Simultaneous Optimization of Multiple Tree Structures for Factor Analyzed HMM-Based Speech Synthesis

Takenori Yoshimura, Kei Hashimoto, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation,
Nagoya Institute of Technology, Nagoya, Japan
{takenori, bonanza, nankaku, tokuda}@sp.nitech.ac.jp

10.21437/Interspeech.2015-303

Abstract

Some speech synthesis approaches are based on an assumption that voice characteristics, *e.g.*, speaker, speaking style, and emotion, are represented in a low-dimensional subspace. In these approaches, the model structures of the basis vectors which span the subspace are typically constructed with decision trees, and are important to synthesize high-quality speech. However, since it is difficult to evaluate all the candidates of the model structures, some strong constraints are usually applied in the model construction to reduce the huge computational complexity. To overcome this problem, this paper presents a new technique that simultaneously construct the model structures with multiple tree structures without the constraints. The proposed technique enables to find the more optimal model structures because the more complex model structure candidates can be evaluated by using some computational complexity reduction algorithms. Experimental results show that the proposed method improves the naturalness of the synthesized speech from the conventional one.

Index Terms: HMM-based speech synthesis, eigenvoice, factor analysis, decision trees, context clustering

1. Introduction

To let machines speak naturally like a human, a hidden Markov model (HMM)-based speech synthesis system has been proposed [1]. This system models spectrum, excitation and duration of speech simultaneously in a unified framework of HMMs, and synthesizes speech using parameter sequences generated from HMMs. One of the main advantages of the system is that the voice characteristics of synthesized speech can be more easily controlled than other systems due to its statistical modeling process. Many approaches using the advantage have been proposed to effectively control the voice characteristics [2, 3].

One of the such approaches is a well-known eigenvoice-based one [4, 5, 6]. The basic idea of the eigenvoice approach [7] is to find a small set of basis vectors (eigenvoices) from a diverse set of speaker's voices by assuming that voice characteristics, *e.g.*, speaker, speaking style, and emotion, can be sufficiently represented as a point in a low-dimensional subspace (eigenspace) rather than in a very high-dimensional model parameter space. Since each voice can be described in terms of a linear combination of the basis vectors span the subspace, new voices with various characteristics are obtained by changing the weights of the basis vectors which correspond to the point of the subspace.

Properly estimating the basis is a critical problem for synthesizing speech with high naturalness and desired voice characteristics. In this paper, we focus on the model structure for the basis to address this problem. The model structure can be seen as a relationship between linguistic contexts, such as lexical stress, pitch accent, tone, and part-of-speech information, and their acoustic realization. In order to take into account detailed acoustic variations, the basis should be modeled considering the contexts because it is well-known that spectral and prosodic features in human speech are affected by the contextual factors. Unfortunately, it is almost impossible to estimate the reliable basis for all the possible combinations of the contexts with a finite set of training data. To solve this problem, some model structure (parameter sharing structure) is required for each basis vectors. The decision tree-based technique, which is widely used in the standard HMM-based speech synthesis, can effectively construct an appropriate parameter sharing structure according to a binary decision tree [8], whose structure strongly depends on the attribute of a training set such as speaker individuality, speaker's dialect, gender, and emotion [9]. However, this technique cannot directly apply to the eigenvoice-based approaches because multiple decision trees are required for the basis vectors rather than a single tree. Although there are some approaches for building multiple decision trees, some strong constraints are usually applied: all the decision trees for the basis vectors have the same structure [4, 6] or while building a decision tree for a basis vector the parameters and the structures of the other basis vectors are held fixed [5, 10, 11]. The constraints can sufficiently reduce the computationally complexity, but bring the suboptimal structures that inappropriately capture the acoustic variations contained in a training set.

This paper describes a technique that simultaneously construct the multiple decision trees for all the basis vectors considering the correlation among the basis vectors in the framework of factor analyzed HMM (FAHMM) [6]. The model structure of FAHMM is based on factor analysis to represent various voice characteristics by controlling its factors, which correspond to the weights of the basis vectors. Due to a huge number of the model structures should be evaluated, enormous computational cost is required to build the decision trees. Fortunately, it is possible to substantially reduce the computational complexity by using some algorithms used in additive structure models [12] because its model structure is similar to that of FAHMM. In the framework of cluster adaptive training (CAT), a similar approach has been proposed [13]. One of the main differences between [13] and this paper is that the proposed method can evaluate all the candidates for model structures rather than the limited ones within reasonable computational time by introducing some computational complexity reduction algorithms. The proposed approach, hence, enables to find the more optimal model structures and to synthesize speech with high quality.

