AUTOMATIC GENERATION OF NON-UNIFORM HMM STRUCTURES BASED ON VARIATIONAL BAYESIAN APPROACH

Takatoshi Jitsuhiro, Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, "Keihanna Science City", Kyoto 619-0288, Japan
{takatoshi.jitsuhiro, satoshi.nakamura}@atr.jp

ABSTRACT

We propose using the Variational Bayesian (VB) approach for automatically creating non-uniform, context-dependent HMM topologies. The Maximum Likelihood (ML) criterion is generally used to create HMM topologies. However, it has an over-fitting problem. Information criteria have been used to overcome this problem, but theoretically they cannot be applied to complicated models like HMMs. Recently, to avoid these problems, the VB approach has been developed in the machine-learning field. We introduce the VB approach to the Successive State Splitting (SSS) algorithm, which can create both contextual and temporal variations for HMMs. We define the prior and posterior probability densities and free energy with latent variables as split and stop criteria. Experimental results show that the proposed method can automatically create a more efficient model and obtain better performance, especially for vowels, than the original method.

1. INTRODUCTION

In creating acoustic models of speech recognition, two kinds of training are needed. One is designing the structures of HMMs and the other is parameter estimation. The latter simply means the EM algorithm. The former includes finding the optimal structure of a model for the training database. A crucial is how to design the optimal model from training data automatically. Phonetic decision tree clustering[1] is generally used as a method of generating tied-state structures of acoustic models. Furthermore, the Maximum Likelihood Successive State Splitting (ML-SSS) algorithm[2] was proposed to create contextual and temporal variations. These methods originally used the Maximum Likelihood (ML) criterion to choose the phonetic question with which each state was split. However, owing to the nature of ML estimation, the ML criterion often causes a model that over-fits the training data. The likelihood value for training data increases as the number of parameters increases. Consequently, it is impossible to stop splitting by only using the ML criterion. Methods based on the ML criterion require heuristic stop criteria, such as the total number of states.

Information criteria such as the Minimum Description Length (MDL) criterion and the Bayesian Information Criterion (BIC) have been introduced as splitting and stop criteria in creating context-dependent Hidden Markov Models (HMM); such methods have used phonetic decision tree clustering[3] or the Successive State Splitting (SSS) algorithm[4]. These methods continue to split states in order to improve the information criteria. Conventional information criteria require some assumptions, e.g., asymptotic normality, but complicated models like neural networks, or HMMs, cannot satisfy such assumptions. Although methods of creating HMM structures using information criteria can be considered rough approximations, they can work well in practical terms.

In the field of machine learning, the Variational Bayesian (VB) method was proposed to avoid over-fitting by ML estimation[5]. It incorporates the variational approximation technique in Bayesian inference. In addition, a general VB framework that can perform optimal model selection was proposed[6], and a method to obtain the optimal model structure by the VB framework was proposed as a way to avoid the local optimal problem for a mixture of experts model[7]. For speech recognition, decision tree clustering with the VB method was proposed[8]. Latent variables are one of the key points in the VB framework. Although they define the VB approach for HMMs including latent variables, their method for making HMM structures does not need latent variables because the alignments of states are given.

We propose an automatic topology creation method using the SSS algorithm with the Variational Bayesian method to estimate topologies more exactly. The SSS algorithm can create contextual and temporal variations. On the other hand, decision tree clustering can only create contextual variations. Furthermore, latent variables should be used in the SSS algorithm because the alignments of phonemes are fixed but those of states are not. Therefore, occupancy probabilities of training samples should be considered by using latent variables to introduce the VB method into the SSS algorithm.

In Section 2, we present the ML-SSS algorithm as the baseline method. Next, the VB approach for the SSS algorithm is described in Section 3. In Section 4, we evaluate the performance of the VB-SSS by segmented phoneme recognition and continuous speech recognition. Finally, we offer our conclusions in Section 5.

2. ML-SSS ALGORITHM

The ML-SSS algorithm combines contextual and temporal splitting to create a complicated structures of shared-state HMMs from a small initial model[2]. First, the contextual splitting and temporal splitting are performed for all states. Second, the gains of both contextual and temporal splitting are calculated. Finally, these expected gains are compared with each other, and the state with the best gain among all states is selected. The ML-SSS needs the total number of states, N_s, and the maximum length of state sequences for allophones, N_p. These parameters must be given before starting the splitting, but it is generally difficult to find their optimal values. Experiments need to be conducted to find the optimal values.

