社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

HMM音声合成のためのクロスバリデーションを用いたベイズ基準によるコンテキストクラスタリング

橋本　佳†　全　炳河†　南角　吉彦†　徳田　恵一†

† 名古屋工業大学大学院 工学研究科 創成シミュレーション工学専攻
〒 466–8555 名古屋市 昭和区 御器所町

あらまし　ベイズ基準はモデルパラメータを確率変数と仮定し，信頼性の高い予測分布を推定することができる統計的手法である．近年，ベイズ基準の近似手法である変分ベイズ法が提案され，様々な統計モデルにベイズ基準を適用することが可能となった．音声認識では音響モデルの学習，コンテキストクラスタリングにベイズ基準が適用されており，その有効性が確認されている．しかし，ベイズ基準では，事前分布が事後分布の推定やモデル構造の選択に大きく影響を与えるため，事前分布を適切に設定する必要がある．この問題に対し，我々はクロスバリデーションに基づいた事前分布設定法を提案し，音声認識におけるコンテキストクラスタリングに適用した．本稿では，クロスバリデーションを用いたベイズ基準によるコンテキストクラスタリングを HMM 音声合成に適用し，その有効性を示す．
キーワード　ベイズ基準，HMM 音声合成，コンテキストクラスタリング，事前分布，クロスバリデーション

Bayesian Context Clustering Using Cross Validation for HMM-Based Speech Synthesis

Kei HASHIMOTO†, Heiga ZEN†, Yoshihiko NANKAKU†, and Keiichi TOKUDA†

† Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya, 466–8555 Japan

Abstract　This paper proposes a prior distribution determination technique using cross validation for HMM-based speech synthesis based on the Bayesian approach. The Bayesian method is a statistical technique for estimating reliable predictive distributions by marginalizing model parameters and its approximate version, the variational Bayesian method has been applied to HMM-based speech synthesis. Since prior distributions representing prior information about model parameters affect the model selection (e.g., decison tree based context clustering), the determination of prior distributions is an important problem. The proposed method can determine reliable prior distributions without tuning parameters and select an appropriate model structure dependently on the amount of training data.
Key words　Bayesian criterion, HMM-based speech synthesis, context clustering, prior distribution, cross validation

1. Introduction

Over the last few years, a statistical parametric speech synthesis system based on hidden Markov models (HMMs) has grown in popularity [1], [2]. In the HMM-based speech synthesis, the maximum likelihood (ML) criterion has been typically used for training HMMs and generating speech parameters. However, the ML criterion produces a point estimate of HMM parameters and the accuracy of estimation may be reduced when little training data is available. The Bayesian approach assumes that a set of model parameters is random variables and reliable predictive distributions are estimated by marginalizing model parameters. However, to estimate the posterior distribution of latent variable models lead to a huge computational cost. The variational Bayesian (VB) method has been proposed as an effective approximation method of the Bayesian approach [3] and it shows good performance in the HMM-based speech recognition [4].

In the HMM-based speech synthesis, context-dependent models are used [5]. 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 is a successful method for context-dependent HMM estimation to deal with the problem of training data insufficiency, not only for robust parameter estimation but also for predicting probability distributions for unseen contexts [6]. 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 question which divides the context using a greedy strategy to maximize some objective function. The ML criterion is inappropriate as a model selection criterion since the ML criterion increases likelihood monotonically as the number of states increases. Some heuristic thresholding are therefore necessary to terminate the dividing nodes in the context clustering. To solve this problem, the minimum description length (MDL) criterion has been employed to select the model structure [7]. However, the MDL criterion is based on an asymptotic assumption, therefore it is ineffective when the amount of training data is small. On the other hand, since the Bayesian approach does not use the asymptotic assumption, unlike the MDL criterion, 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 [4], [8].

