Spectral Modeling with Contextual Additive Structure for HMM-based Speech Synthesis

Shinji Takaki, Yoshihiko Nankaku and Keiichi Tokuda

Department of Computer Science and Engineering,
Nagoya Institute of Technology, Nagoya 466-8555, Japan

Abstract

This paper proposes a spectral modeling technique based on additive structure of context dependencies for HMM-based speech synthesis. Contextual additive structure models can represent complicated dependencies between acoustic features and context labels using multiple decision trees. However, its computational complexity of the context clustering is too high for full context labels of speech synthesis. To overcome this problem, this paper proposes two approaches; covariance parameter tying and a likelihood calculation algorithm using matrix inversion lemma. Experimental results show that the proposed method outperforms the conventional one in subjective listening tests.

Index Terms: Hidden Markov models, Spectral modeing, Decision trees, Context clustering, Additive structure, Distribution convolution

1. Introduction

An HMM-based speech synthesis system has been proposed to enable machines to speak naturally like humans [1, 2]. It is well known that spectral features are affected by contextual factors, and extracting the context dependencies is a critical problem for acoustic modeling. One of the major difficulties in the context dependent modeling is to find an optimum balance between model complexity and the availability of training data. Although increasing model complexity makes it possible to accurately capture variations in spectral features, the reliability of parameter estimation is degraded due to the decrease in the number of training data for each model. Furthermore, since it is difficult to prepare training data covering all context dependent models, there are numerous unseen models that are not observed in the training data but that are required in the synthesis phase.

To avoid this problem, the decision tree based context clustering has been proposed [3]. In the clustering, HMM states of the context dependent models are grouped into• •clusters, • • and all states belonging to the same cluster are assumed to have the same distribution. A binary tree is constructed based on the maximum likelihood criterion by applying a phonetic question to each node and iteratively splitting the cluster into two child clusters. By limiting the number of possible splits using prior knowledge, linguistic and articulatory information can be reflected in the clustering results. Instead of the maximum likelihood criterion, the Minimum Description Length (MDL) criterion can also be adopted to automatically determine the optimal number of clusters without setting a threshold [4].

The context space in the decision tree based context clustering is divided into clusters by contextual factors and the distributions of acoustic features are individually estimated for each cluster. This means that the effects of a particular factor are completely dependent on the other factors within clusters. On the other hand, the linear regression model [5] is another approach to modeling spectral variations in which all the contextual factors independently affect the acoustic features. Since the combination of contextual factors determines the spectral feature, it can efficiently represent the variety of distributions. However, the dependence among contextual factors is ignored and it is difficult to determine those factors that should additively affect acoustic features.

To represent more moderate dependencies between contextual factors and acoustic features, an additive structure of acoustic feature components which have different context dependencies has been proposed. This approach includes intermediate structures of decision tree based context clustering and linear regression models as special cases. Since the output probability distribution is composed of the sum of the mean vectors and covariance matrices of additive components, a number of different distributions can be efficiently represented by a combination of fewer distributions. It is unknown what kinds of contexts have additive dependencies on acoustic features. Then a context clustering algorithm for the additive structure that automatically extracts additive components by simultaneously constructing multiple decision trees has been proposed. Moreover, it can automatically determine an appropriate number of additive components.

In this paper, we apply an additive structure modeling to the spectrum parameter for HMM-based speech synthesis. Huge computational cost is required to extract the additive structure. For this reason, spectral modeling in the additive structure has not been applied to HMM-based speech synthesis. Therefore, we tried to reduce the computational complexity in the training algorithm for estimating parameters when extracting the additive structure. The three main problems with estimating the parameters of additive structure models are as follows: 1) As mean parameters depend on covariance parameters, the mean and covariance parameters should be re-estimated until convergence, 2) A gradient method is required to estimate covariance parameters, and 3) A matrix that depends on the number of leaves in decision trees should be treated when estimating mean parameters. The first and second problems can be solved by covariance parameter tying. Mean parameters are relatively more important than covariance parameters for the quality of HMM-based speech synthesis as investigated by Oura *et al* [10]. By tying covariance parameters, mean parameters become independent of them, and the tied covariance parameter can be estimated analytically. Thus, the impact on speech quality is small and computational complexity is reduced. In the third problem, when splitting the leaf cluster of a decision tree, the influence of statistics in every context are limited and computational complexity is reduced from this by using the matrix inversion lemma.

