Model-based noise suppression using unsupervised estimation of hidden Markov model for non-stationary noise

Masakiyo Fujimoto and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan

{fujimoto.masakiyo, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

Although typical model-based noise suppression including the vector Taylor series-based approach employs a single Gaussian distribution for the noise model, it is insufficient for non-stationary noises which have a complex structured distribution. As a solution to this problem, we have already proposed a method for estimating a Gaussian mixture model (GMM)-based noise model by using a minimum mean squared error (MMSE) estimate of the noise. However, the state transition process of the non-stationary noise is not modeled in the noise GMM. In this paper, we propose a way of modeling the noise with a hidden Markov model (HMM) as an extension of our previous method. The proposed method proves that the HMM-based noise model outperforms a GMM-based noise model composed of the same number of Gaussian components. In addition, we discuss the appropriate topology for the noise HMM, i.e., a left-to-right HMM and an ergodic HMM.

Index Terms: noise suppression, noise modeling, model topology, MMSE estimation

1. Introduction

Noise robustness is one of the most important problems as regards the practical use of automatic speech recognition (ASR). Research on noise robust ASR has a long history and as a result a lot of techniques have been proposed. The front-end processing of ASR employs various techniques including robust feature extraction [1, 2], feature space normalization [3, 4, 5], and noise suppression [6]-[11] are used to reduce the mismatch between noise corrupted feature parameters and acoustic models. On the other hand, the proposed back-end processing techniques include model compensation [12, 13, 14] and model adaptation [15, 16, 17]. For the ASR decoding stage, an uncertainty decoding technique, which exploits the uncertainty of noise signals has also been proposed [18, 19]. To realize further improvement, the integration of front-end and back-end processing has also been proposed [20, 21, 22]. Of these techniques, a vector Taylor series (VTS)-based approach [8] is widely used as a representative method, and various extensions of the VTS-based approach have been proposed in recent years [23, 24, 25].

The VTS-based approach is a statistical model-based technique, and it is applicable to both front and back-end processing, i.e., model-based noise suppression and acoustic model compensation. In this paper, we focus on model-based noise suppression for front-end processing; however, the discussions and the techniques described in this paper can be applied to acoustic model compensation in back-end processing.

The VTS-based approach uses the expectation-maximization (EM) algorithm to optimize statistical model of the observed (noisy speech) signal in the logarithmic mel-filter bank (LMFB) domain or the mel-frequency cepstral coefficient (MFCC) domain. Then, the model of the observed signal is compensated by using models of clean speech and noise based on the mismatch function, which represents the non-linear relation between clean speech and noise in the LMFB or the MFCC domain. Usually, in the VTS-based approach, the clean speech model is trained in advance and its parameters are fixed, hence only the noise model parameters are estimated with the EM algorithm. Here, the VTS-based approach typically employs a single Gaussian distribution for the noise model. If the statistical characteristics of the noise have a uni-modal distribution, a single Gaussian distribution is sufficient for the noise model. However, most of the noises observed in real environments have non-stationary characteristics. In this case, the statistical characteristics of the noise may have a multi-modal distribution or another complex structured distribution. Therefore, a single Gaussian distribution is unsuitable for the noise model due to its poor representation. To ensure robustness against non-stationary noise, a model with a complex structure, for example a Gaussian mixture model (GMM), is suitable for the noise.

Indeed, the GMM-based noise model can be applied to the conventional VTS-based approach in theory. However, the parameter estimation of the noise GMM is computationally intractable due to the following two factors:

(a) Hidden variables, i.e., the occupancy probabilities of each Gaussian component w.r.t. the noise signal, are unknown.

(b) The accurate parameter estimation of each Gaussian component is computationally intractable due to the non-linear relation between clean speech and noise.

As regards these factors, although the first problem can be solved by using marginalization [26], the second problem is difficult to obtain distinct solutions.