The Bayesian approach can use prior information which is represented by the prior distribution. Since prior distributions affect the estimation of posterior distributions and model selection, the determination of prior distributions is an important problem for estimating of appropriate acoustic models. However, prior information is not generally given in many speech synthesis tasks, and determination techniques of prior distribution without using prior information has not been developed. This paper applies the prior distribution determination technique using the cross validation to the context clustering [10]. The Bayesian approach using cross validation can select an appropriate model structure without tuning parameters.

The rest of this paper is organized as follows. Section 2 describes the Bayesian speech synthesis, and Section 3 describes the prior distribution determination technique using cross validation and apply it to the context clustering. In Section 4, 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, $Q_t \in \{1, \ldots, N\}$ 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, 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 little training data is available. 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 as follows [8]:

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

It can be seen that equation (4) directly represents the problem of speech synthesis, that is, generating speech parameter sequence $\boldsymbol{o}$ given training feature sequences with labels and labels to be synthesized. 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 $\boldsymbol{\Lambda}$, $P(\boldsymbol{o}, \boldsymbol{q}|s, \boldsymbol{\Lambda})$ is the likelihood of synthesis data $\boldsymbol{o}$, and $P(\boldsymbol{O}, \boldsymbol{Q}|S, \boldsymbol{\Lambda})$ 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 [3].

2.2 Variational Bayesian method

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

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

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 $P(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}|\boldsymbol{o}, \boldsymbol{O}, s, S)$. 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:

$$Q(\boldsymbol{q}) = C_{\boldsymbol{q}} \exp \langle \log P(\boldsymbol{o}, \boldsymbol{q}|s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{\Lambda})} \,, \tag{8}$$

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

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

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 $\boldsymbol{x}$, 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 $Q(\boldsymbol{Q})$ and $Q(\boldsymbol{\Lambda})$ are independent of $\boldsymbol{o}$. Then, $Q(\boldsymbol{\Lambda})$ is given by

$$Q(\boldsymbol{\Lambda}) \propto P(\boldsymbol{\Lambda}) \exp \langle \log P(\boldsymbol{O}, \boldsymbol{Q}|S, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{Q})} \,. \tag{11}$$

2.3 Prior distribution

In the Bayesian approach, a conjugate prior distribution is widely used as a prior distribution $P(\boldsymbol{\Lambda})$. When the output probability distribution is a Gaussian distribution, the conjugate prior distribution becomes a Gauss-Wishart distribution:

$$P(\boldsymbol{\mu}, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{\mu} \mid \boldsymbol{\nu}, (\xi \boldsymbol{S})^{-1})\, \mathcal{W}(\boldsymbol{S} \mid \eta, \boldsymbol{B}) \,, \tag{12}$$

where $\{\xi, \eta, \boldsymbol{\nu}, \boldsymbol{B}\}$ is a set of hyper-parameters. Using a conjugate prior distribution, a set of parameters of posterior distribution is also represented by the same parameters $\{\bar{\xi}, \bar{\eta}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{B}}\}$. Moreover, a Gaussian distribution is proportional to Gauss-Wishart distribution as follows:

$$\begin{aligned} &\prod_{t=1}^{T} \mathcal{N}(\boldsymbol{O}_t \mid \boldsymbol{\mu}, \boldsymbol{S}^{-1}) \\ &\quad \propto \mathcal{N}(\boldsymbol{\mu} \mid \bar{\boldsymbol{O}}, (T\boldsymbol{S})^{-1}) \mathcal{W}(\boldsymbol{S} \mid T + D, (T\bar{\boldsymbol{C}})\,) \,, \end{aligned} \tag{13}$$

$$\bar{\boldsymbol{O}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{O}_t \,, \tag{14}$$

$$\bar{\boldsymbol{C}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{O}_t \boldsymbol{O}_t^{\top} - \bar{\boldsymbol{O}} \bar{\boldsymbol{O}}^{\top} \,. \tag{15}$$

Thus, the prior distribution can be determined by sufficient statistics of the prior information.