The rest of this paper is organized as follows. Section 2 describes the additive structure models, derivation of the EM algorithm for the proposed model, and the multiple decision tree based context clustering algorithm. In Section 3, the computational complexity reduction in the training algorithm of the additive structure models is shown. The results of experiments are presented in Section 4. Concluding remarks and future plans are presented in the final section.

2. Additive structure models

In the context clustering, all states in the same cluster are assumed to have the same Gaussian distribution. This means that the states have direct dependencies on phonetic contexts. In this paper, we consider a more complex structure, i.e., the additive structure of acoustic feature components. An acoustic feature vector $\boldsymbol{o}_t$ at time t is generated by the sum of additive components:

$$\boldsymbol{o}_t = \sum_{n=1}^{N} \boldsymbol{o}_t^{(n)} \tag{1}$$

where $\boldsymbol{o}_t^{(n)}$ denotes the n-th additive component. If each component is independent and generated according to a Gaussian distribution, the probabilistic density function of acoustic features is represented by the convolution of the additive components [7] so that

$$\begin{aligned} & P\left(\boldsymbol{o}_t \mid c_t, \lambda\right) \\ &= \int \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{o}_t^{(n)} \,|\, \boldsymbol{\mu}_{c_t}^{(n)}, \boldsymbol{\Sigma}_{c_t}^{(n)}) d\boldsymbol{o}_t^{(1)} \cdots \boldsymbol{o}_t^{(N-1)} \\ &= \mathcal{N}(\boldsymbol{o}_t \,|\, \boldsymbol{\mu}_{c_t}, \boldsymbol{\Sigma}_{c_t}) \end{aligned} \tag{2}$$

where $\boldsymbol{\mu}_{c_t}^{(n)}$ and $\boldsymbol{\Sigma}_{c_t}^{(n)}$ are respectively the mean vector and covariance matrix of the n-th component $\boldsymbol{o}_t^{(n)}$ given a context c_t. The output probability distribution is a Gaussian distribution whose mean vector and covariance matrix are respectively given as

$$\boldsymbol{\mu}_{c_t} = \sum_{n=1}^{N} \boldsymbol{\mu}_{c_t}^{(n)}, \;\; \boldsymbol{\Sigma}_{c_t} = \sum_{n=1}^{N} \boldsymbol{\Sigma}_{c_t}^{(n)} \tag{3}$$

Since each additive component $\boldsymbol{o}_t^{(n)}$ has different context dependencies, we assume that each component has a different decision tree that the represents tying structures of model parameters $\boldsymbol{\mu}_{c_t}$ and $\boldsymbol{\Sigma}_{c_t}$.

Although it is unknown which kinds of contexts have additive dependencies on acoustic features in practice, we present an example of a contextual additive structure of triphone HMMs to explain how effective the proposed technique is. Here, we assume that the left, center, and right phones are the contexts of additive components. Figure 1 outlines the generative process for the triphone feature. The generative process for acoustic features is as follows: first, the component of a given monophone (center phone) context is generated from a corresponding distribution obtained by descending the tree. Then, the additive components of left and right contexts are also generated independently from each distribution and then added to the monophone feature.

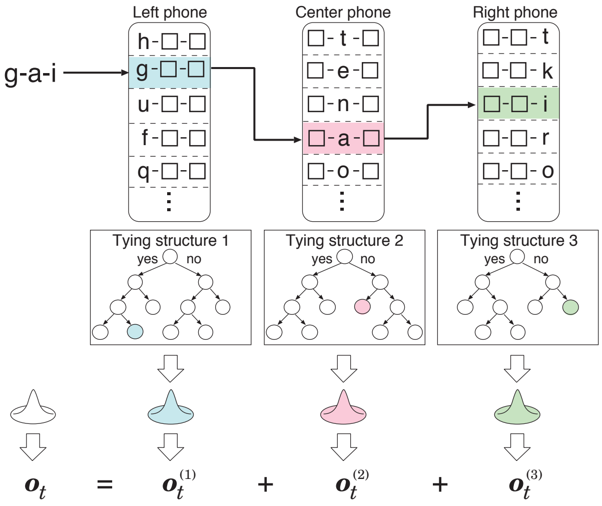


Figure 1: *An example of a contexual additive structure.*

How effective the proposed technique is depends on whether acoustic features really have additive structures for contexts. When acoustic features have additive structure, a number of different distributions can be efficiently represented by a combination of fewer distributions. Furthermore, it is also effective to predict the acoustic features of unseen contexts. Even though in the conventional method, unseen models are assigned to one of the clusters in the decision tree, the proposed method can construct the distribution for unseen contexts, which are different from any distributions of observed contexts.

