# SPEAKING STYLE ADAPTATION USING CONTEXT CLUSTERING DECISION TREE FOR HMM-BASED SPEECH SYNTHESIS

*Junichi Yamagishi, Makoto Tachibana, Takashi Masuko, Takao Kobayashi*

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {Junichi.Yamagishi,Makoto.Tachibana,masuko,takao.kobayashi}@ip.titech.ac.jp

## ABSTRACT

This paper describes an MLLR-based speaking style adaptation technique for HMM-based speech synthesis. Since speaking styles and emotional expressions are characterized by many suprasegmental features as well as segmental features, it is necessary to adapt suprasegmental features for speaking style adaptation. To achieve suprasegmental feature adaptation, we utilize context clustering decision trees, which are constructed in the training stage, for tying of regression matrices. Using this technique, we adapt an initial "reading" style model to "joyful" or "sad" styles. Experimental results show that, using 50 adaptation sentences, speech samples generated from adapted models were judged to be similar to the target speaking styles at rates of 92% and 70% for joyful and sad styles, respectively.

## 1. INTRODUCTION

In recent years, concatenative text-to-speech (TTS) synthesis systems based on large speech corpora have been shown to be able to synthesize natural sounding speech of good quality. However, expression of emotions or speaking styles is still a difficult problem even for the state-of-the-art TTS systems. To overcome this problem, we have shown that an HMM-based speech synthesis system can express several speaking styles and emotional expressions by training HMMs using speech database for each speaking style or emotional expression [1]. However, constructing speech database enough to train HMMs for each desired speaking style or emotional expression needs a lot of labor and time.

On the other hand, we have also proposed speaker adaptation techniques [2][3] based on MLLR (Maximum Likelihood Liner Regression) [4]. Using these techniques, we can change voice characteristics of synthetic speech so as to mimic an arbitrary target speaker's voice with a small amount of speech data uttered by the target speaker.

In this study, we apply these MLLR speaker adaptation techniques to adaptation of speaking style in HMM-based speech synthesis. We refer to this technique as "speaking style adaptation." Here the term "style" stands for one of speaking styles or emotional expressions, and is used throughout this paper. In this technique, a reading style model is used as an initial model, and adapted to that of target speaking style, e.g., "joyful" or "sad" styles, using a small amount of speech data of the target speaking style. Since speaking styles and emotional expressions are characterized by many suprasegmental features as well as segmental features, it is necessary to adapt suprasegmental features in addition to adaptation of segmental features. To achieve suprasegmental feature adaptation, we utilize context clustering decision trees, which are constructed in the training stage, for tying of regression matrices. A set of questions used for context clustering includes a lot of questions related to suprasegmental features such as accent type, length of accentual phrase, and position of mora. As a result, it is thought that not only segmental features but also suprasegmental features can be adapted from a speaking style to another if the context clustering decision trees are constructed appropriately.

## 2. OVERVIEW OF HMM-BASED SPEECH SYNTHESIS WITH MLLR ADAPTATION

In this paper, we use an HMM-based speech synthesis system with an MLLR model adaptation framework which is almost the same as the HMM-based speech synthesis systems used in [2][3] except that speaker adaptation is replaced by speaking style adaptation.

In the training stage, context dependent phoneme HMMs are trained. Spectrum and $F_0$ are modeled by multi-stream HMMs in which output distributions for spectral and $F_0$ parts are modeled using continuous probability distribution and multi-space probability distribution (MSD) [5], respectively. To model variations of spectrum and $F_0$, phonetic and linguistic contextual factors, such as phoneme identity factors, stress related factors and locational factors, are taken into account. Then, a decision tree based context clustering technique [6][7] is separately applied to the spectral and $F_0$ parts of the context dependent phoneme HMMs. Finally, state durations are modeled by multi-dimensional Gaussian distributions, and the state clustering technique is applied to the duration models.