On the basis of the above considerations, we have recently proposed a method for estimating the noise GMM for model-based noise suppression [11]. This method estimates the parameters of the noise GMM by utilizing the minimum mean squared error (MMSE) estimates of noise signals and the alternate EM algorithm. In this work, we reveal that the GMM-based noise model significantly improves ASR accuracy in highly non-stationary noise environments by comparison with a noise model with a single Gaussian distribution. However, the GMM represents the complex structured distribution of the non-stationary noise by considering it as a long-term stationary process. The temporal characteristics of non-stationary noise, namely, the state transition characteristics, are not clearly represented in the GMM-based noise model. Therefore, in this paper, we propose a technique for non-stationary noise modeling with the state transition process by expanding the GMM-based model into a hidden Markov model (HMM)-based model. With this expansion, we employ ASR evaluations in highly

Table 1: Notation list

$\boldsymbol{O}_t$, $\boldsymbol{S}_t$, $\boldsymbol{N}_t$	LMFB vectors of observed signal, clean speech, and noise at frame t
T	Number of frames
m	Model index, $m = O, S, N$
i, j, i',j', i'', j''	State indices for the models $m = O, S, N$
J, J', J''	Number of states for models $m = O, S, N$
k, k', k''	Gaussian indices for models $m = O, S, N$
K, K', K''	Number of Gaussians for models $m = O, S, N$
$\boldsymbol{\lambda}_m$	parameter set for model m given as $\boldsymbol{\lambda}_m = \{a_{m,ij}, w_{m,j,k}, \boldsymbol{\mu}_{m,j,k}, \boldsymbol{\Sigma}_{m,j,k}\}$
$a_{m,ij}$	State transition probability from state i to j for model m
$w_{m,j,k}$	Weight of Gaussian component k in state j for model m
$\boldsymbol{\mu}_{m,j,k}$	Mean vector of Gaussian component k in state j for model m
$\boldsymbol{\Sigma}_{m,j,k}$	Diagonal variance matrix of Gaussian component k in state j for model m
$b_{j,k}(\cdot)$	Likelihood of Gaussian component k in state j

non-stationary noises to show that the HMM-based model outperforms the GMM-based model when the models contain the same number of Gaussian distributions. In addition, we compare the topology of HMM-based noise models, i.e., a left-to-right HMM and an ergodic HMM, and discuss the characteristics of each model topology.

The use of an HMM-based noise model is reported in [12] and [13]. With these methods, the HMM-based noise model greatly improves ASR accuracy in various non-stationary noise environments; however the noise models required in these methods should be trained in advance with some noise signal training data. The proposed method aims to estimate of the noise model parameters without prior knowledge of the noise. This aim is its crucial advantage over the conventional method.

2. Noise suppression with HMMs

This section describes noise suppression approach with HMMs, which is an extension of the conventional model-based noise suppression technique with GMMs [8, 11]. In the method, the feature parameter is an M-dimensional LMFB vector. Before beginning a detailed explanation, we provide a list of the notations used in this paper in Table 1 .

2.1. Definition of models

In our implementation, the clean speech model is given by an ergodic HMM with two internal states ($J' = 2$), i.e., states of silence ($j' = 0$) and speech ($j' = 1$), as shown in Fig. 1a. On the other hand, the noise model is given by a left-to-right HMM (Fig. 1b) or an ergodic HMM (Fig. 1c) with J'' internal states. Each state is modeled by a GMM with K' or K'' Gaussian components in the LMFB domain. Each model has the parameter set $\boldsymbol{\lambda}_S$ or $\boldsymbol{\lambda}_N$. $\boldsymbol{\lambda}_S$ is trained with some clean speech corpora in advance.

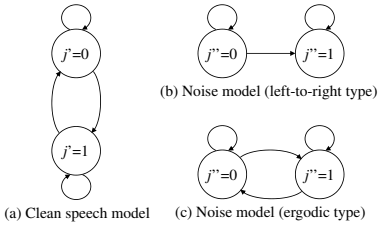
With the models of clean speech and noise, the observed signal model is composed as shown in Fig. 2. The composed model has $J = J'J''$ states and each state has $K = K'K''$ Gaussian components, hence each composed model in Fig. 2 has $J = 4$ states.

In this paper, we change the topology of the noise model with variables J'' and K'', however that of the clean speech model is fixed.