2.1. EM algorithm for additive structure models

The Maximum Likelihood (ML) parameters of additive component distribution can be estimated with the EM algorithm. In the E-step, since the convolved output probability distribution becomes a Gaussian distribution, the standard forward-backward algorithm and the Viterbi algorithm can simply be applied as in standard HMMs. However, there is difficulty in the M-step due to the dependencies among additive component distributions.

Using the statistics obtained by the E-step, the $\mathcal{Q}$-function with respect to the output probability distribution can be written as

$$\begin{aligned} \mathcal{L} &= \sum_{t=1}^{T} \sum_{c \in C} \gamma_t(c) \log P(\boldsymbol{o}_t \mid c_t = c, \lambda) \\ &= -\frac{1}{2} \sum_{c \in C} \tilde{T}_c \Big[K \log 2\pi + \log |\boldsymbol{\Sigma}_c| \\ &\quad + Tr\left\{ \boldsymbol{\Sigma}_c^{-1} \left(\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^{\top} \right) \right\} \Big] \end{aligned} \tag{4}$$

where K is the dimensionality of feature vectors and C denotes all contexts observed in the training data. The statistics with respect to context c are represented by $\tilde{(\cdot)}_c$ and each of the statistics is calculated as follows:

$$\tilde{T}_c = \sum_{t=1}^{T} \gamma_t(c), \;\; \tilde{\boldsymbol{\mu}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c) \boldsymbol{o}_t \tag{5}$$

$$\tilde{\boldsymbol{\Sigma}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c) \left(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c \right) \left(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c \right)^{\top} \tag{6}$$

where $\gamma_t(c)$ is the state occupancy probability and the state index is ignored. For simplicity of notation, $\boldsymbol{\Sigma}_c$ is the diagonal covariance matrix and we focus on a dimension of feature vectors in this section. Then, the covariance parameter is σ_c and the mean parameters of all components is $\boldsymbol{\mu} = [\mu_1, ..., \mu_M]^\top$, where M is the sum of all leaf clusters of all decision trees. To represent the tree structure, function $f^{(n)}(c)$ is introduced that gives the index of the Gaussian distribution (number of leaves in the decision tree) of the n-th additive components for c. Using this function, the mean parameter and the covariance parameter of the convolved distribution are given by

$$\mu_c = \sum_{n=1}^{N} \mu_{f^{(n)}(c)}, \quad \sigma_c = \sum_{n=1}^{N} \sigma_{f^{(n)}(c)} \tag{7}$$

Then, Eq. (4) can be written as

$$\mathcal{L} = -\frac{1}{2} \sum_{c \in C} \tilde{T}_c \left\{ \log 2\pi + \log |\sigma_c| + \frac{\tilde{\sigma}_c + (\mu_c - \tilde{\mu}_c)^2}{\sigma_c} \right\} \tag{8}$$

and, the terms with respect to $\boldsymbol{\mu}$ can be rewritten as

$$\begin{aligned} \mathcal{L} &\propto -\frac{1}{2} \sum_{c \in C} \tilde{T}_c \left(\frac{\mu_c^2 + -2\mu_c\tilde{\mu}_c}{\sigma_c} \right) \\ &= -\frac{1}{2} \sum_{c \in C} \tilde{T}_c \frac{1}{\sigma_c} \left\{ \left(\sum_{n=1}^{N} \mu_{f^{(n)}(c)} \right)^2 - 2 \left(\sum_{n=1}^{N} \mu_{f^{(n)}(c)} \right) \tilde{\mu}_c \right\} \\ &= -\frac{1}{2} \left(\boldsymbol{\mu}^\top \boldsymbol{G} \boldsymbol{\mu} - 2\boldsymbol{\mu}^\top \boldsymbol{k} \right) \end{aligned} \tag{9}$$

