boldsymbol{S}_t$ and $\boldsymbol{N}_t$ are derived as:

$$\hat{\boldsymbol{S}}_t = \mathcal{E}\left\{\boldsymbol{S}_t \,|\boldsymbol{O}_t, \boldsymbol{\lambda}_O, \boldsymbol{\lambda}_{SI}, \boldsymbol{b}\right\} \tag{6}$$

$$\hat{\boldsymbol{N}}_t = \mathcal{E}\left\{\boldsymbol{N}_t \,|\boldsymbol{O}_t, \boldsymbol{\lambda}_O, \boldsymbol{\lambda}_N\right\} , \tag{7}$$

where $\mathcal{E}\{\cdot\}$ denotes the MMSE estimator. Then, with the estimated clean speech $\hat{\boldsymbol{S}}_t$ and noise $\hat{\boldsymbol{N}}_t$, $\boldsymbol{b}$ and $\boldsymbol{\lambda}_N$ are estimated by using EM-based maximum likelihood (ML) estimation based on the following criteria instead of Eq. (5),

$$\hat{\boldsymbol{b}} = \arg\max_{\boldsymbol{b}} Q_{SD}\left(\hat{\boldsymbol{S}}_{0:T-1}, \boldsymbol{\lambda}_{SI}, \boldsymbol{b} | \boldsymbol{b}^{(i)}\right) \tag{8}$$

$$\hat{\boldsymbol{\lambda}}_N = \arg\max_{\boldsymbol{\lambda}_N} Q_N\left(\hat{\boldsymbol{N}}_{0:T-1}, \boldsymbol{\lambda}_N | \boldsymbol{\lambda}_N^{(i)}\right) , \tag{9}$$

where $Q_{SD}(\cdot)$ and $Q_N(\cdot)$ denote the cost functions of the SD clean speech and the noise models, respectively. By iterating these processes with the EM algorithm, $\boldsymbol{b}$ and $\boldsymbol{\lambda}_N$ are successfully optimized.

3.1. Initialization

The initial parameters of the bias vector and noise mixture model with L mixture components are given as

$$\boldsymbol{b}^{(i=0)} = \mathbf{0} \tag{10}$$

$$\boldsymbol{\lambda}_N^{(i=0)} = \left\{w_{N,l}^{(i=0)} = \frac{1}{L}, \boldsymbol{\mu}_{N,l}^{(i=0)} \sim \mathcal{N}\left(\bar{\boldsymbol{\mu}}_N, \bar{\boldsymbol{\Sigma}}_N\right), \boldsymbol{\Sigma}_{N,l}^{(i=0)} = \bar{\boldsymbol{\Sigma}}_N\right\} , \tag{11}$$

where l and $w_{N,l}$ denote Gaussian component index and Gaussian weight of noise muxture model, respectively. $\boldsymbol{\mu}_{N,l}^{(i=0)}$ is initialized by multivariate Gaussian random value given by $\mathcal{N}\left(\bar{\boldsymbol{\mu}}_N, \bar{\boldsymbol{\Sigma}}_N\right)$, where $\bar{\boldsymbol{\Sigma}}_N = diag\left\{\frac{1}{U}\sum_{t=0}^{U-1}\boldsymbol{O}_t\boldsymbol{O}_t^{\mathrm{T}} - \bar{\boldsymbol{\mu}}_N\bar{\boldsymbol{\mu}}_N^{\mathrm{T}}\right\}$.

3.2. E-step

3.2.1. Model compensation

The first stage of **E-step** is the compensation of an observed signal model with the estimated parameters of the previous iteration, $\boldsymbol{b}^{(i)}$ and $\boldsymbol{\lambda}_N^{(i)}$. During the model composition, each state of the SI clean speech model has K Gaussians and the noise model has L Gaussians. Thus, the number of Gaussians contained in each state of the composed model is expanded to $K \times L$. At the i-th iteration, the model parameter set $\boldsymbol{\lambda}_O^{(i)}$ is derived as

$$\boldsymbol{\lambda}_O^{(i)} = \left\{\begin{array}{l} w_{O,j,k,l}^{(i)} = w_{SI,j,k} \cdot w_{N,l}^{(i)} , \\ \boldsymbol{\mu}_{O,j,k,l}^{(i)} = h\left(\boldsymbol{\mu}_{SI,j,k}, \boldsymbol{b}^{(i)}, \boldsymbol{\mu}_{N,l}^{(i)}\right) , \\ \boldsymbol{\Sigma}_{O,j,k,l}^{(i)} = g\left(\boldsymbol{\Sigma}_{SI,j,k}, \boldsymbol{\Sigma}_{N,l}^{(i)}, \boldsymbol{H}_{j,k,l}^{(i)}\right) \end{array}\right\} . \tag{12}$$

