A BAYESIAN APPROACH TO HMM-BASED SPEECH SYNTHESIS

Kei Hashimoto[1], *Heiga Zen*[1*], *Yoshihiko Nankaku*[1], *Takashi Masuko*[2†], *Keiichi Tokuda*[1]

[1]Nagoya Institute of Technology, Department of Computer Science and Engineering, Nagoya, Japan
[2]Tokyo Institute of Technology, Interdisciplinary Graduate School of Science and Engineering, Yokohama, Japan

ABSTRACT

This paper proposes a new framework of speech synthesis based on the Bayesian approach. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters. In the proposed framework, all processes for constructing the system can be derived from one single predictive distribution which represents the basic problem of speech synthesis directly. Using HMM as the likelihood function and assuming some approximations, it can be regarded as an application of the variational Bayesian method to the HMM-based speech synthesis. Experimental results show that the proposed method outperforms the conventional one in a subjective test.

Index Terms— HMM-based speech synthesis, variational Bayesian method, prior distribution, cross validation, context clustering

1. INTRODUCTION

Over the last few years, a statistical parametric speech synthesis system based on hidden Markov models (HMMs) has grown in popularity [1]. In the HMM-based speech synthesis, spectrum, excitation and duration of speech are modeled simultaneously by HMMs, and speech parameter vector sequences are generated from the HMMs themselves. There are a number of contextual factors that affect spectrum, excitation and duration of speech (e.g., phone identity, accent, stress). In the HMM-based speech synthesis, context-dependent models are typically used to capture these factors [2]. Although a large number of context-dependent models can capture variations in speech data, too many model parameters lead to the over-fitting problem. Therefore, maintaining a proper balance between model complexity and the amount of training data is required. The decision tree based context clustering [3] is a successful method for context-dependent HMM estimation to deal with the problem of training data insufficiency, not only for the robust parameter estimation but also for predicting probability distributions for unseen contexts. This method constructs a parameter tying structure which can assign a sufficient amount of training data to each HMM state. A binary tree is grown step by step, by choosing a question which divides the context using a greedy strategy to maximize some objective function.

In the HMM-based speech synthesis, the maximum likelihood (ML) criterion has been typically used for training HMMs and generating speech parameters, and the minimum description length (MDL) criterion has been typically employed to select the model structure [4]. However, the ML criterion produces a point estimate of HMM parameters and accordingly the accuracy of estimation may be reduced when small training data is available, and the MDL criterion is based on the asymptotic assumption, therefore it is ineffective when the amount of training data is small.

This paper proposes a new framework of speech synthesis based on the Bayesian approach. In this framework, all processes for constructing the system can be derived from one single predictive distribution which represents the problem of speech synthesis directly. The Bayesian approach assumes that model parameters are random variables and reliable predictive distributions are estimated by marginalizing model parameters. However, estimation of posterior distributions of latent variables lead to a huge computational cost. The variational Bayesian (VB) method has been proposed as an effective approximation method of the Bayesian approach [5] and it shows a good performance in the HMM-based speech recognition [6]. In the context clustering, since the Bayesian approach does not use an asymptotic assumption, it is available even in the case where the amount of training data is small. In the Bayesian approach, an appropriate model structure can be selected by maximizing the marginal likelihood [6, 7]. The Bayesian approach can use prior information which is represented by prior distributions. Since the prior distributions affect the model selection, the determination of prior distributions is an important problem for estimation of appropriate acoustic models. However, prior information is not generally given in most speech synthesis tasks. This paper applies a prior distribution determination technique using the cross validation to the context clustering [8]. The Bayesian approach using cross validation can select an appropriate model structure without tuning parameters of prior distribution. The rest of this paper is organized as follows. Section 2 describes the new framework for speech synthesis based on the Bayesian approach. In Section 3, subjective listening test results are presented. Concluding remarks and future plans are presented in final section.