In the adaptation stage, the initial seed model is adapted to a new target speaking style using a small amount of speech data of the style. In the following experiment, a reading style model is used as the initial seed model of adaptation. We use MSD-MLLR algorithm [2] for spectrum and $F_0$ adaptation, and extended MLLR algorithm [3], for state duration adaptation, respectively.

In the synthesis stage, texts are transformed into a context dependent label sequence. According to the label sequence, a sentence HMM is constructed by concatenating context dependent HMMs. From the sentence HMM, spectral and $F_0$ parameter sequences are obtained based on ML criterion. Finally, by using MLSA filter, speech is synthesized from the generated mel-cepstral and $F_0$ parameter sequences.

## 3. MLLR-BASED SPEAKING STYLE ADAPTATION

### 3.1. Adaptation of Spectral and $F_0$ parameters

MSD-MLLR [2], which is an extension of MLLR [4] to MSD-HMM, is used for adaptation of spectral and $F_0$ parameters.

Let $\boldsymbol{\mu}_{ig}$ and $\boldsymbol{\Sigma}_{ig}$ be the mean vector and the covariance matrix of the multi-space output probability distribution of state $i$ for space $g$, respectively. The adapted mean vector and covariance matrix $\widehat{\boldsymbol{\mu}}_{ig}$, $\widehat{\boldsymbol{\Sigma}}_{ig}$ are estimated as follows:

$$\widehat{\boldsymbol{\mu}}_{ig} = \boldsymbol{W}_{ig}\boldsymbol{\xi}_{ig}, \tag{1}$$

$$\widehat{\boldsymbol{\Sigma}}_{ig} = \boldsymbol{B}_{ig}^{\top}\boldsymbol{H}_{ig}\boldsymbol{B}_{ig}, \tag{2}$$

where

$$\boldsymbol{\xi}_{ig} = [1, \boldsymbol{\mu}_{ig}^{\top}]^{\top}, \tag{3}$$

$$\boldsymbol{B}_{ig} = \boldsymbol{C}_{ig}^{-1}, \tag{4}$$

$$\boldsymbol{C}_{ig}\boldsymbol{C}_{ig}^{\top} = \boldsymbol{\Sigma}_{ig}^{-1}, \tag{5}$$

and $\cdot^{\top}$ denotes matrix transpose. The regression matrices $\boldsymbol{W}_{ig}$ and $\boldsymbol{H}_{ig}$ for the mean vector and the covariance matrix, respectively, are obtained by solving a maximization problem of logarithm of likelihood for adaptation data using EM algorithm [2][4].

### 3.2. Tying of Regression Matrices

In general, it is impossible to estimate the MLLR regression matrices for each distribution because the amount of adaptation data of a target speaking style is small. Therefore, MLLR makes use of regression class trees to group the distributions in the model, and to tie the regression matrices at each group. Tying of each regression matrices makes it possible to adapt distributions which have no adaptation data.

The regression class tree is constructed based on the distribution distance such as a Euclidean distance measure [8],

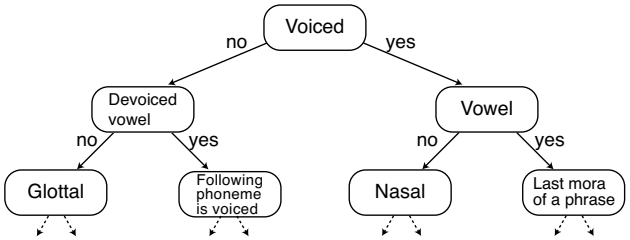



**Fig. 1**. An example of the context clustering decision tree.

and each leaf/terminal node of the tree specifies a particular cluster of distributions in the model. At the regression class tree, nodes in which regression matrices are estimated are determined according to expectation of adaptation data using a top-down approach to traverse regression class tree. We now define a set of distributions belonging to node $l$ as $C_l = \{\mathcal{N}_1, \mathcal{N}_2, \cdots, \mathcal{N}_{I_l}\}$. The expectation of adaptation data of the node $l$ for space $g$ is given by

$$s_{lg} = \sum_{i=1}^{I_l} \sum_{t \in T(\boldsymbol{O}, g)} \gamma_{ig}(t), \tag{6}$$

where $I_l$ is the number of distributions belonging to the node $l$, $T(\boldsymbol{O}, g)$ is a set of time slots for which a set of space indexes of observation vector $\boldsymbol{o}_t$ includes space index $g$ at time $t$, and $\gamma_{ig}(t)$ is a probability that the observation vector at time $t$ is generated in state $i$ for space $g$.

