Transform Mapping Using Shared Decision Tree Context Clustering for HMM-Based Cross-Lingual Speech Synthesis

Daiki Nagahama¹, Takashi Nose², Tomoki Koriyama¹, Takao Kobayashi¹

¹Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
²Graduate School of Engineering, Tohoku University, Japan

nagahama.d.aa@m.titech.ac.jp, tnose@m.tohoku.ac.jp, koriyama@ip.titech.ac.jp, takao.kobayashi@ip.titech.ac.jp

Abstract

This paper proposes a novel transform mapping technique based on shared decision tree context clustering (STC) for HMM-based cross-lingual speech synthesis. In the conventional cross-lingual speaker adaptation based on state mapping, the adaptation performance is not always satisfactory when there are mismatches of languages and speakers between the average voice models of input and output languages. In the proposed technique, we alleviate the effect of the mismatches on the transform mapping by introducing a language-independent decision tree constructed by STC, and represent the average voice models using language-independent and dependent tree structures. We also use a bilingual speech corpus for keeping speaker characteristics between the average voice models of different languages. The experimental results show that the proposed technique decreases both spectral and prosodic distortions between original and generated parameter trajectories and significantly improves the naturalness of synthetic speech while keeping the speaker similarity compared to the state mapping.

Index Terms: HMM-based speech synthesis, cross-lingual TTS, shared decision tree context clustering, bilingual speech corpus

1. Introduction

Cross-lingual speech synthesis is a technique for synthesizing non-native language (output language) speech of a target speaker from an input text using only the speech data of his/her native language (input language). A typical application that has been studied recently is an automatic speech translation system [1–3]. Several techniques using a parametric approach based on HMMs have been proposed for the cross-lingual speech synthesis [4–10]. Lattore et al. proposed polyglot speech synthesis [4] where a language-mixed average voice model is trained using speech data of multiple languages and speakers and is adapted to a target speaker’s data of native language. This polyglot approach has shown the flexibility of the HMM-based speech synthesis, whereas the adaptation performance is not always sufficient because some leaf nodes of the constructed decision tree include only a certain language speech data. Recently, Zen et al. proposed a speaker and language factorization technique to alleviate the problem [10].

There is an alternative approach that uses state mapping based on Kullback-Leibler divergence (KLD) between the acoustic models of the input and output languages [5–7]. Although there are several manners of using the state mapping, in this study, we focus on the transform mapping [7] that has an advantage in terms of the naturalness of the synthetic speech. In this technique, the transformation matrices are estimated using an intra-lingual speaker adaptation process, where the average voice model of the input language is adapted to the target speaker’s data of the same language. Then, the transformation matrices are assigned to the states of the average voice model of the output language, and the cross-lingual speaker adaptation is performed. Although this technique gives better performance than that based on phone mapping, it has been pointed out that language and speaker mismatches degrade the adaptation performance [8, 9].

In this paper, we propose a novel technique of cross-lingual speech synthesis based on shared decision tree context clustering (STC) [11] to alleviate the mismatch problems of languages and speakers. In this technique, the transform mapping is performed by a language-independent decision tree constructed using STC. Unlike the state mapping based on KLD, STC can take both acoustic and contextual information into account in the mapping construction. Under the language-independent decision tree, language-dependent trees are constructed using language-specific questions. This approach uses the property of the decision tree that the language dependency is low near the root node and high near the leaf nodes generally. We also introduce a bilingual speech corpus to train the average voice models having the same speaker characteristics between input and output languages. We examine the performance of the proposed technique through objective and subjective evaluation tests.

2. Cross-lingual speech synthesis based on speaker adaptation

One of the typical approaches to HMM-based cross-lingual speech synthesis is to use cross-lingual speaker adaptation from an average voice model in the model training. In this section, we briefly overview a conventional HMM-based cross-lingual speaker adaptation with state mapping [7], which is used as a baseline system. The outline of the system is shown in Fig. 1. First, we train average voice models of input and output languages. In most of the previous studies, different speaker sets of the two languages were used. From the average voice model and target speaker’s training data of the input language, transformation matrices are estimated using an intra-lingual speaker adaptation framework with linear transformation such as maximum likelihood linear regression (MLLR) [12]. The obtained matrices can be expected to have a function of speaker characteristics conversion from the average voice to the target speaker. Then, the matrices are applied to the average voice model of the output language. When two average voice models are trained separately, there are no explicit mapping rules of the matrices from the input language to the output language. In [7], mapping rules are obtained between the leaf node states of the decision.
trees of the two average voice models based on KLD. By using the obtained state mapping, transformation matrices are applied to the average voice model of the output language, which can be viewed as cross-lingual speaker adaptation.

