INTERSPEECH 2009 BRIGHTON

A Bayesian Approach to Hidden Semi-Markov Model Based Speech Synthesis

Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan

Abstract

This paper proposes a Bayesian approach to hidden semi-Markov model (HSMM) based speech synthesis. Recently, hidden Markov model (HMM) based speech synthesis based on the Bayesian approach was proposed. The Bayesian approach is a statistical technique for estimating reliable predictive distributions by treating model parameters as random variables. In the Bayesian approach, all processes for constructing the system are derived from one single predictive distribution which exactly represents the problem of speech synthesis. However, there is an inconsistency between training and synthesis: although the speech is synthesized from HMMs with explicit state duration probability distributions, HMMs are trained without them. In this paper, we introduce an HSMM, which is an HMM with explicit state duration probability distributions, into the HMM-based Bayesian speech synthesis system. Experimental results show that the use of HSMM improves the naturalness of the synthesized speech.

Index Terms: speech synthesis, HSMM, Bayesian approach

1. Introduction

A statistical speech synthesis system based on hidden Markov models (HMMs) was recently developed. In the HMM-based speech synthesis, spectrum, excitation and duration of speech are modeled simultaneously by HMMs, and speech parameter sequences are generated from the HMMs themselves [1]. 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 employed to select the model structure [2]. However, since the ML criterion produces a point estimate of HMM parameters, the estimation accuracy may be degraded when small training data is available. Because the MDL criterion is based on the asymptotic assumption, it is ineffective when the amount of training data is small.

A framework of speech synthesis based on the Bayesian approach was recently proposed [3]. In this framework, all processes for constructing the system are derived from one single predictive distribution which exactly represents the problem of speech synthesis. The Bayesian approach assumes that model parameters are random variables and reliable predictive distributions are estimated by marginalizing model parameters. However, the estimation of posterior distributions of latent variables lead to a huge computational cost. To overcome this problem, the variational Bayesian method has been proposed as a tractable approximation method of the Bayesian approach [4] and it shows a good performance in the HMM-based speech recognition [5]. In the model selection, since the Bayesian approach does not use an asymptotic assumption as 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 [5, 6].

In the HMM-based speech synthesis, rhythm and tempo are controlled by state duration probability distributions. One of major limitation of HMMs is that they do not provide an adequate representation of the temporal structure of speech. This is because the probability of state occupancy decreases exponentially with time. To overcome this limitation, in the HMM-based speech synthesis system, each state duration probability distribution is explicitly modeled by a single Gaussian distribution. They are estimated from statistics obtained in the last iteration of the forward-backward algorithm, and then clustered by the decision tree-based context clustering [7, 8]. In the synthesis part, we construct a sentence HMM corresponding to an arbitrarily given text and determine state durations which maximize their probabilities. Then, a speech parameter sequence is generated for the given state sequence by the speech parameter generation algorithm [9]. However, there is an inconsistency between training and synthesis: although speech is synthesized from HMMs with explicit state duration probability distributions, HMMs are trained without them. To overcome this inconsistency, hidden semi-Markov model (HSMM) based speech synthesis has been proposed [10]. This framework introduces an HSMM, which is an HMM with explicit state duration probability distributions, into not only for synthesis but also training in the HMM-based speech synthesis system. In this paper, we propose a Bayesian approach to the HSMM-based speech synthesis. Using HSMMs as acoustic models, the proposed method outperforms the HMM-based Bayesian speech synthesis.

The rest of this paper is organized as follows. Section 2 describes the relation between HMM and HSMM. Section 3 describes th Bayesian approach to speech synthesis. HSMM based Bayesian speech synthesis is described in Section 4. In Section 5, subjective listening test results are presented. Concluding remarks and future work are presented in final section.

