Performance Evaluation of Style Adaptation for Hidden Semi-Markov Model Based Speech Synthesis

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

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan

{makoto.tachibana, junichi.yamagishi, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper describes a style adaptation technique using hidden semi-Markov model (HSMM) based maximum likelihood linear regression (MLLR). The HSMM-based MLLR technique can estimate regression matrices for affine transform of mean vectors of output and state duration distributions which maximize likelihood of adaptation data using EM algorithm. In this study, we apply this adaptation technique to style adaptation in HSMM-based speech synthesis. From the results of several subjective tests, we show that the HSMM-based MLLR technique can perform style adaptation with maintaining naturalness of the synthetic speech compared with the conventional HMM-based MLLR technique.

1. Introduction

The recent progress of speech synthesis technology have shown capability of synthesizing speech with high quality, speaker variability, and emotional expressivity. HMM-based speech synthesis is one of such promising approaches, especially in adding speaker variability and emotional expressivity to synthetic speech. It has already been shown that speaking styles and/or emotional expressions are well modeled using the HMM-based speech synthesis framework [1] and a model adaptation technique applied to the style models, called style adaptation technique, enables us to generate speech having a given target style [2] as well as generating an arbitrarily given target speaker’s voice [3] with a small amount of adaptation data.

Recently, an explicit duration modeling has been incorporated into the HMM-based synthesis framework [4]. In the conventional HMM-based synthesis system [5], synthesis unit HMMs are trained without explicit duration models, however, speech parameter generation assumes the use of HMMs with explicit duration models. This inconsistency is resolved by introducing a hidden semi-Markov model (HSMM) [6], which is an HMM with explicit state duration probability distributions. It has been shown that naturalness of the synthetic speech is slightly improved by using HSMM [4]. Furthermore, an extension of a maximum likelihood linear regression (MLLR) [7] model adaptation algorithm for HMM to HSMM has been developed [8]. Although mathematical aspect of adaptation algorithm for state duration was described in [8], clarifying its effectiveness when applied to speech synthesis has been left as a future work.

In this paper, we apply the HSMM-based MLLR adaptation algorithm to the style adaptation and examine its performance with making comparison between HMM-based and HSMM-based systems. In the experiments, we use four styles of read speech — neutral, rough, joyful, and sad styles, which are the same as those used in the style modeling [1]. We choose a neutral reading style model as an initial model and adapt it to that of a target style chosen from remaining three styles using a small amount speech data of the target style. In the MLLR adaptation, we utilize context clustering decision trees for tying of regression matrices in which suprasegmental features are taken into account as well as frame-based features [2].

2. Hidden Semi-Markov Model-based MLLR Adaptation

Here we briefly review the HSMM-based MLLR algorithm [8]. With regard to the HSMM-based speech synthesis, a description of the system can be found in [4].

We assume that each speech synthesis unit is modeled by an N-state HSMM $\lambda$. We also assume that the $i$-th state output vector $b_i(\alpha)$ and duration distributions $p_i(\delta)$ are Gaussian distributions characterized by mean vector $\mu_i$ and diagonal covariance matrix $\Sigma_i$, and mean $m_i$ and variance $\sigma_i^2$, respectively.

$$b_i(\alpha) = N(\alpha; \mu_i, \Sigma_i)$$

$$p_i(\delta) = N(\delta; m_i, \sigma_i^2).$$

In the HSMM-based MLLR, mean vectors of state output and duration distributions are obtained by linearly transforming mean vector of state output and duration distributions of the initial model,

$$b_i(\alpha) = N(\alpha; W_i \xi, \Sigma_i) = N(\alpha; \zeta_i + \xi, \Sigma_i)$$

$$p_i(\delta) = N(\delta; X_i \phi_i, \sigma_i^2) = N(\delta; \chi_i m_i + \nu_i, \sigma_i^2)$$

where $W_i = [\zeta_i, \xi]$ and $X_i = [\chi_i, \nu_i]$ are $n \times (n+1)$ and $1 \times 2$ transformation matrices for state output and duration distributions, respectively, and $\zeta$ and $\epsilon$ are $n \times n$ matrix and $n$-dimensional vector, respectively, and $\xi_i = [\mu_i^T, \Sigma_i^T]$ and $\phi_i = [m_i, 1]^T$ are $(n+1)$-dimensional and 2-dimensional vectors.