2. BAYESIAN SPEECH SYNTHESIS

2.1. Bayesian approach

Let $\boldsymbol{O} = (\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots, \boldsymbol{O}_T)$ be a set of training data of D dimensional feature vectors, and T denotes the number of frames. The output probability of an HMM is defined by:

$$P(\boldsymbol{O}, \boldsymbol{Q} \mid \boldsymbol{\Lambda}) = \prod_{t=1}^{T} a_{Q_{t-1}Q_t} \mathcal{N}(\boldsymbol{O}_t \mid \boldsymbol{\mu}_{Q_t}, \boldsymbol{S}_{Q_t}^{-1}) \,, \tag{1}$$

where $\boldsymbol{Q} = (Q_1, Q_2, \cdots, Q_T)$ is a sequence of HMM states, $z_t \in \{1, \ldots, N\}$ denotes a state at frame t and N is the number of states in an HMM. A set of model parameters $\boldsymbol{\Lambda} = \{a_{ij}, \boldsymbol{\mu}_i, \boldsymbol{S}_i\}_{i,j=1}^{N}$ consists of the state transition probability a_{ij} from state i to state j,

*Heiga Zen is now with the Speech Technology Group, Toshiba Europe Research Ltd. Cambridge Research Laboratory, Cambridge, UK.

†Takashi Masuko is now with the Corporate Research & Development Center, Toshiba Corporation, Kawasaki, JAPAN.

the mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\boldsymbol{S}_i^{-1}$ of a Gaussian distribution $\mathcal{N}(\cdot \mid \boldsymbol{\mu}_i, \boldsymbol{S}_i^{-1})$.

In the HMM-based speech synthesis, the ML criterion has been typically used to train HMMs and generate speech parameters. The optimal model parameters can be obtained by maximizing the likelihood for a given training data as follows:

$$\boldsymbol{\Lambda}_{ML} = \arg\max_{\boldsymbol{\Lambda}} P(\boldsymbol{O} \mid S, \boldsymbol{\Lambda}) \,, \tag{2}$$

where S is a label sequence of training data. Since it is difficult to analytically obtain the model paramter $\boldsymbol{\Lambda}_{ML}$, the model parameter can be obtained using an iterative procedure such as the expectation-maximization (EM) algorithm. In the synthesis part, the speech parameter generation algorithm generates sequences of speech parameter vectors that maximize their output probabilities using model parameters $\boldsymbol{\Lambda}_{ML}$.

$$\boldsymbol{o}_{ML} = \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid s, \boldsymbol{\Lambda}_{ML}) \,, \tag{3}$$

where $\boldsymbol{o} = \left[\boldsymbol{o}_1^\top, \boldsymbol{o}_2^\top, \ldots, \boldsymbol{o}_T^\top\right]^\top$ is a speech parameter sequence and s is a label sequence to be synthesized,

However, the ML estimator produces a point estimate of HMM parameters and the accuracy of estimation may be reduced when the amount of training data is small. The Bayesian approach assumes that a set of model parameters $\boldsymbol{\Lambda}$ is a random variable, while the ML approach estimates constant model parameters. In the Bayesian approach, the speech parameter is generated by the predictive distribution [7] as follows:

$$\begin{aligned} \boldsymbol{o}_{Bayes} &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid s, \boldsymbol{O}, S) \\ &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \,. \end{aligned} \tag{4}$$

It can be seen that equation (4) directly represents the problem of speech synthesis, that is, generating speech feature sequence $\boldsymbol{o}$ given training feature sequences $\boldsymbol{O}$ with labels S and labels to be synthesized s. The marginal likelihood of $\boldsymbol{o}$ and $\boldsymbol{O}$ is defined by

$$\begin{aligned} &P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\ &= \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\ &= \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda}) P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \,, \end{aligned} \tag{5}$$

where $\boldsymbol{q}$ is a sequence of HMM states for a speech parameter sequence $\boldsymbol{o}$, $P(\boldsymbol{\Lambda})$ is a prior distribution for model parameter $\boldsymbol{\Lambda}$, $P(\boldsymbol{o}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda})$ is the likelihood of synthesis data $\boldsymbol{o}$, and $P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda})$ is the likelihood of training data $\boldsymbol{O}$. The model parameters are integrated out in equation (5) so that the effect of over-fitting is mitigated. However, it is difficult to solve the integral and expectation calculations. Especially, when a model includes latent variables, the calculation becomes more complicated. To overcome this problem, the variational Bayesian method has been proposed as a tractable approximation method of the Bayesian approach and it has shown good generalization performance in many applications [5].

