HMM-BASED EXPRESSIVE SPEECH SYNTHESIS —TOWARDS TTS WITH ARBITRARY SPEAKING STYLES AND EMOTIONS

Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi

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

ABSTRACT

This paper describes recent progress in our approach to generating expressive speech. A goal of text-to-speech (TTS) synthesis is to have an ability to generate natural sounding speech with arbitrary speaker's voice characteristics, speaking styles and emotional expressions. To change voice and speaking style and/or emotion of the synthetic speech arbitrarily with maintaining its naturalness, it is required that prosodic features as well as spectral features are controlled properly. Since prosodic features are more or less related to spectral features, it is desirable to control these features simultaneously taking account of the relationship between spectrum and prosody. To resolve this problem, we have proposed several key ideas which include speaking style interpolation and adaptation for HMM-based speech synthesis. This paper focuses on these ideas and provides an overview of our approach. Moreover we show experimental results which show the effectiveness of the approach.

1. INTRODUCTION

Realizing speaker variety with emotional expressivity and speaking style variability is becoming a most important issue in recent speech synthesis research area. This is due to the fact that the latest text-to-speech (TTS) synthesis systems based on large corpus can produce natural sounding speech but hardly change voice quality, speaking style, and emotional expression with maintaining naturalness of the synthetic speech. With regard to generating various speakers' voices in reading style speech, we showed that an HMM-based speech synthesis system [1] can generate synthetic speech which resembles an arbitrarily given target speaker's voice by applying speaker adaptation technique using a small amount of target speaker's speech data [2][3]. In addition, it was also shown that speaker interpolation [4] and eigenvoice [5] techniques have beneficial effects on diversification of speaker's voice.

In this paper, we describe several approaches to realizing emotional expressivity and speaking style variability in HMM-based speech synthesis. To change speaking style and emotional expression of the synthetic speech arbitrarily with maintaining its naturalness, it is required that prosodic features as well as spectral features are controlled properly. Moreover, since prosodic features are more or less related to spectral features, it is undesirable to control these features independently. In our approaches, speaking styles and emotions are fully statistically modeled and generated without using rules controlling prosody and other parameters. Since spectral and prosodic features are modeled simultaneously in each model, we can incorporate the relationship between spectrum and prosody into the speaking style control process implicitly.

We first investigate two methods for modeling speaking styles and emotions, which will be called "style dependent modeling" and "style mixed modeling." In this paper, we will refer to one of speaking styles or emotional expressions as the "style." In the style dependent modeling, each style is individually modeled, whereas in the style mixed modeling, each style is treated as a contextual factor as well as phonetic, prosodic, and linguistic factors, and all styles are modeled by a single acoustic model simultaneously. We choose four styles including "reading," "rough," "joyful," and "sad" styles, and compare the two modeling methods using these styles. Secondly, we investigate a method for synthesizing speech with an intermediate speaking style between two different styles by applying a model interpolation technique. We refer to this approach as "style interpolation approach." In the style interpolation approach, we choose three styles and synthesize speech from models obtained by interpolating two models for every combination of two styles. Thirdly, we present a method for generating speech with a desired style based on model adaptation using a small amount of speech data of the target style. We refer to this approach as "style adaptation." In the style adaptation approach, the reading style model is used as an initial model and adapted to that of target style, e.g., joyful or sad style, based on an MLLR (Maximum Likelihood Liner Regression) model adaptation technique. From results of subjective experiments we show the effectiveness of the proposed approaches.

2. STYLE MODELING FOR HMM-BASED SPEECH SYNTHESIS

2.1. Style Dependent Model and Style Mixed Model

Our first approach to generating speech with changing styles is to make models of various styles which can be utilized for HMM-based speech synthesis. In the HMM-based synthesis, context dependent phoneme HMMs are used as the synthesis units, in which spectrum and F_0 are modeled simultaneously [1]. 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 is separately applied to the spectral and F_0 parts of the context dependent phoneme HMMs.

For the purpose of modeling styles, we investigate two methods called style dependent modeling and style mixed modeling [6]. In the style dependent modeling each style is individually modeled. Then a pseudo root node is added to assemble models for all

Table 1. Evaluation of speech database with four styles.

Reading	Rough	Joyful	Sad
503 (100%)	493 (95%)	499 (98%)	502 (99%)

Table 2. The number of distributions after decision tree based context clustering using MDL criterion.