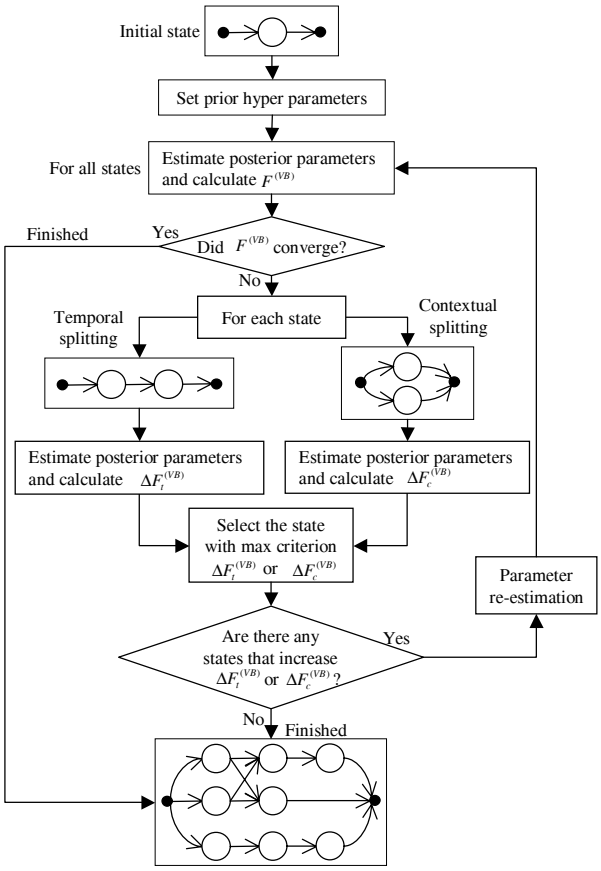

Fig. 1. Flow of the Variational Bayesian SSS algorithm.

3. VARIATIONAL BAYESIAN APPROACH FOR SSS ALGORITHM

3.1. Overview of VB-SSS

Figure 1 shows the flow of the VB-SSS algorithm. First, the topology of an initial model is set and its parameters are estimated. Second, the prior hyper-parameters for each state are set. Next, the posterior hyper-parameters for each state are estimated, and the free energy, $\mathcal{F}^{(VB)}$, is calculated as the baseline energy.

After that, each kind of splitting is done in the same manner as with the ML-SSS algorithm[2]. For each splitting, after two new states are created, the posterior hyper-parameters are estimated, and the energy gains, $\Delta\mathcal{F}_c^{(VB)}$ for the contextual splitting and $\Delta\mathcal{F}_t^{(VB)}$ for the temporal splitting, are calculated. Next, the state splitting with the maximum energy gain is selected. If there is no state that can increase the energy gain, the splitting is stopped. Furthermore, when $\mathcal{F}^{(VB)}$ decreases or converges, the splitting is stopped. Otherwise, the parameters of HMMs are estimated, and these procedures are repeated. In this paper, all of the posterior hyper-parameters are estimated by using all of the data for each test splitting.

3.2. Contextual and Temporal Splitting

The probability density of the HMM Θ, which has N_s states with one Gaussian distribution and N_a transitions for each state for both contextual and temporal splitting, is

$$p(\boldsymbol{O}|\Theta) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{S}_{s_t}^{-1}) a_{s_t r_{t+1}}, \tag{1}$$

where $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_t, \ldots, \boldsymbol{o}_T\}$ is a set of training samples, s_t is the state number at time t, and r_t is the transition arc number at time t. $\boldsymbol{\mu}_{s_t}$ is a mean vector at s_t, $\boldsymbol{S}_{s_t}^{-1}$ is a covariance matrix at s_t, and $a_{s_t r_{t+1}}$ is a transition probability. $\boldsymbol{S}_{s_t}$ is referred to as a precision matrix and is defined as an inverse matrix of a covariance matrix. We use a diagonal matrix as a covariance matrix. The maximum of N_a is N_s, and N_a in this paper can be replaced by N_s. However, this splitting algorithm can use $N_a = 2$ only.

