AN OPTIMIZATION ALGORITHM OF INDEPENDENT MEAN AND VARIANCE PARAMETER TYING STRUCTURES FOR HMM-BASED SPEECH SYNTHESIS

Shinji Takaki, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

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

ABSTRACT

This paper proposes a technique for constructing independent parameter tying structures of mean and variance in HMM-based speech synthesis. Conventionally, mean and variance parameters are assumed to have the same tying structure. However, it has been reported that a clustering technique of mean vectors while tying all variance matrices improves the quality of synthesized speech. This indicates that mean and variance parameters should have different optimal tying structures. In the proposed technique, the decision trees for mean and variance parameters are simultaneously grown by taking into account the dependency on mean and variance parameters. Experimental results show that the proposed technique outperforms the conventional one.

Index Terms— speech synthesis, hidden Markov models, decision trees, context clustering

1. INTRODUCTION

An HMM-based speech synthesis system has been proposed to enable machines to speak naturally like humans [1, 2]. Speech parameters such as spectrum, excitation, and duration depend on a variety of contextual factors such as phoneme identities, accent, parts-of-speech, etc. In the HMM-based speech synthesis system, context dependent models are generally used to capture these contextual factors. If more combinations of these contextual factors are taken into account, we can obtain more accurate models. However, as the number of contextual factors increases, the number of possible combinations also increases exponentially. Consequently, it is difficult to robustly estimate model parameters due to the lack of the training data. Furthermore, it is impossible to cover every possible combination of contextual factors for a finite set of the training data. Various parameter tying techniques have been proposed to prevent this problem. A decision tree based context clustering technique has been widely used [3]. With this technique, top-down clustering is performed to maximize the likelihood of model parameters with respect to the training data by using questions about contexts. Then, parameters of all states belonging to the same leaf node are tied. Unseen models can be generated by traversing the decision trees.

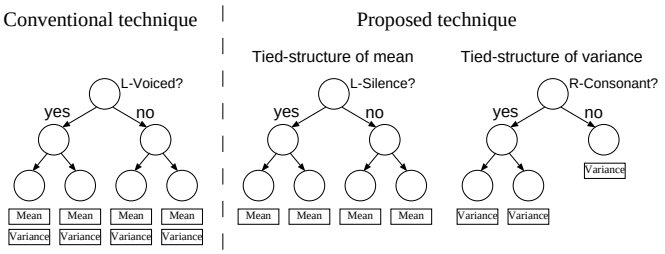

Fig. 1. Example of parameter tying structures constructed with the conventional and proposed techniques.

Conventionally, an HMM stream-level tying structure is constructed in HMM-based speech synthesis, i.e., mean vectors and variance matrices have exactly the same parameter tying structure. However, it may not be always appropriate that mean and variance parameters have the same tying structure. As an example, we confirmed the effectiveness of a technique for context clustering mean vectors while tying all variance matrices [4]. In this technique, the synthesized speech can be expected to improve by constructing different tying structures for both mean and variance parameters. However, some degree of freedom for variance parameters may be necessary for improving the quality of synthesized speech.

In this paper, we assume that both mean and variance parameters have their own tying structure and examine the construction of appropriate parameter tying structures. Figure 1 shows an example of parameter tying structures constructed with the conventional and proposed techniques. In the clustering algorithm, it is necessary to simultaneously construct each parameter tying structure due to the dependency on mean and variance parameters. Although such a context clustering algorithm can be derived by expanding the conventional context clustering algorithm, we derive the algorithm using the fact that simultaneous context clustering of mean and variance parameters can be regarded as a special case of context clustering in additive structure models [5].

The rest of this paper is organized as follows. Section 2 describes a context clustering technique for both mean and variance parameters. The experimental conditions and results are presented in Section 3. Concluding remarks and future work are presented in Section 4.