2. Hidden semi-Markov model

2.1. Likelihood computation of the HMM

The model likelihood of an HMM $\mathbf{\Lambda}$ for an observation vector sequence $\boldsymbol{o} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ can be computed efficiently by the forward-backward algorithm. First, we define partial forward likelihood $\alpha_t(\cdot)$ as follows:

$$\begin{aligned}\alpha_t(j) &= P(\boldsymbol{o}_1, \cdots, \boldsymbol{o}_t, q_t = j \mid \mathbf{\Lambda}) \\ &= \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t), \quad 1 \leq t \leq T, 1 \leq j \leq N \end{aligned} \tag{1}$$

where a_{ij} is a state transition probability from i-th state to j-th state, $b_j(\boldsymbol{o}_t)$ is an output probability of observation vector $\boldsymbol{o}_t$ from j-th state, N is a total number of HMM states. To begin the recursion Eq. (1), we set $\alpha_1(j) = \pi_j b_j(\boldsymbol{o}_1), 1 \leq j \leq N$, where π_j is an initial state probability of j-th state. Secondly,

partial backward likelihood $\beta_t(\cdot)$ is defined as follows:

$$\begin{aligned}\beta_t(i) &= P(\boldsymbol{o}_{t+1}, \cdots, \boldsymbol{o}_T \mid q_{t+1} = i, \boldsymbol{\Lambda}) \\ &= \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(i), \;\; 1 \leq t \leq T, 1 \leq i \leq N \end{aligned} \tag{2}$$

To begin the recursion Eq. (2), we set $\beta_T(i) = 1, 1 \leq i \leq N$. From Eqs. (1) and (2), the model likelihood $P(\boldsymbol{o} \mid \boldsymbol{\Lambda})$ is computed as

$$P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) = \sum_{i=1}^{N} \alpha_t(i) \cdot \beta_t(i). \;\; 1 \leq t \leq T \tag{3}$$

2.2. Likelihood computation of the HSMM

The model likelihood of an HSMM $\boldsymbol{\Lambda}'$ for an observation vector sequence $\boldsymbol{o} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ can be computed efficiently by the generalized forward-backward algorithm. We can compute partial forward likelihood $\alpha'_t(\cdot)$ and partial backward likelihood $\beta'_t(\cdot)$ recursively as follows:

$$\alpha'_0(j) = \pi_j, \tag{4}$$

$$\begin{aligned}\alpha'_t(j) &= \sum_{d=1}^{t} \sum_{i=1, j \neq j}^{N'} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \\ &\quad \times \prod_{s=t-d+1}^{t} b'_j(\boldsymbol{o}_s), \;\; 1 \leq t \leq T \end{aligned} \tag{5}$$

$$\beta'_T(i) = 1, \tag{6}$$

$$\begin{aligned}\beta'_t(i) &= \sum_{d=1}^{T-t} \sum_{j=1, j \neq i}^{N'} a'_{ij} p'_j(d) \\ &\quad \times \prod_{s=t+1}^{t+d} b'_j(\boldsymbol{o}_s) \beta'_{t+d}(j), \;\; 1 \leq t \leq T \end{aligned} \tag{7}$$

where $a'_{ij}, b'_j(\boldsymbol{o}_t), N' p'_j(d)$, and π'_j are a state transition probability from i-th state to j-th state, an output probability of observation vector $\boldsymbol{o}_t$ from j-th state, a total number of HSMM states, a state duration probability of j-th state, and an initial state probability of j-th state, respectively. From above equations, the model likelihood $P(\boldsymbol{o} \mid \boldsymbol{\Lambda}')$ is given by

$$\begin{aligned}P(\boldsymbol{o} \mid \boldsymbol{\Lambda}') &= \sum_{i=1}^{N'} \sum_{j=1, i \neq j}^{N'} \sum_{d=1}^{t} \alpha'_{t-d}(i) a'_{ij} p'_j(d) \\ &\quad \times \prod_{s=t-d+1}^{t} b'_j(\boldsymbol{o}_s) \beta'_t(j). \end{aligned} \tag{8}$$

3. Bayesian approach to speech synthesis

3.1. Bayesian approach

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 [3]:

$$\begin{aligned}\boldsymbol{o}_{Bayes} &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o} \mid s, \boldsymbol{O}, S) \\ &= \arg\max_{\boldsymbol{o}} P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \,. \end{aligned} \tag{9}$$

It can be seen that Eq. (9) 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:

$$\begin{aligned}&P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\ &\quad = \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\ &\quad = \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda}) P(\boldsymbol{O}, \boldsymbol{Q} \mid S, \boldsymbol{\Lambda}) P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \end{aligned} \tag{10}$$

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 Eq. (10) 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 [4].

