Multi-Speaker Modeling with Shared Prior Distributions and Model Structures for Bayesian Speech Synthesis

Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, JAPAN

Abstract

This paper investigates a multi-speaker modeling technique with shared prior distributions and model structures for Bayesian speech synthesis. The quality of synthesized speech is improved by selecting appropriate model structures in HMM-based speech synthesis. Bayesian approach is known to work for such model selection. However, the result is strongly affected by prior distributions of model parameters. Therefore, determination of prior distributions and selection of model structures should be performed simultaneously. This paper investigates prior distributions and model structures in the situation where training data of multiple speakers are available. The prior distributions and model structures which represent acoustic features common to every speakers can be obtained by sharing them between multiple speaker-dependent models.

Index Terms: speech synthesis, Bayesian approach, prior distribution, context clustering, multi-speaker modeling

1. Introduction

A statistical parametric speech synthesis system based on hidden Markov models (HMMs) was recently developed. In HMM-based speech synthesis, the spectrum, excitation, and duration of speech are simultaneously modeled with HMMs, and speech parameter sequences are generated from the HMMs themselves [1]. The maximum likelihood (ML) criterion has typically been used for training HMMs and generating speech parameters. The ML criterion guarantees that the ML estimates approach the true values of the parameters. However, since the ML criterion produces a point estimate of the model parameters, its estimation accuracy may degrade when the amount of training data is insufficient. In the Bayesian approach, all variables introduced when the models are parameterized, such as model parameters and latent variables, are treated as random variables, and their posterior distributions are obtained by the Bayes theorem. The Bayesian approach can generally construct variables introduced when the models are parameterized, such as model parameters and latent variables, are treated as random variables, and their posterior distributions are obtained by the Bayes theorem. The Bayesian approach assumes that a set of model parameters are generated from a predictive distribution as follows.

$$\hat{\alpha}_{Bayes} = \arg \max_{\alpha} P(\alpha \mid I, O, L)$$

$$= \arg \max_{\alpha} P(\alpha, O \mid I, L)$$

(1)

To overcome this problem, we have proposed Bayesian context clustering using cross validation [3]. In this method, prior distributions are determined by using a part of training data, and model structures are evaluated by using the determined prior distribution based on cross validation.

In this paper, we investigate prior distributions and model structures in a situation where training data of multiple speakers can be used. Since there are acoustic features common to every speaker, i.e., acoustic features being independent of speakers, the appropriate multiple speaker-dependent models can be estimated by effectively using training data of multiple speakers. Therefore, the prior distributions and model structures that represent speaker-independent features are obtained by sharing them among multiple speaker-dependent models. Shared tree clustering (STC) has been proposed as the model selection method for multi-speaker modeling [4]. By sharing model structures among multiple speaker-dependent models, canonical model structures can be selected. When prior distributions are shared among multiple speaker-dependent models, speaker-independent prior distributions are obtained by using training data of multiple speakers as the prior data. The speaker-independent prior distribution is also known as a universal background model (UBM) [5] in the field of speaker recognition. In addition, speaker-adaptive prior distributions can be estimated using multi-speaker modeling techniques, such as the speaker adaptive training (SAT) method [6]. This method estimates the canonical prior distributions so that the speaker-dependent models are estimated well. In this paper, these prior distributions and model structures were applied to Bayesian speech synthesis, and multi-speaker modeling with shared prior distributions and model structures were evaluated on a subjective evaluation test.

The rest of this paper is organized as follows. Section 2 describes Bayesian speech synthesis. Section 3 introduces a multi-speaker modeling with shared prior distributions and model structures for Bayesian speech synthesis. Subjective listening test results are presented in Section 4. Concluding remarks and future work are presented in the final section.