where

$$\boldsymbol{G} = \begin{bmatrix} g_{1,1} & \cdots & g_{1,M} \\ \vdots & \ddots & \vdots \\ g_{M,1} & \cdots & g_{M,M} \end{bmatrix}, \quad \boldsymbol{k} = \begin{bmatrix} k_1 \\ \vdots \\ k_M \end{bmatrix} \tag{10}$$

$$g_{m_1,m_2} = g_{m_2,m_1} = \sum_{\substack{c,i,j \\ f^{(i)}(c)=m_1 \\ f^{(j)}(c)=m_2}} \tilde{T}_c \frac{1}{\sigma_c} \tag{11}$$

$$k_{m_1} = \sum_{\substack{c,i \\ f^{(i)}(c)=m_1}} \tilde{T}_c \frac{1}{\sigma_c} \tilde{\mu}_c \tag{12}$$

Since $\boldsymbol{G}$ is a symmetric matrix, the first partial derivative of Eq. (9) with respect to $\boldsymbol{\mu}$ can be written as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} &= -\frac{1}{2} \left\{ \left(\boldsymbol{G} + \boldsymbol{G}^\top \right) \boldsymbol{\mu} - 2\boldsymbol{k} \right\} \\ &= -\boldsymbol{G}\boldsymbol{\mu} + \boldsymbol{k} \end{aligned} \tag{13}$$

By setting Eq. (13) to $\mathbf{0}$, the solution of $\boldsymbol{\mu}$ is given as follows:

$$\boldsymbol{G}\boldsymbol{\mu} = \boldsymbol{k} \tag{14}$$

However, $\boldsymbol{G}$ is typically a singular matrix. Therefore, to solve Eq. (14), we use a Moore-Penrose generalized inverse.

2.2. Context clustering for multiple decision trees

A context clustering algorithm for multiple decision trees has been proposed to automatically extract the additive structure from training data. It is easy to construct a decision tree if the tree structures and the parameters of the other components are fixed. However, as the tree structures of the additive components interact with each other to compose the output probabilities, the multiple decision trees for additive components should be constructed simultaneously. The four steps in the procedure for the proposed clustering algorithm are as follows:

Step 1. Set the number of trees N to one, and create the root node of the first tree and compute its likelihood.

Step 2. Evaluate questions at all leaf nodes of all trees and a root node of a new tree. The likelihood after the node is split is calculated by estimating the ML parameters of all leaf nodes of all trees.

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

Step 4. If the change of likelihood after the node is split is below a predefined threshold, stop the procedure. Otherwise, go to Step 2.

There are some differences from the conventional clustering algorithm in the procedure: first, in Step 2, the ML estimates of all parameters of all trees are required to evaluate questions at a candidate node. In the conventional clustering, the ML parameters of the two nodes that are split can be obtained independently of the other nodes. However, in the proposed model, the change of likelihood before and after a node is split is calculated not only with the parameters of the split nodes but also the parameters of the other trees. For the same reason, the likelihood of a candidate node is affected by splitting other nodes in the additive structure models. Therefore, all questions should be re-evaluated at all leaf nodes after a node is split.

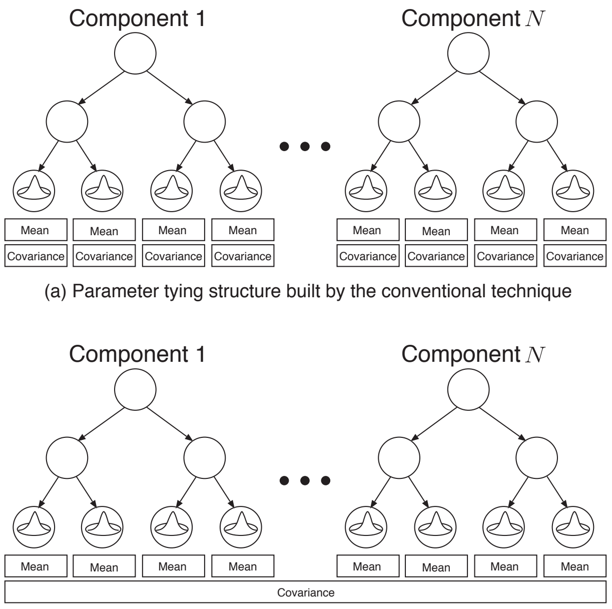
It can be seen that the proposed model, which is restricted to have a single tree, is equivalent to the conventional decision tree based context clustering. If all trees only have two node (one question is applied), the proposed model is equivalent to a linear regression model. Thus, the proposed model can be regarded as an intermediate model between decision tree based context clustering and a linear regression model, and it includes them as special cases. Furthermore, the derived algorithm can extract additive components which independently affect acoustic features and automatically determine an appropriate number of additive components.

