UNSUPERVISED NON-PARAMETRIC BAYESIAN MODELING OF NON-STATIONARY NOISE FOR MODEL-BASED NOISE SUPPRESSION

Masakiyo Fujimoto, Yotaro Kubo, and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Japan
{fujimoto.masakiyo, kubo.yotaro, nakatani.tomohiro}@lab.ntt.co.jp

ABSTRACT

The accurate modeling of non-stationary noise plays an important role in model-based noise suppression for noise robust speech recognition. We have already proposed methods for unsupervised noise modeling with a Gaussian mixture model or a hidden Markov model by using a minimum mean squared error estimate of the noise. However, our previous work fixed the structure of the noise model empirically without any consideration of noise characteristics; thus, optimization of the noise model structure is required if we are to obtain further improvements. Although the Bayesian information criterion (BIC) has been widely used as a conventional approach to model structure estimation, it is not always the optimal criterion. Therefore, this paper presents a way of modeling non-stationary noise with a non-parametric Bayesian approach that estimates the model structure depending on the characteristics of given observations. The proposed method provided improved results for the evaluations of two different speech recognition tasks compared with results obtained using the conventional BIC-based approach.

Index Terms— noise suppression, unsupervised modeling, non-parametric Bayesian model, MMSE estimation

1. INTRODUCTION

With the spread of speech applications including voice search, ensuring the noise robustness of automatic speech recognition (ASR) is becoming a more critical problem. Various techniques have been proposed for the front-end processing of ASR including robust feature extraction [1, 2], feature space normalization [3]-[5], and noise suppression [6]-[13]. Recently, a denoising autoencoder [14], which is based on a deep neural network approach, has attracted attention. On the other hand, back-end processing techniques have been proposed including model compensation [15, 16] and model adaptation [17]-[19] to reduce the mismatch between observed (noisy speech) signals and acoustic models. To exploit the uncertainty of noise, uncertainty decoding techniques have also been proposed [20, 21]. Of these various techniques, we have focused our research on the model-based noise suppression.

As a representative model-based noise suppression technique, a vector Taylor series (VTS)-based approach [8] and its various extensions have been proposed in recent years [16],[22]-[24]. The VTS-based approach compensates the model of the observed signal with models of clean speech and noise. Usually, the parameters of the clean speech model are fixed in the VTS-based approach; hence only the parameters of the noise model are estimated with the EM algorithm and the given observed signals. The typical VTS-based approach employs a single Gaussian distribution for the noise model. Since non-stationary noise has a multi-modal distribution and a temporal structure, a single Gaussian distribution is unsuitable for the model of the non-stationary noise. Therefore, a model with a complex structure, e.g., a Gaussian mixture model (GMM) or a hidden Markov model (HMM) is required to ensure robustness against non-stationary noise. In relation to this problem, we have recently proposed unsupervised estimation methods for a noise model with GMM [11] or HMM [12] by using minimum mean squared error (MMSE) estimates of the noise, and some extensions [13].

In our previous studies, the structure of the noise model, i.e., the numbers of Gaussian components and HMM states, is empirically decided through preliminary experiments, and is fixed regardless of the noise characteristics. Since the characteristics of non-stationary noise change on different occasions, the structure of the noise model should be adequately estimated by depending on the given observed signal. Instead of empirical rules, the Bayesian information criterion (BIC) has been widely used for the model structure selection. However, as described in [25], the BIC is not always the optimal criterion, and should not be strictly applied to a structured model including any latent variables. In addition, since the BIC-based approach requires various models learned with different structures to select a suitable model, its computational cost tends to be high.

To deal with this problem, a non-parametric Bayesian approach, which allows a model structure with infinite components, e.g., an infinite GMM (IGMM) or an infinite HMM (IHMM), has attracted attention in machine learning research. This approach considers that finite data are provided by a generative model with infinite components in the data generation process. Therefore, the optimal number of model components is decided by estimating the latent variable that indicates the generating model components of the given finite data. On the basis of the above considerations, we investigate the application of a non-parametric Bayesian approach to our unsupervised method for noise model estimation. As the first step of this investigation, in this paper, we present a method for unsupervised noise modeling with an IGMM. With the non-parametric Bayesian approach, the IGMM is realized by a Dirichlet process mixture (DPM) [26] which is equivalent to an infinite dimensional Dirichlet distribution. In this paper, we implement the DPM with an efficient sampling-based algorithm, thus we can provide accurate IGMM-based noise modeling at a low computational cost.

