CONTEXTUAL PARTIAL 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, Japan

ABSTRACT

This paper proposes a spectral modeling technique based on a contextual partial additive structure for HMM-based speech synthesis. To represent complicated context dependencies, contextual additive structure models assume multiple independent components which have different context dependencies to form acoustic features. In additive structure models, there is a constraint that a fixed number of additive components are used for generating acoustic features. However, it is natural to assume that the number of components depends on contexts. In the proposed technique, partial additive components affecting arbitrary contextual sub-spaces are created on demand to increase the likelihood. Then, the number of components for each context can be automatically determined with the training data. Experimental results show that the proposed technique outperformed the standard technique in a subjective test.

Index Terms— HMM-based speech synthesis, Decision trees, Context clustering, Contextual additive structure, Distribution convolution

1. INTRODUCTION

Speech parameters, such as spectrum, excitation, and duration, depend on a variety of contextual factors such as phoneme identities, accent, and parts-of-speech, etc. In the HMM-based speech synthesis system [1], context dependent models are generally used to capture these contextual dependencies. One of the major difficulties in context dependent modeling is to find good balance between model complexity and the availability of training data. 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 solve this problem, the decision tree based context clustering has been proposed [2]. In this clustering, HMM states of context dependent models are grouped into *clusters*, and all states belonging to the same cluster share the output probability 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 [3].

In the context clustering, all states in the same cluster share their output probability distribution. This means that the states have direct dependencies of phonetic contexts. To represent more moderate dependencies between contextual factors and acoustic features, an additive structure of acoustic features has been proposed [4]. 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. Additive structure models can robustly represent complicated context dependencies between acoustic features and context labels using multiple decision trees. However, it is unknown what kinds of contexts have additive dependencies on acoustic features. A context clustering algorithm for the additive structure [4] has been proposed to automatically extract additive components from the training data. The algorithm simultaneously constructs multiple decision trees and also can automatically determine an appropriate number of additive components. The effectiveness of this method in HMM-based speech synthesis has been reported [5].

Although the number of components can be automatically determined through the context clustering, there is a constraint that a fixed number of additive components are used for generating acoustic features. However, it is natural to assume that an appropriate number of additive components depends on contexts. That is, it is expected that some context dependent models require many additive components to represent variations in acoustic features and others do not. To represent such context dependencies appropriately, we propose a technique which enable us to extract additive components affecting arbitrary contextual sub-spaces as well as the entire contextual space. In the proposed clustering algorithm, the partial additive components are created on demand at an arbitrary node in the context clustering to increase the likelihood. Therefore, the number of additive components corresponding to each context dependent model is automatically determined from the resultant structure of decision trees. The model structure with various number of additive components yields larger combination of components than the standard additive structure with the same number of parameters. This means that it can effectively represent the context dependencies with a limited amount of the training data.

The rest of this paper is organized as follows. Section 2 describes the standard additive structure models. Section 3 proposes the technique which enable us to extract contextual partial additive components. Section 4 presents the experimental results. Section 5 presents concluding remarks and future research.

2. ADDITIVE STRUCTURE MODELS

In additive structure models, an acoustic feature vector $\boldsymbol{o}_t$ at time t is assumed to be 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 [6] so that

$$P\left(\boldsymbol{o}_t \mid c_t, \lambda\right) = \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{c_t}, \boldsymbol{\Sigma}_{c_t}) \tag{2}$$

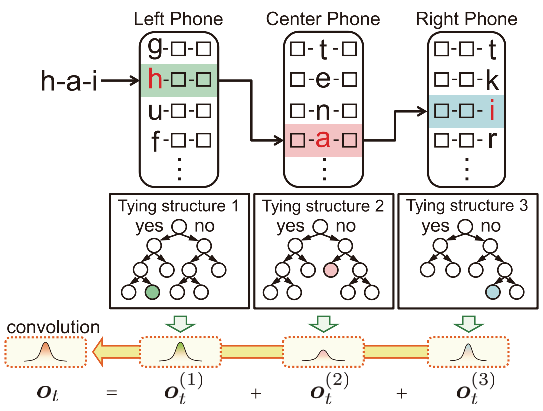

Fig. 1. *Example of a contextual additive structure.*

The output probability distribution is a Gaussian distribution whose mean vector and covariance matrix are respectively given as

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

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

Figure 1 outlines the generative process for the triphone feature in additive structure models. The effectiveness of the proposed technique depends on whether acoustic features actually have an additive structure of contexts. When acoustic features have an 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. Section 3 describes how to extract the additive structure in detail.

2.1. EM algorithm for additive structure models

The Maximum Likelihood (ML) parameters of additive component distribution can be estimated with the EM algorithm. Using the statistics obtained using the E-step, the $\mathcal{Q}$-function with respect to the output probability distribution can be written as

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

where K is the dimensionality of feature vectors, C denotes all contexts observed in the training data, and $\gamma_t(c)$ is the state occupancy probability. In Eq. (4), the state index is ignored for simplicity of notation.