	Dependent					Mixed
	Read.	Rough	Joyful	Sad	Total	
Spec.	891	752	808	926	3377	2796
F_0	1316	1269	1368	1483	5436	4404
Dur.	1070	1272	1057	950	4349	3182

styles into a single acoustic model. One of the advantages of this method, we can easily add a new style by constructing an acoustic model for the new style and adding a path from the pseudo root node to the root node of the decision tree for the style.

In the style mixed modeling, speaking styles and emotional expressions are treated as the contextual factors as well as phonetic and linguistic factors, and all styles are modeled by a single acoustic model. In this method, it is not easy to add new styles because the whole acoustic model should be reconstructed. On the other hand, it is expected that appropriate sharing of similar parameters among some styles would improve accuracy of the shared parameters and lead to a compact model.

2.2. Speech Database

Although there exist a wide variety of speaking styles and emotions in real speech, it is not easy to collect them completely. As the first step toward modeling and synthesis of expressive speech we chose four styles, which are reading, rough, joyful, and sad styles, and constructed a speech database composed of phonetically balanced 503 sentences of ATR Japanese speech database uttered by a male speaker in each style.

In the phonetically balanced 503 sentences, there are a number of sentences whose meaning are unsuitable for several styles. Therefore, first, we evaluated whether the recoded speech samples were perceived as being uttered in the intended styles. Subjects were nine males, presented all sentences of each style, and then asked whether the test speech sounded intended style or not.

Table 1 shows the number and percentage of sentences which are judged to sound the intended style by a majority of the subjects. From the table, it can be seen that almost all of the speech samples in the database sound the intended styles.

2.3. Evaluation of Style Modeling

We used 42 phonemes including silence and pause, and the following contextual factors were taken into account:

- the number of morae in sentence
- position of breath group in sentence
- the number of morae in {preceding, current, succeeding} breath group
- position of current accentual phrase in current breath group
- the number of morae and accent type in {preceding, current, succeeding} accentual phrase
- {preceding, current, succeeding} part-of-speech
- position of current mora in current accentual phrase
- difference between position of current mora and accent type
- {preceding, current, succeeding} phoneme
- style (for style mixed model)

It is noted that these contextual factors except style are the same as [1] in which only reading style is taken into account.

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. 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. Both style dependent and mixed models were trained using 450 sentences for each style.

Table 2 shows the number of distributions of each model after decision tree based context clustering using MDL criterion. The entries for "Dependent" and "Mixed" correspond to the number of distributions for style dependent model and style mixed model, respectively, and "Read.," "Rough," "Joyful," and "Sad" correspond to the numbers of distributions for each style in the style dependent model, respectively. "Total" represents the sum of those numbers. In addition, "Spec.," "F_0," and "Dur." represent the spectrum, F_0, and state duration, respectively. It can be seen that the numbers of distributions for style mixed model are smaller than style dependent model.

We then conducted a classification test for styles of synthesized speech. Subjects were eleven males, and asked which style, namely, reading, rough, joyful, and sad, the test speech sounded. It it noted that test speech was classified into "Other" when it was thought to be classified into none of the above four styles. For each subject, eight test sentences were chosen at random from 53 test sentences which were not contained in training data[1]. Table

Table 3. Subjective evaluations of reproduction of styles.
(a) Style Dependent Model

Synthetic Speech	Classification (%)				
	Read.	Rough	Joyful	Sad	Other
Read.	98.3	0.6	0.0	0.0	1.1
Rough	6.9	82.3	0.0	0.0	10.8
Joyful	1.1	0.0	94.9	0.0	4.0
Sad	0.6	1.1	0.0	94.9	3.4

(b) Style Mixed Model

Synthetic Speech	Classification (%)				
	Read.	Rough	Joyful	Sad	Other
Read.	98.9	0.0	0.0	0.0	1.1
Rough	2.8	89.8	0.0	1.1	6.3
Joyful	0.6	0.0	96.0	0.0	3.4
Sad	0.0	0.6	0.0	96.0	3.4

[1]Several speech samples used in the test are available from http://sp-www.ip.titech.ac.jp/research/demo/.

3 shows the classification rates. In the table, (a) is the result for speaker dependent model, and (b) is that for speaker mixed model. It can be seen from the results that both modeling methods have almost the same performance, and that it is possible to synthesize speech with similar styles to those of the recorded speech.