### 3.3. Speaking Style Adaptation Using Context Clustering Decision Tree

In this paper, we apply the MLLR-based adaptation techniques to adaptation of speaking style in HMM-based speech synthesis. However, the tying method of regression matrices based on the regression class tree has several problems to achieve appropriate speaking style adaptation. One of significant problems is that the regression class tree has an ability to adapt just segmental level features, in other words, it is difficult to adapt suprasegmental level features. This is because the regression class tree is constructed using the distributions distance and does not reflect connections between the distributions in the model on the time axis. However, it is obvious that speaking styles and emotional expressions are characterized by many suprasegmental features as well as segmental features. Therefore, to adapt an initial speaking style to another, we have to determine the matrix tying structure taking account of the suprasegmental phonetic and linguistic features. To overcome this problem, we utilize context clustering decision trees constructed in the training stage for the tying of the regression matrices instead of regression class trees.

The context clustering decision tree is a binary tree, and each non-terminal node of the decision tree has a question related to phonetic and linguistic contextual factor and each leaf/terminal node of the decision tree is associated with a distribution in the model. The set of questions includes a lot of questions related to suprasegmental features such as accent type, length of accentual phrase, and position of mora. Therefore, the use of context clustering decision tree for the tying of the regression matrices makes it possible to adapt not only segmental features but also suprasegmental features if the context clustering decision trees are constructed appropriately.

Figure 1 shows an example of the context clustering decision tree. In this tree, there is a node whose question is "Does the phoneme belong to the last mora of a phrase?" If regression matrices are tied at the "yes" node to the question, and the target speaking style has a particular change at the end of phrases, it is expected that synthesized speech from adapted models also has a similar characteristic to the target speaking style, that is, the particular change at the end of phrases.

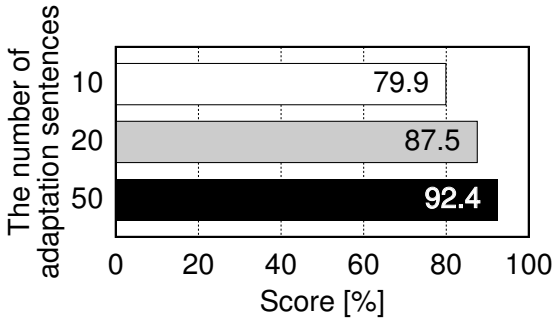
## 4. EXPERIMENTS

### 4.1. Experimental Conditions

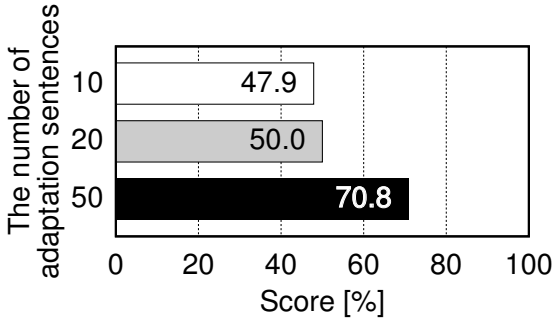
We used three speaking styles, that is, "reading," "joyful," and "sad." Speech database [1] contains a set of phonetically balanced 503 sentences of ATR Japanese speech database uttered by a male speaker for each speaking style. We used 42 phonemes including silence and pause, and phoneme labels and linguistic informations were taken into account. Details are given in [1].

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis [9]. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients.