By focusing on a dimension of feature vectors and equating the first partial derivative of Eq. (4) to $\mathbf{0}$ with respect to the mean parameters of all components $\boldsymbol{\mu} = [\mu_1, ..., \mu_M]^\top$, the solution of $\boldsymbol{\mu}$ is given by a set of linear equations [5], where M is the sum of all leaf clusters of all decision trees. In this paper, independent tying structures of covariance parameters are constructed [7]. Covariance parameter tying is also applied to reduce the computational cost and the tied covariance parameter can be estimated analytically [5].

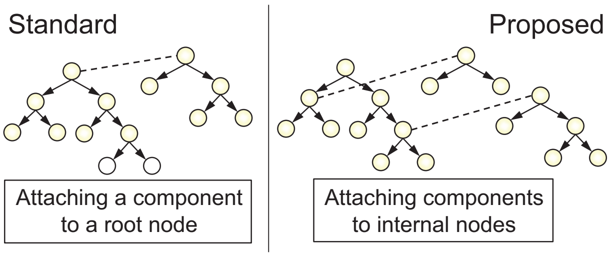
Standard
Proposed
Attaching a component to a root node
Attaching components to internal nodes

Fig. 2. *Examples of standard and partial additive structures.*

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 the training data [4]. It is easy to construct a decision tree if the tree structures and 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 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 representing 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 based on the ML criterion, 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.

Since in this technique an appropriate splitting of a leaf node or a root node representing a new tree is selected based on the ML criterion in STEP 2. A splitting of a root node is equivalent to a extracting a new component. Thus, an appropriate number of components can be automatically determined with context clustering for multiple decision trees based on the ML criterion.

3. CONTEXTUAL PARTIAL ADDITIVE STRUCTURE

Although additive structure models can automatically determine the number of components, there is a constraint that a fixed number of additive components are used for generating acoustic features. However, it is natural to assume that an appropriate number of additive components depends on contexts. That is, is, it is expected that some context dependent models require many additive components to represent variations in acoustic features and others not. To represent such context dependencies, we introduce partial additive components affecting arbitrary contextual sub-spaces.

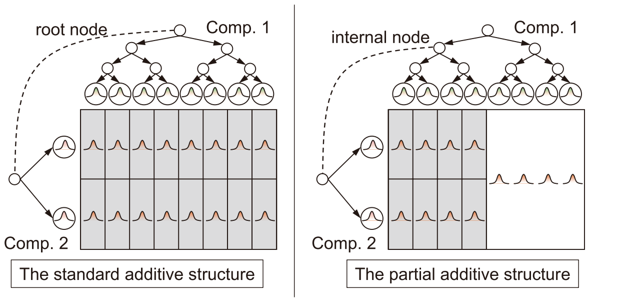
root node
Comp. 1
Comp. 2
The standard additive structure
internal node
Comp. 1
Comp. 2
The partial additive structure

Fig. 3. *The effect of an partial additive structure in distribution modeling of acoustic features.*

In the proposed technique, a partial additive component is represented by a decision tree attached to an internal node of another decision tree. Figure 2 shows examples of the standard and partial additive structure. The standard technique extracts additive components for the only entire contextual space corresponding to a root node. The proposed technique can attach the additive component to an arbitrary node including internal nodes as well as root nodes. Figure 3 shows the effect of a partial additive structure in distribution modeling of acoustic features. The gray regions represent the contextual spaces affected by the second additive component "Comp. 2". The second component of the partial additive structure affects the contextual sub-space corresponding to the internal node of the first component, even though the second component of the standard structure always divides the entire contextual space. The proposed model structure yields larger combination of components than the standard additive structure with the same number of parameters.

Considering the relation between the standard and partial additive structure models, an arbitrary partial additive structure can be converted to a global additive component, because a partial decision tree can be expanded to a global decision tree by copying the upper structure of the parent decision tree. In this case, the partial decision tree is represented as a sub-tree at the internal node of the copied tree and the other nodes are assumed to have zero mean and variance. Therefore, the proposed structure can be regarded as special case of the standard additive structure. This means there is no advantage of the proposed technique in the representation of decision trees. However, the proposed technique provides an efficient representation for partial context dependencies with a smaller number of model parameters. Furthermore, if there exists an optimal structure representing partial context dependencies, it is difficult to extract an equivalent global additive structure by using the context clustering algorithm described in Section 2.2, due to the greedy strategy. Therefore, an explicit representation of partial context dependencies and a context clustering algorithm for extracting partial additive structures are required.