3.2.2. Expectation of cost function

When $\boldsymbol{O}_{0:T-1}$ is given, the expectation of the cost function related to the parameter set $\boldsymbol{\lambda}_O^{(i)}$ is derived as follows:

$$\begin{aligned} Q_O\left(\boldsymbol{O}_{0:T-1}, \boldsymbol{\lambda}_O | \boldsymbol{\lambda}_O^{(i)}\right) &= \sum_{t,j,k,l} P_{t,j}^{(i)} P_{t,j,k,l}^{(i)} \\ &\times \left(\log w_{O,j,k,l}^{(i)} + \log \mathcal{N}\left(\boldsymbol{O}_t; \boldsymbol{\mu}_{O,j,k,l}^{(i)}, \boldsymbol{\Sigma}_{O,j,k,l}^{(i)}\right)\right) , \end{aligned} \tag{13}$$

where $\mathcal{N}(\cdot)$, $P_{t,j}^{(i)}$, and $P_{t,j,k,l}^{(i)}$ denote the probability density function of a Gaussian distribution and *a posteriori* (occupancy) probabilities with respect to j, k, and l, respectively.

3.3. M-step

3.3.1. MMSE estimation of noise signal

The MMSE estimate of $\boldsymbol{N}_t^{(i)}$ defined by Eq. (7) is derived as

$$\begin{aligned} \hat{\boldsymbol{N}}_t^{(i)} &= P_{t,j=1}^{(i)} \boldsymbol{O}_t + P_{t,j=2}^{(i)} \left(\boldsymbol{O}_t + \mathcal{E}\left\{\boldsymbol{G}_t^{(i)}\right\}\right) \\ &= \boldsymbol{O}_t + P_{t,j=2}^{(i)} \sum_{k,l} P_{t,j=2,k,l}^{(i)} \left(\boldsymbol{\mu}_{N,l}^{(i)} - \boldsymbol{\mu}_{O,j=2,k,l}\right) , \end{aligned} \tag{14}$$

where $\boldsymbol{G}_t^{(i)}$ denotes the Wiener filter for extraction of the noise MMSE estimate in the LMFB domain. Since the state $j = 1$ is the silence (speech absence) state, $\boldsymbol{N}_t^{(i)}$ is obtained as observed signal $\boldsymbol{O}_t$. With the speech activity state, $j = 2$, $\boldsymbol{N}_t^{(i)}$ is obtained by using the MMSE estimation with $P_{t,j=2,k,l}^{(i)}$. This method implicitly involves the voice activity detection for the noise estimation [5].

3.3.2. Noise mixture model estimation with MMSE estimates

With the MMSE estimate $\boldsymbol{N}_t^{(i)}$, $\boldsymbol{\lambda}_N^{(i)}$ is estimated by using a nested EM-based ML estimation with the following cost function:

$$\begin{aligned} &Q_N\left(\hat{\boldsymbol{N}}_{0:T-1}^{(i)}, \boldsymbol{\lambda}_N^{(i+1)} | \boldsymbol{\lambda}_N^{(i')}\right) \\ &= \sum_{t,l} P_{t,l}^{(i')} \left(\log w_{N,l}^{(i')} + \log \mathcal{N}\left(\hat{\boldsymbol{N}}_t^{(i)}; \boldsymbol{\mu}_{N,l}^{(i')}, \boldsymbol{\Sigma}_{N,l}^{(i')}\right)\right) , \end{aligned} \tag{15}$$

where i' and $P_{t,l}^{(i')}$ denote the iteration index of the $\boldsymbol{\lambda}_N^{(i+1)}$ estimation and the *a posteriori* probability with respect to l, respectively. Then, $\boldsymbol{\lambda}_N^{(i+1)}$ is estimated as follows:

$$\boldsymbol{\lambda}_N^{(i'+1)} = \left\{ \begin{array}{l} w_{N,l}^{(i'+1)} = \dfrac{\sum_t P_{t,l}^{(i')}}{\sum_{t,l} P_{t,l}^{(i')}} , \\ \boldsymbol{\mu}_{N,l}^{(i'+1)} = \dfrac{\sum_t P_{t,l}^{(i')} \hat{\boldsymbol{N}}_t^{(i)}}{\sum_t P_{t,l}^{(i')}} , \\ \boldsymbol{\Sigma}_{N,l}^{(i'+1)} = \dfrac{\sum_t P_{t,l}^{(i')} \hat{\boldsymbol{N}}_t^{(i)} \hat{\boldsymbol{N}}_t^{(i)\mathrm{T}}}{\sum_t P_{t,l}^{(i')}} - \boldsymbol{\mu}_{N,l}^{(i'+1)} \boldsymbol{\mu}_{N,l}^{(i'+1)\mathrm{T}} \end{array} \right\} . \tag{16}$$

Finally, $\boldsymbol{\lambda}_N^{(i+1)} = \boldsymbol{\lambda}_N^{(i'+1)}$ is given by iterating Eqs. (15) and (16) until convergence.

3.3.3. MMSE estimation of clean speech signal

The MMSE estimate of $\boldsymbol{S}_t^{(i)}$ defined by Eq. (6) is derived as

$$\begin{aligned} \hat{\boldsymbol{S}}_t^{(i)} &= \boldsymbol{O}_t + \mathcal{E}\left\{\boldsymbol{F}_t^{(i)}\right\} \\ &= \boldsymbol{O}_t + \sum_{j,k,l} P_{t,j}^{(i)} P_{t,j,k,l}^{(i)} \left(\boldsymbol{\mu}_{SI,j,k} + \boldsymbol{b}^{(i)} - \boldsymbol{\mu}_{O,j,k,l}\right) , \end{aligned} \tag{17}$$

where $\boldsymbol{F}_t^{(i)}$ denotes the Wiener filter for extraction of the speech MMSE estimate in the LMFB domain.

3.3.4. Bias vector estimation with MMSE estimates

With the MMSE estimate $\boldsymbol{S}_t^{(i)}$, $\boldsymbol{b}^{(i)}$ is also estimated by using a nested EM-based ML estimation with the following cost function:

$$\begin{aligned} &Q_{SD}\left(\hat{\boldsymbol{S}}_{0:T-1}^{(i)}, \boldsymbol{\lambda}_{SI}, \boldsymbol{b}^{(i+1)} | \boldsymbol{b}^{(i'')}\right) = \sum_{t,j,k} P_{t,j}^{(i'')} P_{t,j,k}^{(i'')} \\ &\times \left(\log w_{SI,j,k} + \log \mathcal{N}\left(\hat{\boldsymbol{S}}_t^{(i)}; \boldsymbol{\mu}_{SI,j,k} + \boldsymbol{b}^{(i'')}, \boldsymbol{\Sigma}_{SI,j,k}\right)\right) \end{aligned} \tag{18}$$

where i'' and $P_{t,j,k}^{(i'')}$ denote the iteration index of the $\boldsymbol{b}^{(i+1)}$ estimation and the *a posteriori* probability with respect to j and k, respectively. Then, $\boldsymbol{b}^{(i+1)}$ is estimated as follows:

$$\boldsymbol{b}^{(i''+1)} = \left(\sum_{t,j,k} \boldsymbol{B}_{t,j,k}\right)^{-1} \sum_{t,j,k} \boldsymbol{B}_{t,j,k} \left(\hat{\boldsymbol{S}}_t^{(i)} - \boldsymbol{\mu}_{SI,j,k}\right) , \tag{19}$$

where $\boldsymbol{B}_{t,j,k} = P_{t,j}^{(i'')} P_{t,j,k}^{(i'')} \boldsymbol{\Sigma}_{SI,j,k}^{-1}$.

Finally, $\boldsymbol{b}^{(i+1)} = \boldsymbol{b}^{(i''+1)}$ is given by iterating Eqs. (18) and (19) until convergence.