2.4 Speech parameters generation

We assume that a speech parameter vector $\boldsymbol{o}_t$ consists of a static feature vector $\boldsymbol{c}_t$ and its first and second order dynamic feature vectors, that is

$$\begin{aligned} \boldsymbol{o} &= \boldsymbol{W}\boldsymbol{c} \\ &= \left[(\boldsymbol{W}\boldsymbol{c})_1^{\top}, (\boldsymbol{W}\boldsymbol{c})_2^{\top}, \ldots, (\boldsymbol{W}\boldsymbol{c})_T^{\top} \right]^{\top} \end{aligned} \tag{16}$$

$$(\boldsymbol{W}\boldsymbol{c})_t = \left[\boldsymbol{c}_t^{\top}, \Delta \boldsymbol{c}_t^{\top}, \Delta^2 \boldsymbol{c}_t^{\top} \right]^{\top} \tag{17}$$

where $\boldsymbol{W}$ is a window matrix to calculate dynamic features from static features [9]. In the synthesis part, a static feature vector sequence $\boldsymbol{c}$ is generated. By the variational method, the lower boud $\mathcal{F}$ approximates the log marginal likelihood $P(\boldsymbol{W}\boldsymbol{c}, \boldsymbol{O} \mid s, S)$. Therefore, the optimal speech parameter sequence is generated by maximizing the lower boud $\mathcal{F}$:

$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \boldsymbol{c}} &= \frac{\partial}{\partial \boldsymbol{c}} \langle \log P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\Lambda}) P(\boldsymbol{q} \mid s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{q})Q(\boldsymbol{\Lambda})} \\ &= \left\langle \frac{\partial}{\partial \boldsymbol{c}} \log P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{q})Q(\boldsymbol{\Lambda})} = \boldsymbol{0} \,. \end{aligned} \tag{18}$$

Under the condition in equation (16), the optimal static feature sequence which maximizes equation (18) can be determined by solving the following set of linear equations:

$$\boldsymbol{W}^\top \langle \boldsymbol{S} \rangle \boldsymbol{W} \boldsymbol{c}_{Bayes} = \boldsymbol{W}^\top \langle \boldsymbol{S}\boldsymbol{\mu} \rangle \,, \tag{19}$$

where $\langle \boldsymbol{S} \rangle$ and $\langle \boldsymbol{S}\boldsymbol{\mu} \rangle$ represent the expectation value of $\boldsymbol{S}$ and $\boldsymbol{S}\boldsymbol{\mu}$, respectively, and they are defined as follows:

$$\begin{aligned}\langle \boldsymbol{S} \rangle &= \sum_{\boldsymbol{q}} \int Q(\boldsymbol{q}) Q(\boldsymbol{\Lambda}) \boldsymbol{S} \, d\boldsymbol{\Lambda} \\ &= \mathrm{diag}\left[\langle \boldsymbol{S}_{q_1} \rangle, \langle \boldsymbol{S}_{q_2} \rangle, \ldots, \langle \boldsymbol{S}_{q_T} \rangle \right] \,, \end{aligned} \tag{20}$$

$$\begin{aligned}\langle \boldsymbol{S}\boldsymbol{\mu} \rangle &= \sum_{\boldsymbol{q}} \int Q(\boldsymbol{q}) Q(\boldsymbol{\Lambda}) \boldsymbol{S}\boldsymbol{\mu} \, d\boldsymbol{\Lambda} \\ &= \left[\langle \boldsymbol{S}_{q_1}\boldsymbol{\mu}_{q_1} \rangle^\top, \langle \boldsymbol{S}_{q_2}\boldsymbol{\mu}_{q_2} \rangle^\top, \ldots, \langle \boldsymbol{S}_{q_T}\boldsymbol{\mu}_{q_T} \rangle^\top \right]^\top . \end{aligned} \tag{21}$$

In equation (20) and (21), each element of the matrices independently is derived.