The rest of this paper is organized as follows. Section 2

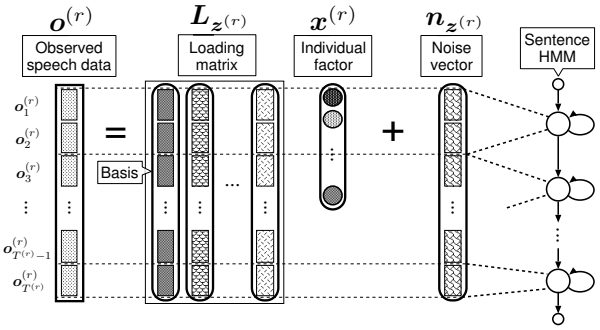


Figure 1: *The process of generating observation sequence in FAHMM.*

gives an overview of FAHMM. The multiple decision trees-based context clustering algorithm with the computational complexity reduction techniques is proposed in Section 3. Results on speaker adaptation task are presented in Section 4. Finally, Section 5 describes conclusions and future work.

2. Factor analyzed hidden Markov model

FAHMM [6] incorporates the idea of eigenvoice [7] into the model structure of HMM [1]. In this framework, the basis is constructed by training HMMs iteratively via a kind of expectation-maximization (EM) algorithm rather than by applying principle component analysis (PCA) to supervectors derived by stacking the parameters of speaker-dependent HMMs. Figure 1 shows the overview of the model structure of FAHMM. In this approach, an observation sequence $\boldsymbol{o}^{(r)}$ of a class[1] r is generated based on factor analysis:

$$\boldsymbol{o}^{(r)} = \boldsymbol{L}_{\boldsymbol{z}^{(r)}}\boldsymbol{x}^{(r)} + \boldsymbol{n}_{\boldsymbol{z}^{(r)}} \quad (1)$$

where $\boldsymbol{x}^{(r)}$ is a low-dimensional latent variable vector which is called factor, $\boldsymbol{L}_{\boldsymbol{z}^{(r)}}$ denotes a loading matrix whose column vector is called basis vector, $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ denotes a noise vector, and $\boldsymbol{z}^{(r)}$ is a state sequence of a context-dependent HMM sequence corresponding to an arbitrarily given text. The parameters $\boldsymbol{L}_{\boldsymbol{z}^{(r)}}$ and $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ are composed by concatenating the parameters of context-dependent HMMs according to the state sequence $\boldsymbol{z}^{(r)}$. The factor $\boldsymbol{x}^{(r)}$ and the noise vector $\boldsymbol{n}_{\boldsymbol{z}^{(r)}}$ are typically given by the following Gaussian distributions:

$$\boldsymbol{x}^{(r)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad (2)$$

$$\boldsymbol{n}_{\boldsymbol{z}^{(r)}} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}^{(r)}}, \boldsymbol{\Sigma}_{\boldsymbol{z}^{(r)}}) \quad (3)$$

where $\boldsymbol{\mu}_{\boldsymbol{z}^{(r)}}$ is a noise mean vector and $\boldsymbol{\Sigma}_{\boldsymbol{z}^{(r)}}$ is a noise diagonal covariance matrix. As factors are shared within a class while loading matrices and noise vectors are shared in all classes, various voice characteristics are controlled by the factor.

The output distribution of $\boldsymbol{o}_t^{(r)}$ ($\boldsymbol{o}^{(r)}$ at a frame t) for a context m is given by

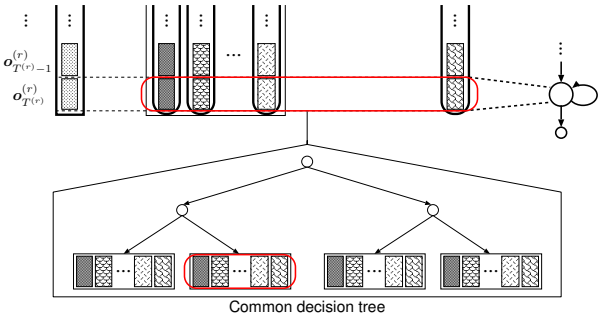
$$p\left(\boldsymbol{o}_t^{(r)} \middle| m, \boldsymbol{x}^{(r)}, \boldsymbol{\Lambda}\right) = \mathcal{N}\left(\boldsymbol{o}_t^{(r)} \middle| \boldsymbol{W}_m\boldsymbol{\xi}^{(r)}, \boldsymbol{\Sigma}_m\right) \quad (4)$$

