ICSP2008 Proceedings

Analysis of Stream-Dependent Tying Structure for HMM-based Speech Synthesis

Zhi-Peng Yu, Yi-Jian Wu, Heiga Zen, Yoshihiko Nankaku, Keiichi Tokuda
Nagoya Institute of Technology, Japan
{shixin33,yjwu,zen,nankaku}@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

In conventional HMM-based speech synthesis framework, spectral features are modeled in one stream, and stream-dependent tree-based clustering was then applied for tying the model parameters. In this paper, we investigate several different stream-dependent tying structures for spectral features by splitting the feature vector into several streams. One splitting approach is to split each feature dimension into each stream. Another one is to split the static and dynamic features into different streams. Although splitting spectral features into different streams would ignore the correlation of context dependency between them, the number of model parameters can be optimized for each stream after stream-dependent clustering. From the experimental results, both splitting approaches can improve the quality of synthesized speech. However, the quality of synthesized speech became worse when we combined these two splitting approaches.

Index Terms: HMM-based speech synthesis, stream-dependent tying structure

1. Introduction

HMM-based speech synthesis (HTS) has been proposed for a decade [1]. In this method, spectrum, pitch and duration are modeled simultaneously in a unified framework of HMMs [2], where continuous probability distributions are used for spectral modeling and multi-space probability distributions (MSD) [3] are used for F0 modeling. In synthesis, the speech parameter sequence is generated by maximizing the likelihood of HMMs related to the parameter sequence under the constraint between static and dynamic features [4]. Due to its trainable framework and stable synthesized speech, HMM-based speech synthesis approach has been widely adopted over recent years.

In the conventional framework of HMM-based speech synthesis, the feature vector for HMM modeling consists of two streams, where one stream is used for modeling of spectral parameters and another one is for modeling of F0 parameters. In the model training, the context dependent models are firstly trained, and then the stream-dependent tying structure are built by using decision-tree based context clustering [5]. Since only one stream is used for modeling of spectral parameters, the tying structure may not be optimized for each dimension of spectral parameters.

In this paper, we investigate several stream-dependent tying structures for spectral parameters by splitting the spectral feature vector into several streams. Two splitting approach are adopted, where one approach is to split each feature dimension into each stream, and another one is to split the static and dynamic features into different streams. We analyze the effect of different stream splitting way by considering the number of model parameters after tree-based clustering and the quality of synthesized speech.

The rest of paper is organized as follows. In Section 2, the conventional stream-dependent tying structure is introduced. In Section 3, we present the stream-dependent tying structures after splitting spectral features into different streams. In Section 4, the experiments and listening test to evaluate the effect of new stream-dependent tying structure are described. In Section 5, we give the conclusions and future work.

2. Stream-dependent tying structure

2.1. Stream structure

In current HMM-based speech synthesis framework, the speech feature vector used for HMM modeling consists of spectrum and pitch part. The spectrum part includes the mel-cepstral coefficients [6] and their dynamic features (delta and delta-delta coefficients). The pitch part includes logarithm of F0 and its dynamic features. The stream structure is show in Fig. 1(a). In our system, 25th mel-cepstral coefficients including c(0) are used.

In the spectrum part, there is only one stream for modeling all of the dimensions of spectrum feature vector. Considering correlation between different dimensions, we combine them together to train HMMs and perform tree-based context clustering.

2.2. Tree-based clustering

When we construct context dependent models with the combinations of contextual factors such as mora count, stress and part-of-speech in addition to current, preceding and succeeding phonemes, the model parameter could be trained with high accuracy if enough data is available.

