Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices

Xianglin Peng, Keiichiro Oura, Yoshihiko Nankaku, Keiichi Tokuda
Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan
{pengxl,uratec,nankaku}@sp.nitech.ac.jp, tokuda@nitech.ac.jp

***Abstract*—This paper proposes an improved cross-lingual speaker adaptation technique with considering the differences between language-dependent average voices in a Speech-to-Speech Translation system. A state mapping based method had been introduced for cross-lingual speaker adaptation in HMM-based speech synthesis. In this method, the transforms estimated from the input language are applied to average voice models of the output language according to the state mapping information. However, the differences between average voices in the input and output language may degrade the adaptation performance. To reduce the differences, we apply a global linear transform to output average voice models, which minimizes the symmetric Kullback-Leibler divergence between two average voice models. From the experimental results, our approach could not obtain a better result than the original state mapping based method. This is because the global transform affects not only speaker characteristics but also language identity in acoustic features, and this degrades the synthetic speech quality. Therefore, it becomes clear that a technique which separate speaker and language identities is required.**

***Index Terms*—HMM, speech synthesis, cross-lingual speaker adaptation, average voice**

I. Introduction

Researches on cross-lingual speaker adaptation [1-6] for Speech-to-Speech Translation (S2ST) system have been conducted to enable the output speech sounds like the target speaker (input speaker). To realize such a S2ST systems, the HMM-based speech synthesis technique [7-8] is suitable for cross-lingual speaker adaptation, because it provides flexible speaker adaptation, and a small amount of adaptation data is required. Figure 1 shows an overview of the S2ST system and cross-lingual speaker adaptation. A state mapping based method [1] had been proposed for cross-lingual speaker adaptation in the S2ST system using HMM-based speech synthesis. The system is based on Constrained Maximum Likelihood Linear Regression (CMLLR) [9-10], and the transforms estimated from the input language are applied to average voice models of the output language according to the state mapping information. However the difference between average voices of input and output language are not considered. Consequently, this may degrade adaptation performance.

In order to alleviate this issue, we propose an approach for cross-lingual speaker adaptation considering differences between language-dependent average voices. In this approach, a global linear transform from average voice models of the input and output language was estimated and applied to average voice model of the output language.

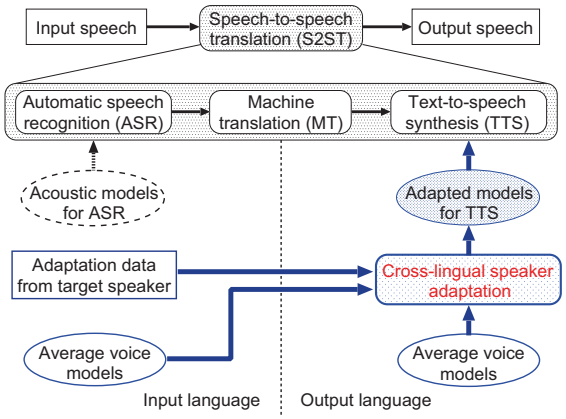

Fig. 1. Overview of the system.

This paper is organized as follows. In section II, the state mapping based method is briefly reviewed. The details of cross-lingual speaker adaptation considering differences between language-dependent average voices are presented in section III. The experiments were conducted to evaluate the performance of the proposed approach, and the experimental conditions and results are shown in section IV. Finally, conclusions and suggestions for future work are presented in section V.

II. State mapping based method

The basic idea of the state mapping based method [1-2] for cross-lingual speaker adaptation is shown in Fig. 2. First, two average voice models of both the input and output language are constructed. The average voice models are trained using multiple speaker's data and speaker adaptive training (SAT) [11] is applied. Then the state mapping between these two average voice models is established by finding the state in the input language which minimizes the symmetric Kullback-Leibler divergence (KLD) for each state in the output language.

The symmetric KLD between two states is calculated as follows

$$D_{\mathrm{KL}}\left(G_i^{(O)}, G_j^{(I)}\right) \approx D_{\mathrm{KL}}\left(G_j^{(I)} \| G_i^{(O)}\right) + D_{\mathrm{KL}}\left(G_i^{(O)} \| G_j^{(I)}\right), \quad (1)$$