The proposed method was evaluated on two ASR tasks; the AURORA2 task [27] and the Japanese large vocabulary task. The evaluation results reveal that the proposed IGMM-based noise model successfully improves the ASR accuracies of both tasks in results obtained by conventional model selection based on the BIC.

2. REVIEW OF MODEL-BASED NOISE SUPPRESSION

This section briefly reviews our previous work [11]-[13].

2.1. Definition of statistical models

In our implementation, the clean speech model is given by an ergodic HMM with two internal states, i.e., states of silence ($j = 1$) and speech ($j = 2$), where j denotes the state index. Each state is modeled in advance by a GMM with K Gaussian components in the D-dimensional log mel-filter bank (LMFB) domain, and

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

has model parameters that consist of the mixture weight $w_{S,j,k}$, the mean vector $\boldsymbol{\mu}_{S,j,k} \triangleq \{\mu_{S,j,k,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{S,j,k} \triangleq \mathrm{diag}\{\sigma_{S,j,k,d}\}_{d=0}^{D-1}$. k and d denote the indices of the Gaussian component and the element of a vector or a diagonal component of a matrix. On the other hand, the noise model is given by a GMM with L Gaussian components in the LMFB domain, and has model parameters that consist of the mixture weight $w_{N,l}$, the mean vector $\boldsymbol{\mu}_{N,l} \triangleq \{\mu_{N,l,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{N,l} \triangleq \mathrm{diag}\{\sigma_{N,l,d}\}_{d=0}^{D-1}$. l denotes the Gaussian index.

2.2. Mismatch function and model composition

At the t-th frame, the LMFB vector of the observed signal $\boldsymbol{O}_t \triangleq \{O_{t,d}\}_{d=0}^{D-1}$ is derived by using the following mismatch function with the LMFB vectors of the clean speech $\boldsymbol{S}_t \triangleq \{S_{t,d}\}_{d=0}^{D-1}$ and the noise $\boldsymbol{N}_t \triangleq \{N_{t,d}\}_{d=0}^{D-1}$.

$$O_{t,d} = S_{t,d} + \log\left(1 + \exp\left(N_{t,d} - S_{t,d}\right)\right) \equiv h\left(S_{t,d}, N_{t,d}\right) \quad (1)$$

Based on Eq. (1), the parameters of the observed signal model, which consist of the mixture weight $w_{O,j,k,l}$, the mean vector $\boldsymbol{\mu}_{O,j,k,l} \triangleq \{\mu_{O,j,k,l,d}\}_{d=0}^{D-1}$, and the diagonal variance matrix $\boldsymbol{\Sigma}_{O,j,k,l} \triangleq \mathrm{diag}\{\sigma_{O,j,k,l,d}\}_{d=0}^{D-1}$ are derived as

$$w_{O,j,k,l} = w_{S,j,k,l} \cdot w_{N,l} \quad (2)$$

$$\mu_{O,j,k,l,d} = h\left(\mu_{S,j,k,d}, \mu_{N,l,d}\right) \quad (3)$$

$$\sigma_{O,j,k,l,d} \simeq H_{j,k,l,d}^2 \cdot \sigma_{S,j,k,d} + \left(1 - H_{j,k,l,d}\right)^2 \cdot \sigma_{N,l,d}\ , \quad (4)$$

with the Jacobian $H_{j,k,l,d} = \partial h\left(\mu_{S,j,k,d}, \mu_{N,l,d}\right) / \partial \mu_{S,j,k,d}$.