2.2. Mismatch function and model composition

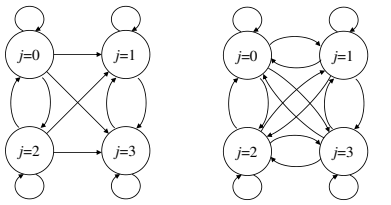
In the LMFB domain, the mismatch function between the observed signals $\boldsymbol{O}_t$ and speech signals $\boldsymbol{S}_t$ is derived as follows:

$$\begin{aligned}\boldsymbol{O}_t &= \boldsymbol{S}_t + \mathbf{log}\left(\mathbf{1} + \mathbf{exp}\left(\boldsymbol{N}_t - \boldsymbol{S}_t\right)\right) \\ &= h\left(\boldsymbol{S}_t, \boldsymbol{N}_t\right) ,\end{aligned} \tag{1}$$



(a) Clean speech model (b) Noise model (left-to-right type) (c) Noise model (ergodic type)

Figure 1: Examples of clean speech model and noise models. Each model has $J' = 2$ or $J'' = 2$ states.



(a) Observed signal model with left-to-right type noise model (b) Observed signal model with ergodic type noise model

Figure 2: Examples of composed observed signal models. Each model has $J = 4$ states.

where the operations $\mathbf{log}(\cdot)$ and $\mathbf{exp}(\cdot)$ are independently applied to each vector element, and $\mathbf{1} = \{1, \cdots, 1\}^{\mathrm{T}}$.

Based on Eq. (1), the model composition is derived as

$$\boldsymbol{\lambda}_O = \Psi\left(\boldsymbol{\lambda}_S, \boldsymbol{\lambda}_N\right) = \left\{ \begin{array}{r} a_{O,ij} = a_{S,i'j'} \cdot a_{N,i''j''} \\ w_{O,j,k} = w_{S,j',k'} \cdot w_{N,j'',k''} \\ \boldsymbol{\mu}_{O,j,k} = h\left(\boldsymbol{\mu}_{S,j',k'}, \boldsymbol{\mu}_{N,j'',k''}\right) \\ \boldsymbol{\Sigma}_{O,j,k} \simeq \boldsymbol{H}_{S,j,k}\boldsymbol{\Sigma}_{S,j',k'}\boldsymbol{H}_{S,j,k}^{\mathrm{T}} \\ \qquad + \boldsymbol{H}_{N,j,k}\hat{\boldsymbol{\Sigma}}_{N,j'',k''}\boldsymbol{H}_{N,j,k}^{\mathrm{T}} \end{array} \right\} , \tag{2}$$

with the Jacobian matrices,

$$\boldsymbol{H}_{S,j,k} = diag\left\{\partial h\left(\boldsymbol{\mu}_{S,j',k'}, \boldsymbol{\mu}_{N,j'',k''}\right) / \partial \boldsymbol{\mu}_{S,j',k'}\right\} \tag{3}$$

$$\boldsymbol{H}_{N,j,k} = diag\left\{\partial h\left(\boldsymbol{\mu}_{S,j',k'}, \boldsymbol{\mu}_{N,j'',k''}\right) / \partial \boldsymbol{\mu}_{N,j'',k''}\right\} , \tag{4}$$

where the indices of the states and the Gaussian components are derived as $i = i'J'' + i''$, $j = j'J'' + j''$, and $k = k'K'' + k''$, respectively. Then, with the LMFB vector sequence of the observed signal $\boldsymbol{O}_{0:T-1}$, the model parameter $\boldsymbol{\lambda}_N$ is estimated as the parameter that maximizes cost function $Q(\cdot|\cdot)$.