3. Computational complexity reduction in the training algorithm

In the context clustering for multiple decision trees, the ML parameters of all leaf nodes need to be simultaneously estimated. Moreover, all questions should be re-evaluated at all leaf nodes after a node is split. Therefore, it is necessary to use an enormous amount of computational complexity when extracting the additive structure.

3.1. Computational complexity reduction by covariance parameters tying

In the additive structure models, mean parameters can be analytically estimated. However, as it is difficult to analytically solve



(a) Parameter tying structure built by the conventional technique

(b) Parameter tying structure built by the proposed technique

Figure 2: *Examples of parameter tying structures built with the conventional and the proposed technique.*

the update of covariance parameters, a gradient method is applied to each covariance parameter. Furthermore, as Eqs. (11) and (12) indicate that mean parameters depend on covariance parameters, the mean and covariance parameters should be re-estimated until convergence. Therefore, huge computational cost is involved when extracting additive structures.

In this paper, the covariance parameter tying is applied to the additive structure models. It has been reported that mean parameters are relatively more important than covariance parameters for the quality of HMM-based speech synthesis [10]. The impact on speech quality in the additive structure models caused by the covariance parameter tying would also be small. Figure 2 shows examples of parameter tying structures built with the conventional technique (Figure 2(a)) and the proposed technique (Figure 2(b)). By tying covariance parameters, the mean parameters can be updated independently of the covariance parameters and iterative updates are not required. Using the tied covariance parameter $\boldsymbol{\Sigma_g}$, the $\mathcal{Q}$-function with respect to the output probability distribution (Eq. (4)) can be rewritten as

$$\begin{aligned}\mathcal{L} \;=\; -\frac{1}{2}\sum_{c\in C}\tilde{T}_c\Bigg[&\, K\log 2\pi + \log|N\boldsymbol{\Sigma}_g| \\ &+Tr\left\{(N\boldsymbol{\Sigma}_g)^{-1}\left(\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^\top\right)\right\}\Bigg]\end{aligned} \tag{15}$$

The first partial derivative of Eq. (15) with respect to $\boldsymbol{\Sigma}_g$ can be written as

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma}_g} \;=\; -\frac{1}{2}\sum_{c\in C}\tilde{T}_c\Bigg[&\,\boldsymbol{\Sigma}_g^{-1} - N^{-1}\boldsymbol{\Sigma}_g^{-1} \\ &\left\{\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^\top\right\}\boldsymbol{\Sigma}_g^{-1}\Bigg]\end{aligned} \tag{16}$$

By setting Eq. (16) to $\mathbf{0}$, $\boldsymbol{\Sigma}_g$ is analytically calculated as follows:

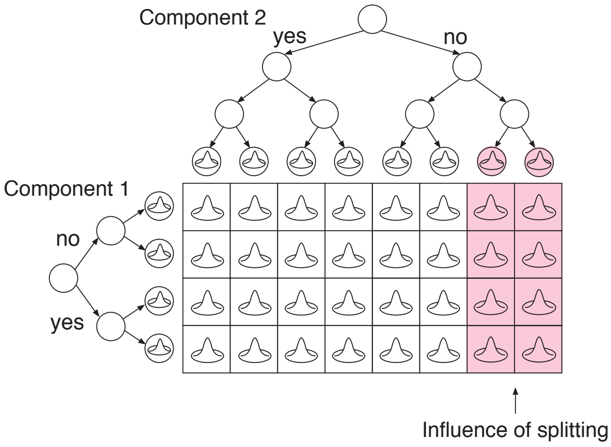


Figure 3: *Example of splitting a leaf cluster of a tree.*

$$\begin{aligned}\boldsymbol{\Sigma}_g = N^{-1}&\left(\sum_{c\in C}\tilde{T}_c\right)^{-1} \\ &\cdot\sum_{c\in C}\tilde{T}_c\left\{\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^\top\right\}\end{aligned} \tag{17}$$

the log likelihood $\mathcal{L}$ after the parameters are estimated can be written as

$$\mathcal{L} \;=\; -\frac{1}{2}\sum_{c\in C}\tilde{T}_c\left\{K\log 2\pi + \log|N\boldsymbol{\Sigma}_g| + K\right\} \tag{18}$$