2.3. Parameter estimation of noise model with MMSE estimates

The parameters of the noise model are estimated by using the EM algorithm [13] with the MMSE estimate of the noise $\tilde{\boldsymbol{N}}_t$ derived as:

$$\tilde{\boldsymbol{N}}_t = \boldsymbol{O}_t + \sum_{j,k,l} P_{t,j,k,l}\left(\boldsymbol{\mu}_{N,l} - \boldsymbol{\mu}_{O,j,k,l}\right)\ , \quad (5)$$

with posterior probability

$$P_{t,j,k,l} = \frac{w_{O,j,k,l}\mathcal{N}\left(\boldsymbol{O}_t \left| \boldsymbol{\mu}_{O,j,k,l}, \boldsymbol{\Sigma}_{O,j,k,l}\right.\right)}{\sum_{j,k,l} w_{O,j,k,l}\mathcal{N}\left(\boldsymbol{O}_t \left| \boldsymbol{\mu}_{O,j,k,l}, \boldsymbol{\Sigma}_{O,j,k,l}\right.\right)}\ , \quad (6)$$

where $\mathcal{N}(\cdot|\cdot)$ denotes the probability density function (PDF) of the Gaussian distribution.

2.4. MMSE-MAP estimation of clean speech

We employ an MMSE-maximum *a posteriori* (MAP) estimation for noise suppression [12]. This method first estimates the clean speech $\tilde{\boldsymbol{\mu}}_{S,t,j,k,l}$ for each Gaussian component k and l in state j at frame t with the MAP criterion. With the MAP estimates $\tilde{\boldsymbol{\mu}}_{S,t,j,k,l}$, the clean speech $\tilde{\boldsymbol{S}}_t$ is estimated with the MMSE manner of Eq. (7).

$$\tilde{\boldsymbol{S}}_t = \sum_{j,k,l} P_{t,j,k,l} \cdot \tilde{\boldsymbol{\mu}}_{S,t,j,k,l} \quad (7)$$

3. INFINITE GAUSSIAN MIXTURE MODEL

3.1. Conventional finite GMM with Dirichlet distribution

The data generation process of a finite GMM (FGMM) with fixed L Gaussian components is derived as Eqs. (8) and (9). The parameters of Gaussian components $\boldsymbol{\Theta}_N \triangleq \{\boldsymbol{\theta}_{N,l}\}_{l=1}^{L}$, where $\boldsymbol{\theta}_{N,l} \triangleq \left\{\boldsymbol{\mu}_{N,l}, \boldsymbol{\Sigma}_{N,l}\right\}$ and mixture weights $\boldsymbol{w}_N \triangleq \{w_{N,l}\}_{l=1}^{L}$ are drawn from the base measure G_0 and the Dirichlet distribution $\mathcal{D}\left(\cdot\,|\gamma/L\right)$ with the concentration parameter γ, respectively. The latent variable z_t, which indicates the generating Gaussian component of the data $\boldsymbol{N}_t$, is decided by the multinomial distribution $\mathcal{M}\left(\cdot\,|\boldsymbol{w}_N\right)$, then, $\boldsymbol{N}_t$ is generated from the Gaussian component $\mathcal{N}\left(\cdot\,|\boldsymbol{\theta}_{N,l=z_t}\right)$.

$$\boldsymbol{\Theta}_N \sim G_0\ ,\ \boldsymbol{w}_N \sim \mathcal{D}\left(\cdot\,|\gamma/L\right) \quad (8)$$

$$z_t\,|\boldsymbol{w}_N \sim \mathcal{M}\left(\cdot\,|\boldsymbol{w}_N\right)\ ,\ \boldsymbol{N}_t\,|\boldsymbol{\theta}_{N,l=z_t} \sim \mathcal{N}\left(\cdot\,|\boldsymbol{\theta}_{N,l=z_t}\right) : \forall t \quad (9)$$

G_0 is given by the following Gaussian-Gamma distribution.

$$G_0\left(\boldsymbol{\Theta}_N|\boldsymbol{\Theta}_N^{(0)}\right) = \prod_d \mathcal{N}\left(\mu_{N,l,d}\left|\mu_{N,d}^{(0)}, \sigma_{N,d}/\xi^{(0)}\right.\right) \times \mathcal{G}\left(\sigma_{N,l,d}^{-1}\left|\eta^{(0)}, r_{N,d}^{(0)}\right.\right)\ , \quad (10)$$

where the parameter $\boldsymbol{\Theta}_N^{(0)}$ consists of the mean $\mu_{N,d}^{(0)}$, the precision $\xi^{(0)}$, the shape $\eta^{(0)}$, and the scale $r_{N,d}^{(0)}$. $\mathcal{G}(\cdot|\cdot)$ denotes the PDF of the Gamma distribution.