2.3. MAP-based MMSE estimation of clean speech

We employ a maximum *a posteriori* (MAP)-based MMSE estimation for noise suppression. For Gaussian component k in state j at frame t, the MAP estimate of clean speech $\tilde{\boldsymbol{\mu}}_{S,t,j,k}$ is derived as:

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{S,t,j,k} &= \underset{\tilde{\boldsymbol{\mu}}_{S,t,j,k}}{\arg\max} \left\{p\left(\tilde{\boldsymbol{\mu}}_{S,t,j,k}|\boldsymbol{\lambda}_S\right) b_{j,k}(\boldsymbol{O}_t)\right\} \\ &= \boldsymbol{\mu}_{S,j',k'} + \boldsymbol{\Sigma}_{S,j',k'}\boldsymbol{H}_{S,j,k}^{\mathrm{T}}\boldsymbol{\Sigma}_{O,j,k}^{-1}\left(\boldsymbol{O}_t - \boldsymbol{\mu}_{O,j,k}\right) ,\end{aligned} \tag{5}$$

with

$$p\left(\tilde{\boldsymbol{\mu}}_{S,t,j,k}|\boldsymbol{\lambda}_S\right) = \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{S,t,j,k}|\boldsymbol{\mu}_{S,j',k'}, \boldsymbol{\Sigma}_{S,j',k'}\right) \tag{6}$$

$$b_{j,k}(\boldsymbol{O}_t) = \mathcal{N}\left(\boldsymbol{O}_t|\boldsymbol{\mu}_{O,j,k}, \boldsymbol{\Sigma}_{O,j,k}\right) , \tag{7}$$

where $p\left(\tilde{\boldsymbol{\mu}}_{S,t,j,k}|\boldsymbol{\lambda}_S\right)$ and $\mathcal{N}(\cdot|\cdot)$ denote the prior distribution of the clean speech signal and the probability density function of Gaussian distribution, respectively. $p\left(\tilde{\boldsymbol{\mu}}_{S,t,j,k}|\boldsymbol{\lambda}_S\right)$ is given by a Gaussian component contained in the clean speech model.

With the MMSE manner of Eq. (8), the LMFB vector of the clean speech is estimated with the occupancy probability $\zeta_{O,t,j,k}$ which is obtained with the Baum-Welch algorithm described in Sec.3.

$$\tilde{\boldsymbol{S}}_t = \sum_{j,k} \zeta_{O,t,j,k} \tilde{\boldsymbol{\mu}}_{S,t,j,k} \tag{8}$$

3. Parameter estimation based on Baum-Welch algorithm

The parameter set $\boldsymbol{\lambda}_N$ is estimated by using the Baum-Welch algorithm.

3.1. Parameter initialization

If the noise model is given as the left-to-right HMM, the state transition probabilities are initialized as

$$\hat{a}_{N,i''j''} = \begin{cases} 0.6 & \text{if } i'' = j'' \\ 0.4 & \text{if } i''+1 = j'' \\ 0.0 & \text{otherwise} \end{cases} . \tag{9}$$

Otherwise, with the ergodic HMM, they are given as

$$\hat{a}_{N,i''j''} = \begin{cases} 0.6 & \text{if } i'' = j'' \\ 0.4/(J''-1) & \text{otherwise} \end{cases} . \tag{10}$$

The other initial parameters of the noise model are given as

$$\hat{w}_{N,j'',k''} = 1/K'' \tag{11}$$

$$\hat{\boldsymbol{\mu}}_{N,j'',k''} \sim \mathcal{N}\left(\cdot|\bar{\boldsymbol{\mu}}_N, \bar{\boldsymbol{\Sigma}}_N\right) \tag{12}$$

$$\hat{\boldsymbol{\Sigma}}_{N,j'',k''} = \bar{\boldsymbol{\Sigma}}_N \ , \tag{13}$$

where $\mathcal{N}\left(\cdot|\bar{\boldsymbol{\mu}}_N, \bar{\boldsymbol{\Sigma}}_N\right)$ is a multivariate Gaussian random value with $\bar{\boldsymbol{\mu}}_N = \frac{1}{U}\sum_{t=0}^{U-1} \boldsymbol{O}_t$ and $\bar{\boldsymbol{\Sigma}}_N = diag\left\{\frac{1}{U}\sum_{t=0}^{U-1} \boldsymbol{O}_t \boldsymbol{O}_t^{\mathrm{T}} - \bar{\boldsymbol{\mu}}_N \bar{\boldsymbol{\mu}}_N^{\mathrm{T}}\right\}$.