3.4. Noise suppression

The noise is suppressed using a Mel-scaled Wiener filter $\boldsymbol{W}_t^{Mel} = \exp\left(\boldsymbol{F}_t\right)$ as described in our previous work [5]. By applying a third order spline interpolation, $\boldsymbol{W}_t^{Mel}$ can be transformed into a linear-scaled filter $\boldsymbol{W}_t^{Lin}$. The noise suppressed signal $\hat{s}_\tau$ is obtained by applying $\boldsymbol{W}_t^{Lin}$ and an inverse fast Fourier transform to the complex spectrum of the observed signal.

3.5. Processing flow

The following algorithm summarizes the proposed method, and is applied to each utterance.

Algorithm 1 Parameter estimation and noise suppression

1: Initialize $\boldsymbol{\lambda}_N^{(i=0)}$ and $\boldsymbol{b}^{(i=0)}$ (See Sec. 3.1.)
2: **repeat**
3: Model compensation (See Sec. 3.2.1.)
4: Compute expectation of cost function (See Sec. 3.2.2.)
5: Estimate $\hat{\boldsymbol{N}}_t^{(i)}$ for all t (See Sec. 3.3.1.)
6: Update $\boldsymbol{\lambda}_N^{(i)}$ with EM-based ML estimation (See Sec. 3.3.2.)
7: Estimate $\hat{\boldsymbol{S}}_t^{(i)}$ for all t (See Sec. 3.3.3.)
8: Update $\boldsymbol{b}^{(i)}$ with EM-based ML estimation (See Sec. 3.3.4.)
9: **until** convergence is achieved
10: Apply the noise suppression (See Sec. 3.4.)

4. EXPERIMENTS

4.1. Experimental setup

The experimental materials were 100 utterances spoken by 23 Japanese males that were taken from the Information-technology Promotion Agency (IPA)-98-TestSet. The speaking style of the speech data is read speech. Three types of highly non-stationary noises, i.e., airport lobby noise, platform noise, and street noise, were artificially added to clean speech signals by changing the SNR at three levels; 10, 5, and 0 dB. The sampling frequency of the speech data and noises was 16 kHz.

The feature parameters for the noise suppression were 24 LMFBs that were extracted by using a Hamming window with a 20 msec frame length and a 10 msec frame shift length. Each state of the SI clean speech model had $K = 128$ Gaussians. The number of Gaussians of the noise model was set at $L = 1, 2, 3, 4$. The training materials for the SI speech model were 33,820 phonetically balanced sentences spoken by 180 Japanese males. We also trained SD speech models whose corresponding training materials were 155 sentences per a speaker. The parameter U was set at 10.

Table 1. ASR results in WER (%). The bias vector update was only applied to the SI clean speech model.

Method	Bias vector update	Noise model update	Airport lobby noise			Platform noise			Street noise			Avg.
			10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	
w/o noise suppression	—	—	26.1	59.1	87.1	27.2	55.1	79.0	11.5	28.7	61.0	48.3
VTS	—	✓	17.0	39.5	72.0	24.0	43.9	70.7	7.3	14.5	29.9	35.4
	✓	✓	16.1	34.5	65.8	24.6	43.3	69.8	7.7	13.2	28.2	33.7
	SD model	✓	*13.8*	*34.0*	*62.8*	*21.2*	*43.7*	*66.2*	*6.5*	*11.8*	*24.9*	*31.7*
Proposal ($L = 1$)	—	✓	17.1	34.2	63.1	23.1	42.4	64.5	8.4	15.6	28.8	33.0
	✓	✓	**11.2**	28.8	59.9	19.1	39.5	57.8	7.2	11.9	25.0	28.9
	SD model	✓	*11.3*	*25.3*	*52.7*	*18.6*	*35.6*	*54.8*	*7.2*	*12.3*	*22.7*	*26.7*
Proposal ($L = 2$)	—	✓	13.1	28.8	59.9	20.4	37.5	60.2	6.9	13.2	26.6	29.6
	✓	✓	11.7	27.8	57.0	19.0	35.7	57.6	6.7	11.2	**24.0**	27.9
	SD model	✓	*9.8*	*23.1*	*50.0*	*15.1*	*28.5*	*49.8*	*6.7*	*10.0*	*20.2*	*23.7*
Proposal ($L = 3$)	—	✓	13.1	29.8	60.1	17.8	37.1	59.1	7.0	12.4	26.1	29.2
	✓	✓	11.5	**27.7**	**56.8**	15.9	**32.9**	57.6	**6.2**	11.2	**24.0**	27.1
	SD model	✓	*9.5*	*22.0*	*50.1*	*12.0*	*27.2*	*47.7*	*6.6*	*9.1*	*19.3*	*22.6*
Proposal ($L = 4$)	—	✓	12.9	29.7	60.0	17.1	35.4	59.3	7.6	13.3	26.9	29.1
	✓	✓	11.3	27.8	57.3	**15.8**	33.2	**55.6**	6.7	**10.4**	24.1	**26.9**
	SD model	✓	*8.9*	*21.7*	*47.7*	*13.6*	*25.2*	*44.3*	*5.8*	*9.1*	*18.3*	*21.6*