With the FGMM, L is decided by an empirical rule or an information criterion, e.g., the BIC, as follows:

$$L = \arg\min_L \left\{-2\sum_t \log p\left(\boldsymbol{N}_t\,|\boldsymbol{w}_N, \boldsymbol{\Theta}_N\right) + q\log T\right\}\ , \quad (11)$$

where $p\left(\boldsymbol{N}_t\,|\boldsymbol{w}_N, \boldsymbol{\Theta}_N\right)$ denotes the likelihood of the model with a certain number of Gaussian components. $q \propto L$ and T denote the number of model parameters and data, respectively.

3.2. Infinite GMM with Dirichlet process mixture

The IGMM employs the DPM, which is equivalent to a infinite dimensional Dirichlet distribution for the prior distribution of mixture weights. In the data generation process with the DPM, although the IGMM consists of infinite Gaussian components, the generated data are restricted to a finite number. Thus, the number of Gaussian components that generate the finite data must also be finite. By solving the inverse problem of this data generation process, the number of Gaussian components L is decided flexibly by depending on the characteristics of the given finite data.

The DPM is often implemented by using the Chinese restaurant process (CRP), because it has the potential to avoid a local solution with Gibbs sampling. The CRP has exchangeability, namely the joint distribution of the latent variable z_t $P\left(z_0, \cdots, z_{T-1}\right) = P\left(z_0\right)P\left(z_1\,|z_0\right)\cdots P\left(z_{T-1}\,|z_0, \cdots, z_{T-2}\right)$ is invariant even if z_t is exchanged on any t. With this property, we can easily implement the CRP-based IGMM estimation with the Gibbs sampling.

The data generation process of the CRP-based IGMM is derived as Eqs. (12) and (13). The latent variable z_t is drawn from the posterior distribution $P(\cdot\,|\boldsymbol{z}_{\setminus t})$, where $\boldsymbol{z}_{\setminus t} \triangleq \{z_i : \forall i, i \neq t, \}$.

$$\boldsymbol{\theta}_{N,l} \sim G_0 : l \in \{1, \cdots, \infty\} \quad (12)$$

$$z_t \sim P(\cdot\,|\boldsymbol{z}_{\setminus t})\ ,\ \boldsymbol{N}_t \sim \mathcal{N}\left(\cdot\,|\boldsymbol{\theta}_{N,l=z_t}\right) : \forall t \quad (13)$$

With the above definitions, the posterior probability of $z_t = l$, given all data $\boldsymbol{N} \triangleq \{\boldsymbol{N}_t\}_{t=0}^{T-1}$, $\boldsymbol{z}_{\setminus t}$, and $\boldsymbol{\Theta}_N^0$, which is required for Gibbs sampling, is derived as

$$P\left(z_t = l\left|\boldsymbol{N}, \boldsymbol{z}_{\setminus t}, \boldsymbol{\Theta}_N^{(0)}\right.\right) \propto P\left(z_t = l\,|\boldsymbol{z}_{\setminus t}\right)p\left(\boldsymbol{N}_t\left|\boldsymbol{N}_{\setminus t}, z_t = l, \boldsymbol{\Theta}_N^{(0)}\right.\right)\ , \quad (14)$$

where $\boldsymbol{N}_{\setminus t} \triangleq \{\boldsymbol{N}_i : \forall i, i \neq t\}$.

With the FGMM, the posterior probability $P\left(z_t = l\left|\boldsymbol{z}_{\setminus t}\right.\right)$ of Eq. (14) is derived as

$$P\left(z_t = l|\boldsymbol{z}_{\setminus t}\right) = \int P\left(z_t = l\,|\boldsymbol{w}_N\right)p\left(\boldsymbol{w}_N\,|\boldsymbol{z}_{\setminus t}\right)d\boldsymbol{w}_N = \left(n_l + \gamma/L\right)/\left(T + \gamma\right)\ , \quad (15)$$