3.2. Variational Bayesian method

The variational 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:

$$\begin{aligned}&\log P(\boldsymbol{o}, \boldsymbol{O} \mid s, S) \\ &\quad = \log \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S) d\boldsymbol{\Lambda} \\ &\quad = \log \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}) \frac{P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\ &\quad \geq \sum_{\boldsymbol{q}} \sum_{\boldsymbol{Q}} \int Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}) \log \frac{P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} d\boldsymbol{\Lambda} \\ &\quad = \left\langle \log \frac{P(\boldsymbol{o}, \boldsymbol{q}, \boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{\Lambda} \mid s, S)}{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} \right\rangle_{Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda})} \\ &\quad = \mathcal{F} \end{aligned} \tag{11}$$

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.

$$Q(\boldsymbol{q}, \boldsymbol{Q}, \boldsymbol{\Lambda}) = Q(\boldsymbol{q})\, Q(\boldsymbol{Q})\, Q(\boldsymbol{\Lambda}) \tag{12}$$

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

$$Q(\boldsymbol{q}) = C_{\boldsymbol{q}} \exp \langle \log P(\boldsymbol{o}, \boldsymbol{q}|s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{\Lambda})} \;, \tag{13}$$

$$Q(\boldsymbol{Q}) = C_{\boldsymbol{Q}} \exp \langle \log P(\boldsymbol{O}, \boldsymbol{Q}|S, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{\Lambda})} \;, \tag{14}$$

$$\begin{aligned}Q(\boldsymbol{\Lambda}) = C_{\boldsymbol{\Lambda}} P(\boldsymbol{\Lambda}) &\exp \langle \log P(\boldsymbol{o}, \boldsymbol{q}|s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{q})} \\ &\times \exp \langle \log P(\boldsymbol{O}, \boldsymbol{Q}|S, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{Q})} \;, \end{aligned} \tag{15}$$

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.

4. HSMM based Bayesian speech synthesis

4.1. Optimization of poterior distributions

In the HSMM based Bayesian speech synthesis, the optimizations using Eqs. (13, 14, 15) can be effectively performed by iterative calculations as the expectation maximization (EM) algorithm, which increases the value of objective function $\mathcal{F}$ at each iteration until convergence. The normalization term $C_{\boldsymbol{q}}$ of an HSMM can be computed efficiently by the generalized forward-backward algorithm for the variational Bayes method.

$$\begin{aligned} C_{\boldsymbol{q}}^{-1} &= \sum_{\boldsymbol{q}} \exp \langle \log P(\boldsymbol{o}, \boldsymbol{q} \mid s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{\Lambda})} \\ &= \sum_{i=1}^{\hat{N}} \sum_{j=1, i \neq j}^{\hat{N}} \sum_{d=1}^{t} \hat{\alpha}_{t-d}(i) \exp \langle \log a_{ij} \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \exp \langle \log p_j(d) \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \prod_{s=t-d+1}^{t} \exp \langle \log b_j(\boldsymbol{o}_s) \rangle_{Q(\boldsymbol{\Lambda})} \hat{\beta}_t(j). \end{aligned} \tag{16}$$