The probability for the complete data set to which the latent variables are introduced is

$$p(\boldsymbol{O}, Z|\Theta) = \prod_{t=1}^{T} \prod_{i=1}^{N_s} \prod_{j=1}^{N_a} \{\mathcal{N}(\boldsymbol{o}_t; \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1}) a_{ij}\}^{z_{ij}^t}, \tag{2}$$

where $Z = \{z_{ij}^t\}_{i=1,j=1,t=1}^{N_s,N_a,T}$ is the set of latent variables.

When the ith state with the HMM parameter Θ_i is split into the i_1th state and the i_2th state, and the parameter $\hat{\Theta}_i$ is estimated for the current splitting, the splitting criterion can be represented by using the objective function $\mathcal{F}$ as follows.

$$\Delta\mathcal{F}_{n+1}^{(VB)} = \mathcal{F}_{n+1}^{(VB)}(\hat{\Theta}_i) - \mathcal{F}_n^{(VB)}(\Theta_i), \tag{3}$$

where n is the iteration number. $\mathcal{F}^{(VB)}$'s definition is given in Section 3.5 briefly.

3.3. Prior of Parameters

We assume that the probability of parameters can be factorized as follows.

$$p(\Theta) = p(N_s, N_a) p(\boldsymbol{a}|N_s, N_a) p(\boldsymbol{S}|N_s) p(\boldsymbol{\mu}|\boldsymbol{S}, N_s). \tag{4}$$

We also assume the prior of $\boldsymbol{a} = \{a_{ij}\}_{i=1,j=1}^{N_s,N_a}, a_{ij} \geq 0, \sum_{j=1}^{N_a} a_{ij} = 1$ is a *Dirichlet* distribution, and the prior of $\{\boldsymbol{\mu}, \boldsymbol{S}\} = \{\{\boldsymbol{\mu}_i\}_{i=1}^{N_s}, \{\boldsymbol{S}_i\}_{i=1}^{N_s}\}$ is a *normal-Gamma* distribution.

$$p(\boldsymbol{a}|N_s, N_a) = \prod_{i=1}^{N_s} \mathcal{D}(\{a_{ij}\}_{j=1}^{N_a}; \phi_0) \propto \prod_{i=1}^{N_s} \prod_{j=1}^{N_a} a_{ij}^{\phi_0 - 1} \tag{5}$$

$$p(\boldsymbol{\mu}, \boldsymbol{S}|N_s) = \prod_{i=1}^{N_s} \prod_{k=1}^{D} \mathcal{N}(\mu_{ik}; \nu_{0k}, \xi_0^{-1} s_{ik}^{-1}) \mathcal{G}(s_{ik}; \eta_0/2, b_{0k}/2), \tag{6}$$

where D is the order of parameters. μ_{ik} and s_{ik} are the kth elements of $\boldsymbol{\mu}_i$ and $\boldsymbol{S}_i$, respectively. $\mathcal{N}()$ is the Gaussian distribution, and $\mathcal{G}()$ is the Gamma distribution. ϕ_0, ν_{0k}, ξ_0, η_0, and b_{0k} are prior hyper-parameters. The Gamma distribution is

$$\mathcal{G}(s; \eta, \lambda) = \frac{\lambda^\eta}{\Gamma(\eta)} s^{\eta-1} \exp(-\lambda s), \quad s > 0, \tag{7}$$

where $\Gamma()$ is the Gamma function.