where

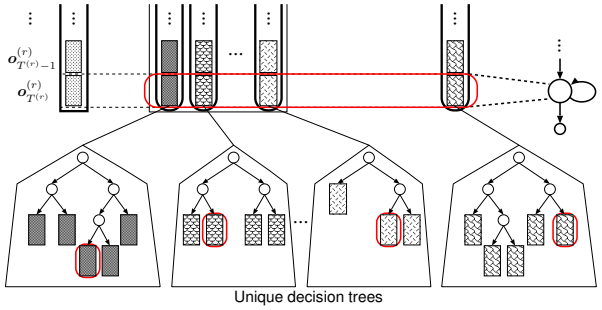
$$\boldsymbol{W}_m = \left[\begin{array}{cc} \boldsymbol{\mu}_m & \boldsymbol{L}_m \end{array}\right], \quad (5)$$

$$\boldsymbol{\xi}^{(r)} = \left[\begin{array}{cc} 1 & \boldsymbol{x}^{(r)\mathsf{T}} \end{array}\right]^{\mathsf{T}}, \quad (6)$$

[1] A class is an arbitrary group that share a factor among all the models associated in the group. For instance, every speaker is a class in the experiment described in this paper.



(a) *with the constraint that all decision trees for the basis vectors have the same structure*



(b) *without the above constraint*

Figure 2: *The model structure of FAHMM.*

and $\boldsymbol{\Lambda}$ is a set of model parameters. In order to estimate reliable parameters for all possible contexts, some parameter sharing structure for $\boldsymbol{W}_m$ is required. The decision tree-based context clustering [8] is widely used to construct a parameter sharing structure in the standard HMM-based speech synthesis. The technique clusters acoustically similar models into the same cluster and ties the model parameters among all the models associated to the same cluster. A binary decision tree (parameter sharing structure) is built by applying a phonetic question to a cluster and iteratively splitting the cluster into two child clusters. This technique, however, cannot be directly applied to FAHMM because each basis vector requires a parameter sharing structure and depends on each other. One approach to solve this problem is to assume that all the decision trees for the basis vectors have the same structure [6]. The model structure is illustrated in Figure 2a. This assumption allows to directly use the decision tree-based context clustering. Another approach is to assume that each basis vector does not depend on each other: when building a model structure for a basis vector, the other model structures and their model parameters are held fixed [10]. Figure 2b shows the model structure. However, the constraints prevent to appropriately capture acoustic variations.

3. Multiple decision trees-based context clustering for FAHMM

The optimal parameter sharing structure for a basis vector seems to differ each other. To build the model structures, simultaneously clustering strategy is essential rather than sequentially one due to their dependency. In terms of the output probability

of FAHMM, the auxiliary function should be maximized is

$$\mathcal{Q} = \sum_{r,m,t} \gamma_{m,t}^{(r)} \left\langle \log \mathcal{N}\left(\boldsymbol{o}_t^{(r)} \,\middle|\, \boldsymbol{W}_m \boldsymbol{\xi}^{(r)}, \boldsymbol{\Sigma}_m\right)\right\rangle \quad (7)$$

where $\gamma_{m,t}^{(r)}$ is the posterior probability of the context m generating the observation $\boldsymbol{o}_t^{(r)}$ given the current model parameters and $\langle\cdot\rangle$ is the expectation of the posterior distribution of the factor $\boldsymbol{x}^{(r)}$. The HMM state index is omitted to simplify description.

All the basis vectors (including the noise mean vector) must be simultaneously optimized to consider the dependency among them. Let $\boldsymbol{w}$ be a vector that concatenates all the basis vectors:

$$\boldsymbol{w} = \left[\begin{array}{cccc} \boldsymbol{w}_1^\mathsf{T} & \boldsymbol{w}_2^\mathsf{T} & \cdots & \boldsymbol{w}_V^\mathsf{T} \end{array}\right]^\mathsf{T} \quad (8)$$

where V is the sum of all the leaf nodes of all decision trees and $\boldsymbol{w}_v$ is a basis vector in a leaf node v. By differentiating Eq. (7) with respect to $\boldsymbol{w}$ and equating it to zero, a set of linear equations to determine $\boldsymbol{w}$ is obtained as

