State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis

Yi-Jian Wu, Yoshihiko Nankaku, Keiichi Tokuda

Nagoya Institute of Technology, Japan

yjwu@sp.nitech.ac.jp, nankaku@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

A phone mapping-based method had been introduced for cross-lingual speaker adaptation in HMM-based speech synthesis. In this paper, we continue to propose a state mapping based method for cross-lingual speaker adaptation, where the state mapping between voice models in source and target languages is established under minimum Kullback-Leibler divergence (KLD) criterion. We introduce two approaches to use the established mapping information for cross-lingual speaker adaptation, including data mapping and transform mapping approaches. From the experimental results, the state mapping based method outperformed the phone mapping based method. In addition, the data mapping approach achieved better speaker similarity, and the transform mapping approach achieved better speech quality after cross-lingual speaker adaptation.

Index Terms: Speech synthesis, HMM, speaker adaptation, minimum generation error, linear regression

1. Introduction

Spoken language translation (SLT) systems have been under development for many years. In a recently started European FP7 project – Effective Multilingual Interaction in Mobile Environments (EMIME) [1] – we are developing methods to personalize such SLT systems. In particular, the synthesized speech in the target language should sound like the input speaker, even though that speaker can not speak the target language. This problem has been previously explored in the TC-Star project using cross-lingual voice conversion techniques [2].

The HMM-based speech synthesis [3, 4] was adopted in our framework. One of the unique capabilities of this method is the ability to change the characteristics of the synthesized speech by modifying the HMM parameters using model adaptation technique. In this study, we investigate a cross-lingual speaker adaptation technique for HMM-based speech synthesis, where a source voice model for a source language (English) is transformed into a speaker-specific model using adaption data from the target speaker in a target language (Japanese). The adapted model can be used to synthesize English, with the speaker characteristics of the target speaker. To realize such cross-lingual speaker adaption, a phone mapping based method [5] had been previously introduced. However, the phone mapping is not accurate enough to characterize the acoustic similarities of the phonetic units in the source and target language, and such an inaccurate mapping may reduce the adaptation performance.

In order to alleviate the issue of inaccurate phone mapping, we propose a state mapping based method for cross-lingual speaker adaptation. In this method, we first train two Average Voice models in source and target languages, respectively, and then establish the state mapping between these two models under a minimum Kullback-Leibler divergence (KLD) criterion. We introduce two approaches to use the established mapping information for cross-lingual speaker adaptation, including data mapping and transform mapping approaches. In the data mapping approach, the mapping information is used to attach the adaptation data in the target language (Japanese) to the source voice model in the source langauge (English), the usual speaker adaptation for the English voice model is conducted by regarding the Japanese adaptation data as English adaptation data. This procedure is similar to the phone mapping based method. In the transform mapping approach, we first conduct the speaker adaptation for the Japanese voice model using Japanese adaptation data, and get the transforms. Then the state mapping information is used to attach the transforms of the Japanese voice model to the English voice model. Finally, we apply these transforms to the English voice model, and get the adapted model. This approach has a similar concept to the cross-lingual speaker adaptation method proposed in [6];

The rest of this paper is organized as follows. In section 2, we first briefly review the concept of cross-lingual speaker adaptation and a phone mapping based method. In section 3, we present the details of the state mapping based method for cross-lingual speaker adaptation, including data mapping and transform mapping approaches. In section 4, we describe the experiments used to evaluate the performance of the proposed state mapping based method and present the results. Finally, our conclusions are given in section 5.

2. Cross-lingual speaker adaptation

2.1. From intra-lingual to cross-lingual speaker adaptation

Intra-lingual speaker adaptation (usually just called "speaker adaptation"), transforms a source model to a target speaker using a limited amount of speech data from the target speaker. Initially developed for use in HMM-based speech recognition, many model adaptation algorithms, including MAP, MLLR, CMLLR, etc., have been proposed [7]. In HMM-based speech synthesis, speaker adaptation techniques are used to adapt the source model using speech data from target speaker, and thus make the speech synthesized from the adapted model sound like the target speaker. Several adaptation algorithms have been borrowed from speech recognition and further developed for HMM-based speech synthesis [8]. It has been demonstrated that speaker adaptation of an "Average Voice" model [9] is superior to speaker adaptation of a speaker-dependent model.