3.2. Expectation of cost function

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

$$\begin{aligned} &Q\left(\boldsymbol{O}_{0:T-1}, \boldsymbol{\lambda}_O | \hat{\boldsymbol{\lambda}}_O\right) = \\ &\sum_{t,i,j,k} \chi_{O,t,i,j,k} \log \hat{a}_{O,ij} + \sum_{t,j,k} \zeta_{O,t,j,k} \log b_j(\boldsymbol{O}_t) \ , \end{aligned} \tag{14}$$

with the likelihood of each state,

$$b_j(\boldsymbol{O}_t) = \sum_k \hat{w}_{O,j,k} b_{j,k}(\boldsymbol{O}_t), \tag{15}$$

and the occupancy probabilities,

$$\chi_{O,t,i,j,k} = \frac{\alpha_{i,t}\hat{a}_{O,ij}\hat{w}_{O,j,k}b_{j,k}(\boldsymbol{O}_{t+1})\beta_{t+1,j}}{\sum_{i,j,k}\alpha_{i,t}\hat{a}_{O,ij}\hat{w}_{O,j,k}b_{j,k}(\boldsymbol{O}_{t+1})\beta_{t+1,j}} \tag{16}$$

$$\zeta_{O,t,j,k} = \sum_i \chi_{O,t-1,i,j,k} \ , \tag{17}$$

where $\alpha_{t,j}$ and $\beta_{t,j}$ denote the forward and backward probabilities, respectively.

3.3. Occupancy probabilities of noise model

To estimate the state transition probabilities and Gaussian weights of the noise model, we need the occupancy probabilities w.r.t. the noise model. The occupancy probabilities of the noise model are obtained by marginalizing those of the observed signal model around the parameters of the clean speech model as follows:

$$\zeta_{N,t,j'',k''} = \sum_{j',k'} \zeta_{O,t,j=j'J''+j'',k=k'K''+k''} \tag{18}$$

$$\gamma_{N,t,i'',j''} = \sum_{i',j',k} \chi_{O,t,i=i'J''+i'',j=j'J''+j'',k} \ . \tag{19}$$

3.4. Parameter update

3.4.1. Update of state transition probabilities and Gaussian weights

With the maximum likelihood (ML) manner, each parameter contained in $\boldsymbol{\lambda}_N$ is given by differentiating the cost function $Q(\cdot|\cdot)$ w.r.t. the corresponding parameter and equating it to zero. Thus, $\hat{a}_{N,i''j''}$ and $\hat{w}_{N,j'',k''}$ are given by the following equations:

$$\hat{a}_{N,i''j''} = \frac{\sum_t \gamma_{N,t,i'',j''}}{\sum_{t,j''} \gamma_{N,t,i'',j''}} \tag{20}$$

$$\hat{w}_{N,j'',k''} = \frac{\sum_t \zeta_{N,t,j'',k''}}{\sum_{t,k''} \zeta_{N,t,j'',k''}} \ . \tag{21}$$

3.4.2. Update of mean vectors and variance matrices based on VTS

Since $Q(\cdot|\cdot)$ is a non-linear function, we cannot obtain the closed form solution of $\hat{\boldsymbol{\mu}}_{N,j'',k''}$ and $\hat{\boldsymbol{\Sigma}}_{N,j'',k''}$. With this problem, the first order VTS is usually applied to $\boldsymbol{\mu}_{O,j,k}$ as shown in Eq. (22). Then, the VTS-based linear approximation gives the closed form solution of $\hat{\boldsymbol{\mu}}_{N,j'',k''}$ by expanding around the old parameter $\boldsymbol{\mu}_{N,j'',k''}$ [8].

$$\hat{\boldsymbol{\mu}}_{O,j,k} = \boldsymbol{\mu}_{O,j,k} + \boldsymbol{H}_{N,j,k}\left(\hat{\boldsymbol{\mu}}_{N,j'',k''} - \boldsymbol{\mu}_{N,j'',k''}\right) \tag{22}$$

Even with the VTS approximation, the closed form solution of $\hat{\boldsymbol{\Sigma}}_{N,j'',k''}$ is still unavailable. Thus, $\hat{\boldsymbol{\Sigma}}_{N,j'',k''}$ is given by an approximation approach or a gradient-based approach, e.g., the Newton's method or the L-BFGS algorithm [27]. In this paper, we obtain the ML estimate of $\hat{\boldsymbol{\Sigma}}_{N,j'',k''}$ with the approximation described in [23].