The problem of the HSMM-based MLLR can be written as follows:

$$\lambda = \arg \max_{\lambda} P(O|\lambda, \Lambda)$$

where $W = \{W_i\}_{i=1}^n, X = \{X_i\}_{i=1}^n$ and $\Lambda = (W, X)$. The HSMM-based MLLR estimates these matrices so as to

† Currently with Corporate Research & Development Center, Toshiba Corporation.
maximize the likelihood of the adaptation data. The re-
estimation formulas based on EM algorithm of the transforma-
tion matrices $\Lambda$ are given by

$$\begin{align*}
\pi_i^t &= G_i^{-1} y_i^t, \\
X_z^t &= \left( \sum_{l,r,d} \frac{\gamma^d_{l}(r)}{\sigma^2} \Phi^r \right) \left( \sum_{l,r,d} \frac{\gamma^d_{l}(r)}{\sigma^2} \Phi^r \Phi^r\top \right)^{-1}
\end{align*}$$

where $(n+1) \times (n+1)$ matrix $G_i$ and $n \times (n+1)$ matrix $Y$ are given by

$$\begin{align*}
G_i &= \sum_{l,r,d} \frac{\gamma^d_{l}(r)}{\sigma^2}\epsilon_{l}\epsilon_{l}\top, \\
Y &= \sum_{l,r,d} \gamma^d_{l}(r) \Sigma^{-1}_r \sum_{s=t-d+1}^t o_s \epsilon_{r}\top.
\end{align*}$$

$w_l$ and $y_t$ are the $l$-th row vectors of $W_z$ and $Y$ respectively, and $\Sigma_r(l)$ is the $l$-th diagonal element of $\Sigma_r$. Note that $W_z$ and $X_z$ are tied across $R$ distributions, and $\gamma^d_{l}(i)$ is state occupancy probability for HSMM [6].

### 3. Experiments

#### 3.1. Speech Database and Experimental Conditions

In the following experiments, we used four styles of read speech — neutral, joyful, sad, and rough styles. Speech database [1] contains a set of phonetically balanced 503 sentences taken from the ATR Japanese speech database. All the sentences were uttered by a male speaker MMI in all the styles. In the modeling of synthesis units, we used 42 phonemes including silence and pause and took the phonetic and linguistic contexts [1] 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 [9]. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency (F0), and their delta and delta-delta coefficients.

We used 5-state left-to-right HMMs/HSMMs and trained the style-dependent model [1] using 450 sentences for each style. A decision-tree-based context clustering was applied using the MDL criterion.

**Table 1:** Classification results of synthesized speech using style-dependent models.

<table>
<thead>
<tr>
<th>Classification (%)</th>
<th>Neutral</th>
<th>Joyful</th>
<th>Sad</th>
<th>Rough</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>96.2</td>
<td>1.2</td>
<td>1.9</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Joyful</td>
<td>3.8</td>
<td>95.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Sad</td>
<td>0.6</td>
<td>0.0</td>
<td>99.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Rough</td>
<td>6.9</td>
<td>0.6</td>
<td>0.6</td>
<td>91.2</td>
<td>0.6</td>
</tr>
<tr>
<td>(b) HSMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>93.1</td>
<td>0.0</td>
<td>2.5</td>
<td>1.2</td>
<td>3.1</td>
</tr>
<tr>
<td>Joyful</td>
<td>1.9</td>
<td>97.5</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Sad</td>
<td>0.6</td>
<td>0.0</td>
<td>99.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Rough</td>
<td>2.5</td>
<td>0.6</td>
<td>0.6</td>
<td>96.9</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Figure 1:** Preference scores of naturalness of synthesized speech using style-dependent models.

speech samples synthesized using HMM and HSMM in random order and then asked which speech sounded more natural. For each subject, five test sentences were chosen at random from 53 test sentences.