where n_l denotes the number of data assigned to the cluster l. On the other hand, with the CRP, $P\left(z_t = l|\boldsymbol{z}_{\setminus t}\right)$ is given by taking the limit of L, i.e., $L \to \infty$ in Eq. (15). In addition, $P\left(z_t = l|\boldsymbol{z}_{\setminus t}\right)$ is separately computed for an existing cluster, i.e., $n_l > 1$ and for a new cluster, i.e., $n_l = 0$. Thus, $P\left(z_t = l|\boldsymbol{z}_{\setminus t}\right)$ is derived as

$$P\left(z_t = l|\boldsymbol{z}_{\setminus t}\right) = \begin{cases} n_l/(T+\gamma) & \text{if } z_i = l, \text{ for } \exists i \neq t \\ \gamma/(T+\gamma) & \text{if } z_i \neq l, \text{ for } \forall i \neq t \end{cases}. \quad (16)$$

The posterior probability $p\left(\boldsymbol{N}_t|\boldsymbol{N}_{\setminus t}, z_t = l, \boldsymbol{\Theta}_N^{(0)}\right)$ of Eq. (14) is also computed separately as follows:

$$p\left(\boldsymbol{N}_t|\boldsymbol{N}_{\setminus t}, z_t = l, \boldsymbol{\Theta}_N^{(0)}\right) = \begin{cases} \int p\left(\boldsymbol{N}_t|\boldsymbol{\theta}_{N,l}\right) p\left(\boldsymbol{\theta}_{N,l}|\boldsymbol{N}_{\setminus t}^{(l)}, \boldsymbol{\Theta}_N^{(0)}\right) d\boldsymbol{\theta}_{N,l} \\ = \prod_d \mathcal{T}\left(N_{t,d}\,|\mu_{N,l,d}, \xi_l, \eta_l, r_{N,l,d}\right) \\ \quad \text{if } z_i = l, \text{ for } \exists i \neq t \\ \int p\left(\boldsymbol{N}_t|\boldsymbol{\Theta}_N\right) G_0\left(\boldsymbol{\Theta}_N|\boldsymbol{\Theta}_N^{(0)}\right) d\boldsymbol{\Theta}_N \\ = \prod_d \mathcal{T}\left(N_{t,d}\left|\mu_{N,d}^{(0)}, \xi^{(0)}, \eta^{(0)}, r_{N,d}^{(0)}\right.\right) \\ \quad \text{if } z_i \neq l, \text{ for } \forall i \neq t \end{cases}, \quad (17)$$

where $\boldsymbol{N}_{\setminus t}^{(l)} \triangleq \{\boldsymbol{N}_i : \forall i, i \neq t, z_i = l\}$. $\mathcal{T}(\cdot|\cdot)$ denotes the PDF of the Student's t-distribution given by Eq. (18). $\mu_{N,l,d}$, ξ_l, η_l, and $r_{N,l,d}$ denote the mean, the precision, the shape, and the scale of the posterior distribution of the cluster l, respectively.

$$\mathcal{T}\left(x\,|\mu, \xi, \eta, r\right) \equiv \left(\frac{\xi}{2\pi r(\xi+1)}\right)^{\frac{1}{2}} \frac{\Gamma\left(\eta+\frac{1}{2}\right)}{\Gamma\left(\eta\right)} \left(1 + \frac{\xi\left(x-\mu\right)^2}{2r(\xi+1)}\right)^{-(\eta+\frac{1}{2})} \quad (18)$$

4. CRP-BASED NOISE MODEL ESTIMATION

We attempt to estimate the IGMM-based noise model with the CRP.

4.1. Initialization

The parameter of the base measure $\boldsymbol{\Theta}_N^{(0)}$ is given empirically as:

$$\boldsymbol{\Theta}_N^{(0)} = \left\{ \left\{\mu_{N,d}^{(0)}\right\}_{d=0}^{D-1}, \xi^{(0)}, \eta^{(0)}, \left\{r_{N,d}^{(0)}\right\}_{d=0}^{D-1} \right\}, \quad (19)$$

where $\mu_{N,d}^{(0)} = \frac{1}{U}\sum_{t=0}^{U-1} O_{t,d}$, $\xi^{(0)} = 1$, $\eta^{(0)} = 1$, $r_{N,d}^{(0)} = \eta^{(0)} \cdot \sigma_{N,d}^{(0)}$, $\sigma_{N,d}^{(0)} = \frac{1}{U}\sum_{t=0}^{U-1}\left(O_{t,d} - \mu_{N,d}^{(0)}\right)^2$, and $U = 10$.