2.2. Variational Bayesian method

The variatonal Bayesian method maximizes a lower bound of log marginal likelihood $\mathcal{F}$ instead of the true marginal likelihood. A lower bound $\mathcal{F}$ is defined by using Jensen's inequality:

$$\begin{aligned} &\log P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\ &= \log \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\ &= \log \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\ &\geq \left\langle \log \frac{P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} \right\rangle_{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} \\ &= \mathcal{F} \end{aligned} \tag{6}$$

where, $\langle \cdot \rangle_Q$ denotes a calculation of expectation with respect to Q, and $Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})$ is an approximate distribution of the true posterior distribution $P(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid \boldsymbol{W}\boldsymbol{x}, \boldsymbol{O}, s, S)$. The variational Bayesian method uses the assumption that probabilistic variables associated with $\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}$ are statistically independent of the other variables.

$$Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}) = Q(\boldsymbol{q}) Q(\boldsymbol{Q}) Q(\boldsymbol{\Lambda}) \tag{7}$$

In the VB method, VB posterior distributions $Q(\boldsymbol{Q})$, $Q(\boldsymbol{q})$ and $Q(\boldsymbol{\Lambda})$ are introduced to approximate the true posterior distributions. The optimal VB posterior distributions can be obtained by maximizing the objective function $\mathcal{F}$ with the variational method as follows:

$$Q(\boldsymbol{q}) = C_{\boldsymbol{q}} \exp \left\langle \log P(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{\Lambda})} \,, \tag{8}$$

$$Q(\boldsymbol{Q}) = C_{\boldsymbol{Q}} \exp \left\langle \log P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{\Lambda})} \,, \tag{9}$$

$$\begin{aligned} Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) &\exp \left\langle \log P(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{q})} \\ &\times \exp \left\langle \log P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{Q})} \,, \end{aligned} \tag{10}$$

where $C_{\boldsymbol{q}}$, $C_{\boldsymbol{Q}}$ and $C_{\boldsymbol{\Lambda}}$ are the normalization terms of $Q(\boldsymbol{q})$, $Q(\boldsymbol{Q})$ and $Q(\boldsymbol{\Lambda})$, respectively. These optimizations can be effectively performed by iterative calculations as the EM algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence.

However, in the above algorithm, the optimal posterior distributions depend on synthesized speech parameter $\boldsymbol{o}$, i.e., the posterior distributions given a label sequence of synthesis speech are estimated. Consequently, it leads to a huge computational cost in the synthesis part. To avoid this problem, this paper assumes that $Q(\boldsymbol{\Lambda})$ is independent of speech parameter $\boldsymbol{o}$. Then, $Q(\boldsymbol{\Lambda})$ is given by

$$Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{Q}} P(\boldsymbol{\Lambda}) \exp \left\langle \log P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{Q})} \,. \tag{11}$$

2.3. Bayesian context clustering

The decision tree based context clustering is a top-down clustering method to optimize the state tying structure for robust model parameter estimation. A leaf of the decision tree corresponds to a set of HMM states to be tied. The decision tree growing process begins with a root node which has all HMM states to be tied. Then, a question which divides the set of states into two subsets assigned respectively to two child nodes, "Yes" node and "No" node, is chosen so as to maximize the value of objective function. The decision tree is grown in a greedy fashion, successively splitting nodes by selecting the pair of a question and node which maximize the gain of objective function at each step. In the HMM-based speech synthesis, model parameters of spectrum, excitation, and duration are clustered separately because they have their own influential contextual factors.

In the Bayesian approach, an optimal model structure can be selected by maximizing the objective function $\mathcal{F}$ [6]. When a node is split into "Yes" node and "No" node by the question q, the gain $\Delta\mathcal{F}_q$ is defined as the difference of $\mathcal{F}$ before and after splitting:

$$\Delta\mathcal{F}_q = \mathcal{F}_q^y + \mathcal{F}_q^n - \mathcal{F}_q^p, \tag{12}$$

where $\mathcal{F}_q^y$ and $\mathcal{F}_q^n$ are the value of objective function $\mathcal{F}$ of split nodes by a question q, and $\mathcal{F}_q^p$ is the value before splitting. The question $\hat{q}$ for splitting a node is chosen from the question set as follows:

$$\hat{q} = \arg\max_q \Delta\mathcal{F}_q. \tag{13}$$

By splitting nodes until $\Delta\mathcal{F}_{\hat{q}} \leq 0$, the decision tree which maximizes the objective function $\mathcal{F}$ is obtained.