978-1-4244-5899-8/10/$26.00 ©2010 IEEE

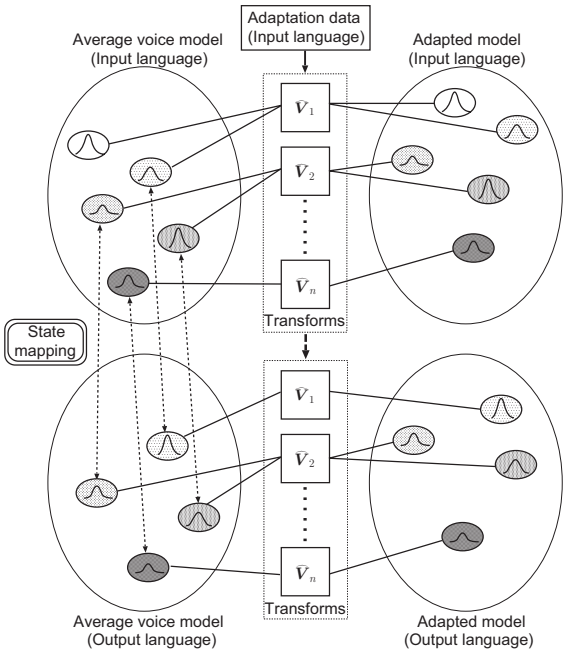

Fig. 2. Outline of state mapping based method.

$$D_{\mathrm{KL}}\left(G_j^{(I)} \| G_i^{(O)}\right) = \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Sigma}_i^{(O)}\right|}{\left|\boldsymbol{\Sigma}_j^{(I)}\right|}\right) - \frac{D}{2} + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{(O)^{-1}}\boldsymbol{\Sigma}_j^{(I)}\right) + \frac{1}{2}\left(\boldsymbol{\mu}_i^{(O)} - \boldsymbol{\mu}_j^{(I)}\right)^{\top}\boldsymbol{\Sigma}_i^{(O)^{-1}}\left(\boldsymbol{\mu}_i^{(O)} - \boldsymbol{\mu}_j^{(I)}\right), \quad (2)$$

where $G_j^{(I)}$ $\left(j = 1, \ldots, N^{(I)}\right)$ and $G_i^{(O)}$ $\left(i = 1, \ldots, N^{(O)}\right)$ denote the states in the input language models $\lambda^{(I)}$ and the output language models $\lambda^{(O)}$ respectively, $\boldsymbol{\mu}_j^{(I)}$ and $\boldsymbol{\mu}_i^{(O)}$ represent the mean vectors, and $\boldsymbol{\Sigma}_j^{(I)}$ and $\boldsymbol{\Sigma}_i^{(O)}$ represent the covariance matrices of the Gaussian pdf associated with state $G_j^{(I)}$ and $G_i^{(O)}$.

Based on the above KLD measurement, the nearest state $G_{f(i)}^{(I)}$ in the input language for each state $G_i^{(O)}$ in the output language is found as

$$f(i) = \arg\min_j D_{\mathrm{KL}}\left(G_i^{(O)}, G_j^{(I)}\right). \quad (3)$$

Next, the transforms for average voice models $\lambda^{(I)}$ of the input language is estimated using the adaptation data in the following way. A set of linear transforms $\widehat{\boldsymbol{\Lambda}}$ for the input language models $\lambda^{(I)}$ is calculated as

$$\begin{aligned}\widehat{\boldsymbol{\Lambda}} &= \left(\widehat{\boldsymbol{V}}_1, \ldots, \widehat{\boldsymbol{V}}_{N^{(I)}}\right) \\ &= \arg\max_{\boldsymbol{\Lambda}} P\left(\boldsymbol{O} \mid \lambda^{(I)}, \boldsymbol{\Lambda}\right) P(\boldsymbol{\Lambda}),\end{aligned} \quad (4)$$

where $\boldsymbol{V}_j$ denotes a linear transform for state $G_j^{(I)}$, and $\boldsymbol{O}$ denotes the adaptation data. $P(\boldsymbol{\Lambda})$ denotes the prior distribution of the linear transforms, which is a uniform distribution for MLLR [9] and CMLLR [10].

Finally, cross-lingual speaker adaptation is achieved by applying these transforms estimated in the average voice models of the input language to the average voice models of the output language according to the state mapping information.

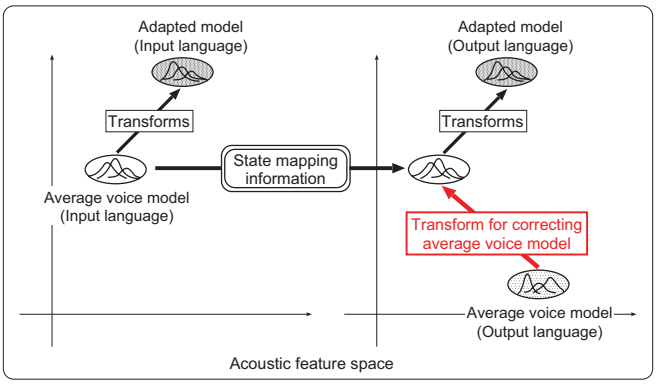

Fig. 3. Cross-lingual speaker adaptation using a transform for correcting average voice model.