3.5. Parameter update with MMSE estimates

To cope with the problem described in Sec. 3.4.2, we have already proposed a parameter estimation with MMSE estimates of the noise signals [11]. The use of MMSE estimates makes it possible to estimate accurate parameters for the noise model without any approximations. The MMSE estimates of $\boldsymbol{N}_t$ are given as

$$\begin{aligned} \tilde{\boldsymbol{N}}_t &= \mathcal{E}\left\{\boldsymbol{O}_t | \boldsymbol{\lambda}_O, \boldsymbol{\lambda}_N\right\} \\ &= \boldsymbol{O}_t + \sum_{j,k} \zeta_{O,t,j,k}\left(\boldsymbol{\mu}_{N,j'',k''} - \boldsymbol{\mu}_{O,j,k}\right) \ , \end{aligned} \tag{23}$$

where $\mathcal{E}\{\cdot|\cdot\}$ denotes the MMSE estimator for the noise signal in the LMFB domain.

With $\tilde{\boldsymbol{N}}_t$ and occupancy probability $\zeta_{N,t,j'',k''}$, $\hat{\boldsymbol{\mu}}_{N,j'',k''}$ and $\hat{\boldsymbol{\Sigma}}_{N,j'',k''}$ are given as

$$\hat{\boldsymbol{\mu}}_{N,j'',k''} = \frac{\sum_t \zeta_{N,t,j'',k''}\tilde{\boldsymbol{N}}_t}{\sum_t \zeta_{N,t,j'',k''}} \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}_{N,j'',k''} = \frac{\sum_t \zeta_{N,t,j'',k''}\tilde{\boldsymbol{N}}_t\tilde{\boldsymbol{N}}_t^{\mathrm{T}}}{\sum_t \zeta_{N,t,j'',k''}} - \hat{\boldsymbol{\mu}}_{N,j'',k''}\hat{\boldsymbol{\mu}}_{N,j'',k''}^{\mathrm{T}} \ . \tag{25}$$

4. Experiments

4.1. Processing flow

Algorithm 1 summarizes the proposed method.

4.2. Experimental setup

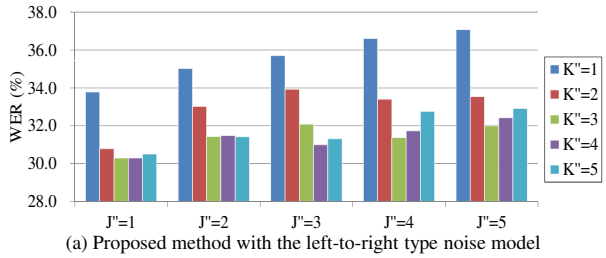
The experimental materials were 100 utterances of Japanese read speech spoken by 23 males. These utterances were taken from the Information-technology Promotion Agency (IPA)-98-TestSet. Three types of highly non-stationary noises, i.e., airport lobby noise, platform noise, and street noise, were artificially added to clean speech signals by changing the signal to

Table 2: ASR results of the overall conditions in WER (%)

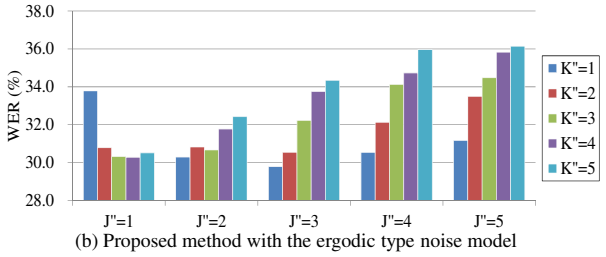
Method	Noise model		Airport lobby noise			Platform noise			Street noise			Avg.
	Type	Topology	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	
No processing	—	—	26.1	59.1	87.1	27.2	55.1	79.0	11.5	28.7	61.0	48.3
SS [6]	—	—	25.8	46.1	78.5	30.0	50.5	75.9	12.2	21.3	36.9	41.9
VTS [8]	Left-to-right	$J''=1, K''=3$	**13.6**	36.6	67.9	22.9	37.1	65.8	9.7	14.8	28.5	33.0
	Ergodic	$J''=1, K''=3$	**13.6**	36.6	67.9	22.9	37.1	65.8	9.7	14.8	28.5	33.0
Proposal	Left-to-right	$J''=1, K''=3$	13.8	**28.5**	62.0	19.4	37.0	61.0	9.2	**13.9**	28.0	30.3
	Ergodic	$J''=3, K''=1$	14.4	30.5	**59.9**	**18.1**	**35.8**	**59.3**	**8.7**	14.4	**26.9**	**29.8**