2.4. Bayesian context clustering using cross validation

In the Bayesian approach, prior distributions are usually determined heuristically. However, hyper-parameters (parameters of prior distributions) affect the model selection as tuning parameters. Therefore, to automatically select an appropriate model structure, a determination technique of prior distribution is required. One possible approach is to optimize the hyper-parameters using training data so as to maximize the marginal likelihood. However, it still needs tuning parameters which control influences of prior distributions, and often leads to the over-fitting problem as the ML criterion. To overcome this problem, the prior distribution determination technique using cross validation has been proposed [8]. In this paper, we apply it to the context clustering for the HMM-based speech synthesis.

Let $\boldsymbol{O} = \{\boldsymbol{O}^{(1)}, \boldsymbol{O}^{(2)}, \cdots, \boldsymbol{O}^{(k)}, \cdots, \boldsymbol{O}^{(K)}\}$ be a set of training data and $\boldsymbol{O}^{(k)}$ be a partition for K-fold cross validation. For the k-th evaluation, $\boldsymbol{O}^{(\bar{k})} = \{\boldsymbol{O}^{(j)} \mid j \neq k\}$ is used for the determination of prior distributions and $\boldsymbol{O}^{(k)}$ is used for the estimation of posterior distributions. Then, the Bayesian approach using cross validation calculates the log marginal likelihood $\log P(\boldsymbol{o}, \boldsymbol{O}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}, s, S)$. Using Jensen's inequality, the lower bound of log marginal likelihood $\mathcal{F}^{(k)}$ is defined as equation (6):

$$\log P(\boldsymbol{o}, \boldsymbol{O}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}, s, S) \geq \mathcal{F}^{(k)}\,. \tag{14}$$

For the k-th evaluation, the optimal VB posterior distributions of model parameters can be obtained by maximizing $\mathcal{F}^{(k)}$ with respect to $Q(\boldsymbol{\Lambda}^{(k)})$ with the variational method as follows: For the k-th evaluation, the optimal VB posterior distributions of model parameters can be obtained by maximizing $\mathcal{F}^{(k)}$ with respect to $Q(\boldsymbol{\Lambda}^{(k)})$ with the variational method as equation (11):

$$\begin{aligned} Q(\boldsymbol{\Lambda}^{(k)}) = {} & C_{\boldsymbol{\Lambda}^{(k)}} P(\boldsymbol{\Lambda}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \\ & \times \left\langle \log P(\boldsymbol{O}^{(k)}, \boldsymbol{Q}^{(k)} \mid \boldsymbol{\Lambda}^{(k)}) \right\rangle_{Q(\boldsymbol{Q}^{(k)})}\,, \end{aligned} \tag{15}$$

where $P(\boldsymbol{\Lambda}^{(k)} \mid \boldsymbol{O}^{(\bar{k})})$ is a prior distribution which represents prior information $\boldsymbol{O}^{(\bar{k})}$ and $C_{\boldsymbol{\Lambda}^{(k)}}$ is a normalization term.

In the Bayesian approach, a conjugate prior distribution is widely used as a prior distribution. When the output probability distribution is a Gaussian distribution, the conjugate prior distribution becomes a Gauss-Wishart distribution:

$$P(\boldsymbol{\mu}, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\nu}, (\xi\boldsymbol{S})^{-1})\, \mathcal{W}(\boldsymbol{S} \mid \eta, \boldsymbol{B})\,, \tag{16}$$

where $\{\xi, \eta, \boldsymbol{\nu}, \boldsymbol{B}\}$ is a set of hyper-parameters. Moreover a Gaussian distribution is proportional to Gauss-Wishart distribution as follows:

$$\begin{aligned} & \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{O}_t \mid \boldsymbol{\mu}, \boldsymbol{S}^{-1}) \\ & \qquad \propto \mathcal{N}(\boldsymbol{\mu} \mid \bar{\boldsymbol{O}}, (T\boldsymbol{S})^{-1})\, \mathcal{W}(\boldsymbol{S} \mid T + D, (T\bar{\boldsymbol{C}}))\,, \end{aligned} \tag{17}$$

where $\bar{\boldsymbol{O}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{O}_t$ and $\bar{\boldsymbol{C}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{O}_t\boldsymbol{O}_t^\top - \bar{\boldsymbol{O}}\bar{\boldsymbol{O}}^\top$ are sufficient statistics of training data. Thus, the prior distribution can be determined by sufficient statistics of the prior information. The prior distribution of the k-th cross validation model parameters $P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} \mid \boldsymbol{O}^{(\bar{k})})$ is obtained from equation (17):