2. INDEPENDENT TYING STRUCTURES FOR MEAN AND VARIANCE PARAMETERS

In this section, we describe a context clustering technique for both mean and variance parameters. First, we explain the additive structure models that have multiple decision trees. Next, an optimization algorithm of independent mean and variance parameter tying structures is shown as a special case of the additive structure models.

2.1. Additive Structure Models

In additive structure models, 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 [6] so that

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

where $\boldsymbol{\mu}_{c_t}^{(n)}$ and $\boldsymbol{\Sigma}_{c_t}^{(n)}$ are respectively the mean vector and variance 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, we assume that each component has a different decision tree that represents tying structures of model parameters $\boldsymbol{\mu}_{c_t}$ and $\boldsymbol{\Sigma}_{c_t}$.

2.2. Proposed Model Structure

In additive structure models, an acoustic feature vector is generated by the sum of additive components. In this paper, an acoustic feature vector $\boldsymbol{o}_t$ is generated by the sum of two components, i.e., $\boldsymbol{o}_t^{(m)}$ and $\boldsymbol{o}_t^{(v)}$:

$$\boldsymbol{o}_t = \boldsymbol{o}_t^{(m)} + \boldsymbol{o}_t^{(v)}. \tag{3}$$

If each component is independent and generated according to a Gaussian distribution, each component usually has mean and variance parameters. In this paper, it is assumed that $\boldsymbol{o}_t^{(m)}$ is generated from a Gaussian distribution that has only a mean parameter and zero variance and $\boldsymbol{o}_t^{(v)}$ is generated from one that has only a variance parameter and zero mean. In this case, the probabilistic density function of the acoustic feature is represented by the convolution of these two components so that

$$\begin{aligned} \boldsymbol{o}_t^{(m)} &\sim \mathcal{N}(\boldsymbol{\mu}_{c_t}, \boldsymbol{0}), & (4) \\ \boldsymbol{o}_t^{(v)} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{c_t}), & (5) \\ P\left(\boldsymbol{o}_t \mid c_t, \lambda\right) &= \mathcal{N}(\boldsymbol{o}_t \mid \boldsymbol{\mu}_{c_t}, \boldsymbol{\Sigma}_{c_t}). & (6) \end{aligned}$$

Assuming that each component has a different decision tree, independent parameter tying structures of mean and variance can be represented.

2.3. Parameter Estimation for the proposed technique

In this model structure, the Maximum Likelihood (ML) parameters can be estimated with the Expectation Maximization (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.

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

$$\begin{aligned} \mathcal{Q} &= \sum_{t=1}^{T} \sum_{c \in C} \gamma_t(c) \log P(\boldsymbol{o}_t \mid c_t = c, \lambda) \\ &= -\frac{1}{2} \sum_{c \in C} \tilde{T}_c \Big[K \log 2\pi + \log |\boldsymbol{\Sigma}_c| \\ &\quad + Tr\left\{ \boldsymbol{\Sigma}_c^{-1} \left(\tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^\top \right) \right\} \Big], \end{aligned} \tag{7}$$

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:

$$\tilde{T}_c = \sum_{t=1}^{T} \gamma_t(c), \;\; \tilde{\boldsymbol{\mu}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c) \boldsymbol{o}_t, \tag{8}$$

$$\tilde{\boldsymbol{\Sigma}}_c = \frac{1}{\tilde{T}_c} \sum_{t=1}^{T} \gamma_t(c) \left(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c\right) \left(\boldsymbol{o}_t - \tilde{\boldsymbol{\mu}}_c\right)^\top, \tag{9}$$

where $\gamma_t(c)$ is the state occupancy probability and the state index is ignored for simplicity of notation.