Algorithm 1 HMM-based noise model estimation

1: Feature extraction of $\boldsymbol{O}_t$ for all t
2: Initialize $\hat{\boldsymbol{\lambda}}_N$ (See Sec. 3.1.)
3: **repeat**
4: Model composition with function $\Psi(\cdot)$ (See Sec. 2.2.)
5: Compute expectation of cost function $Q(\cdot|\cdot)$ (See Sec. 3.2.)
6: Compute $\zeta_{N,t,j'',k''}$ and $\gamma_{N,t,i'',j''}$ (See Sec. 3.3.)
7: Estimate $\tilde{\boldsymbol{N}}_t$ for all t (See Sec. 3.5.)
8: Update $\hat{\boldsymbol{\lambda}}_N$ (See Secs. 3.4.1 and 3.5.)
9: **until** convergence is achieved
10: Apply noise suppression (See Sec. 2.3.)



(a) Proposed method with the left-to-right type noise model



(b) Proposed method with the ergodic type noise model

Figure 3: ASR results with the average WER for all the noise conditions and SNRs with the variable J'' and K''

noise ratio (SNR) at three levels; 10, 5, and 0 dB. The sampling frequency 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 clean speech model had $K' = 128$ Gaussian components. With the noise model, the number of states and the number of Gaussian components per a state were set at $J'' = 1, 2, 3, 4, 5$ and $K'' = 1, 2, 3, 4, 5$, respectively. The training materials for the clean speech HMMs were 33,820 phonetically balanced sentences spoken by 180 Japanese males. The parameter U was set at 10. In the Baum-Welch algorithm, the number of iterations was fixed at 100.

The ASR was carried out by employing a weighted finite state transducer-based decoder [28]. We used speaker independent HMMs trained by clean speech with the same training materials as those employed for the clean speech HMMs used in the noise suppression. The HMMs were trained with a variational Bayesian approach [29]. 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 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 %.

4.3. Experimental results

Fig. 3 shows the results obtained with the proposed method described in Sec. 3.5. Each result is given with the variable J'' and K''. In the figure, when the noise model is given by $J'' = 1$, the results with each method are equivalent to the results obtained with the GMM-based noise model. In this case, the parameters of the left-to-right model are exactly the same as those of the ergodic model.

When the left-to-right model is used for the noise model, the occupancy count of each noise model state tends to over-concentrate strongly in the final state of the noise model. In this case, the accuracy of the noise model largely depends on simply the parameter estimation accuracy of the final state, thus the state transition structure of the noise model becomes meaningless. This tendency may arise as a result of an insufficient quantity of training data i.e., a given one utterance. On the other hand, the ergodic model exhibits no such tendency, and accurately represents the state transition process of the noise. As seen in Fig. 3b, when the number of Gaussian components contained in the noise model $J''K''$ is the same, the WERs tend to improve by increasing the number of state J''. This tendency proves that the proposed method with the ergodic model can accurately model the state transition characteristics of the noise.

Table 2 shows the detailed ASR results provided by the best performance with each method. As seen in the table, the proposed method with the ergodic model improves the WERs with the HMM-based noise model of $J'' = 3$ states and $K'' = 1$ Gaussian component per state. Other results are given by the noise models of the $J'' = 1$ state and $K'' = 3$ Gaussian components per state, namely these models are equivalent to GMMs with $K'' = 3$ Gaussian components[1]. With these results, we can confirm the effectiveness of the proposed non-stationary noise modeling approach with ergodic HMMs.