978-1-4244-2179-4/08/$25.00 ©2008 IEEE

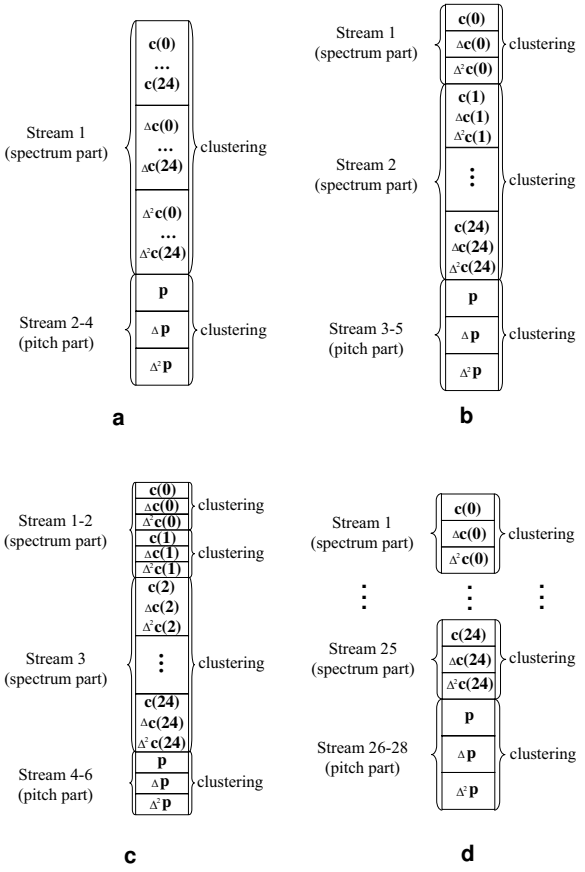

Figure 1: *(a) Conventional stream structure; (b) The structure separating c(0) of mel-cepstral's vector; (c) The structure separating c(0) and c(1) of mel-cepstral's vector; (d) The structure separating every dimension*

However, as contextual factors increase, their combinations also increase exponentially. Therefore, the model parameters with sufficient accuracy cannot be estimated with limited training data. Furthermore, it is impossible to prepare speech database which includes all combinations of contextual factors. To overcome this problem, a decision-tree based context clustering technique is applied for tying the model parameters for spectrum, F0 and duration. Since spectrum, F0 and duration are affected by different contextual contextual factors, the model parameters for spectrum, F0 and duration are clustered independently.

3. Stream-splitting approaches

Due to the correlation of context dependency, we combined all dimensions of spectral feature vector and their dynamic features into one stream. However, we may ignore the model diversity for each dimension or between static and dynamic features. In order to investigate these diversities, we try to split the spectral features into independent streams as follows.

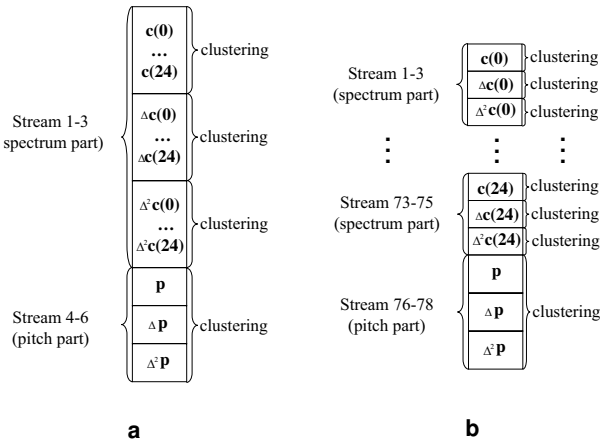

Figure 2: *(a) The structure separating dynamic features from static features; (b) The structure Splitting all features*

3.1. Splitting feature dimensions

3.1.1. Structure separating $c(0)$

Here we adopt mel-cepstral coefficients with 25-th order including c(0) as a spectrum feature vector. As we have known, the lower dimensions of mel-cepstral coefficients contains much more independent information and complexity than the higher dimensions. Accordingly, it could be separated as an independent stream for tree-based context clustering. We firstly split one spectrum feature vector stream into two streams by separating $c(0)$ from other dimensions. The stream structure after splitting is shown in Fig. 1(b).

3.1.2. Structure separating $c(0)$ and $c(1)$

We continue to involve more lower dimensions of mel-cepstral coefficients, and split the spectral feature into more streams by separating $c(1)$ by the same way. As a result, the stream for spectrum feature vector is split into three streams, and the related stream structure is shown in Fig. 1(c).

3.1.3. Structure separating every dimension

In order to observe independent influence by each dimension of mel-cepstral feature, we split each dimension of feature vector into independent streams for tree-based context clustering. The structure with 25 streams for the spectrum feature vector is shown in Fig. 1(d).

3.2. Splitting static and dynamic features

Not only large diversities may exist in different mel-cepstral feature dimensions, but also there are differences between the static and dynamic features. From the results in [7], the static features contains more complexity than

the dynamic features. In order to figure out the difference between static and dynamic features, we split them into three independent streams, which is shown in Fig. 2(a).