In cross-lingual speaker adaptation, a source voice model for a source language is transformed into a speaker-specific model using adaption data from the target speaker in a target language. The adapted model are still used to synthesize the source language, but with the speaker characteristics of the target speaker. Note that only the speech data in the target language are required for the target speaker.

2.2. Phone mapping based method

A phone mapping based method [5] had been previously introduced for cross-lingual (English–Chinese) speaker adaptation. In this method, firstly the context labels are mapped from the target language into the source language, i.e., the Chinese labels are mapped into English labels. Then the mapped Chinese adaptation data is regarded as English adaptation data, and the model adaption technique is applied in a similar way to intra-lingual speaker adaptation.

Since "full context" labels used in HMM-based speech synthesis, include the phonetic and prosodic information, both phonetic and prosodic mapping between source and target languages are needed. For the phonetic mapping, two mapping rules between Chinese Initials/Finals and English phonemes are manually designed by considering the phonetic definition of these units in the IPA and their acoustic realizations. However, it is extremely hard to design a prosodic feature mapping between different languages. In order to avoid using a prosodic feature mapping, an ingenious adaptation procedure [10] was adopted, in which the regression classes and transform matrices are built for triphone models, and then applied to full context models. More details of phone mapping based cross-lingual speaker adaptation can be found in [5].

3. State mapping based method

The phone mapping is not accurate enough to characterize the acoustic similarities of the phonetic units in the source and target languages. In order to alleviate this issue, we propose a state mapping based method for cross-lingual speaker adaptation. In this method, we firstly train two Average Voice models in source and target languages, respectively, and then establish the state mapping between these two models. Finally, we conduct cross-lingual speaker adaptation based on the established mapping information.

3.1. KLD-based state mapping

We adopt the Kullback-Leibler divergence (KLD) for establishing the state mapping in this study, which is similar to that used for bi-lingual speech synthesis [11]. Let us denote Ω_k^s $(k = 1, \ldots, N^s)$ as the state models in the model space of source language, where N^s is the total number of state models. Since the single Gaussian mixture is used here, the parameters of each state model Ω_k^s include a self-transition probability a_k^s, a mean vector $\boldsymbol{\mu}_k^s$ and a covariance matrix $\boldsymbol{\Sigma}_k^s$. Similarly, we denote Ω_j^g $(j = 1, \ldots, N^g)$ as the state models in the model space of target language, and the corresponding self-transition probability, mean vector and covariance matrix are a_j^g, $\boldsymbol{\mu}_j^g$ and $\boldsymbol{\Sigma}_j^g$, respectively.

For each state model Ω_j^g in the target language, we want to find a nearest state model Ω_k^s in the source language, which has the minimum KLD with Ω_j^g. In the case of single Gaussian mixture, the upper bound of KLD [12] between two state models is calculated as

$$D_{KL}(\Omega_j^g, \Omega_k^s) \leqslant \frac{D_{KL}\left(G_k^s || G_j^g\right)}{1 - a_k^s} + \frac{D_{KL}\left(G_j^g || G_k^s\right)}{1 - a_j^g} + \frac{(a_k^s - a_j^g)\log(a_k^s / a_j^g)}{(1 - a_k^s)(1 - a_j^g)} \quad (1)$$

where G_k^s denote the Gaussian distribution related to the state model Ω_k^s, which includes the mean vector $\boldsymbol{\mu}_k^s$ and covariance matrix $\boldsymbol{\Sigma}_k^s$, and the KLD between two Gaussian distributions is calculated as

$$D_{KL}\left(G_k^s || G_j^g\right) = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_j^g|}{|\boldsymbol{\Sigma}_k^s|}\right) - \frac{D}{2} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_j^{g-1}\boldsymbol{\Sigma}_k^s\right) + \frac{1}{2}\left(\boldsymbol{\mu}_j^g - \boldsymbol{\mu}_k^s\right)^\top \boldsymbol{\Sigma}_j^{g-1}\left(\boldsymbol{\mu}_j^g - \boldsymbol{\mu}_k^s\right) \quad (2)$$