$$\boldsymbol{G}\boldsymbol{w} = \boldsymbol{k} \quad (9)$$

where

$$\boldsymbol{G} = \left[\begin{array}{ccc} \boldsymbol{G}_{1,1} & \cdots & \boldsymbol{G}_{1,V} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{V,1} & \cdots & \boldsymbol{G}_{V,V} \end{array}\right], \quad \boldsymbol{k} = \left[\begin{array}{c} \boldsymbol{k}_1 \\ \vdots \\ \boldsymbol{k}_V \end{array}\right], \quad (10)$$

$$\boldsymbol{G}_{v_1,v_2} = \sum_{\substack{p,q,m \\ f_p(m)=v_1 \\ f_q(m)=v_2}} \boldsymbol{\Sigma}_m^{-1} \sum_{r,t} \gamma_{m,t}^{(r)} \left\langle \xi_p^{(r)} \xi_q^{(r)} \right\rangle, \quad (11)$$

$$\boldsymbol{k}_{v_1} = \sum_{\substack{p,m \\ f_p(m)=v_1}} \boldsymbol{\Sigma}_m^{-1} \sum_{r,t} \gamma_{m,t}^{(r)} \left\langle \xi_p^{(r)} \right\rangle \boldsymbol{o}_t^{(r)}, \quad (12)$$

$\xi_p^{(r)}$ is the p^{th} element of $\boldsymbol{\xi}^{(r)}$, and $f_p(m)$ is the function that gives a leaf node index in the p^{th} tree for the context m. The update formula for the noise covariance matrix is derived by differentiating Eq. (7) with respect to $\boldsymbol{\Sigma}_m$ and setting its result to zero:

$$\boldsymbol{\Sigma}_m = \operatorname{diag}\left(\sum_{r,t} \gamma_{m,t}^{(r)} \left(\boldsymbol{o}_t^{(r)} \boldsymbol{o}_t^{(r)\mathsf{T}} - 2\boldsymbol{o}_t^{(r)} \left\langle \boldsymbol{\xi}^{(r)\mathsf{T}} \right\rangle \boldsymbol{W}_m^\mathsf{T} + \boldsymbol{W}_m \left\langle \boldsymbol{\xi}^{(r)} \boldsymbol{\xi}^{(r)\mathsf{T}} \right\rangle \boldsymbol{W}_m^\mathsf{T}\right)\right) \Bigg/ \sum_{r,t} \gamma_{m,t}^{(r)}. \quad (13)$$

The following steps are the procedure of the proposed clustering algorithm.

Step 1. Create root nodes for every basis vectors.

Step 2. Evaluate all questions at all the leaf nodes of all trees using Eqs. (9) and (13).

Step 3. Select a set of a node and a question that gives the maximum likelihood (ML), and then split the selected node into two by applying the selected question. The model parameters of all the leaf nodes are updated by the ML parameters.

Step 4. If the likelihood gain falls below a predefined threshold, stop this procedure. Otherwise, go to **Step 2**.

Step 2 requires extremely high computational cost due to solving Eqs. (9) and (13) many times over. However, the computational reduction techniques used in additive structure models [12] are available because its model structure is almost similar to that of FAHMM. The following subsection shows the techniques briefly.

3.1. Computational complexity reduction by tying noise covariance matrices

Since the basis vectors and the noise covariance matrix depend on each other as shown in Eqs. (11) through (13), they should be iteratively updated until a convergence is reached. To reduce the computational cost, all the noise covariance matrices are tied globally while context clustering:

$$\forall_m \left(\boldsymbol{\Sigma}_m = \boldsymbol{\Sigma}_g\right) \quad (14)$$

where $\boldsymbol{\Sigma}_g$ is a globally-tied noise covariance matrix. It has been reported that covariance parameters are relatively less important than mean parameters for the quality in HMM-based speech synthesis [14]. The impact on speech quality caused by tying the noise covariance parameters in FAHMM also may be small. Using this technique, the requirement of the iteratively updating is eliminated because the term $\boldsymbol{\Sigma}_m^{-1}$ is canceled out in Eqs. (11) and (12). In addition, the more computational reduction is achieved because $\boldsymbol{G}$ becomes independent of the dimension of observation.