The ASR was carried out by employing a weighted finite state transducer-based decoder [12]. We used SI triphone hidden Markov models (HMMs) trained by clean speech. The HMM was trained with a variational Bayesian approach [13]. The HMM topology was a three state left-to-right HMM and there were 2,364 HMM states. Each state had 16 Gaussians. The feature parameters for the ASR consisted of 12 MFCCs and the log energy with their first and second order derivatives. Cepstral mean normalization was applied to each utterance. The training materials for the HMMs were the same as those for the SI speech model used in the noise suppression.

The language model was a back-off tri-gram with Witten-Bell discounting. It was trained using 75 months' worth of Japanese newspaper articles. The vocabulary size was 20k words. The evaluation criterion for ASR was the word error rate (WER). The WER of a clean speech signal was 3.9 %.

4.2. Experimental results

Table 1 shows the ASR results for each method. With the results of the VTS and the proposed method with only the bias vector update, the noise model consists of the initial parameters, $\bar{\boldsymbol{\mu}}_N$ and $\bar{\boldsymbol{\Sigma}}_N$.

In the table, the results of "Proposal ($L = 1$)" with both bias vector and noise model updates show significant improvements in WER compared with the results of "VTS" under the same condition. In each result, the number of Gaussian distributions for the noise model is set at $L = 1$, thus the essential differences between the two methods are the introduction of an MMSE estimators and the parameter estimation criteria. This result proves the effectiveness of the proposed method, namely the parameter estimation scheme, when utilizing the MMSE estimates of $\boldsymbol{S}_t$ and $\boldsymbol{N}_t$. When L exceeds two, the results of the proposed method with both bias vector and noise model updates improve further. These results also prove the importance of considering noise mixture models for non-stationary noise with a multi-peak distribution. As seen in the table, the optimum number of Gaussian distributions for a noise model depends on the noise conditions and SNRs. Thus, a scheme for optimizing the noise model topology is a crucial factor in the proposed method.

Under all the conditions, the results for the SD clean speech model are clearly superior to those of a bias vector update with the SI speech model. Thus, the improvement of the speaker adaptation scheme is also a crucial factor in the proposed method.

5. CONCLUSIONS

This paper presented a joint unsupervised approach for estimating a bias vector for speaker adaptation and a noise mixture model with MMSE estimates of the clean speech and the noise. The evaluation results show that further improvement was realized by integrating the speaker adaptation and the estimation of the noise mixture model. We plan to investigate the improvement of the speaker adaptation scheme and the selection of an adaptive model topology.

6. REFERENCES

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. 32, pp. 1109–1121, Dec. 1984.

[4] P. J. Moreno, *et al.*, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP '96*, vol. II, pp. 733–736, May 1996.

[5] M. Fujimoto, *et al.*, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," in *Proc. of Interspeech '09*, pp. 1235–1238, Sept. 2009.

[6] M. Fujimoto, *et al.*, "A robust estimation method of noise mixture model for noise suppression," in *Proc. of Interspeech '11*, pp. 697–700, Aug. 2011.

[7] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. on SAP*, vol. 4, no. 5, pp. 352–359, May 1996.

[8] K. K. Chin, *et al.*, "Rapid joint speaker and noise compensation for robust speech recognition," in *Proc. of ICASSP '11*, pp. 5500–5503, May 2011.

[9] C. L. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[10] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, vol. 4, no. 1, pp. 19–30, Jan. 1996.

[11] C. H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[12] T. Hori, *et al.*, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Trans. on ASLP*, vol. 15, no. 4, pp. 1352–1365, May 2007.

[13] S. Watanabe, *et al.*, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. on SAP*, vol. 12, no. 4, pp. 365–381, July 2004.