Although it was shown that this approach outperformed the phone-mapping-based approach, there still remain some problems in this approach. There are two crucial mismatches between the two average voice models in the model training. The first is a mismatch of language characteristics and the second is a mismatch of speaker characteristics. We will describe the problems in detail in Sect. 3.

3. Cross-lingual speaker adaptation using STC with a bilingual corpus

In the state mapping technique described in the previous section, the mismatch of language characteristics affects the mapping performance of transformation matrices because only the acoustic features are taken into account in the KLD-based mapping construction. To improve the mapping performance, we use not only acoustic features but also contextual factors when constructing the transform mapping. By using contextual factors, we can also take articulation manners and suprasegmental features into account for the mapping construction. In this section, we propose a novel transform mapping technique based on shared decision tree context clustering (STC) for cross-lingual speaker adaptation, which can reduce the influences of the speaker and language mismatches between average voice models of input and output languages.

3.1. Shared decision tree context clustering (STC)

STC [11] was originally proposed to avoid generating speaker-biased leaf nodes in the tree construction of an average voice model. In the conventional decision-tree-based context clustering for the average voice model, each leaf node does not always have the training data of all speakers, and some leaf nodes have only a few speakers’ training data. The experimental results have shown that such speaker-biased leaf nodes degrade the naturalness of the speech synthesized from the adapted model. On the other hand, in STC, we only use the questions which can be applied to all speakers. As a result, every node of the decision tree has the training data of all speakers, which leads to a speaker-unbiased average voice model.

3.2. Transform mapping based on language-independent decision tree using STC

To use contextual information in the transform mapping between different languages, we must consider the language dependency of decision trees. In general, near the root node of the decision trees, there are language-independent properties between the two languages in terms of basic articulation manners such as vowel, consonant, and voiced/unvoiced sound. On the other hand, near the leaf nodes, there frequently appear language-dependent properties because some nodes are split using language-specific questions, e.g., “Is the current phoneme diphthong?” To alleviate the language mismatch in the transform mapping between the average voice models, we generate a transform mapping based on a language-independent decision tree constructed by STC. Specifically, we use both average voice models of input and output languages in the context clustering, and the transformation matrices for the two average voice models are explicitly mapped to each other in the leaf nodes of the language-independent decision tree. Constructing the tree, we split nodes from the root using only the questions that can be applied to all speakers of both languages. In this study, we control the tree size by introducing a weight into stopping criterion based on the minimum description length (MDL) [13]. To avoid the effect of the language dependency, a smaller tree is constructed compared with that based on MDL. Since the node splitting is based on the acoustic parameters of each node, the transform mapping is conducted using both the acoustic and contextual information, which is more desirable than the conventional state mapping based on KLD. An appropriate size of the tree is experimentally examined in Sect. 4.3.

3.3. Cross-lingual speaker adaptation based on two-stage context clustering

As described in Sect. 3.2, the transform mapping using acoustic and contextual information is achieved using STC for cross-lingual speaker adaptation. However, the constructed tree is unsuitable for practical purpose of tying model parameters of the average voice models because the tree structure is language-independent and language-specific acoustic characteristics is not modeled. In this context, we employ two-stage context clustering where language-dependent trees are constructed under the language-independent tree structure using language-specific questions such as unique phonemes of both languages. It is noted that we also use intra-lingual STC when constructing the language-dependent trees to reduce the speaker bias in the leaf nodes.

Figure 2 illustrates the two-stage context clustering and transform mapping. First, the language-independent decision
Table 1: Different conditions of speaker sets for the average voice models of input and output languages. AVM2 is fully bilingual speaker sets.

<table>
<thead>
<tr>
<th>AVM1</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>EJF02, EJF04</td>
<td>EJF02, EJF06</td>
<td>EJF02, EJF04</td>
</tr>
</tbody>
</table>

Figure 3: Effect of a bilingual speech corpus in speaker similarity.