We also compared the naturalness of synthesized speech generated from style dependent and mixed models by a paired comparison test. It was found that the style mixed modeling is almost equal to the style dependent modeling in naturalness of synthesized speech, though the number of output distributions of style mixed model is smaller than that of style dependent model. From the result, it can be thought that style mixed modeling is more efficient for modeling speech with several styles than style dependent modeling.

3. INTERPOLATION OF STYLE MODELS

3.1. Model Interpolation for HMMs

It has been shown in [4] that it is possible to synthesize speech with intermediate voice characteristics between two speakers by interpolating two speakers' models. In our second approach, we apply a model interpolation technique to style models to synthesize speech with an intermediate style between representative styles.

Among three interpolation methods proposed in [4], we adopt the simplest method. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be models of N representative styles $S_1, S_2, \ldots, S_N$, and $\tilde{\lambda}$ be a model of a style $\tilde{S}$ obtained by interpolating N representative style models. When an observation vector $\tilde{\boldsymbol{o}}$ of the style $\tilde{S}$ is obtained by linearly interpolating observation vectors $\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_N$ of the representative styles as follows:

$$\tilde{\boldsymbol{o}} = \sum_{k=1}^{N} a_k \boldsymbol{o}_k \tag{1}$$

where $\sum_{k=1}^{N} a_k = 1$, and a mean vector $\tilde{\boldsymbol{\mu}}$ and a covariance matrix $\tilde{\boldsymbol{U}}$ of a Gaussian output pdf $p(\tilde{\boldsymbol{o}}) = \mathcal{N}(\tilde{\boldsymbol{o}}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{U}})$ is calculated by

$$\tilde{\boldsymbol{\mu}} = \sum_{k=1}^{N} a_k \boldsymbol{\mu}_k, \quad \tilde{\boldsymbol{U}} = \sum_{k=1}^{N} {a_k}^2 \boldsymbol{U}_k, \tag{2}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{U}_k$ are the mean vector and the covariance matrix of an output pdf of speaking style S_k, respectively.

If the models λ_k $(1 \leq k \leq N)$ of representative styles have a tying structure common to all models, it is possible to obtain the interpolated model $\tilde{\lambda}$ by interpolating λ_k directly. In general, however, the models λ_k have different structure from each other when the context clustering is independently performed for each style model in the training stage. Consequently, it is difficult to obtain $\tilde{\lambda}$ by interpolating λ_k taking account of model structure. To avoid this problem, in the synthesis stage, we first generate N pdf sequences from λ_k independently, and then obtain a pdf sequence corresponding to $\tilde{\lambda}$ by interpolating these N pdf sequences. Finally, a speech parameter sequence is generated from the interpolated pdf sequence.

3.2. Evaluation of Style Interpolation

We used three speaking styles, which are reading, joyful, and sad styles. Although style mixed modeling has been shown to be able to reduce the total number of output distributions, we adopt style dependent modeling in this paper because it is easy to add or remove styles without retraining the whole model. We obtained interpolated speech samples for three combinations, namely, reading and joyful, reading and sad, and joyful and sad. Here we denote the new speaking style interpolated between styles A and B as (A, B). Interpolation ratio is set to $1:1$, i.e., $(a_A, a_B) = (0.5, 0.5)$ for all cases.

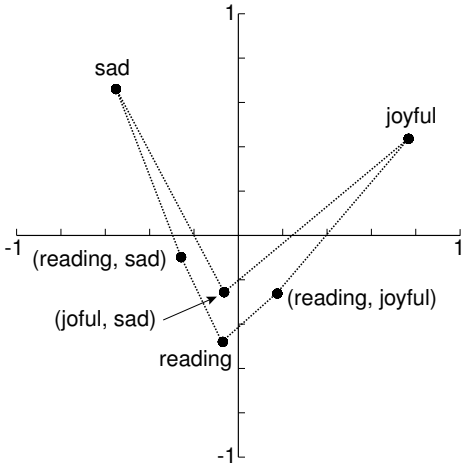

Fig. 1. Evaluation of similarity between speaking styles

We carried out a subjective test to evaluate similarity of synthetic speech samples between the following representative and interpolated styles: reading, joyful, sad, (reading, joyful), (reading, sad), and (joyful, sad). Subjects were presented two speech samples chosen from the above six styles in random order, and asked to evaluate similarity of each pair in a five-point scale in which "5" means very similar and "1" means quite different. For each subject, four test sentences were chosen at random from 53 test sentences which were not included in the training data. Subjects were eight males.