$$\begin{aligned} & P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \\ & \qquad = \mathcal{N}(\boldsymbol{\mu}^{(k)} \mid \bar{\boldsymbol{O}}^{(\bar{k})}, (T^{(\bar{k})}\boldsymbol{S}^{(k)})^{-1}) \\ & \qquad\quad \times \mathcal{W}(\boldsymbol{S}^{(k)} \mid T^{(\bar{k})} + D, (T^{(\bar{k})}\bar{\boldsymbol{C}}^{(\bar{k})}))\,, \end{aligned} \tag{18}$$

where $\bar{\boldsymbol{O}}^{(\bar{k})}$ and $\bar{\boldsymbol{C}}^{(\bar{k})}$ are sufficient statistics of a subset of training data $\boldsymbol{O}^{(\bar{k})}$. The cross valid prior distribution can be determined without tuning parameters. In the HMM-based speech synthesis, the multi-space probability distribution HMMs (MSD-HMMs) [10] have been used to model excitation. However, the cross valid prior distributions for the MSD-HMMs can be determined by using sufficient statistics of each space as equation (18).

The objective function of the Bayesian approach using cross validation $\mathcal{F}^{(CV)}$ is obtained by summing $\mathcal{F}^{(k)}$ for each fold:

$$\mathcal{F}^{(CV)} = \sum_{k=1}^{K} \mathcal{F}^{(k)}\,. \tag{19}$$

An optimal model structure can be selected by maximizing the objective function $\mathcal{F}^{(CV)}$ instead of $\mathcal{F}$. As equation (13), the question which maximizes the gain of the objective function $\Delta\mathcal{F}_q^{(CV)}$ is selected. By splitting nodes until $\Delta\mathcal{F}_{\hat{q}}^{(CV)} \leq 0$, the decision tree which maximizes the objective function $\mathcal{F}^{(CV)}$ is obtained.

3. EXPERIMENTS

3.1. Experimental conditions

To evaluate the performance of the proposed method, speech synthesis experiments were performed. In these experiments, the ATR Japanese speech database [9] B-set which consits of the phonetically balanced 503 sentences were used. The first 450 of the 503 sentences, uttered by one male speaker (MHT), were used for training. The remaining 53 sentences were used for evaluations. Speech signals were sampled at a rate of 16 kHz and windowed at a 5 ms frame rate using a 25 ms Blackman window. Feature vectors consisted of spectrum and F_0 parameter vectors. The spectrum parameter vectors consisted of 24 mel-cepstral coefficients excepting the zero-th coefficients and their delta and delta-delta coefficients. The F_0 parameter vectors consisted of log F_0, its delta and delta-delta. A left-to-right, five-state, MSD-HMM with no skip structure was used. Each state output PDF was composed of spectrum and F_0 streams. The spectrum stream was modeled by single multi-variate Gaussian distributions with diagonal covariance matrices. The F_0 stream was modeled by a multi-space probability distribution consisting of a Gaussian distribution for voiced frames and a discrete distribution for unvoiced frames. Each state duration PDF was modeled by a one-dimensional Gaussian distribution.

The decision tree-based context clustering technique was separately applied to distributions of spectrum, F_0, and state duration.

In these experiments, the following four approaches were compared.

- "ML-MDL" : HMMs were trained by the ML criterion and model structures were selected by the MDL criterion.
- "Bayes-Bayes" : HMMs were trained by the Bayesian criterion and model structures were selected by the Bayesian criterion with cross validation.

Table 1. Number of states of selected model structure by the conventional and proposed methods.

	mel-cepstram	F_0	duration
ML-MDL	956	1,151	280
Bayes-Bayes	9,070	12,836	4,005
Bayes-MDL	1,941	565	47
ML-Bayes	15,077	8,844	3,185

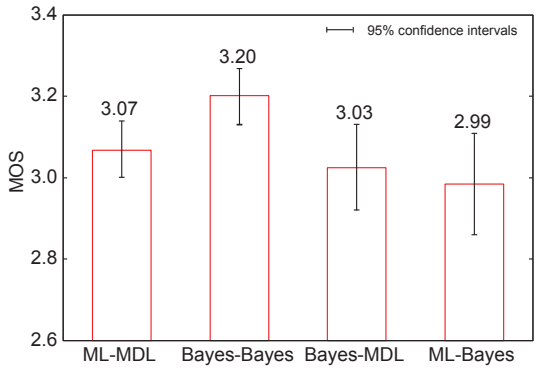

Fig. 1. Mean opinion scores of synthsized speech by the conventional and proposed methods. Error bars show 95% confidence intervals.