3.3. Splitting all features

Finally, we split each feature dimensions into one stream by combining the above two stream-splitting approaches, i.e. 75-dimension spectral feature vector is split into 75 streams. The stream structure is shown in Fig. 2(b). In this case, the correlations of context dependency between each dimension of feature vector and that between static and dynamic features are ignored in the stream dependent tree-based clustering.

4. Experiments

4.1. Experimental setups

In this experiment, we used the phonetically balanced 503 sentences from ATR Japanese speech database (B-set, speaker: myi), where the first 450 sentences were used as training data, and the remaining 53 sentences were used for evaluation. Speech signal were sampled at a rate of 16kHz. On the condition of 5ms frame shift, F0 was extracted by TEMPO [8], and mel-cepstral coefficients (mcp) which describe the spectrum of acoustic features were extracted by SPTK Toolkit [9]. Feature vector consists of static features, including 25-th mel-cepstral coefficients and logarithm of F0, and their delta and delta-delta coefficients. The system for training and synthesis was built using HTS-2.1, which is a hidden-semi Markov model (HSMM) based speech synthesis system [10]. In synthesis, the Mel Log Spectrum Approximation (MLSA) filter [11] was used to synthesize the speech waveform.

4.2. Parameter complexity

As described in Sec. 3, we compared the results of three stream-splitting approaches. Fig. 3 shows the tree size after clustering in the baseline and the approach one by splitting feature dimension into different streams. In baseline, we combined every dimension feature together, each dimension have the same number of Gauss distribution after tree-based clustering. In approach one, we separate $c(0)$, $c(0)$ & $c(1)$ and each dimension of mel-cepstral coefficients, respectively. From the results, we confirmed that the lower dimensions of spectrum feature contain much more independent information and complexity than the higher dimensions. The tree size after clustering for another splitting approach is shown in Fig. 4. From the figure, it can be seen that the static features contains more complexity than the dynamic features. The similar result can be found in Fig. 5 for the case by combining these two splitting approaches to split all the dimensions of features.

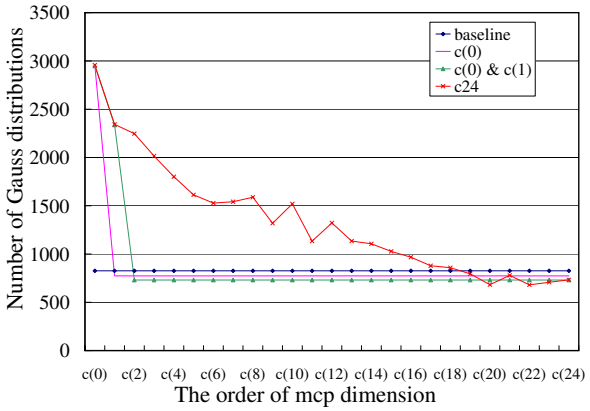

Figure 3: *The tree size of baseline, separating c(0), separating c(0) and c(1), and separating every dimension.*

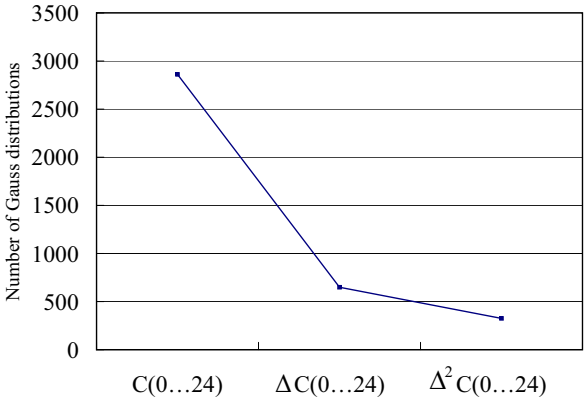

Figure 4: *The tree size of static and dynamic features.*

4.3. Listening test

We conducted a formal subjective listening test to evaluate five types of stream-dependent structure, which includes the baseline (all dimensions of mcp and their dynamic features are combined in one stream), c0 (only $c(0)$ is separated as one stream), c24 (each dimension of mel-cepstral feature is split into an independent stream), c+delta (three streams consist of static and dynamic features) and c24+delta (75 streams which include separated each dimension of mel-cepstral and their dynamic features). Ten Japanese listeners participated in the test. Each listener evaluated 15 sets of samples, where each set includes five synthesized speech from the above five systems, and gave a score from 1(bad) to 5(good) on the quality of synthesized speech. The speech samples were randomly selected for each listener from the 53 test sentences of each set.