The context clustering algorithm for the partial additive structure can be derived by modifying STEP. 2 in the standard context clustering algorithm for multiple decision trees as follows:

STEP 2. Evaluate questions at all leaf nodes of all trees and a root node representing a new tree. In addition, all candidate root nodes representing partial additive components are also evaluated at all internal nodes. The likelihood after the node splitting is calculated by estimating the ML parameters of all leaf nodes of all trees.

The difference with the standard context clustering algorithm for multiple decision trees is to explicitly evaluate all questions at all internal nodes for constructing a new tree representing a partial additive component. The number and position of additive components corresponding to each context dependent model are automatically determined on demand to increase the likelihood based on the ML criterion. Thus, the proposed technique can effectively represent the context dependencies with a limited amount of the training data. For an unseen context, the corresponding distribution can be found by answering the question from the top-node as the standard decision tree. However, if there is an attached decision tree at the current node, the number of components for the current context is increased and the corresponding distributions must be searched for in both the parent and attached decision trees.

3.1. Related model structures

The additive structure models include different model structures as special cases. If the additive structure is restricted to having a single decision tree, it becomes the conventional decision tree (tree regression). Linear regression models [8] can also be represented by additive structure models, which consist of additive components with only one contextual question. Therefore, additive structure models can be regarded as intermediate models between tree regression and linear regression. Partial decision trees in the proposed technique inherit this property. Constrained Tree Regression (CTR) [9] also has a strong relation to the proposed model structure. CTR has an additive component corresponding to a contextual question at each intermediate node, and feature vectors are predicted by adding all additive components from the top-node to leaf-node. Although CTR can also represent a variable number of additive components, similar to the proposed structure, only a sub-set of standard additive structure models can be represented by CTR because it integrates the structures of tree regression and linear regression into a single tree structure. As mentioned above, partial additive structure models have the same ability in the representing model structures as standard additive structure models.

4. EXPERIMENT

4.1. Experimental conditions

Subjective listening tests were conducted to evaluate the effectiveness of the proposed technique. From the phonetically balanced 503 sentences from the ATR Japanese speech database B-set, uttered by male speaker MHT, 450 sentences were used for training. The remaining 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 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 a diagonal covariance matrix was used for the hidden semi-Markov model. Additive structure modeling was applied to only the spectrum parameters, and the excitation parameters were modeled with conventional multi-space probability distribution HMMs [10]. The tying structures for excitation parameters were constructed with the conventional decision tree based context clustering.

Three techniques were compared; *CONV*: the conventional decision tree, *ADD*: the standard additive structure models, and *PADD*:

Table 1. Number of decision trees in each state. The number of decision trees In *PADD* consists of that attached to the root node and internal nodes.

	CONV	*ADD*	*PADD*
State 1	1	5	10 (root 4 + internal 6)
State 2	1	7	13 (root 2 + internal 11)
State 3	1	7	14 (root 3 + internal 11)
State 4	1	5	13 (root 3 + internal 10)
State 5	1	6	9 (root 3 + internal 6)
Total	5	30	59 (root 15 + internal 44)

the proposed partial additive structure models. Covariance parameter tying [5] was applied to *ADD* and *PADD* for reducing computational cost.

The minimum description length (MDL) criterion [3] was used to select splitting a node in all techniques. In the proposed technique, the increase in the the number of parameters of splitting a leaf node and extracting a new component differs. The increase in the number of parameters by extracting a new component doubles compared with that by splitting a leaf node. Penalty terms of the description length then grows large in extracting a new component. The MDL criterion was used to determine the size of the decision trees.

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). The experiment was carried out using headphones in a soundproof room .

4.2. Experimental results

Table 1 lists the number of decision trees in each HMM state and total number of decision trees in each technique. The number of decision trees in *PADD* consists of that attached to the root node and internal nodes corresponding to the global and partial additive components respectively. It can be seen from Table 1 that the additive structure models constructed multiple trees for each state in the context clustering, even though they can select single tree structures. This results suggest that there is an additive structure in the training data. Furthermore, *PADD* created decision trees at internal nodes as well as the root node. This means that the proposed clustering algorithm extracted partial additive components to efficiently represent context dependencies in the training data. Table 2 lists the number of leaf nodes, the total number of parameters and the average likelihoods per frame of training (450 sentences) and test data (53 sentences). Note that *CONV* has double number of parameters in each leaf node compared with *ADD* and *PADD*, because the covariance parameter tying was applied to *ADD* and *PADD*. In Table 2, the likelihood of *CONV* in the training and test data was the highest of the three techniques. This is because covariance parameter tying was not applied to *CONV* and the total number of parameters was larger than other two techniques. It can also be seen from Table 2 that *ADD* and *PADD* have almost the same number of parameters and there is not the significant difference in the likelihood of *ADD* and *PADD*.