3.4. Posterior of Parameters

The posterior probability densities can be derived from the Variational Bayesian EM algorithm. The posterior probability of transition probabilities is

$$q(\boldsymbol{a}|\boldsymbol{O}, N_s, N_a) = \prod_{i=1}^{N_s} \mathcal{D}(\{a_{ij}\}_{j=1}^{N_a}; \{\phi_{ij}\}_{j=1}^{N_a}) \propto \prod_{i=1}^{N_s}\prod_{j=1}^{N_a} a_{ij}^{\phi_{ij}-1}, \tag{8}$$

$$\phi_{ij} = \phi_0 + \bar{N}_{ij},\ \bar{N}_{ij} = \sum_{t=1}^{T} \bar{z}_{ij}^t,\ \bar{z}_{ij}^t =< z_{ij}^t >_{q(Z)}\ .$$

The joint probability of mean vectors and covariance matrices is

$$\begin{aligned} & q(\boldsymbol{\mu}, \boldsymbol{S}|\boldsymbol{O}, N_s) \\ &= \prod_{i=1}^{N_s}\prod_{k=1}^{D} \mathcal{N}(\mu_{ik}; \nu_{ik}, \xi_i^{-1} s_{ik}^{-1}) \mathcal{G}(s_{ik}; \eta_i/2, b_{ik}/2), \end{aligned} \tag{9}$$

$$\bar{N}_i = \sum_{t=1}^{T} \bar{z}_i^t,\ \bar{z}_i^t =< z_i^t >_{q(Z)},$$

$$\nu_{ik} = \frac{\bar{N}_i \bar{o}_{ik} + \xi_0 \nu_{0k}}{\bar{N}_i + \xi_0},\ \xi_i = \xi_0 + \bar{N}_i,\ \eta_i = \eta_0 + \bar{N}_i,$$

$$b_{ik} = b_{0k} + \bar{c}_{ik} + \frac{\bar{N}_i \xi_0}{\bar{N}_i + \xi_0} (\bar{o}_{ik} - \nu_{0k})^2,$$

$$\bar{\boldsymbol{o}}_i = \frac{1}{\bar{N}_i} \sum_{t=1}^{T} \bar{z}_i^t \boldsymbol{o}_t,\ \bar{c}_{ik} = \sum_{t=1}^{T} \bar{z}_i^t (o_{tk} - \bar{o}_{ik})^2.$$

The variational posterior probability of latent variables is also derived in the same manner as the unknown parameters.

3.5. Objective Function

Hereafter, $p(\cdot|N_s, N_a)$ is simplified to $p(\cdot)$, i.e., $p(\Theta|N_s, N_a) \rightarrow p(\Theta)$. The objective function is

$$\mathcal{F}^{(VB)} = \int q(Z) q(\Theta) \ln \frac{p(\boldsymbol{O}, Z|\Theta) p(\Theta)}{q(Z) q(\Theta)} dZ d\Theta.$$

This should be derived using the prior and posterior probabilities.

4. EXPERIMENTS

4.1. Experimental Conditions

In this section, we evaluated our proposed method by both segmented phoneme recognition and conventional continuous speech recognition. The segmented phoneme recognition is the classification test for segments that are divided into phonemes in order to evaluate each phoneme model's performance.

We compared our proposed method, the VB-SSS, to the ML-SSS and the MDL-SSS algorithms. For the ML-SSS, two models with different maximum state lengths, 3 or 4, were created. These two models are the baseline models. For the MDL-SSS, we used the same criteria as [4]. The scaling factors, C_c and C_t, were set as $C_c = 2$ and $C_t = 20$ in the experiments.

For the acoustic training set, we used Japanese dialog speech from the ATR travel arrangement task (TRA) database uttered by 166 males. The total speech period was 2.1 hours. For testing, we used dialog speech including 213 sentences from the TRA database uttered by a different set of 17 males. These utterances were separated into individual phoneme periods by forced alignments. In this paper, we used the VB approach only for the splitting and stopping criteria. Additionally, we evaluated performance by word-based continuous speech recognition for the same evaluation set. Multi-class composite bigram models [9] were used, and the vocabulary size was 5,000.

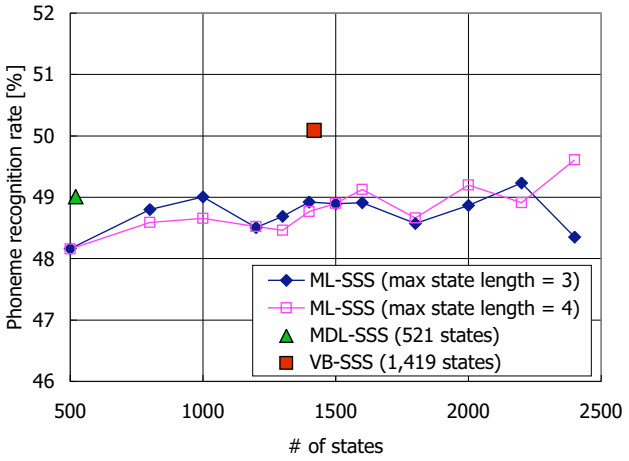

Fig. 2. Average phoneme recognition rates for vowels.