Based on Eqs. (2) to (4), the parameters of the observed signal model are composed with $\mu_{N,l=1,d} = \mu_{N,d}^{(0)}$, $\sigma_{N,l=1,d} = \sigma_{N,d}^{(0)}$, and $w_{N,l=1} = 1$. Then, $\tilde{\boldsymbol{N}}_t$ is estimated by the MMSE manner of Eqs. (5) and (6). With $\tilde{\boldsymbol{N}}_t$, zero-th, first, and second order sufficient statistics of the cluster $l = 1$ are given as

$$s_{0,l=1} = T, \; s_{1,l=1,d} = \sum_t \tilde{N}_{t,d}, \; s_{2,l=1,d} \sum_t \tilde{N}_{t,d}^2. \quad (20)$$

4.2. Gibbs sampler

First, $\tilde{\boldsymbol{N}}_t$ is cancelled from the sufficient statistics of the cluster $l = z_t$. Here, although the sufficient statistics are computed by using $\tilde{\boldsymbol{N}}_t$ at the previous iteration, $\tilde{\boldsymbol{N}}_t$ is updated in each iteration of the Gibbs sampler. Thus, $\tilde{\boldsymbol{N}}_t$ at the previous iteration is kept in $\tilde{\boldsymbol{N}}_{old,t}$, and it is cancelled from the sufficient statistics based on Eq. (21).

$$s_{0,l=z_t} \leftarrow s_{0,l=z_t} - 1, \; s_{1,l=z_t,d} \leftarrow s_{1,l=z_t,d} - \tilde{N}_{old,t,d}$$
$$s_{2,l=z_t,d} \leftarrow s_{2,l=z_t,d} - \tilde{N}_{old,t,d}^2 \quad (21)$$

With the suffcient statistics, the posterior statistics of the cluster l are updated by Eqs. (22) to (25), and the posterior probability of Eq. (14) is computed by using them.

$$\mu_{N,l,d} = \left(\xi^{(0)} \cdot \mu_{N,d}^{(0)} + s_{1,l,d}\right)/\xi_l \quad (22)$$
$$\xi_l = \xi^{(0)} + s_{0,l} \quad (23)$$
$$\eta_l = \eta^{(0)} + s_{0,l}/2 \quad (24)$$
$$r_{N,l,d} = r_{N,d}^{(0)} + \left(s_{2,l,d} + \xi^{(0)} \cdot \left(\mu_{N,d}^{(0)}\right)^2 - \xi_l \cdot \mu_{N,l,d}^2\right)/2 \quad (25)$$

The latent variable z_t is sampled by Eq. (26), then $\tilde{N}_{t,d}$ is added to the sufficient statistics of the cluster $l = z_t$ as shown in Eq. (27)

$$z_t \sim P\left(z_t = l\left|\boldsymbol{N}, \boldsymbol{z}_{\setminus t}, \boldsymbol{\Theta}_N^{(0)}\right.\right) \quad (26)$$
$$s_{0,l=z_t} \leftarrow s_{0,l=z_t} + 1, \; s_{1,l=z_t,d} \leftarrow s_{1,l=z_t,d} + \tilde{N}_{t,d}$$
$$s_{2,l=z_t,d} \leftarrow s_{2,l=z_t,d} + \tilde{N}_{t,d}^2 \quad (27)$$

4.3. Parameter update

The parameter of the noise model $\boldsymbol{\Theta}_N$ is updated by taking the expectation of the posterior distribution $p\left(\boldsymbol{\theta}_{N,l}|\boldsymbol{N}_{\setminus t}^{(l)}, \boldsymbol{\Theta}_N^{(0)}\right)$ of Eq. (17) derived as Eq. (28). The mixture weight is also updated by Eq. (29). With the updated $\boldsymbol{\Theta}_N$ and $\boldsymbol{w}_N$, $\tilde{\boldsymbol{N}}_t$ is updated with the MMSE manner of Eqs. (5) and (6) after the model composition of Eqs. (2) to (4).

$$\boldsymbol{\theta}_{N,l} = \left\{\left\{\mu_{N,l,d}\right\}_{d=0}^{D-1}, \mathrm{diag}\left\{r_{N,l,d}/\eta_l\right\}_{d=0}^{D-1}\right\} \quad (28)$$
$$w_{N,l} = s_{0,l}/T \quad (29)$$

4.4. Processing flow

Algorithm 1 summarizes the proposed method.