Tree is constructed using model parameters of both languages. The node splitting threshold \( d(i) \) at the \( i \)-th state is given by [11]

\[
d(i) = cK \sum_{l=1}^{2} \sum_{n=1}^{N} \log \Gamma_{lni} \tag{1}
\]

where \( c \) is a weight for controlling the size of the language-independent tree. \( K \) is the dimensionality of the observation vector, \( N \) is the number of speakers, and \( \Gamma_{lni} \) represents the state occupancy count of speaker \( n \) in language \( L_l \) at state \( i \). The resultant trees of input and output languages have identical structure. In Fig. 2, three transformation matrices, \( W_1 - W_3 \), are mapped from input to output languages in the leaf nodes of the language-independent trees.

Next, language-dependent trees are constructed under the language-independent tree. The node splitting threshold \( d_l(i) \) of language \( L_l \) is given by

\[
d_l(i) = c'K \sum_{n=1}^{N} \log \Gamma_{lni} \tag{2}
\]

where \( c' \) is a weight for controlling the size of language-dependent trees.

3.4. Training average voice models using a bilingual speech corpus

By introducing the language-independent tree based on STC into the transform mapping, the language mismatches are expected to be reduced. However, when using a different speaker sets for the training of the average voice models of input and output languages, it is inevitable that speaker characteristics of the two models are acoustically different. Even when transformation matrices are mapped from the input language to the output language using the language-independent tree, there is no guarantee of converting speaker characteristics to those of the target speaker in the cross-lingual speaker adaptation. To relax this problem, we use a bilingual speech corpus where the training utterances of the input and output languages are spoken by the same speakers.

4. Experiments

4.1. Experimental conditions

We used bilingual speech data included in [14] and set English and Japanese as the input and output languages, respectively. The training speakers of the average voice models were chosen from three females (EJF02, EJF04, and EJF06) and the training data of the average voice models consisted of 20 minutes utterances per speaker and language. The target speaker was a female speaker (EJF101), and the speaker adaptation was conducted using 20 minutes speech data. The test data were 50 sentences of the target speaker, and these sentences were not included neither in the training nor adaptation data. Speech signals were sampled at a rate of 16 kHz and the frame shift was 5 ms. We used STRAIGHT analysis [15] for acoustic feature extraction and extracted spectral envelop, log F0, and aperiodicity features. The feature vector consisted of 40 mel-cepstral coefficients including the zeroth coefficient, log F0, five-band aperiodicity values, and their delta and delta-delta coefficients.

The total dimensionality was 138. We used five-state no-skip left-to-right HMM with diagonal covariance matrices. According to the conclusion of the state mapping technique [7], the adaptation of duration models did not make much sense in current cross-lingual speaker adaptation framework, and hence we only adapted the parameters of spectral, F0 and aperiodicity models. The constrained structural MAPLR [16] was used for the model adaptation. In this study, we used the same \( c' \) for both languages for simplicity and Eq. (2) for the stopping criterion of node splitting in the state mapping technique.

4.2. Effectiveness of a bilingual speech corpus

First, we conducted subjective evaluation to examine the effectiveness of using a bilingual speech corpus in the model training of average voice models for the cross-lingual speaker adaptation. We evaluated the performance of the cross-lingual speaker adaptation using two different speaker sets as shown in Table 1. AVM1 used the different speaker sets for the training of the average voice model of input and output languages, whereas AVM2 used the same speaker sets. The amount of training data of each speaker was twenty minutes. In this experiment, we fixed the weight \( c \) and \( c' \) to 4.0 and 1.0 for constructing language-independent and dependent trees, respectively in an ad-hoc manner. The appropriate values of these weights are examined in Sect. 4.3. Seven sentences were randomly chosen from the 50 test sentences for each participant. For each sentence, eight participants listened to a pair of reference and synthetic samples in this order and rated the speaker similarity of them in a five point scale, i.e., 1: very different, 2: different, 3: slightly similar, 4: similar, 5: very similar. As the reference, synthetic speech of the output language of the target speaker was generated using intra-lingual speaker adaptation with the same amount of adaptation data in the cross-lingual case. The result is shown in Fig. 3 with 95% confidence intervals. It is apparent that the average voice model with bilingual corpus gave better performance than that without bilingual corpus in speaker similarity.