2. Bayesian approach

2.1. Bayesian speech synthesis

The Bayesian approach assumes that a set of model parameters is a random variable, while the ML approach estimates constant model parameters. In the Bayesian approach, the speech parameters are generated from a predictive distribution as follows.

$$\hat{\alpha}_{Bayes} = \arg \max_{\alpha} P(\alpha \mid I, O, L)$$

$$= \arg \max_{\alpha} P(\alpha, O \mid I, L)$$

(1)
where \( o, l, O \) and \( L \) are respectively synthesis data, a label sequence of the synthesis data, training data, and a label sequence of the training data, respectively. It can be seen that Eq. (1) directly represents the problem of speech synthesis; that is, the speech feature sequence \( o \) is generated from given training feature sequences \( O \). When the latent variable \( m \), which represents a model structure, is introduced into this predictive distribution, Eq. (1) can be represented as

\[
P(o, O \mid l, L) = \sum_m P(o, O \mid l, L, m) P(m)
\]

(2)

where \( P(m) \) is a prior distribution for model structures. In Eq. (2), the predictive distribution is obtained by marginalizing the all possible model structures. However, the predictive distribution can be approximated by selecting the single model structure which obtains a maximum posterior probability. Since the posterior distribution of model structures \( P(m \mid o, O, l, L) \) is proportional to the marginal likelihood \( P(o, O \mid l, L, m) \) when it is assumed that the prior distribution of model structures \( P(m) \) is a uniform distribution, the model structure that maximizes the posterior distribution is obtained as follows.

\[
m = \arg \max_m P(o, O \mid l, L, m)
\]

(3)

If the posterior distribution of model structures is a single peak and a narrow distribution, the predictive distribution is approximated well by using a single model structure.

\[
P(o, O \mid l, L) \approx P(o, O \mid l, L, m)
\]

(4)

In this work, the predictive distribution with a single model structure is used. Although the latent variable of a model structure \( m \) is omitted for a simple representation in the rest of this paper, a single model structure is used in the predictive distribution.

When it is assumed that a HMM is used as an acoustic model, the predictive distribution is represented as

\[
P(o, O \mid l, L) = \sum_z \sum_Z \int P(o, z \mid l, A) \times P(O, Z \mid L, A) P(A) dA
\]

(5)

where \( z \) and \( Z \) are a sequence of HMM states for synthesis and training data respectively, and \( A \) is a set of model parameters. In addition, \( P(o, z \mid l, A) \) is a likelihood of synthesis data \( o \), \( P(O, Z \mid L, A) \) is a likelihood of training data \( O \), and \( P(A) \) is a prior distribution of model parameter \( A \). The model parameters are integrated out in Eq. (5) so that the effect of over-fitting is mitigated. However, solving the integral and expectation calculations is difficult. The calculations become more complicated when a model includes latent variables. The variational Bayesian method has been proposed as a tractable approximation to overcome this problem, and it has good generalization performance in many applications [7].

### 2.2. Variational Bayesian method

The variational Bayesian method estimates an approximate posterior distribution \( Q(z, Z, A) \) by maximizing the lower bound of the log marginal likelihood \( F \) instead of the true marginal likelihood. The lower bound \( F \) is defined by using Jensen’s inequality.

\[
\log P(o, O \mid l, L)
\]

\[
= \log \sum_z \sum_Z \int P(o, z, O, Z, A \mid l, L) dA
\]

\[
= \log \sum_z \sum_Z \int \frac{P(o, z, O, Z, A \mid l, L)}{Q(z, Z, A)} Q(z, Z, A) dA
\]

\[
\geq \langle \log P(o, z, O, Z, A \mid l, L) \rangle_{Q(z, Z, A)} = F
\]

(6)

where \( \langle \cdot \rangle_Q \) denotes a calculation of the expectation with respect to \( Q \), and \( Q(z, Z, A) \) is an approximate distribution of the true posterior distribution \( P(z, Z, A \mid o, O, l, L) \). The VB method assumes that the probabilistic variables associated with \( z, Z, A \) are statistically independent of the other variables.

\[
Q(z, Z, A) = Q(z)Q(Z)Q(A)
\]

(7)

The optimal posterior distribution of each factor can be obtained by maximizing the objective function \( F \) with the variational method. These posterior distributions can be updated by using iterative calculations similar to those of the EM algorithm in the ML approach. The VB method guarantees that the log marginal likelihood is approximated well by the lower bound \( F \). Consequently, an optimal model structure can be selected by maximizing the objective function \( F \).

### 3. Multi-speaker modeling with shared prior distributions and model structures

#### 3.1. Multi-speaker modeling for Bayesian speech synthesis

A speaker-dependent model is conventionally estimated from only the training data of a single speaker. However, there are acoustic features common to every speaker, i.e., acoustic features being independent of speakers. Therefore, the appropriate multiple speaker-dependent models can be estimated by effectively using training data of multiple speakers.