5. Conclusions

This paper presented an unsupervised HMM parameter estimation method for the sufficient modeling of non-stationary noise. The evaluation results show that the proposed method further improves the performance of ASR. We plan to investigate an automatic optimization scheme for the model topology w.r.t. the numbers of states and Gaussian components.

[1] Since the results of VTS with each model are given by the GMM-based noise model, they all have the same value.

6. References

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

[2] K. Ishizuka and T. Nakatani, "A feature extraction method using subband based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1447–1457, November 2006.

[3] J. C. Segura, M.C. Benítez, A. de la Torre, A. M. Peinado, and A. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proc. of ICASSP '02*, May 2002, vol. I, pp. 401–404.

[4] V. Digalakis, D. Ritischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. on SAP*, vol. 3, no. 5, pp. 357–366, September 1995.

[5] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. of Interspeech '06*, September 2006, pp. 2286–2289.

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

[7] 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, December 1984.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. of ICASSP '96*, May 1996, vol. II, pp. 733–736.

[9] ETSI ES 202 050 v.1.1.4, *Speech processing, transmission and quality aspects (STQ), advanced distributed speech recognition; front-end feature extraction algorithm; compression algorithms*, November 2006.

[10] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A study of mutual front-end processing method based on statistical model for noise robust speech recognition," in *Proc. of Interspeech '09*, September 2009, pp. 1235–1238.

[11] M. Fujimoto, S. Watanabe, and T. Nakatani, "A robust estimation method of noise mixture model for noise suppression," in *Proc. of Interspeech '11*, August 2011, pp. 697–700.

[12] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. of ICASSP '90*, April 1990, pp. 845–848.

[13] 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.

[14] R. C. van Dalen and M. J. F Gales, "Extended VTS for noise-robust speech recognition," *IEEE Trans. on SAP*, vol. 19, no. 4, pp. 733–743, May 2011.

[15] 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, April 1995.

[16] O. Siohan, T. Myrvoll, and C. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, January 2002.

[17] S. Watanabe, A. Nakamura, and B. H. Juang, "Bayesian linear regression for hidden Markov model based on optimizing variational bounds," in *Proc. of MLSP '11*, December 2011, pp. 1–6.

[18] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. of ICASSP '02*, May 2002, pp. 57–60.

[19] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Communication*, vol. 50, no. 4, pp. 265–277, April 2008.

[20] D. Povey and G. Saon, "Feature and model space speaker adaptation with full covariance Gaussians," in *Proc. of Interspeech '06*, September 2006, pp. 1145–1148.

[21] S. Komeiji, T. Arakawa, and T Koshinaka, "A noise-robust speech recognition method composed of weak noise suppression and weak vector Taylor series adaptation," in *Proc. of SLT '12*, December 2012.

[22] S. J. Hahm, A. Ogawa, M. Delcroix, M. Fujimoto, T. Hori, and Nakamura. A, "Feature and model space adaptation based on variational Bayesian linear regression," in *Proc. of ICASSP '13*, May 2013.

[23] Y. Zhao and B. H. Juang, "A comparative study of noise estimation algorithms for VTS-based robust speech recognition," in *Proc. of Interspeech '10*, September 2010, pp. 2090–2093.

[24] K. K. Chin, H. Xu, M. J. F. Gales, C. Breslin, and K. Knill, "Rapid joint speaker and noise compensation for robust speech recognition," in *Proc. of ICASSP '11*, May 2011, pp. 5500–5503.

[25] Y.-Q. Wang and M. J. F. Gales, "Speaker and noise factorisation on the AURORA4 task," in *Proc. of ICASSP '11*, May 2011, pp. 4584–4587.

[26] M. Fujimoto and T. Nakatani, "A reliable data selection for model-based noise suppression using unsupervised joint speaker adaptation and noise model estimation," in *Proc. of ICSPCC '12*, August 2012, pp. 4713–4716.

[27] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming B*, vol. 45, pp. 503–528, 1989.

[28] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "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.

[29] S. Watanabe, Y. Minimi, A. Nakamura, and N. Ueda, "Variational Bayesian estimation and clustering for speech recognition," *IEEE Trans. on SAP*, vol. 12, no. 4, pp. 365–381, July 2004.