3.2. Computational complexity reduction with matrix inversion lemma

Since the size of $\boldsymbol{G}$ depends on the sum of all leaf nodes of all trees in Eq.(14), the computational complexity to solve the linear equations becomes enormous. However, when a leaf cluster of a tree is split in the additive structure models, the statistics only change in contexts related to newly created nodes by splitting. Figure 3 shows an example of how a leaf cluster of a tree is split. Since $\boldsymbol{G}$ only becomes dependent on $\tilde{T}_c$ due to covariance parameter tying, many elements of $\boldsymbol{G}$ do not change at the same node even if a different question is applied. The computational complexity can significantly be reduced by using this property.

Assuming that $\boldsymbol{G}'$ is obtained with one question, and $\boldsymbol{G}''$ is obtained with another question at the same node, $\boldsymbol{G}''$ can be represented by using $\boldsymbol{G}'$ as follows:

$$\boldsymbol{G}'' = \boldsymbol{G}' + \boldsymbol{G}^{(d)} \tag{19}$$

where $\boldsymbol{G}^{(d)}$ is a symmetric matrix and can be written as

$$\boldsymbol{G}^{(d)} = \left[\begin{array}{cccccc} \mathbf{0} & & g^{(d)}_{1,m} & g^{(d)}_{1,m+1} & & \mathbf{0} \\ & & \vdots & \vdots & & \\ g^{(d)}_{m,1} & \cdots & g^{(d)}_{m,m} & g^{(d)}_{m,m+1} & \cdots & g^{(d)}_{m,M} \\ g^{(d)}_{m+1,1} & \cdots & g^{(d)}_{m+1,m} & g^{(d)}_{m+1,m+1} & \cdots & g^{(d)}_{m+1,M} \\ & & \vdots & \vdots & & \\ \mathbf{0} & & g^{(d)}_{M,m} & g^{(d)}_{M,m+1} & & \mathbf{0} \end{array}\right] \tag{20}$$

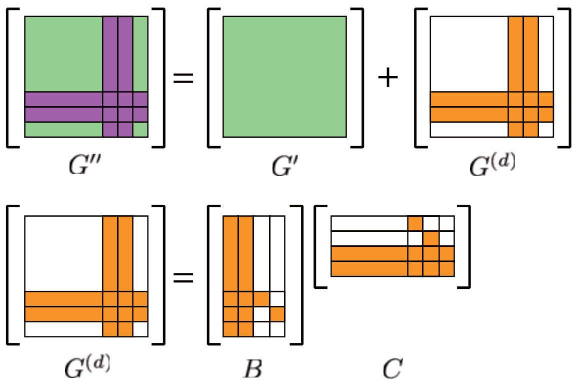


Figure 4: *Relations between G' and G''.*

where m and $m+1$ are the number of leaf nodes created by splitting. $G^{(d)}$ is represented by $M \times 4$ and $4 \times M$ matrices as follows:

$$G^{(d)} = BC \tag{21}$$

$$\begin{aligned} B &= \left[\begin{array}{cccc} B_1 & B_2 & B_3 & B_4 \end{array}\right] \\ &= \left[\begin{array}{cccc} g^{(d)}_{1,m} & g^{(d)}_{1,m+1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ g^{(d)}_{m-1,m} & g^{(d)}_{m-1,m+1} & 0 & 0 \\ \frac{1}{2} g^{(d)}_{m,m} & \frac{1}{2} g^{(d)}_{m,m+1} & 1 & 0 \\ \frac{1}{2} g^{(d)}_{m+1,m} & \frac{1}{2} g^{(d)}_{m+1,m+1} & 0 & 1 \\ g^{(d)}_{m+2,m} & g^{(d)}_{m+2,m+1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ g^{(d)}_{M,m} & g^{(d)}_{M,m+1} & 0 & 0 \end{array}\right] \end{aligned} \tag{22}$$

$$C = \left[\begin{array}{cccc} B_3 & B_4 & B_1 & B_2 \end{array}\right]^{\top} \tag{23}$$