$$\begin{aligned}\langle \boldsymbol{S}_{q_t} \rangle &= \sum_{i=1}^{N} \langle q_t^i \rangle \int Q(\boldsymbol{\Lambda}) \boldsymbol{S}_i \, d\boldsymbol{\Lambda} \\ &= \sum_{i=1}^{N} \langle q_t^i \rangle \bar{\eta}_i \bar{\boldsymbol{B}}_i^{-1} \,, \end{aligned} \tag{22}$$

$$\begin{aligned}\langle \boldsymbol{S}_{q_t}\boldsymbol{\mu}_{q_t} \rangle &= \sum_{i=1}^{N} \langle q_t^i \rangle \int Q(\boldsymbol{\Lambda}) \boldsymbol{S}_i\boldsymbol{\mu}_i \, d\boldsymbol{\Lambda} \\ &= \sum_{i=1}^{N} \langle q_t^i \rangle \bar{\eta}_i \bar{\boldsymbol{B}}_i^{-1} \bar{\boldsymbol{\nu}}_i \,, \end{aligned} \tag{23}$$

where

$$\langle q_t^i \rangle = \sum_{\boldsymbol{q}} Q(\boldsymbol{q}) q_t^i \,. \tag{24}$$

The equation (19) can be solved efficiently using the Cholesky or QR decomposition [9].

2.5 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 clustered. 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 Bayesian approach, an optimal model structure can be selected by maximizing the objective function $\mathcal{F}$ [4]. When a node is split into "Yes" node and "No" node by the question r, the gain $\Delta\mathcal{F}_r$ is defined as the difference of $\mathcal{F}$ before and after splitting:

$$\Delta\mathcal{F}_r = \mathcal{F}_r^y + \mathcal{F}_r^n - \mathcal{F}_r^p, \tag{25}$$

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

$$\hat{r} = \arg\max_r \Delta\mathcal{F}_r. \tag{26}$$

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

3. 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 [10]. In this paper, we apply it to the context clustering for the HMM-based speech synthesis.

3.1 Bayesian approach using cross validation

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:

$$\mathcal{L}^{(k)}(\boldsymbol{O}) = \log P(\boldsymbol{O}^{(k)} \mid \boldsymbol{O}^{(\bar{k})}) \,. \tag{27}$$

Using Jensen's inequality, the lower bound of log marginal likelihood $\mathcal{F}^{(k)}$ is defined as equation (6). For the k-th evaluation, the optimal VB posterior distributions of model parameters can be obtained by maximizing $\mathcal{F}^{(k)}$ with respect

to $Q(\mathbf{\Lambda}^{(k)})$ with the variational method as follows:

$$Q(\mathbf{\Lambda}^{(k)}) = C_{\mathbf{\Lambda}^{(k)}} P(\mathbf{\Lambda}^{(k)} | \boldsymbol{O}^{(\bar{k})}) \exp\left\{ \sum_{\boldsymbol{Q}^{(k)}} Q(\boldsymbol{Q}^{(k)}) \log P(\boldsymbol{O}^{(k)}, \boldsymbol{Q}^{(k)} \mid \mathbf{\Lambda}^{(k)}) \right\}, \quad (28)$$

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

The prior distribution of the k-th cross validation model parameters $P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} | \boldsymbol{O}^{(\bar{k})})$ is obtained from equation (13):

$$\begin{aligned} P(\boldsymbol{\mu}^{(k)}, \boldsymbol{S}^{(k)} | \boldsymbol{O}^{(\bar{k})}) &= \mathcal{N}(\boldsymbol{\mu}^{(k)} | \bar{\boldsymbol{O}}^{(\bar{k})}, (T^{(\bar{k})} \boldsymbol{S}^{(k)})^{-1}) \\ &\quad \times \mathcal{W}(\boldsymbol{S}^{(k)} | T^{(\bar{k})} + D, (T^{(\bar{k})} \bar{\boldsymbol{C}}^{(\bar{k})})), \end{aligned} \quad (29)$$

$$\bar{\boldsymbol{O}}^{(\bar{k})} = \frac{1}{T^{(\bar{k})}} \sum_{t}^{T^{(\bar{k})}} \boldsymbol{O}_t^{(\bar{k})}, \quad (30)$$

$$\bar{\boldsymbol{C}}^{(\bar{k})} = \frac{1}{T^{(\bar{k})}} \sum_{t}^{T^{(\bar{k})}} \boldsymbol{O}_t^{(\bar{k})} \boldsymbol{O}_t^{(\bar{k})\top} - \bar{\boldsymbol{O}}^{(\bar{k})} \bar{\boldsymbol{O}}^{(\bar{k})\top}, \quad (31)$$

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.