The result of listening test is shown in Fig. 6. It can be seen from the figure that the quality of synthesized speech was improved when we separate the dimension of mel-

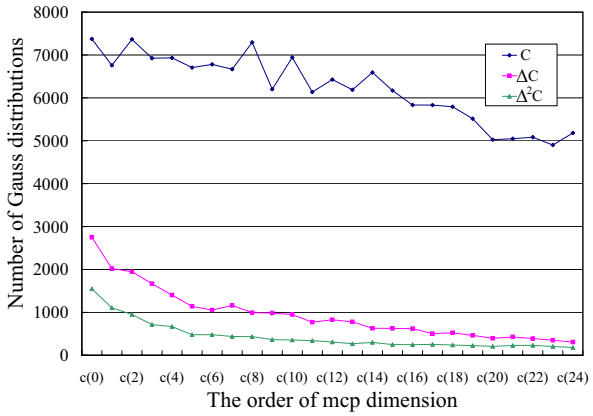

Figure 5: *The tree size of every dimension's static and dynamic features.*

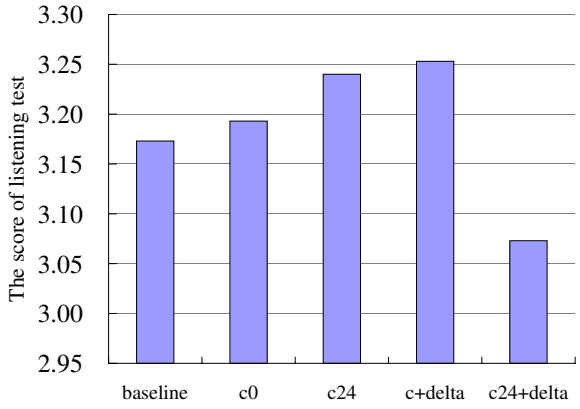

Figure 6: *the result of listening test.*

cepstral coefficients, or when we separate the static and dynamic features. Although splitting the spectral feature into different streams may ignore the correlation of context dependency between them, the model complexities can be optimized for each spectral features in tree-based clustering. The improvement of synthesized speech is benefited from this. However, the quality of synthesized speech became much worst when we combined these two splitting approach to split every dimension into an independent stream and separate their static feature from dynamic features. This indicates the ignorance of context dependency between each dimension result in a serious problem in such extreme case. Therefore, we should make a tradeoff by considering the model complexity of each features and the correlations between them.

5. Conclusions and future work

In this paper, we adopt three stream-splitting approaches to split the spectral feature vector into different streams, and investigate the effect on the tree-based context clustering and the model training. The experimental results show that separating each dimension of mel-cepstral coefficients or separating static features and dynamic features into different streams can improve the quality of synthesized speech. However, the results became much worse when separating all of the feature dimensions by combining both splitting approach, which indicates that we need to make a tradeoff between the model complexity and the correlation of context dependency for each spectral features.

Future work is to explore better stream-dependent tying structures by involving another spectrum feature such as line spectral pairs.

6. References

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

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of ICASSP*, 1999, vol. 5, pp. 2347–2350.

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

[4] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proc. of ICASSP*, 1995, pp. 660–663.

[5] J. J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," in *PhD dissertation*,Cambridge Universit,1995.

[6] T. Fukada, K. Tokuda, Kobayashi T., and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137C140.

[7] H. Zen, K. Tokuda and T. Kitamura, "Decision tree based simultaneous clustering of phonetic contexts, dimensions, and state positions for acoustic modeling, " in *Proc. of Eurospeech*, 2003, pp.3189-3192.

[8] Kawahara, H. et al., "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *IProc. Eurospeech99*, 1999, pp.2781-2784.

[9] http://sp-tk.sourceforge.net/

[10] Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, Tadashi Kitamura, "A hidden semi-Markov model-based speech synthesis system, " in *IEICE Trans. Inf and Syst*, 2007, vol.E90-D, No.5, pp.825-834.

[11] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*,1983, pp. 93C96.