Figure 4 shows relations between G' and G''. Assuming that G'^{-1} is given, G''^{-1} can be calculated as follows:

$$\begin{aligned} G''^{-1} &= (G' + BC)^{-1} \\ &= G'^{-1} - G'^{-1} B \Psi C G'^{-1} \end{aligned} \tag{24}$$

where $\Psi = (CG'^{-1}B + I)^{-1}$ and I is the identity matrix. Eq. (24) is derived using the following matrix inversion lemma.

$$\begin{aligned} & (G'^{-1} - G'^{-1} B \Psi C G'^{-1})(G' + BC) \\ = {} & I + G'^{-1}BC - G'^{-1}B\Psi C - G'^{-1}B\Psi CG'^{-1}BC \\ = {} & I + G'^{-1}B\left\{C - \Psi(I + CG'^{-1}B)C\right\} \\ = {} & I + G'^{-1}B\left(C - \Psi\Psi^{-1}C\right) \\ = {} & I \end{aligned} \tag{25}$$

The size of matrix Ψ is 4×4 in Eq. (24); therefore, it can significantly reduce the computational complexity in comparison with directly calculating the inverse of G''.

In the context clustering, this algorithm can be applied to the likelihood calculation of questions at the same leaf node. The matrix G'^{-1} is calculated from the first question using the Moore-Penrose inverse, and the likelihood of other questions can then be calculated by using Eq. (24) with lower computational complexity.

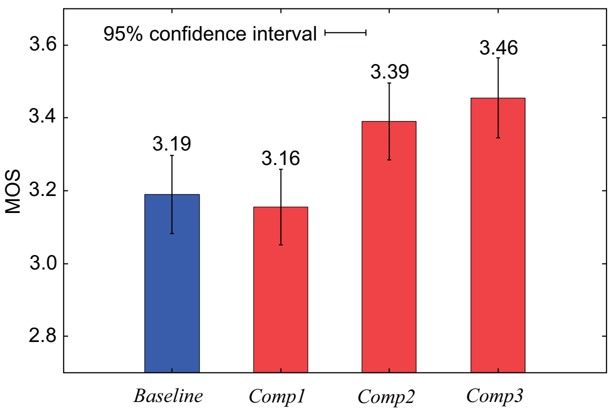


Figure 5: *Mean opinion scores for synthesized speech with 95% confidence intervals otained by conventional and proposed methods.*

4. Experiment

4.1. Experimental conditions

Subjective listening tests were conducted to evaluate the effectiveness of the proposed method. The 200 sentences of the phonetically balanced 503 sentences from the ATR Japanese speech database B-set, uttered by male speaker MHT, were used for training. The 53 sentences were used for evaluation. The speech data was down-sampled from 20 to 16 kHz and windowed at a frame rate of 5-ms using a 25-ms Blackman window.

The feature vectors consisted of spectral and F_0 feature vectors. The mel-cepstral coefficients were obtained from STRAIGHT spectra [11]. The spectrum parameter vectors consisted of 39 STRAIGHT mel-cepstral coefficients including the zero coefficient and their delta and delta-delta coefficients. The excitation parameter vectors consisted of log F_0 and its delta and delta-delta.

A five-state, left-to-right, no-skip structure with diagonal covariance matrix was used for the hidden semi-Markov model. We applied additive structure modeling to only the spectrum parameters, and the excitation parameters were modeled with conventional multi-space probability distributions HMMs [12]. The proposed and the conventional methods has the same tying structures for the excitation parameters. The MDL criterion was used to determine the size of the decision trees. The maximum number of decision trees in each state was varied from one to three. When using one additive component, the proposed method has only one decision tree for each HMM state the same as the conventional method. However, it is still different from the conventional one because of covariance parameter tying.

Ten subjects participated in these listening tests. Twenty sentences were randomly selected from the 53 sentences for each subject. The subjects were asked to rate the naturalness of the synthesized speech on a scale from one (completely unnatural) to five (natural). All experiments were carried out using headphones in a soundproof room .