We used 5-state left-to-right HMMs. The initial seed model for speaking style adaptation was trained using 450 sentences of reading style. We set joyful and sad style as target speaking styles, and adapted the initial seed model, namely reading style, to the target speaking styles using 10, 20, or 50 sentences which were not included in the test sentences. In the adaptation, thresholds for traversing regression class tree or context clustering decision tree were set to 1000 for the spectral part, 150 for the $F_0$ part, and 200 for state duration distributions, respectively. For comparison, we also trained target speaking style models using 450 sentences for each style. Subjects of the following listening tests were nine males.




(a) Joyful




(b) Sad

**Fig. 2**. Results of the ABX tests.

### 4.2. Evaluation of Speaking Style Adaptation

We conducted ABX listening tests to evaluate the performance of speaking style adaptation using the proposed technique with decision trees. In the ABX tests, A and B were synthesized speech generated from the initial reading style model and the target speaking style model, respectively. And X was synthesized speech generated from adapted model. Subjects were presented synthesized speech in the order of A, B, X or B, A, X, and asked to select first or second speech as being similar to X. For each subject, three test sentences were chosen at random from 53 test sentences which were not included in the training data.

Figure 2 shows the average percentages that synthesized speech from adapted models are judged to be similar to speech from target models. Figure 2(a) shows the results for "Joyful" style, and (b) shows the results for "sad" style. In the figure, white, gray, and black bars represents the results for adapted models using 10, 20, and 50 sentences, respectively. The results show that using 50 adaptation sentences, more than 70% of speech samples generated from adapted models were judged to be similar to the target speaking styles. Although one of characteristics of sad style is slow speaking rate, the speaking rate of adaptation data for sad style was much faster than the average speaking rate of whole sad style speech data. This results in lower performance for

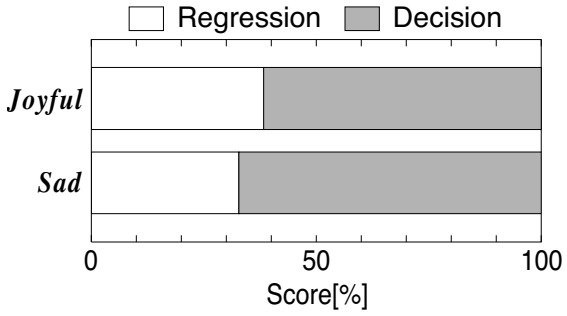



**Fig. 3**. Result of the paired comparison test.

sad style than joyful style.

### 4.3. Comparison of Regression Class Tree and Decision Tree

We compared the speaking styles of synthetic speech generated from adapted models using decision trees and regression trees by a paired comparison test. Subjects were presented a pair of the target style speech samples synthesized from those two models in random order, and then asked which synthesized speech was perceived as the intended style. For each subject, four test sentences were chosen at random from 53 test sentences which were not included in the training data.

Figure 3 shows the preference scores. It can be seen from this figure that speaking styles of synthetic speech using proposed technique are perceived better as the target styles than the conventional technique. These results shows that context clustering decision tree is more efficient for determining the regression matrix tying structure of MLLR than regression class tree.

## 5. CONCLUSION

We have described an adaptation technique of speaking styles for HMM-based speech synthesis using speaking style adaptation. The proposed adaptation technique is based on MLLR adaptation using context clustering decision tree to reflect an influence of suprasegmental features. From the results of subjective tests, we have shown that speaking styles of synthetic speech generated from the adapted model using the proposed technique with a small amount of target data resemble the target speaking styles. Future work will focus on improvement of adaptation technique using context clustering decision tree.

## 6. REFERENCES

[1] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. EUROSPEECH 2003*, Sept. 2003, pp. 2461–2464.

[2] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP 2001*, May 2001, pp. 805–808.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. EUROSPEECH 2001*, Sept. 2001, pp. 345–348.

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

[5] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP-99*, Mar. 1999, pp. 229–232.

[6] S. J. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. ARPA Human Language Technology Workshop*, Mar. 1994, pp. 307–312.

[7] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 79–86, Mar. 2000.

[8] S. Young, G. Everman, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.1*, Dec. 2001.

[9] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.