3.2. Computational complexity reduction with matrix inversion lemma

Since the size of $\boldsymbol{G}$ depends on V, the computational complexity for inverting $\boldsymbol{G}$ to solve Eq. (9) is increased exponentially with growing the trees. This means that the proposed technique is infeasible within reasonable computational time. However, when a leaf node is split, the statistics only change in contexts related to newly created nodes by the split, *i.e.*, the almost elements of $\boldsymbol{G}$ are unchanged at the same node if a different question is applied[2]. The computational complexity can be significantly reduced without approximation by using this property and matrix inversion lemma. Consequently, inversion $V \times V$ becomes inversion 4×4. The detail is omitted due to space limitations (see [12]).

4. Experiments

4.1. Experimental setups

The task to evaluate the proposed method was speaker adaptation. A Japanese speech database, which was constructed by our research group, was used for this experiment. The database contains sets of 503 phonetically balanced sentences uttered by more than 100 college students. We chose 22 male and 8 female speakers for training and 5 male speakers for test. The total number of training sentences were 2000 and the number of adaptation sentences per speaker was 2 or 10. Forty test sentences, which were included in neither the training nor the adaptation data, were used for the evaluation. The speech signals were sampled at a rate of 48 kHz and windowed by a 25-ms Blackman window with a 5-ms shift. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, log-fundamental frequency ($\log F_0$), and their first and second time derivatives. A five-state left-to-right context-dependent factor analyzed MSD-HSMM [15, 16] without skip paths was used. The number of basis vectors was 30 including the noise mean vector. The proposed method was applied to only spectral parameter and the conventional method shown in Figure 2a was used for $\log F_0$ and duration. The decision trees were constructed based on minimum description length (MDL) criterion [17]. The following methods were compared:

[2]Note that the statistics of factor are held fixed and the noise covariance matrices are globally tied while context clustering.

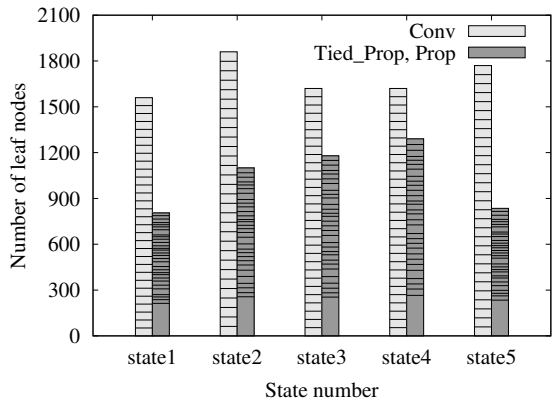


Figure 3: *The Number of leaf nodes for each HMM state.*

- **Conv**: Construct decision trees based on the constraint that all decision trees have the same structure. The parameter sharing structure for noise covariance parameters is same as that of noise mean vector. This is the conventional method [6] (see Figure 2a).
- **Tied_Prop**: Construct decision trees based on the proposed algorithm using the computational complexity reduction algorithms described in Subsection 3.1 and 3.2 (see Figure 2b).
- **Prop**: Construct decision trees based on the proposed algorithm. After the context clustering, the globally-tied noise covariance is untied and then they are tied according to the parameter sharing structure of noise mean vector.

An objective evaluation and a subjective listening test were conducted. Mel-cepstrum distance (MCD) was used for the objective measure in the objective evaluation. In the subjective listening test, the naturalness of synthesized speech were evaluated by a preference test (AB test). After the subjects had listened to a pair of speech samples they were asked which sample sounds better in terms of naturalness. Ten subjects evaluated 15 sentences, which were randomly chosen from 200 (40 test sentences $\times$ 5 test speakers) sentences. The speech samples were generated by the models adapted using 10 adaptation sentences.

4.2. Experimental results

Figure 3 shows the number of leaf nodes for each HMM state, where the bottom bar represents the number of leaf nodes for the noise mean vector, the above bar denotes that for the first basis vector, and so on. **Prop** had the different-sized trees while **Conv** had the same-sized ones. It can be seen that the importances of basis vectors are different from each other. **Prop** also had the smaller number of model parameters than **Conv**. This indicates that a compact acoustic model is derived due to various loading matrices composed by well-estimated basis vectors.

The results of the objective test for the training speakers and the test speakers are shown in Tables 1 and 2, respectively. In these tables, the second column denotes the results without adaptation. In the training speakers, **Prop** and **Tied_Prop** showed smaller MCDs than **Conv** in both with and without adaptation. This result suggests that the proposed method can more appropriately model spectrum than the conventional one. **Prop** obtained slightly smaller MCDs than **Tied_Prop**. This is because the globally-tied noise covariance causes inappropriate parameter generation [18]. The similar tendency was observed in the test speakers. This implicitly supports appropriate speaker characteristics. The MCDs were insensitive to the amount of adaptation sentences. This is because the number of free parameters for adaptation is relatively small. Further improvement may be achieved by increasing training speakers or incorporating with other adaptation techniques such as MLLR [2] and MAP [3].