Since we only focus on the distribution of the state model, we ignore the effect of transition probabilities, and calculate the KLD between two state models as

$$D_{KL}(\Omega_k^s, \Omega_j^g) \approx D_{KL}\left(G_k^s || G_j^g\right) + D_{KL}\left(G_j^g || G_k^s\right) \quad (3)$$

Based on the above KLD measurement, the nearest state model $\Omega_{k'}^s$ in source language for each state model Ω_j^g in target language is calculated as

$$k'_j = \arg\min_k D_{KL}(\Omega_j^g, \Omega_k^s). \quad (4)$$

Finally, we map all state models in the target language to the state models in the source language, which can be formulated as

$$\Omega_j^g \Rightarrow \Omega_{k'_j}^s, \quad j = 1, \ldots, N^g. \quad (5)$$

Here we establish the state mapping from the model space of the target language to the model space of the source language. In this case, all the state models in the target language have a mapped state model in the source language. However, it should be noted that not all the state models in the source language have a corresponding state model in the target language.

3.2. Approaches to use mapping information

Here we propose two approaches to use the mapping information for speaker adaptation, including data mapping and transform mapping approaches.

3.2.1. Data mapping approach

In the data mapping approach, the state mapping information is used to attach the adaptation data in the target language to the voice model in the source language. The procedure of data mapping approach is as follows:

a) Train two Average Voice models in both source and target languages.
b) Establish state mappings between the source state model and the target state models, i.e. find a nearest state model in **source language** for each state model in **target language**.
c) Attach the adaptation data in the target language to the voice model in the source langauge based on the state mapping information.
d) Regard the adaptation data as the data in the source language and conduct the intra-lingual speaker adaptation for the voice model in source language.

Comparing to the previous phone mapping based method, this data mapping approach of state mapping based method is an extended method of previous one. The only difference between them is that we build the connection between the source voice model and the adaptation data in different levels, i.e., the phone level and state level, respectively. It should be noted that the mapping between the data and the voice model does not reflect the exact correspondence between them, i.e., it does not include the information about the affiliated state model for each frame

of the data. The mapping information only reflect the mapped state sequence related to each adaptation utterance. The exact affiliation between data and models are determined by the forward-backward algorithm in adaptation training.

3.2.2. Transform mapping approach

In the transform mapping approach, the state mapping information is used to attach the transforms of the voice model in target language to the voice model in source language. The procedure of transform mapping approach is as follows:

a) Train two Average Voice models in both source and target languages.
b) Establish state mappings between the source state model and the target state models, i.e. find a nearest state model in **target language** for each state model in **source language**.
c) Train the transforms for the Average Voice model in target language using the adaptation data.
d) Adapt each state model in source language using the transform attached to the mapped state model in target language.

In this approach, we do not build a direct connection between the voice model in the source language and the adaptation data in the target language. The established state mapping information is used to map the transforms in the target language to the source language, and then the state models in source language can be adapted using the mapped transforms.

The underline assumption of this transform mapping approach is that the model space in source language is similar to the the model space in target language. For example, if we train the voice models for both source and target languages from the bilingual speech database uttered by the same speakers, it basically satisfy this assumption. However, it is too difficult to get such a bilingual speech database uttered by the same speakers in practical. Although we can minimize the mismatch between the training data by using two speech database with the similar number of speakers and similar distribution of genders for source and target languages, there are still differences between these two voice model spaces.

3.3. Discussion

In cross-lingual speaker adaptation, both speaker characteristic and language identity of adaptation data are different from the source voice model. We want to keep the language identity of source voice model while adapting the speaker characteristic to the target speaker, which means we need to avoid the influence of the language identity of adaptation data in cross-lingual speaker adaptation.