4.3. Objective evaluation

In the tree-based clustering of model training, it is known that the size of the decision tree affects the quality of the synthetic speech. As described in Sect. 3.3, the proposed technique has two parameters of \( c \) and \( c' \) for controlling the sizes of the language-independent and dependent trees. In this experiment, we examined the appropriate sizes of the language-independent and dependent trees using objective measures. We used speech data of all three females for training the average voice models. The total amount of data was about one hour for each language.

Figures 4 and 5 show the mel-cepstral distances and RMS errors of log F0 between generated and original parameter sequences for different values of \( c \) and \( c' \). In the figures, the re-
Table 4. Mel-cepstral distance [dB] and Log F0 RMS error [cent] when changing the sizes of language-independent and dependent trees by the weight parameters $c$ and $c'$, respectively.

<table>
<thead>
<tr>
<th>$c$</th>
<th>Mel-cepstral distance [dB]</th>
<th>Log F0 RMS error [cent]</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5.35</td>
<td>0.2</td>
</tr>
<tr>
<td>10</td>
<td>5.40</td>
<td>0.3</td>
</tr>
<tr>
<td>20</td>
<td>5.45</td>
<td>0.4</td>
</tr>
<tr>
<td>40</td>
<td>5.50</td>
<td>0.5</td>
</tr>
<tr>
<td>80</td>
<td>5.55</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Figure 4. Mel-cepstral distance when changing the sizes of language-independent and dependent trees by the weight parameters $c$ and $c'$, respectively.

Figure 5. RMS error of log F0 when changing the sizes of language-independent and dependent trees by the weight parameters $c$ and $c'$, respectively.

Results of state mapping are also shown. It is noted that the same bilingual training data was used for training the average voice models of the state mapping and the proposed technique. For the mel-cepstrum, $c = 20$, $c' = 0.4$ gave the best performance in the proposed technique, which was better than the state mapping. For the log F0, $c = 640$, $c' = 0.7$ gave the best performance, and the appropriate size of the language-independent tree was much smaller than that for the mel-cepstrum. This result is acceptable because the prosodic variation between different languages are often much larger than the phonetic variation. It is also seen that the effectiveness of the proposed technique compared to the state mapping is more prominent in log F0 than in mel-cepstrum.

4.4 Subjective evaluation

Finally, we compared the performance of the state mapping and the proposed technique in terms of the speaker similarity and speech naturalness using Differential Mean Opinion Score (DMOS) and Mean Opinion Score (MOS) tests, respectively. In both tests, the number of participants was six, and fifteen sentences were chosen randomly from the 50 test sentences for each participant. For the state mapping and the proposed technique, we used the weights $c$ and $c'$ that gave the best performance in the objective evaluation of Sect. 4.3. The numbers of leaf nodes of mel-cepstrum and log F0 in the state mapping are 2376 and 2226, and those in the proposed technique are 1089 and 2381, respectively. We used a parameter generation algorithm considering global variance [17]. First we conducted a DMOS test on speaker similarity. Participants listened to reference and synthetic speech samples in the order and rated the speaker similarity of the two samples in the same five point scale described in Sect. 4.2. Figure 6 shows the result with confidence intervals of 95%. As can be seen from the figure, the speaker similarity of the synthetic speech of the proposed technique is comparable to that of the state mapping. Next, we conducted a MOS test on speech naturalness. Participants listened to synthetic speech samples and rated the speech naturalness in a five point scale, i.e., 1: bad, 2: poor, 3: fair, 4: good, 5: excellent. Figure 7 shows the result with confidence intervals of 95%. It can be shown that there is a statistically significant difference at a 1% significance level and the proposed technique improved the naturalness of the synthetic speech. This is because of the use of the contextual information, especially suprasegmental features, for generating transform mapping, and the result is consistent with that of objective evaluation shown in Fig. 5.

5. Conclusions

In this paper, we proposed an HMM-based cross-lingual speech synthesis technique using transform mapping based on shared decision tree context clustering (STC). In the proposed technique, we introduced a language-independent decision tree constructed by STC and a bilingual speech corpus, and represented the average voice models using language-independent and dependent tree structures. The experimental results have shown that the proposed technique decreased both mel-cepstral distances and RMS errors of log F0 between original and generated parameter sequences and significantly improved the naturalness of synthetic speech while keeping the speaker similarity compared to the state mapping. In future work, we will investigate the effect of different speakers and languages.

6. Acknowledgments

Part of this work was supported by JSPS Grant-in-Aid for Scientific Research 23700195 and 24300071.
7. References