Table 1: *Average of training speaker's MCD [dB].*

# of adaptation sentences	0	2	10
Conv	5.46	4.38	4.34
Tied_Prop	5.35	4.35	4.31
Prop	5.34	4.29	**4.25**

Table 2: *Average of test speaker's MCD [dB].*

# of adaptation sentences	0	2	10
Conv	5.34	4.74	4.70
Tied_Prop	5.25	4.65	4.60
Prop	5.22	4.60	**4.55**

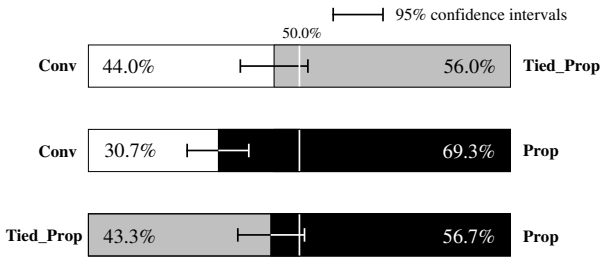


Figure 4: *The results of preference test of speech quality with 95% confidence intervals.*

Figure 4 illustrates the results of the preference test for speech quality. There is a clear preference for **Prop** over **Conv**. This clearly shows that the proposed model structure generates better synthesized speech in terms of naturalness. The multiple decision trees differ from each other are effective for spectrum modeling. Comparing **Prop** with **Tied_Prop**, **Prop** achieved a slightly better score in spite of the small difference of MCDs between the methods. Therefore, appropriately estimating the noise covariance with its reasonable parameter sharing structure is important to generate high-quality speech.

5. Conclusions

This paper proposes a clustering technique that simultaneously construct the multiple parameter sharing structures in the framework of FAHMM. The proposed technique enables to find the more optimal model structures because the more complex model structures can be evaluated by using some computational complexity reduction algorithms. Objective and subjective experimental results show that the proposed model structures outperform the conventional one. Our future work includes comparison with a sequential strategy, investigation of an appropriate parameter sharing structure for noise covariance parameters, and to apply the proposed method to F_0 parameters.

6. Acknowledgements

This research was partly funded by Core Research for Evolutionary Science and Technology (CREST) from Japan Science and Technology Agency (JST), and the Hori Sciences and Arts Foundation.

7. References

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[2] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[3] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[4] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," *Proceedings of ICSLP 2002*, pp. 1269–1272, 2002.

[5] J. Latorre, V. Wan, M. J. F. Gales, L. Chen, K. Chin, K. Knill, and M. Akamine, "Speech factorization for HMM-TTS based on cluster adaptive training," *Proceedings of Interspeech 2012*, 2012.

[6] K. Kazumi, Y. Nankaku, and K. Tokuda, "Factor analyzed voice models for HMM-based speech synthesis," *Proceedings of ICASSP 2010*, pp. 4234–4237, 2010.

[7] R. Kuhn, P. N. J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contlini, "Eigenvoices for speaker adaptation," *Proceedings of ICSLP 1998*, pp. 1771–1774, 1998.

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

[9] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E88–D, no. 3, pp. 502–509, 2005.

[10] V. Wan, J. Latorre, K. Chin, L. Chen, M. J. F. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," *Proceedings of Interspeech 2012*, 2012.

[11] K. Yu and H. Xu, "Cluster adaptive training with factorized decision trees for speech recognition," *Proceedings of Interspeech 2013*, pp. 1243–1247, 2013.

[12] S. Takaki, Y. Nankaku, and K. Tokuda, "Contextual additive structure for HMM-based speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 229–238, 2014.

[13] H. Zen and N. Braunschweiler, "Context-dependent additive log F0 model for HMM-based speech synthesis," *Proceedings of Interspeech 2009*, pp. 2091–2094, 2009.

[14] K. Oura, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "A covariance-tying technique for HMM-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E93–D, no. 3, pp. 595–601, 2010.

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

[16] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Transactions on Information and Systems*, vol. E90–D, no. 5, pp. 825–834, 2007.

[17] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Journal of the Acoustical Society of Japan*, vol. 21, no. 2, pp. 79–86, 2000.

[18] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," *Proceedings of ICASSP 1996*, pp. 389–392, 1996.