From the result of similarity evaluation, we placed six styles in a 2-dimensional space according to the similarities between styles by using the Hayashi's quantification theory type IV [7]. Figure 1 shows the relative similarity distance between speaking styles. In this figure, it is thought that the horizontal axis corresponds to the degree of pleasure, and the vertical axis corresponds to intensity of emotional expression. It can be seen that interpolated speaking styles (reading, joyful) and (reading, sad) are placed in between representative styles. From this result, it is thought that we can synthesize speech with a speaking style in between two representative speaking styles by using model interpolation technique. It can also be thought that since interpolated style (joyful, sad) is placed near the reading style, joyful and sad styles have opposite features putting reading style between these two styles in the model parameter space.

4. ADAPTATION OF STYLE MODELS

4.1. Style Adaptation Using Context Clustering Decision Tree

Our third approach to generating speech with a desired style is based on model adaptation. We apply here an MLLR-based model adaptation technique [2], to style models. It is obvious that speaking styles and emotional expressions are characterized by many

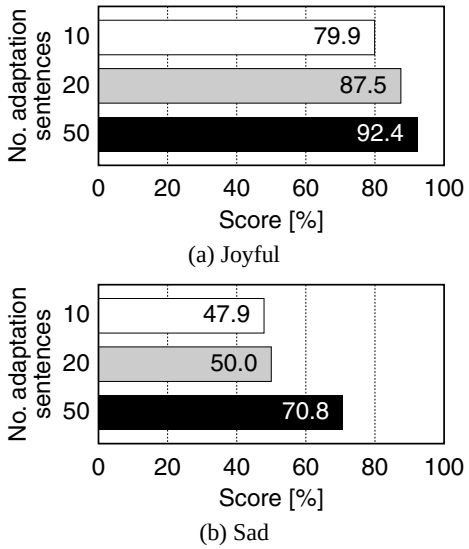

Fig. 2. Results of the ABX tests.

segment-based features as well as frame-based features. Therefore, to adapt an initial style to another, we have to determine the matrix tying structure taking account of the segment-based 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 segment-based 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 make it possible to adapt not only frame-based features but also segment-based features if the context clustering decision trees are constructed appropriately.

4.2. Evaluation of Style Adaptation

We used the following three styles: reading, joyful, and sad. The initial seed model for style adaptation was trained using 450 sentences of the reading style. We set the joyful and sad style as target speaking styles, and adapted the initial seed models, namely the 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 the target style models using 450 sentences for each style. Subjects of the following listening tests were nine males.

We conducted ABX listening tests to evaluate the performance of the style adaptation using the proposed technique with the 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 the adapted model. Subjects were presented synthesized speech in the order of A, B, X or B, A, X, and asked to select the 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 scores that synthesized speech from adapted models were judged to be similar to speech from target models. Figure 2(a) shows the results for the joyful style, and (b) shows the results for the sad style. It is shown that using 50 adaptation sentences, more than 70% of speech samples generated from the adapted models were judged to be similar to the target speaking styles. Although one of the characteristics of the sad style is slow speaking rate, the speaking rate of the adaptation data for the sad style was much faster than the average speaking rate of the whole sad style speech data. This results in lower performance for the sad style than the joyful style.

5. CONCLUSION

We have presented several promising approaches to realizing emotional expressivity and speaking style variability in HMM-based speech synthesis. First, we described two methods for modeling speaking styles and emotions. Secondly, we investigated a method for synthesizing speech with an intermediate speaking style between two different styles by applying a model interpolation technique. Thirdly, we presented a method for generating speech with a desired style based on model adaptation using a small amount of speech data of the target style based on an MLLR model adaptation technique. From results of subjective experiments we have shown the effectiveness of the proposed approaches. Our future work will focus on investigation using other styles and simultaneous adaptation of speaker and style.

6. REFERENCES

[1] 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. EUROSPEECH-99*, Sept. 1999, pp. 2374–2350.

[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] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation for hmm-based speech synthesis system," *J. Acoust. Soc. Jap. (E)*, vol. 21, pp. 199–206, Apr. 2000.

[5] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, and T. Masukoand T. Kobayashi, "Eigenvoices for hmm-based speech synthesis," in *Proc. ICSLP-2002*, Sept. 2002, pp. 1269–1272.

[6] 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.

[7] C. Hayashi, "On the prediction of phenomena from mathematicostatistical point of view," *Annals of the Institute of Statistical Mathematics*, vol. 3, pp. 69–98, 1952.