By setting the first partial derivative of $\mathcal{Q}$ function with respect to an arbitrary mean vector or variance matrix, the ML parameters are given as follows:

$$\boldsymbol{\mu}_{n^{(m)}} = \left(\sum_{c \in \phi_{n^{(m)}}} \tilde{T}_c \boldsymbol{\Sigma}_c^{-1} \right)^{-1} \sum_{c \in \phi_{n^{(m)}}} \tilde{T}_c \boldsymbol{\Sigma}_c^{-1} \tilde{\boldsymbol{\mu}}_c, \quad (10)$$

$$\boldsymbol{\Sigma}_{n^{(v)}} = \left(\sum_{c \in \phi_{n^{(v)}}} \tilde{T}_c \right)^{-1} \cdot \sum_{c \in \phi_{n^{(v)}}} \tilde{T}_c \left\{ \tilde{\boldsymbol{\Sigma}}_c + (\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)(\boldsymbol{\mu}_c - \tilde{\boldsymbol{\mu}}_c)^\top \right\}, \quad (11)$$

where $n^{(m)}, n^{(v)}$ are respectively the number of clusters in mean and variance parameter trees, and $\phi_{n^{(\cdot)}}$ denotes the contexts included in the $n^{(\cdot)}$-th cluster.

It can be seen from the Eqs. (10) and (11) that the update of $\boldsymbol{\mu}_{n^{(m)}}$ and $\boldsymbol{\Sigma}_{n^{(v)}}$ requires the parameters of the other clusters. Hence, all parameters of all trees have dependencies on each other to compose the output probabilities; therefore, all parameters of all trees should be estimated simultaneously. Thus, iterative updates are needed for estimating mean and variance parameters until a convergence.

2.4. Simultaneous Context Clustering for Mean and Variance Parameters

In the context clustering, the optimal parameter tying structures are given by maximizing Eq. (7). However, it is necessary to simultaneously construct each parameter tying structure due to the dependency on mean and variance parameters. Since this problem corresponds to a problem of estimating parameter tying structures of additive components $\boldsymbol{o}_t^{(m)}$ and $\boldsymbol{o}_t^{(v)}$, appropriate parameter tying structures of mean and variance parameters are constructed with simultaneous context clustering in additive structure models [5]. The procedure for the proposed context clustering algorithm is as follows.

Step 1. The root nodes of the two trees of mean and variance parameters are created.

Step 2. Questions at all leaf nodes of two trees are evaluated. The likelihood after the node is split is calculated by estimating the ML parameters of all leaf nodes of all trees.

Step 3. The pair of a node and question that gives the maximum likelihood is selected, and the node is split 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.

The decision trees of mean and variance parameters can be simultaneously constructed with this technique. Furthermore, we can independently control the size of mean and variance decision trees with the the proposed technique by adjusting the weights in the MDL criterion. Thus, the proposed context clustering would construct more appropriate parameter tying structures than the conventional one.

3. EXPERIMENTS

3.1. Experimental conditions

The first 450 sentences of the phonetically balanced 503 sentences the ATR Japanese speech database B-set, uttered by male speaker MHT, 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 [7] including the zero coefficient, their delta and delta-delta coefficients. The excitation parameter vectors consisted of log F_0, its delta and delta-delta.

A five-state, left-to-right, no-skip structure with diagonal covariance matrices was used for the hidden semi-Markov model. We applied the proposed context clustering technique for mean and variance parameters to only the spectrum parameters. The conventional and proposed techniques have the same tying structures for the excitation parameters. The MDL criterion was used to control the size of the tree of the conventional technique and the mean parameter tree of the proposed technique. We changed the heuristic weight for the penalty term (Eq. (18) in [3]) to construct the variance parameter tree of the proposed technique. The weights used here were 4.0, 2.0, and 1.0. In addition, we compared the proposed technique with a technique for tying variance parameters in each state of HMMs as conventional one[1].