Algorithm 1 IGMM-based noise model estimation with CRP

```
Feature extraction of O_t for all t
Estimate Θ_N^(0) (Eq. (19))
Model composition (Eqs. (2) to (4))
Estimate Ñ_t for all t (Eqs. (5) and (6))
Ñ_{old,t} = Ñ_t for all t
Initialize sufficient statistics (Eq. (20))
for i = 1 to Ite do
    for t = shuffle (0, ··· , T − 1) do
        Cancel Ñ_{old,t} from cluster z_t (Eq. (21))
        Update posterior parameters (Eqs. (22) to (25))
        Compute posterior probability (Eq. (14))
        Sample z_t (Eq. (26))
        Add Ñ_t to cluster z_t (Eq. (27))
    end for
    Ñ_{old,t} = Ñ_t for all t
    Update Θ_N and w_n (Eqs. (28) and (29))
    Model composition (Eqs. (2) to (4))
    Estimate Ñ_t for all t (Eqs. (5) and (6))
end for
Apply noise suppression (Eq. (7))
```

5. EXPERIMENTS

The proposed method was evaluated in two ASR tasks. For these evaluations, we mainly compared the previous FGMM [11], the FGMM with the BIC, and the proposed IGMM. These evaluations essentially compared each technique for deciding the number of Gaussian components L, i.e., the empirical rule, the BIC, and the non-parametric Bayesian approach.

Table 1. ASR results of the AURORA2 task with the clean acoustic model in the average WER (%)

Data set	Set A (development set)							Set B (evaluation set)						
SNR	20dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	20dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline	2.2	7.1	25.5	58.9	81.1	88.2	43.8	1.7	5.1	17.8	47.9	76.2	87.2	39.3
VTS	1.6	3.6	8.8	20.9	44.8	72.9	25.4	1.4	3.3	7.8	20.1	43.8	72.0	24.7
FGMM ($L = 2$)	1.2	2.3	6.6	17.3	41.4	70.8	23.3	**1.0**	2.1	5.6	16.4	40.2	69.2	22.4
FGMM w/ BIC	**1.1**	**2.2**	6.5	17.1	41.4	71.0	23.2	1.1	2.1	6.1	16.8	40.4	69.2	22.6
IGMM	1.3	2.4	**5.9**	**16.1**	**39.4**	**68.9**	**22.3**	**1.0**	**2.0**	**5.3**	**15.0**	**37.5**	**67.3**	**21.4**

Table 2. ASR results of the large vocabulary task with the clean acoustic model in WER (%)

Noise type	Airport lobby noise			Platform noise			Street noise			Avg.
SNR	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	
Baseline	26.1	59.1	87.1	27.2	55.1	79.0	11.5	28.7	61.0	48.3
VTS	17.3	38.6	71.7	23.4	44.2	70.0	**7.4**	15.5	30.0	35.3
FGMM ($L = 2$)	12.5	**27.9**	62.5	18.9	36.7	60.2	7.6	13.3	27.8	29.7
FGMM w/ BIC	12.8	29.2	61.0	**17.0**	34.0	60.4	8.2	14.1	25.9	29.2
IGMM	**12.2**	29.3	**60.8**	17.3	**31.0**	**57.4**	8.1	**12.3**	**25.4**	**28.2**

5.1. Experiments with small vocabulary task

5.1.1. Experimental setup

We firstly evaluated the proposed method on the AURORA2 task [27]. AURORA2 consists of three evaluation sets, i.e., set A (four types of known additive noises), set B (four types of unknown additive noises), and set C (one known and one unknown additive noises with the different channel characteristic). In this evaluation, sets A and B were used for the development set and the evaluation set, respectively.

The feature parameters for the noise suppression were 24 LMFBs that were extracted by using a Hamming window with a 25 ms frame length and a 10 ms frame shift. The clean speech model was trained by using AURORA2 clean training data. Each state of the model had $K = 256$ Gaussian components.