4.2. Experimental results

Figure 5 plots the experimental results. In the figure, *Baseline* is the conventional method and *Comp1* to *Comp3* respectively represent the proposed method with one to three additive components. First, it can be seen from the figure that *Baseline* and *Comp1* obtained almost the same score. This indicates

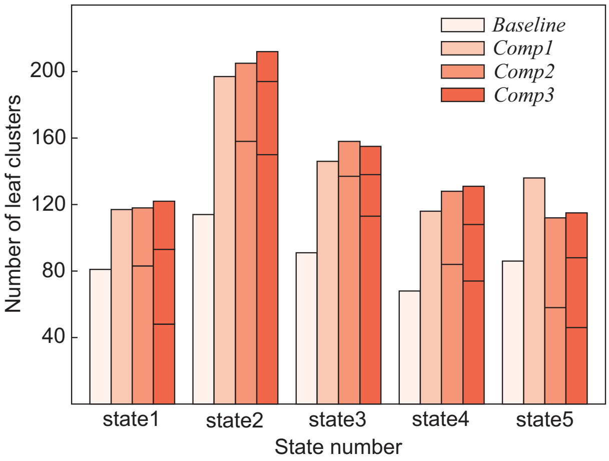


Figure 6: *Number of leaf clusters for each state.*

that the impact of speech quality by tying covariance parameters is small for HMM-based speech synthesis. Next, it can also be seen that *Comp2* and *Comp3*, i.e., additive structure models, achieved better subjective scores than the other methods *Baseline* and *Comp1*. In addition, the scores tend to increase with an increase in additive components. This means that the additive structure is more appropriate than the conventional method and it can represent complicated context dependencies.

Figure 6 is a bar chart of the number of leaf clusters for each state. The total number of leaf nodes in all trees is shown since the additive structure models have multiple decision trees. When the conventional and proposed methods have the same number of leaf clusters, the proposed method only has half the number of parameters because of covariance parameter tying. Figure 6 shows that *Comp1* has more leaf clusters than *Baseline*. This means that decision trees with respect to the mean parameter are constructed by tying covariance parameters. Similar to *Comp1*, the number of leaf clusters increases in the additive structure models with multiple decision trees. This is because the MDL criterion was used to determine the size of decision trees and decision trees ware constructed to represent variations in acoustic features by only using mean parameters in the additive structure models. Although the size of decision trees differs among additive components, all decision trees are split. This suggests that there is an additive structure in the training data.

5. Conclusions

In this paper, we proposed a spectral modeling technique based on the additive structure of context dependencies representing complicated context dependences. Assuming that an acoustic feature is generated by the sum of additive components, we estimated model parameters and extracted additive structures. It is difficlut to apply this method to HMM-based speech synthesis due to its computational complexity. Tying of covariance parameters in each state and using the matrix inversion lemma allowes us to reduce the amount of computational complexity. Then, spectral modeling which has contextual additive structure for HMM-based speech synthesis was accomplished. In the experiments, the proposed method outperformed than the conventional method. Experiments on larger datasets will be a future work.

6. Acknowledgements

The research leading to these results was partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project) and the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication, Japan.

7. References

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

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

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

[4] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J.Acoust. Soc. Jpn. (E), vol. 21, no. 2, pp. 76–86, 2000.

[5] Y. Abe and K. Nakajima, "Speech Recognition Using Dynamic Transformation of Phoneme Templates Depending of Acoustic/Phonetic Environments," Proc. ICASSP, pp. 326–329, 1992.

[6] N. Iwahashi and Y. Sagisaka, "Statistical Modeling of Speech Segment Duration by Constrained Tree Regression," Proc. IEICE trans, vol. E83–D, no. 7, pp. 1550–1559, 2000.

[7] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," Proc. ICASSP, pp. 137–140, 1992.

[8] Y. Nankaku, K. Nakamura, H. Zen, and K. Tokuda, "Acoustic modeling with contextual additive structure for HMM-based speech recognition," Proc. ICASSP, pp. 4469–4472, 2008.

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

[10] K. Oura, H. Zen, Y. Nankaku, A, Lee, and K. Tokuda, "A Covariance-Tying Technique for HMM-based Speech Synthesis," IEICE, vol. E93–D, no.3, pp.595–601, 2010.

[11] H. Kawahara, M. K. Ikuyo, and A. Cheneigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneousfrequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, 27, pp.187–207, 1999.

[12] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling," Proc. ICASSP, pp. 229–232, 1999.