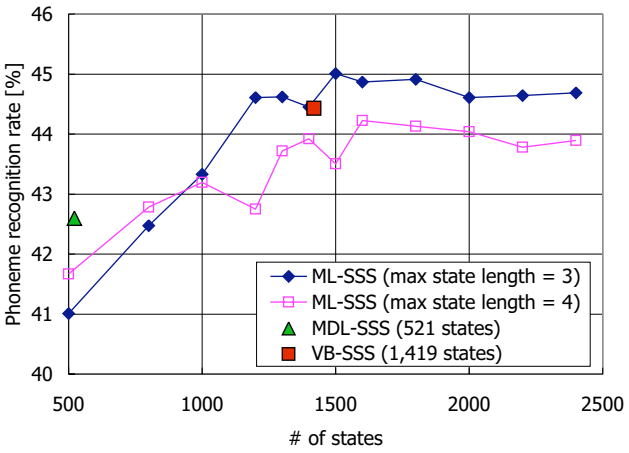

Fig. 3. Average phoneme recognition rates for consonants.

The sampling frequency was 16 kHz, the frame length was 20 ms, and the frame shift was 10 ms. We used 12-order MFCC, ΔMFCC, and Δ log power as feature parameters. Cepstrum mean subtraction was applied to each utterance. We used 26 kinds of phonemes and one silence. Three states were used as the initial model for each phoneme. One Gaussian distribution for each state was used during topology training. A silence model with three states was built separately from the phoneme models. The number of transitions at each state, N_a, was fixed at two.

In these experiments, we used $\phi_0 = 1.0$, $\xi_0 = 1.0$, $\eta_0 = 2.0$ for initial prior hyper-parameters of the VB-SSS. ν_{0k} and b_{0k} were set from the element values of the mean vectors and the precision matrices.

4.2. Segmented Phoneme Recognition

Figures 2 and 3 show the average phoneme recognition rates for vowels and consonants, respectively. "Phoneme recognition rate"

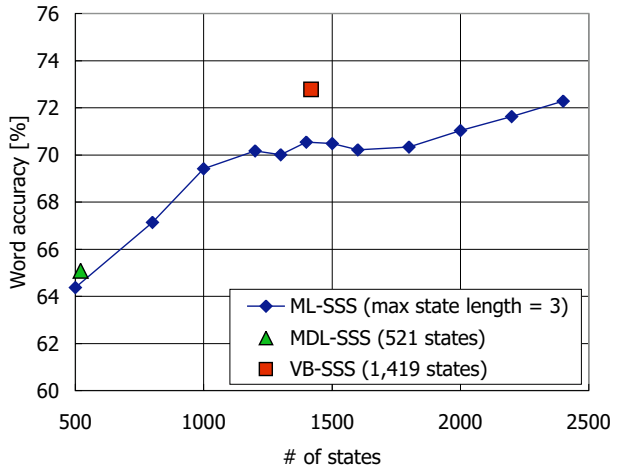

Fig. 4. Word accuracy rates for 5k-word continuous speech recognition.

Table 1. Word accuracy rates [%] and # of states in parentheses for several initial prior parameters.

$\xi_0 = 0.1$	$\phi_0 = 1.0$	$\phi_0 = 10$	$\phi_0 = 100$
$\eta_0 = 0.2$	68.87 (766)	69.14 (764)	68.76 (775)
$\eta_0 = 2.0$	72.12 (1,361)	71.30 (1,252)	67.73 (1,246)
$\xi_0 = 1.0$	$\phi_0 = 1.0$	$\phi_0 = 10$	$\phi_0 = 100$
$\eta_0 = 0.2$	68.87 (760)	68.92 (761)	68.87 (749)
$\eta_0 = 2.0$	72.77 (1,419)	72.55 (1,425)	72.17 (1,433)
$\xi_0 = 10$	$\phi_0 = 1.0$	$\phi_0 = 10$	$\phi_0 = 100$
$\eta_0 = 0.2$	70.01 (771)	68.60 (757)	69.09 (780)
$\eta_0 = 2.0$	71.30 (1,315)	72.06 (1,358)	66.00 (1,211)