III. Cross-lingual speaker adaptation considering differences between language-dependent average voices

The underlying assumption of the state mapping based method is that speaker characteristics of acoustic features in the output language appear similarly in the input language. Therefore, the speaker characteristics of average voice models in the input and output language should be the same prior to adaptation. If average voice models of both the input and output language are constructed from a bilingual speech database uttered by bilingual speakers, it satisfies this assumption. However, it is usually difficult to obtain such a database with a large amount fo data in practice. Consequently, the differences of speaker characteristics are included in two average voice models.

Since these differences between the input and output language average voices are not considered in the state mapping based method, this may degrade the adaptation performance. To reduce the differences between average voice models, we propose a new approach for cross-lingual speaker adaptation. In the proposed method, a global linear transform is applied to reduce the differences of speaker characteristic between two average voices as shown in Fig. 3.

The linear transform is applied to each mean vector $\boldsymbol{\mu}_i^{(O)}$ of the output average voice models $\lambda^{(O)}$ as

$$\begin{aligned}\widehat{\boldsymbol{\mu}}_i^{(O)} &= \boldsymbol{A}\boldsymbol{\mu}_i^{(O)} + b \\ &= \boldsymbol{W}\bar{\boldsymbol{\mu}}_i^{(O)},\end{aligned} \quad (5)$$

where $\widehat{\boldsymbol{\mu}}_i^{(O)}$ is the transformed mean vector, and

$$\bar{\boldsymbol{\mu}}_i^{(O)} = \left[\boldsymbol{\mu}_i^{(O)}\ 1\right]^{\top}, \quad (6)$$

$$\boldsymbol{W} = [\boldsymbol{A}\ b]. \quad (7)$$

The differences between two average voice models can be diminished by estimating the transform $\boldsymbol{W}$ appropriately. In the proposed method, the transform $\boldsymbol{W}$ is estimated based on KLD, which is similar to the state mapping based method. The symmetric KLD between two average voice models is

calculated as

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(\widehat{G}_i^{(O)}, G_{f(i)}^{(I)}\right) \\
&= \frac{1}{2}\left\{-2D + \mathrm{tr}\left(\boldsymbol{\Sigma}_{f(i)}^{(I)\ -1}\boldsymbol{\Sigma}_i^{(O)}\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_i^{(O)-1}\boldsymbol{\Sigma}_{f(i)}^{(I)}\right)\right\} \\
&\quad + \frac{1}{2}\left\{\left(\boldsymbol{W}\bar{\boldsymbol{\mu}}_i^{(O)} - \boldsymbol{\mu}_{f(i)}^{(I)}\right)^{\top}\left(\boldsymbol{\Sigma}_{f(i)}^{(I)\ -1} + \boldsymbol{\Sigma}_i^{(O)-1}\right)\left(\boldsymbol{W}\bar{\boldsymbol{\mu}}_i^{(O)} - \boldsymbol{\mu}_{f(i)}^{(I)}\right)\right\}, \quad (8)
\end{aligned}
$$

where $\widehat{G}_i^{(O)}$ represents transformed state $G_i^{(O)}$. The optimal transform can be given by

$$
\boldsymbol{W} = \arg\min_{\boldsymbol{W}} \sum_i D_{\mathrm{KL}}\left(\widehat{G}_i^{(O)}, G_{f(i)}^{(I)}\right). \quad (9)
$$

Here, we present $\boldsymbol{W}_d$ as the elemental vector of the transform $\boldsymbol{W}$ in the d-th row. The partial derivative of $\boldsymbol{W}_d$ can be derived as

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{W}_d}\sum_d\sum_i D_{\mathrm{KL}}\left(\widehat{\Omega}_i^{(O)}, \Omega_{f(i)}^{(I)}\right) \\
&\qquad = \sum_i \boldsymbol{Z}_{id}\boldsymbol{\mu}_{f(i)d}^{(I)}\bar{\boldsymbol{\mu}}_i^{(O)\top} - \sum_i \boldsymbol{Z}_{id}\boldsymbol{W}_d\bar{\boldsymbol{\mu}}_i^{(O)}\bar{\boldsymbol{\mu}}_i^{(O)\top}, \quad (10)
\end{aligned}
$$

where

$$
\boldsymbol{Z}_{id} = \boldsymbol{\Sigma}_{f(i)d}^{(I)\ -1} + \boldsymbol{\Sigma}_{id}^{(O)-1}. \quad (11)
$$