3. 2 Bayesian context clustering using cross validation

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

$$\mathcal{F}^{(CV)} = \sum_{k=1}^{K} \mathcal{F}^{(k)}. \quad (32)$$

In the proposed method, an optimal model structure can be selected by maximizing the objective function $\mathcal{F}^{(CV)}$ instead of $\mathcal{F}$. As equation (26), the question which maximizes the gain of the objective function $\Delta\mathcal{F}_r^{(CV)}$ is selected. By splitting nodes until $\Delta\mathcal{F}_{\hat{r}}^{(CV)} \leqq 0$, the decision tree which maximizes the objective function $\mathcal{F}^{(CV)}$ is obtained. Using cross validation, an appropriate model structure can be selected without tuning parameters.

4. Experiments

4. 1 Experimental conditions

To evaluate the performance of the proposed method, speech synthesis experiments were performed. In these experiments, the ATR Japanese speech database [11] B-set which consits of the phonetically balanced 503 sentences, uttered by six male and four female speakers, 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, multi-space probability distribution HMM (MSD-HMM) [12] 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.

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

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

4. 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 12 Japanese graduate students. Twenty sentences were randomly chosen from the evaluation

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

	mel-cepstram	F_0	duration
MDL	956	1151	280
CVB	9,070	12,836	4,005
CVB-MDL	1,941	565	47
ML-CVB	15,077	8,844	3,185

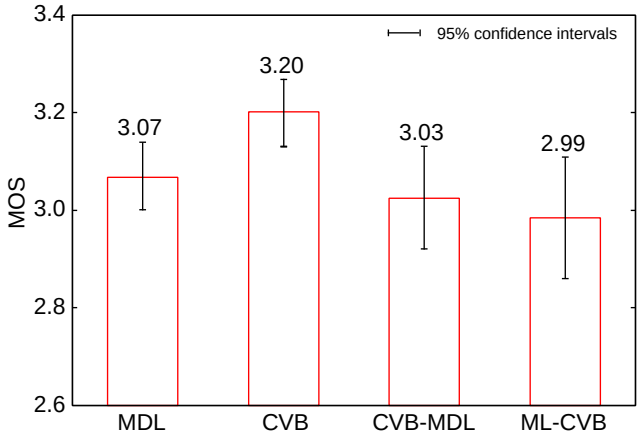

Fig. 1 Mean opinion scores of synthsized speech by the conventional and proposed methods. Error bars show 95% confidence intervals.

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 "CVB" achieved a better subjective score than the conventional method "MDL." Moreover, although "ML-CVB" have the similar number of states as "CVB," the subjective score of "ML-CVB" was worse than "CVB," and although "CVB-MDL" is trained by the Bayesian criterion, the subjective score of "CVB-MDL" was worse than "CVB." These results cleary show the effectiveness of the proposed method in both the model training and model structure selection.

5. Conclusion

This paper proposed a Bayesian approach to the HMM based speech synthesis using the determination technique of prior distribution based on cross validation. The results on the MOS test demonstrated that proposed method outperform the conventional one. The proposed method could determine prior distributions without tuning parameters, and select the model structure which accurately predict acoustic features for each HMM state.

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.

Acknowledgement 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.

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, pp.2347–2350, 1999.

[2] A.W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in Proc. ICASSP, pp.1229–1232, 2007.

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

[4] 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.

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

[6] 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.

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

[8] 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).

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in Proc. ICASSP, pp.1315–1318, 2000.

[10] 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.

[11] 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.

[12] 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.