We can compute partial forward likelihood $\hat{\alpha}_t(\cdot)$ and partial backward likelihood $\hat{\beta}_t(\cdot)$ recursively as follows:

$$\begin{aligned} \hat{\alpha}_t(j) &= \sum_{d=1}^{t} \sum_{i=1, j \neq j}^{\hat{N}} \hat{\alpha}_{t-d}(i) \exp \langle \log a_{ij} \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \exp \langle \log p_j(d) \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \prod_{s=t-d+1}^{t} \exp \langle \log b_j(\boldsymbol{o}_s) \rangle_{Q(\boldsymbol{\Lambda})}, \end{aligned} \tag{17}$$

$$\begin{aligned} \hat{\beta}_t(i) &= \sum_{d=1}^{T-t} \sum_{j=1, j \neq i}^{\hat{N}} \exp \langle \log a_{ij} \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \exp \langle \log p_j(d) \rangle_{Q(\boldsymbol{\Lambda})} \\ &\quad \times \prod_{s=t+1}^{t+d} \exp \langle \log b_j(\boldsymbol{o}_s) \rangle_{Q(\boldsymbol{\Lambda})} \hat{\beta}_{t+d}(j). \end{aligned} \tag{18}$$

Because the Bayesian approach assumes that a set of model parameters $\boldsymbol{\Lambda}$ is a random variable, model parameters are represented by the expectation values. The normalization term $C_{\boldsymbol{Q}}$ can be computed as like Eq. (16). Although the computational cost is increased by using HSMMs, the Bayesian approach requires almost the same computational cost with the ML criterion.

4.2. Prior distribution for duration distribution

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

$$P(\boldsymbol{\mu}, \boldsymbol{S}) = \mathcal{N}(\boldsymbol{\mu} \mid \nu, (\xi \Sigma)^{-1}) \, \mathcal{G}\left(\Sigma \mid \frac{\eta}{2}, \frac{B}{2}\right), \tag{19}$$

where $\{\xi, \eta, \nu, B\}$ is a hyper-parameter set. Using a conjugate prior distribution, a set of parameters of posterior distribution is also represented by the same parameter set $\{\bar{\xi}, \bar{\eta}, \bar{\boldsymbol{\nu}}, \bar{\boldsymbol{B}}\}$.

4.3. Speech parameters generation

In the synthesis part, first an arbitrarily given text to be synthesized is converted to a context-dependent label sequence and a sentence HSMM is constructed by concatenating context-dependent HSMMs according to the label sequence. Secondly, state durations $\boldsymbol{d}$ of the sentence HSMM $\boldsymbol{\Lambda}$ are determined as follows:

$$\boldsymbol{d}_{max} = \arg\max_{\boldsymbol{d}} \langle \log P(\boldsymbol{d} \mid \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{\Lambda})} \tag{20}$$

Thirdly, a speech parameter sequence is generated for a given state sequence. 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

$$\begin{aligned} \boldsymbol{o} &= \boldsymbol{W}\boldsymbol{c} \\ &= \left[(\boldsymbol{W}\boldsymbol{c})_1^\top, (\boldsymbol{W}\boldsymbol{c})_2^\top, \ldots, (\boldsymbol{W}\boldsymbol{c})_T^\top\right]^\top \end{aligned} \tag{21}$$

$$(\boldsymbol{W}\boldsymbol{c})_t = \left[\boldsymbol{c}_t^\top, \Delta \boldsymbol{c}_t^\top, \Delta^2 \boldsymbol{c}_t^\top\right]^\top \tag{22}$$

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 Bayesian method, the lower bound $\mathcal{F}$ approximates the log marginal likelihood $\log P(\boldsymbol{W}\boldsymbol{c}, \boldsymbol{O} \mid s, S)$. Therefore, the optimal speech parameter sequence $\hat{\boldsymbol{c}}$ is generated by maximizing the lower bound $\mathcal{F}$:

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \boldsymbol{c}} &= \frac{\partial}{\partial \boldsymbol{c}} \langle \log P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\Lambda}) P(\boldsymbol{q} \mid s, \boldsymbol{\Lambda}) \rangle_{Q(\boldsymbol{q})Q(\boldsymbol{\Lambda})} \\ &= \left\langle \frac{\partial}{\partial \boldsymbol{c}} \log P(\boldsymbol{W}\boldsymbol{c} \mid \boldsymbol{q}, \boldsymbol{\Lambda}) \right\rangle_{Q(\boldsymbol{q})Q(\boldsymbol{\Lambda})} = \boldsymbol{0}. \end{aligned} \tag{23}$$

Under the condition in Eq. (21), the optimal static feature sequence $\hat{\boldsymbol{c}}$ can be determined by solving the following set of linear equations:

$$\boldsymbol{W}^\top \langle \boldsymbol{S} \rangle \boldsymbol{W} \hat{\boldsymbol{c}} = \boldsymbol{W}^\top \langle \boldsymbol{S}\boldsymbol{\mu} \rangle, \tag{24}$$

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. Eq. (24) can be solved efficiently using the Cholesky or QR decomposition [9].