By using speaker index \( r \), the synthesis and training data can be represented as \( o = \{o^{(1)}, \ldots, o^{(R)}\} \) and \( O = \{O^{(1)}, \ldots, O^{(R)}\} \). Here, \( R \) is the number of speakers included in the synthesis and training data. The log marginal likelihood of each speaker and the lower bound are represented as

\[
\sum_r \log P(o^{(r)}, O^{(r)} \mid l^{(r)}, L^{(r)}, m^{(r)}) \geq \sum_r F^{(r)}
\]

(8)

where \( F^{(r)} \) is a lower bound of speaker \( r \). The multiple speaker-dependent models are estimated by maximizing \( \sum_r F^{(r)} \). In this work, we investigates prior distributions and model structures in a situation where training data of multiple speakers can be used.

#### 3.2. Model structures for multi-speaker modeling

By sharing model structures among multiple speaker-dependent models, the model structures that represent acoustic features common to every speakers can be selected. In this method, the model structure is selected by maximizing the sum of the lower
by maximizing the sum of the lower bounds $\sum_r F(r)$, as like a SAT method.

$$\hat{\theta} = \arg\max_{\theta} \sum_r \log P(\alpha^{(r)}, O^{(r)} \mid l^{(r)}, L^{(r)})$$

$$\approx \arg\max_{\theta} \sum_r F(r)$$ (14)

This method estimates the canonical prior distributions so that the speaker-dependent models are estimated well. The hyper-parameters $\xi$ and $\eta$ are determined by using the tuning parameter $T$ as Eqs. (12) and (13).

4. Experiments

We used 2,250 utterances, which are recorded by five male speakers in our research group, as training data. The contexts of the data were the same as the B-set of the ATR Japanese speech database [8]. Speech signals were sampled at a 16-kHz rate and windowed at a 5-ms frame rate using a 25-ms Hamming window. Feature vectors consisted of spectrum and $F_0$ parameter vectors. The spectrum parameter vectors consisted of 24 mel-cepstral coefficients and their delta and delta-delta coefficients. The $F_0$ parameter vectors consisted of log $F_0$ and its delta and delta-delta. A five-state left-to-right MSD-HSMM [9, 10] without skip transitions was used. Each state output distribution 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 distribution was modeled by a one-dimensional Gaussian distribution. The decision tree-based context clustering technique was separately applied to distributions of the spectrum, $F_0$, and state duration.

A subjective listening test was conducted to evaluate the quality of the synthesized speech. The naturalness of the synthesized speech was assessed by the mean opinion score (MOS) test method. The subjects were 10 Japanese students belonging to our research group. Twenty sentences were chosen at random from the test sentences, and a speaker was chosen at random from five male speakers for each test sentence. Speech samples were presented in random order for each test sentence. In the MOS test, after listening to each test sample, the subjects were asked to assign the sample a five-point naturalness score (5: natural – 1: poor). In this subjective listening test, the following five systems were compared.

- **SD**: The model structures were selected for each speaker-dependent model. The prior distributions were determined from the training data of each speaker.
- **Tree**: The model structures were shared among all speaker-dependent models. The prior distributions were determined from the training data of each speaker.
- **Prior**: The model structures were selected for each speaker-dependent model. The prior distributions were determined from the training data of all speakers. Note that the prior distributions were not shared.
- **Tree-Prior**: The model structures and prior distributions were shared among all speaker-dependent models. The speaker-independent prior distributions were determined from the training data of all speakers.
The selection of speakers that have similar characteristics should be investigated for sharing prior distributions and model structures.

5. Conclusion

In this paper, a multi-speaker modeling technique with shared prior distributions and model structures for Bayesian speech synthesis was investigated. By sharing prior distributions and model structures among multiple speaker-dependent models, training data of multiple speakers can be used effectively and appropriate acoustic models can be estimated. The results of an MOS test demonstrated that the multi-speaker modeling with shared prior distributions and model structures outperforms the baseline method significantly. Our future work will include experiments using the training data of more speakers and investigation of model structures of prior distributions.

6. 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) and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan. A part of this research was supported by Japan Society for the Promotion of Science (JSPS) Research Fellowships for Young Scientists.

7. References


Figure 1: Results of MOS test comparing five models: SD, Tree, Prior, Tree-Prior, and Tree-SAT.