3.2. Experimental results

Table 1 lists the number of leaf nodes and the total number of parameters for each technique. In this table, *Baseline* is the conventional technique, *TieVar* is the technique for tying variance parameters in each state of HMMs, and *MDL4.0*, *MDL2.0*, and *MDL1.0* respectively represent the proposed technique with 4.0, 2.0, and 1.0 weights of the MDL criterion. Although leaf nodes have mean and variance parameters in *Baseline*, in the other techniques leaf nodes have only parameters of either. First, it can be seen from the table that *MDL1.0* has more mean parameters and less variance parameters than *Baseline*. This indicates that the proposed technique constructs decision trees that are appropriately sized for both mean and variance parameters. Next, *MDL2.0* and *MDL4.0*

[1] In [4], variance parameters are tied to one in all states of HMMs. In this paper, we assume that the technique with the enough big weight of the MDL criterion in the proposed technique is the conventional one.

Table 1. Number of leaf nodes and total number of parameters.

	Number of leaf nodes		The total number of parameters
	Mean	Variance	
Baseline	809	809	194160
TieVar	1316	5	158520
MDL4.0	1255	147	168240
MDL2.0	1249	247	179520
MDL1.0	1235	403	196560

have less variance parameters and slightly more mean parameters in the proposed technique. This means that the mean parameter decision tree was constructed to compensate for less variance parameters.

A subjective listening test was conducted to evaluate quality of synthesized speech. The subjects were asked to rate the naturalness of the synthesized speech on a scale from one (completely unnatural) to five (natural). The subjects were 10 Japanese. Twenty sentences were randomly chosen from the evaluation sentences. Figure 2 plots the experimental results. In this figure, although *TieVar* and *MDL4.0* obtained almost the same score, the proposed technique with the small weight of MDL criterion achieved better subjective scores than the conventional one. This indicates that the proposed technique constructed the optimal tying structures for each of mean and variance parameters. It can be seen from the table 1 that although the total number of parameters is almost the same in *Baseline* and *MDL1.0*, their balance between the number of mean and variance parameters are different. Even though this indicates that mean parameters are relatively more important than variance parameters, some degree of freedom for variance parameters is necessary for improving the quality of synthesized speech.

4. CONCLUSIONS

In this paper, we proposed an optimization algorithm of independent mean and variance parameter tying structures for HMM-based speech synthesis. The proposed technique constructed simultaneously tying structures for both mean and variance parameters using context clustering algorithm in additive structure models. In the experiments, the proposed technique outperformed the conventional one. Investigation of the appropriate size of the trees will be future work.

5. 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.

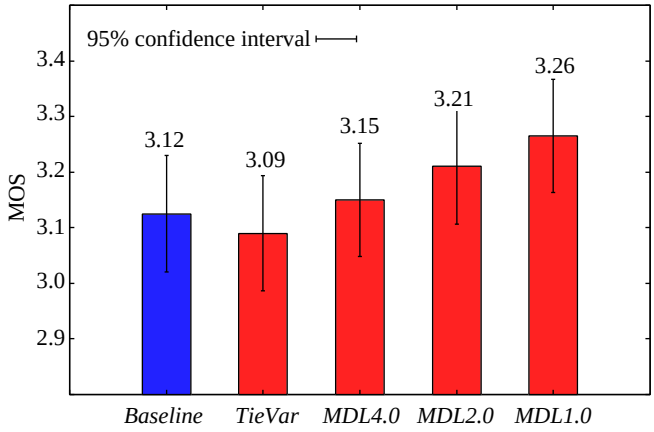

Fig. 2. *Mean opinion scores for synthesized speech obtained by the conventional and proposed techniques.*

6. 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] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," Proc. of J. Acoust. Soc. Jpn. (E), vol. 21, pp. 76–86, 2000.

[4] K. Oura, H. Zen, Y. Nankaku, A, Lee, and K. Tokuda, "A Covariance-tying technique for HMM-based speech synthesis," Proc. of IEICE, vol. E93–D, no. 3, pp. 595–601, 2010.

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

[6] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," Proc. of ICASSP, pp. 137–140, 1992.

[7] H. Kawahara, M. K. Ikuyo, and A. Cheneigne, "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," Proc. of Speech Communication, 27, pp. 187–207, 1999.

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

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

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

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

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