Figure 4 shows the subjective listening results. In Figure 4, *ADD* and *PADD* achieved better subjective scores than *CONV* that has larger number of parameters. This means that additive structure models could represent complicated context dependencies. It can be seen from Figure 4 that *PADD* achieved better subjective scores than *ADD*. These results mean that the proposed technique can represent appropriate context dependencies with the contextual partial additive structure, even though *ADD* and *PADD* have almost the same number of parameters. Moreover, the proposed technique could automatically determine the number of components affecting contextual sub-spaces as well as the entire contextual space and effectively represent the context dependencies with the training data.

Table 2. Number of leaf clusters, total number of parameters and average likelihood per frame of training and test data.

	CONV	*ADD*	*PADD*
Number of leaf nodes	814	1391	1446
Total number of parameters	195,360	166,920	173,520
Ave. likelihood (training)	138.65	132.15	132.36
Ave. likelihood (test)	136.10	130.07	130.27

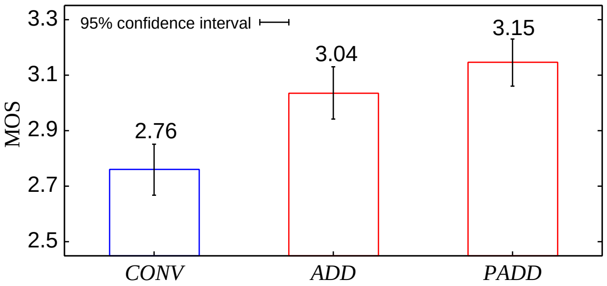

Fig. 4. *Mean opinion scores for synthesized speech obtained by conventional, standard and proposed techniques.*

5. CONCLUSIONS

This paper proposed a spectral modeling technique based on the contextual partial additive structure. In the standard additive structure models, it is difficult to extract partial additive components which affects arbitrary contextual sub-spaces. The proposed technique can extract the contextual partial additive structure. Furthermore, the number of partial additive components as well as standard global additive components can be automatically determined with the training data. In the experiment, the proposed technique outperformed the conventional technique and the standard additive structure models. Experiments on other datasets including style, emotions, etc, will be a future work.

6. ACKNOWLEDGEMENTS

The research leading to these results was partly funded by the Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).

7. REFERENCES

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

[2] J. Odell, “The use of context in large vocabulary speech recognition,” *PhD dissertation, Cambridge University*, 1995.

[3] 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.

[4] Y. Nankaku, K. Nakamura, H. Zen, and T. Tokuda, “Acoustic modeling with contextual additive structure for HMM-based speech recognition,” *Proceedings of ICASSP*, pp. 4469–4472, 2008.

[5] S. Takaki, Y. Nankaku, and K. Tokuda, “Spectral modeling with contextual additive structure for HMM-based speech synthesis,” *Proceedings of 7th ISCA Speech Synthesis Workshop*, pp. 100–105, 2010.

[6] S. Matsoukas and G. Zavaliagkos, “Convolutional density estimation in hidden Markov models for speech recognition,” *Proceedings of ICASSP*, pp. 113–116, 1999.

[7] S. Takaki, K. Oura, Y. Nankaku, and K. Tokuda, “An optimization algorithm of independent mean and variance parameter tying structures for hmm-based speech synthesis,” *Proceedings of ICASSP*, pp. 4100–4103, 2011.

[8] Y. Abe and K. Nakajima, “Speech recognition using dynamic transformation of phoneme templates depending of acoustic/phonetic environments,” *Proceedings of ICASSP*, pp. 326–329, 1989.

[9] N. Iwahashi and Y. Sagisaka, “Statistical modeling of speech segment duration by constrained tree regression,” *Proceedings of IEICE trans*, vol. E83–D, no. 7, pp. 1550–1559, 2000.

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Multi-space probability distribution HMM,” *IEICE Transactions on Information & Systems*, vol. E85-D, no. 3, pp. 455–464, 2002.

[11] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, pp. 187–207, 1999.

[12] K. F. Lee, “Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599–609, 1990.

[13] S. Young, J. Odell, and P. Woodland, “Tree-based state tying for high accuracy acoustic modelling,” *Proceedings of ARPA Workshop on Human Language Technology*, pp. 307–312, 1994.

[14] K. Oura, H. Zen, A. Lee, and K. Tokuda, “A covariance-tying technique for HMM-based speech synthesis,” *Proceedings of IEICE*, vol. E93–D, no. 3, pp. 595–601, 2010.

[15] H. Zen and N. Braunschweiler, “Context-dependent additive log f0 model for HMM-based speech synthesis,” *Proceedings of Interspeech*, pp. 2091–2094, 2009.

[16] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of japanese,” *J. Acoust. Soc. Jpn. (E)*, vol. 5, no. 4, 1984.