The above data mapping and transform mapping approaches for cross-lingual speaker adaptation have their own advantage and disadvantage. For the data mapping approach,

a) Advantage: Since we directly use the adaptation data for speaker adaptation, the speaker characteristic of the source voice model can be totally adapted to the target speaker.
b) Disadvantage: Although we regarded them as the data in the source language after mapping, the target language identity still exist in the adaptation data, which means the language identity of the source voice model will be adapted or partially adapted to the target language.

For the transform mapping approach,

a) Advantage: Since the transform are trained in the target language under an intra-lingual adaptation way, the influence of different language identity in adaptation data is avoided.
b) Disadvantage: If the voice models in source and target languages are not trained from the speech database uttered by the same bilingual speakers, there would be a mismatch of speaker characteristic between these two voice models. Such mismatch will make the speaker characteristic of synthesized speech after adaptation different with the target speaker.

As we mentioned in Sec. 3.2.1, the previous phone mapping based method are basically a phone-level data mapping approach. Therefore, this method inherit the same advantages and disadvantages of data mapping approach. However, this method has another disadvantage that the phone mapping between the phonetic sets of two languages is usually not accurate enough, especially for some language pair which has much phonetic differences, e.g. English-Chinese, which may reduce the adaptation performance. In addition to such a phone-level data mapping approach, we could have a corresponding phone-level transform mapping approach, where the transforms are mapped in phone level. However, the phone-level mapping structure may be incompatible with the state-level regression classes for the voice models. For example, two transforms may related to the same state of one phoneme.

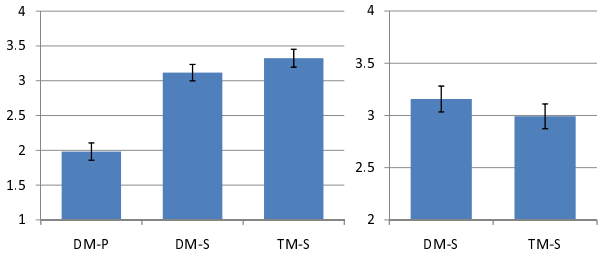
4. Experiments

4.1. Experimental conditions

In the experiments, the source language is English, and the target language is Japanese. For the English Average Voice model training, we adopted the speech data from the CMU-ARCTIC English database [13] – about 1 hour of speech data from each of 2 males (awb, bdl) and 2 females (clb, slt). For the Japanese Average Voice model training, we adopted NIT Japanese database, which includes about 1 hour of speech data from each of 2 males (mai, mat) and 2 females (fky, fss). The Japanese speech data including 50 utterances from a male (mhk) was used as the target adaptation data. All speech waveforms were sampled at a rate of 16KHz. The acoustic features, including F0 and mel-cepstral coefficients, were extracted with a 5ms shift. The feature vector consists of static features, including 25-th order mel-cepstral coefficients, log F0, their delta and delta-delta coefficients. A 5-state left-to-right no-skip HMM was used to model each English and Japanese phoneme, and MSD-HMMs [14] were used for F0 modeling. For synthesis, a Mel Log Spectrum Approximation (MLSA) filter [16] was used to generate the speech waveform.

From the results on phone mapping based method [], we already concluded that the adaptation of duration models did not make much sense in current cross-lingual speaker adaptation framework. In this experiment, we only adapted the parameters of spectral and F0 models. The standard regression trees and classes generated by HTK tools are used, and the CMLLR-based method was adopted for model adaptation. In the experiment, we investigated the performances of the following approaches for cross-lingual speaker adaptation:

a) DM-P: phone mapping based method, i.e. data mapping approach in phone level;
b) DM-S: data mapping approach in state level;
c) TM-S: transform mapping approach in state level;



(a) MOS (b) DMOS

Figure 1: MOS and DMOS scores of synthesized speech from the adapted models using different cross-lingual speaker adaptation approaches