Figure 1 shows the preference scores. A confidence interval of 95% is also shown in the figure. It can be seen that naturalness of the synthesized speech from the style-dependent models using HSMM were comparable to or slightly better than the models using HMM. This result is consistent with that of [4], where only neutral reading style case was examined.

#### 3.2. Subjective Evaluation of Style Modeling

We first conducted a classification test and a paired comparison test for styles of speech generated from the style-dependent models using HMM and HSMM.

In the classification test, ten subjects were asked to classify eight test sentences chosen at random from 53 test sentences not included in the training data as being neutral, joyful, sad, or rough, depending on the style of speech. Speech samples that were not assigned by the subjects to one of these groups were classified as “other”.

Table 1 shows the classification results for synthesized speech. In the table, (a) shows the result for the style-dependent models using HMM and (b) shows that for the models using HSMM. It can be seen from the results that the style-dependent models using HSMM had almost the same reproduction performance as using HMM.

In the paired comparison test, we compared the naturalness of the synthesized speech. Subjects were presented a pair of

Figure 1: Preference scores of naturalness of synthesized speech using style-dependent models.

3.3. Subjective Evaluation of Style Adaptation

We next conducted the classification test and paired comparison test for generated speech from the models using style adaptation by HMM-based MLLR [3] and the proposed HSMM-based MLLR techniques. We chose the neutral reading style model obtained in 3.2 as an initial model. We then adapted it to that of one target style chosen from joyful, sad, and rough styles. The adaptation data was a 50-sentence set of the target style taken from training sentences. Since adaptation performance would be much affected by the choice of adaptation data, in particular, when the amount of adaptation data is limited, we used two sets of adaptation data. The ATR database sentences consists of ten subsets — subset A, B, ..., and J — and we chose subsets A
Table 2: Classification results of synthesized speech using style adaptation with subset A.

(a) HMM-based MLLR

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Joyful</th>
<th>Sad</th>
<th>Rough</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyful</td>
<td>7.5</td>
<td>85.0</td>
<td>0.0</td>
<td>5.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Sad</td>
<td>25.6</td>
<td>0.6</td>
<td>70.6</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>Rough</td>
<td>8.1</td>
<td>2.5</td>
<td>0.6</td>
<td>83.1</td>
<td>5.6</td>
</tr>
</tbody>
</table>

(b) HSMM-based MLLR

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Joyful</th>
<th>Sad</th>
<th>Rough</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyful</td>
<td>5.0</td>
<td>92.5</td>
<td>0.0</td>
<td>1.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Sad</td>
<td>27.5</td>
<td>0.6</td>
<td>71.2</td>
<td>0.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Rough</td>
<td>8.1</td>
<td>1.9</td>
<td>0.6</td>
<td>89.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 3: Classification results of synthesized speech using style adaptation with subset I.

(a) HMM-based MLLR

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Joyful</th>
<th>Sad</th>
<th>Rough</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyful</td>
<td>2.5</td>
<td>91.9</td>
<td>0.0</td>
<td>3.1</td>
<td>2.5</td>
</tr>
<tr>
<td>Sad</td>
<td>0.0</td>
<td>0.0</td>
<td>88.1</td>
<td>0.0</td>
<td>11.9</td>
</tr>
<tr>
<td>Rough</td>
<td>1.9</td>
<td>0.0</td>
<td>0.0</td>
<td>95.6</td>
<td>2.5</td>
</tr>
</tbody>
</table>