The HTK version 3.4.1 [28] was used for the training and evaluation. Sixteen-state left-to-right word HMMs were trained by using AURORA2 clean training data. Each state had 20 Gaussian components. The feature parameters for the ASR consisted of 13 MFCCs (including the zero-th MFCC) and their first and second order derivatives. Cepstral mean normalization (CMN) was applied to each utterance. The evaluation criterion was the word error rate (WER).

5.1.2. Experimental results

Table 1 shows the average WER of each set and each method in the AURORA2 task. The parameters of each method were adjusted by using the development set. The results of a previous method "FGMM," were obtained with $L = 2$. With "FGMM w/ BIC," the optimal number of L for each utterance was selected from the noise models with $L = 1, \cdots, 10$ based on the criterion of Eq. (11). The parameters Ite and γ used in the proposed method "IGMM" were set at 10 and 0.04, respectively.

As seen in the table, the proposed method showed the best results of the development set for come conditions and of the evaluation set for all conditions. In the results of the BIC-based approach, no improvements were obtained compared with the previous method due to the inapplicability of the BIC to the structured model described in [25].

The computational costs of the BIC-based approach were much greater than those of the proposed method, because various structured models must be learned for the model section with the BIC. In contrast, the proposed method is able to decide the model structure uniquely without considering the possibility of various model structures. Therefore, the computational cost of the proposed method is much less than that of the BIC-based approach.

From the viewpoints of both the WER and the computational cost, we can confirm the effectiveness of the proposed non-stationary noise modeling technique with the non-parametric Bayesian approach.

5.2. Experiments with large vocabulary task

5.2.1. Experimental setup

The second evaluation employed a large vocabulary task with 20k words. The evaluation data were the IPA-98-TestSet which consists of 100 Japanese utterances spoken by 23 males. Three types of highly non-stationary noise, namely, airport lobby, platform, and street noise, were artificially added to clean speech with three SNR 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 ms frame length and a 10 ms frame shift. The training data for the clean speech model were 33,820 phonetically balanced sentences spoken by 180 Japanese males. Each state of the model had $K = 256$ Gaussian components.

The ASR was carried out by employing a weighted finite state transducer-based decoder [29]. The three-state left-to-right triphone HMMs were trained by clean speech with the same training data employed for the clean speech model used for the noise suppression. There were 2,364 states in total. Each state had 16 Gaussian components. The feature parameters for the ASR consisted of 12 MFCCs and the log energy with their first and second order derivatives. CMN was applied to each utterance. The language model was a back-off tri-gram with Witten-Bell discounting. The model was trained using 75 months' worth of Japanese newspaper articles. The WER of a clean speech signal was 3.9 %.

5.2.2. Experimental results

Table 2 shows the detailed ASR results of the large vocabulary task. The parameters of each method were the same in Sec. 5.1. As seen in the table, the proposed method also achieved the best average WER even in a different task. These results prove that the proposed method is robust for unknown noise environments, because it is insensitive to changes in the ASR task and the noise environment.

6. CONCLUSIONS

This paper presented an unsupervised model estimation method based on a non-parametric Bayesian approach. The proposed method consists of the MMSE estimation of noise and a CRP-based IGMM estimation, and it automatically decides the number of Gaussian components depending on the characteristics of the given data. The evaluation results showed that the proposed method achieves superior performance to the conventional technique of model structure selection with the BIC. In future, we plan to investigate the expansion of this approach to an IHMM-based noise model with the hierarchical Dirichlet process [30].

7. 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] M. Fujimoto and T. Nakatani, "Model-based noise suppression using unsupervised estimation of hidden Markov model for non-stationary noise," in *Proc. of Interspeech '13*, August 2013, pp. 2982–2986.

[13] 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.

[14] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. of Interspeech '13*, August 2013, pp. 436–440.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] S. Watanabe, "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory," *Journal of Machine Learning Research*, vol. 11, pp. 3571–3591, December 2010.

[26] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, March 1973.

[27] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," in *Proc. ISCA ITRW ASR'00*, September 2000, pp. 18–20.

[28] The hidden Markov model toolkit: http://htk.eng.cam.ac.uk/

[29] 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.

[30] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, December 2006.