4.2. Experimental results

In the experiment, two formal subjective listening tests were conducted. The first test evaluated the quality of synthesized speech using the MOS score, and the second one evaluated the speaker similarity between the target speech and the synthesized speech from adapted models using DMOS score. 40 English sentences, which were not included in the training data, were synthesized from the adapted models using different cross-lingual speaker adaptation approaches (DM-P, DM-S, TM-S). In the MOS test, each listener evaluated 15 sets of samples consisting of three synthesized speech samples, and gave the MOS scores for each sample. In order to remove the influence of the vocoder in the DMOS test, we used synthetic speech from a speaker-dependent (SD) model, rather than natural speech from the target speaker. Since no English speech data for the target Japanese speaker were available, we trained a SD model using 1-hour Japanese speech data from the target speaker. Therefore, Japanese utterances were compared to English utterances in the DMOS test. Each listener was presented with 15 pairs of synthesized speech samples (firstly one utterance from the speaker-dependent Japanese model and then one utterance from the adapted English models) and asked to give a DMOS score to each English speech sample. Eight listeners participated in the test, and the speech samples were randomly selected for each listener from the 40 test sentences.

The results are shown in Fig. 1, with the vertical line indicating the 95% confidence intervals. In this figure, it can be seen that the quality of synthesized speech using the proposed state mapping based adaptation method (including both data mapping and transform mapping approaches) is significantly better than that using the phone mapping based method. Furthermore, from the MOS and DMOS scores of data mapping and transform mapping approaches, the transform mapping approach achieved better quality of synthesized speech after adaptation, and data mapping approach has better speaker similarity. This result is coincident with the advantages and disadvantages of these two mapping approaches discussed in Sec 3.3. The problem of introducing target Japanese language identity to the synthesized English speech in data mapping approach will reduce the speech quality. In the transform mapping approach, the difference between the voice models in source and target languages reduce the similarity of the synthetic speech to the target speaker.

5. Conclusions

In this paper, we introduce a state mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. Two approaches to use the established mapping information for cross-lingual speaker adaptation are presented, including data mapping and transform mapping approaches. From the experimental results, the state mapping based method outperformed the phone mapping based method. In addition, the data mapping approach achieved better speaker similarity, and the transform mapping approach achieved better speech quality after cross-lingual speaker adaptation. Future work is to apply a model space normalization for transform mapping approach.

6. Acknowledgements

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

7. References

[1] EMIME project: http://www.emime.org

[2] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney and J. Hirschberg, "TC-Star: Cross-language voice conversion revisited," in *Proc. of the TC-Star Workshop 2006*, Spain, 2006.

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

[4] 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*, vol. 5, pp. 2347-2350, 1999.

[5] Y.-J. Wu, S. King and K. Tokuda, "Cross-Lingual Speaker Adaptation for HMM-based Speech Synthesis," in *Proc. of ISCSLP*, pp. 9-12, 2008.

[6] Y.-N. Chen, Y. Jiao, Y. Qian and F.K. Soong, "State mapping for cross-language speaker adaptation in TTS," in *Proc. of ICASSP*, 2009.

[7] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," in *Computer Speech and Language*, vol. 12, no. 2, pp. 75-98, 1998.

[8] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273-276, 1998.

[9] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," in *IEICE Trans. of Fundamentals*, vol. E86-A, no. 8, pp. 1956-1963, 2003.

[10] S. King, K. Tokuda, H. Zen and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. of Interspeech*, pp. 1869-1872, 2008.

[11] H. Liang, Y. Qian, F. Soong and G. Liu, "A cross-language state mapping approach to bilingual (Mandarin-English) TTS," in *Proc. of ICASSP*, pp. 4641-4644, 2008.

[12] P. Liu, F.K. Soong and J.-L. Zhou, "Divergence-based similarity measure for spoken document retrieval," in *Proc. of ICASSP*, pp. 89-92, 2007.

[13] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177, http://festvox.org/cmu_arctic/, 2003.

[14] 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*, pp. 229-232, 1999.

[15] http://hts.sp.nitech.ac.jp/

[16] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. of ICASSP*, pp. 93-96, 1983.

[17] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. of Interspeech*, pp. 2801-2804, 2005.