(b) HSMM-based MLLR

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Joyful</th>
<th>Sad</th>
<th>Rough</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joyful</td>
<td>5.6</td>
<td>91.9</td>
<td>0.0</td>
<td>1.9</td>
<td>0.6</td>
</tr>
<tr>
<td>Sad</td>
<td>0.0</td>
<td>0.0</td>
<td>98.8</td>
<td>0.6</td>
<td>0.6</td>
</tr>
<tr>
<td>Rough</td>
<td>5.6</td>
<td>0.6</td>
<td>0.0</td>
<td>91.9</td>
<td>1.9</td>
</tr>
</tbody>
</table>

and I from these subsets.

In the MLLR adaptation, we utilized context clustering decision trees for tying of regression matrices in which suprasegmental features are taken into account as well as frame-based features [2]. In the context clustering decision tree, those nodes having state occupancy count below a given threshold are placed in the same regression class as that of their parent node. The thresholds were set to 1000 for the spectral part, 150 for the F0 part, and 200 for state duration distributions, respectively.

3.4. Comparison of HMM-based MLLR and HSMM-based MLLR

Tables 2 and 3 show the classification rates for synthesized speech using adapted models with the adaptation data subsets A and I, respectively. The subjects were the same as those in 3.2. In each table, (a) shows the result for synthesized speech using style adaptation with the conventional HMM-based MLLR and (b) shows that with the proposed HSMM-based MLLR. It can be seen that, in both cases, most speech samples generated from the adapted model were classified into the target style.

Then we compared the naturalness of the synthesized speech generated by the HMM-based and HSMM-based systems. Figure 2 shows the preference scores. We can see that the naturalness of the speech samples from the adapted models by using HSMM-based MLLR were significantly better at the 95% confidence level than using HMM-based MLLR in all cases. We have observed that the synthesized speech by using the HMM-based MLLR sometimes gives unnatural prosody in duration.

The most prominent difference appeared in the case of adaptation to the sad style with the subset I. In this case, we observed that the duration of preceding and succeeding phonemes of a geminated consonant became extraordinarily long when using HMM-based MLLR.

3.5. Discussion

One of reasons for generating unnatural phoneme duration is that the conventional HMM-based MLLR adaptation estimates the transformation matrices for state duration distributions approximately. When phoneme durations of the target style are much different from the initial model, it provides inappropriate transformations, and, as a result, decreases naturalness of not only duration but also spectrum and fundamental frequency. On the other hand, the HSMM-based MLLR can perform the simultaneous adaptation of output distribution and state duration rigorously, and it maintains naturalness of the speech under all adaptation data.

Figure 3 shows the distributions of phoneme duration of all preceding vowels of the geminated consonant contained in the test sentences used for the above subjective tests. In the figure, (a) shows the histogram for the initial neutral model using HSMM, (b) for the target sad style-dependent model using HSMM, (c) for the adapted model to the sad style from the neutral model using HMM-based MLLR, and (d) for the adapted model to the sad style from the neutral model using HSMM-based MLLR, respectively. The values of the mean and the standard deviation are also shown in each histogram. Note that one frame corresponds to 5ms in duration. From this figures, we can clearly see that the duration for the HMM-based MLLR adapted model becomes different with those of the style-dependent and HSMM-based adapted models. Similarly, Fig. 4 shows the distributions of phoneme duration of the same vowels for joyful and rough styles. It is seen again that HSMM-based MLLR provides closer duration distributions to the style dependent model than HMM-based MLLR.
4. Conclusion

We have described performance evaluation of a style adaptation technique for hidden semi-Markov model based speech synthesis. The HSMM-based adaptation technique can perform the simultaneous linear transform of output distribution and state duration. From the results of subjective tests, we have shown that the HSMM-based MLLR technique can improve the naturalness of synthesized speech compared with conventional HMM-based MLLR technique. Future work will focus on investigation using other emotional expressions and speaking styles.

5. References