- "Bayes-MDL" : HMMs were trained by the Bayesian criterion and model structures were selected by the Bayesian criterion with cross validation. In the context clustering, splitting nodes was performed by the Bayesian criterion with cross validation, and stopping criterion was adjusting a threshold to make model structures which have the similar number of states with "ML-MDL."
- "ML-Bayes" : HMMs were trained by the ML criterion and model structures were selected by the MDL criterion using threshold. In the context clustering, splitting nodes was performed by the MDL criterion, and stopping criterion was adjusting a threshold to make model structures which have the similar number of states with "Bayes-Bayes."

In "Bayes-Bayes" and "Bayes-MDL," each context is regarded as 1-fold of the cross validation. The number of states for each method is "ML-MDL":2,491 , "Bayes-Bayes":25,911 , "Bayes-MDL":2,553 , "ML-Bayes":27,106. Table 1 represents the details of the number of states.

3.2. Experimental results

A subjective listeing test was conducted to evaluate quality of synthesized speech. The test compared the naturalness of converted speech by the mean opinion score (MOS) test method. The subjects were 10 Japanese graduate students. Twenty sentences were randomly chosen from the evaluation sentences. Samples were presented in a random order for each test sentence. In the MOS test, after listening to each test sample, the subjects were asked to assign it a five-point naturalness score (5: natural – 1: poor).

Figure 1 plots the experimental results. It can be seen from the figure that the proposed method "Bayes-Bayes" achieved a better subjective score than the conventional method "ML-MDL." Moreover, although "Bayes-MDL" is trained by the Bayesian criterion, the subjective score of "Bayes-MDL" was worse than "Bayes-Bayes," and although "ML-Bayes" has the similar number of states as "Bayes-Bayes," the subjective score of "ML-Bayes" was worse than "Bayes-Bayes." Because the model structure of "ML-Bayes" is too big for the ML training, "ML-Bayes" leads to the over-fitting problem. Thus, the error bar of "ML-Bayes" in Figure 1 are larger than others. These results cleary show the effectiveness of the proposed method in both the model training and model structure selection. Most of the subjects observed that the proposed method improved the naturalness in spectrum and excitation.

4. CONCLUSION

This paper proposed the new framework of speech synthesis based on the Bayesian approach. In the proposed framework, all processes for constructing the system could be derived from one single predictive distribution which represents the problem of speech synthesis directly. The results on the MOS test demonstrated that the proposed method outperform the conventional one.

Future works include applying the Bayesian approach to hidden semi-Markov model (HSMM) based speech synthesis and research of the relation between the quality of synthesized speech and the size of model structure.

5. ACKNOWLEDGMENT

The authors would like to thank Dr. Akinobu Lee for him helpful comments and discussions. This work was partly supported by the FP7 EMIME project.

6. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in Proc. Eurospeech, 1999, pp.2347–2350.

[2] J. J. Odell, "The use of context in large vocabulary speech recognition," PhD dissertation, Cambridge University, 1995.

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

[4] K. Shinoda and T. Watanabe, "Acoustic Modeling Based on the MDL Criterion for speech recognition," in Proc. Eurospeech, pp.99–102, 1997.

[5] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in Proc. UAI 15, 1999.

[6] S. Watanabe, Y. Minami, A. Nakamura and N. Ueda "Variational Bayesian estimation and clustering for speech recognition," IEEE Trans. SAP, vol.12, pp.365–381, 2004.

[7] Y. Nankaku, H. Zen, K. Tokuda, T. Kitamura, and T. Masuko, "A Bayesian Approach to HMM-Based Speech Synthesis," in Proc. TECHNICAL REPORT OF IEICE, vol.99, pp.19–24, 2003, (in Japanese).

[8] K. Hashimoto, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Bayesian context clustering using cross valid prior distribution for HMM-based speech recognition," in Proc. Interspeech, pp.936–939, 2008.

[9] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," Speech Commun., vol.9, pp.357–363, 1990.

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in Proc. ICASSP, pp.229–232, 1999.