5. Experiments

5.1. Experimental conditions

To evaluate the performance of the proposed method, the speech synthesis experiment was performed. In this experiment, the following four models were compared.

- "ML-HMM" : HMMs trained by the ML criterion were used as the acoustic models. Model structures were selected by the MDL criterion.
- "ML-HSMM" : HSMMs trained by the ML criterion were used as the acoustic models. Model structures were selected by the MDL criterion.
- "Bayes-HMM" : HMMs trained by the Bayesian method were used as the acoustic models. Model structures were selected by the Bayesian criterion with cross validation [3].
- "Bayes-HSMM" : HSMMs trained by the Bayesian method were used as the acoustic models. Model structures were selected by the Bayesian criterion with cross validation.

Table 1 represents the details of the number of states.

In this experiment, the ATR Japanese speech database [11] B-set which consists of the phonetically balanced 503 sentences

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

	mel-cepstram	F_0	duration
ML-HMM	1,115	2,267	275
ML-HSMM	1,128	2,272	283
Bayes-HMM	9,532	16,044	3,005
Bayes-HSMM	9,485	16,130	3,490

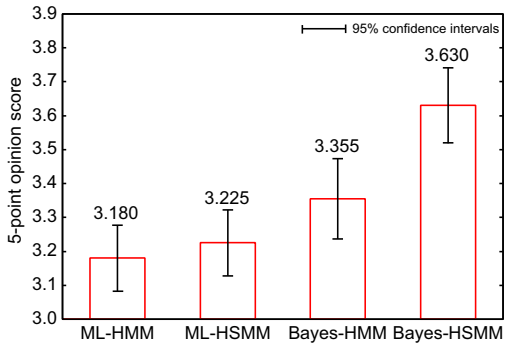

Figure 1: Mean opinion scores of synthesized speech by the conventional and proposed methods. Error bars show 95% confidence intervals.

was 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 five-state, left-to-right MSD-HSMM and MSD-HMM [12] without skip transition 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.

5.2. Experimental results

A subjective listening test was conducted to evaluate the 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 students in our research group. Twenty sentences were chosen at random 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 model "Bayes-HSMM" achieved a better subjective score than the conventional model "Bayes-HMM," and the subjective score of "ML-HSMM" was better than "ML-HMM." Consequently, the speech quality is improved by using HSMMs as the acoustic models. Moreover, the proposed model "Bayes-HSMM" outperformed the model "ML-HSMM." These results clearly show the effectiveness of the proposed model. The number of states of "Bayes-HMM" and "Bayes-HSMM" was considerable larger than "ML-HMM" and "ML-HSMM." Although the large model structure alleviated the over-smoothihg problem, the ML training leads to the over-fitting problem. However, the Bayesian approach avoided the over-fitting problem because the posterior distributions of model parameters were used. Therefore, the Bayesian approach overcame the over-fitting and over-smoothing problems simultaneously. Consequently, most of the subjects observed that the proposed model improved the naturalness in spectrum and excitation.

6. 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 work includes investigation of the relation between the speech quality and the size of model structure.

7. Acknowledgements

The research leading to these results was partly funded from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project http://www.emime.org).

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

[3] K. Hashimoto, H. Zen, Y. Nankaku, T. Masuko, and K. Tokuda, "A Bayesian approach to HMM-based speech synthesis," in Proc. ICASSP, pp.4029–4032, 2009.

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

[5] 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.

[6] 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.

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

[8] 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.

[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.936–939, 2000.

[10] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in Proc. ICSLP, pp.1185–1180, 2004.

[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.