means the rate of the number of correctly classified segments in a phoneme. It includes the results by the ML-SSS with maximum state length of 3 or 4, by the MDL-SSS, and by the VB-SSS. The VB-SSS obtained the best results among these three methods. In addition, the topology created by the MDL-SSS is too small to obtain performance comparable with the other methods. The MDL criterion generally does not work for a small amount of training data. Our work on the MDL-SSS[4] shows that the MDL-SSS can automatically obtain almost the same performance as the ML-SSS. However, the total amount of training data in this paper is much smaller than that in the reference[4]. Checking the results in detail, the VB-SSS obtained better results for many phonemes than the ML-SSS. However, the VB-SSS obtained worse results for a few consonants, such as /f/ and /z/. The amount of training data for such phonemes is usually very small.

4.3. Continuous Speech Recognition

We evaluated the same evaluation set by word-based continuous speech recognition with 5k words. The same single-Gaussian acoustic models as described in the previous section were evaluated. Figure 4 shows the word accuracy rates for three types of models. The performance of the MDL-SSS was again worse than the baseline, the ML-SSS, because of the small amount of training data. On the other hand, with about 60% of the ML-SSS states, the VB-SSS achieved a comparable recognition rate.

Next, we analyzed the dependencies of the prior hyper-parameters. Table 1 shows word accuracy rates and the number of states for several initial prior parameters for the 5k-CSR task. The trend of results for segmented phoneme recognition is almost the same. ϕ_0's dependency is low because it only relates to transition probabilities. Although we also evaluated models with $\eta_0 = 20$, the parameters of posteriors could not be obtained in some phonemes because the parameters diverged and no model could be obtained. Therefore, nearly optimal values of ξ_0 and η_0 are limited to a certain range of vales. Furthermore, nearly optimal priors and models can be selected by the maximum $\mathcal{F}^{(VB)}$ after models are trained for a few priors.

5. CONCLUSIONS

We proposed the Variational Bayesian approach for automatically creating non-uniform, context-dependent HMM topologies. We introduced the VB approach to the SSS algorithm to create contextual and temporal variations for HMMs and then defined posterior probability densities and free energy as split and stop criteria. Experimental results using segmented phoneme recognition show that the proposed method can automatically create an appropriate model and obtain better performance, especially for vowels, than the original method. We also evaluated the proposed method for word-based continuous speech recognition. With about 60% of the ML-SSS states, the VB-SSS achieved comparable performance. We found that performance depends on the initial prior parameters but also that appropriate values do exist. Furthermore, we are planning to evaluate the proposed method by using a large amount of training data.

6. ACKNOWLEDGMENT

This research was supported in part by the Telecommunications Advancement Organization of Japan.

7. REFERENCES

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. of the ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[2] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," in *Computer Speech and Language*, 1997, vol. 11, pp. 17–41.

[3] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. of EUROSPEECH'97*, 1997, pp. 99–102.

[4] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform context-dependent HMM topologies based on the MDL criterion," in *Proc. of EUROSPEECH'03*, 2003, vol. 4, pp. 2721–2724.

[5] S. R. Waterhouse, D. MacKay, and A. J. Robinson, "Bayesian methods for mixture of experts," in *Advances in Neural Information Processing Systems (NIPS)*, 1996, vol. 8.

[6] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. of Uncertainty in Artificial Intelligence*, 1999.

[7] N. Ueda and Z. Ghahramani, "Optimal model inference for Bayesian mixture of experts," in *Proc. of IEEE Neural Networks for Signal Processing (NNSP)*, 2000, pp. 145–154.

[8] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Application of the variational bayesian approach to speech recognition," in *NIPS2002*, 2002.

[9] H. Yamamoto and Y. Sagisaka, "Multi-class composite n-gram based on connection direction," in *Proc. of ICASSP'99*, 1999, vol. 1, pp. 533–536.