By setting Eq. (10) to 0, $\boldsymbol{W}_d$ can be estimated as

$$
\boldsymbol{W}_d = \left(\sum_i \boldsymbol{Z}_{id}\boldsymbol{\mu}_{f(i)d}^{(I)}\bar{\boldsymbol{\mu}}_i^{(O)\top}\right)\left(\sum_i \boldsymbol{Z}_{id}\bar{\boldsymbol{\mu}}_i^{(O)}\bar{\boldsymbol{\mu}}_i^{(O)\top}\right)^{-1}. \quad (12)
$$

The speaker characteristic in the output language can be modified to that in the input language by applying the global transform $\boldsymbol{W}$ to average voice models of the output language. Moreover, a new state mapping can be estimated after applying the transform $\boldsymbol{W}$ practically so that we can obtain a more accurate state mapping.

IV. Experiments

A. Experimental conditions

We performed experiments on cross-lingual speaker adaptation for HMM-based speech synthesis, in which the input and output languages are English and Japanese respectively. Although, unsupervised cross-lingual speaker adaptation [2-4] has been investigated in recent work, we used supervised cross-lingual speaker adaptation in our experiments. For text-to-speech (TTS), we adopted the WSJ0 database (15 hours of speech, 7.2k sentences uttered by 42 male and 42 female speakers) for English and JNAS database (19 hours of speech, 10k sentences uttered by 43 male and 43 female speakers) for Japanese as the training data. Speech signals were sampled at 16kHz and windowed by a 25ms Hamming window with a 5-ms shift, and 5-state left-to-right context-dependent multi-stream MSD-HSMMs were used. TTS feature vectors are comprised of 138-dimensions: 39-dimension STRAIGHT [12] mel-Cepstral coefficients, log F_0, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. One male and one female American English speaker were chosen from the "long term" subset of the WSJ0 database as target speakers. The adaptation data comprised 50 sentences were selected arbitrarily from the 2.3k sentences available for each target speaker.

B. Experimental results

In the experiments, we investigated the performance of the following approaches for cross-lingual speaker adaptation:

- No-adaptation: average voice models in the target language
- Baseline: speaker adaptation based on the state mapping without global transform
- Proposed-1: the global transform estimated from the initial state mapping
- Proposed-2: the global transform iteratively updated with the state mapping

Firstly, we performed an experiment with adaptation data of a male target speaker. We conducted a subjective listening test to evaluate the speaker similarity between the target speech and the synthesized speech using DMOS score at a 5 point psychometric response scale. The subjects were 12 Japanese native listeners. Each subject was presented with 12 sets of synthetic speech samples selected randomly from the 50 Japanese translated sentences: the first sample in each set was a reference English utterance of the target speaker and the others were synthetic Japanese speech utterances generated using four cross-lingual speaker adaptation approaches.

Figure 4 shows the sum of KLD in iterative updates of the global transform in the proposed method. It can be seen that the sum of KLD is significantly reduced in Proposed-1 compared with Baseline. It can also be confirmed that the sum of KLD monotonically decreases by iterative updates of the state mapping and global transform. Figure 5 shows the average DMOS and their 95% confidence intervals. From the result, we can see that Proposed-1 and Proposed-2 outperformed No-adaptation. However they underperformed Baseline, and the Proposed-2 has a tendency to degrade compared with Proposed-1. Listening to the speech samples, background noise in Proposed-1 was louder than that in Baseline. We guess this occurred because of the difference of recording environments between two speech databases.

To confirm the effectiveness of the proposed method, we performed an experiment on a special condition: a new Japanese average voice model was trained with only 43 male speaker's data in JNAS database, and a female speaker's data was used as the adaptation data. This is an extreme case that the difference of the average voice models was large, and the spectral features of adaptation data were also very different from the average voice models in the output language. Figure 6 shows the results of a DMOS test, and it can be seen that the proposed approach obtained a similar performance with Baseline and the degradation of Proposed-2 was not seen in this case.

The state mapping was obtained by minimizing the KLD in the proposed approach. However, the acoustical differences between the states may include not only the differences of the speaker characteristics but also the phonological differences between two languages. Therefore, the phonological differences were also contained in the transform which is expected

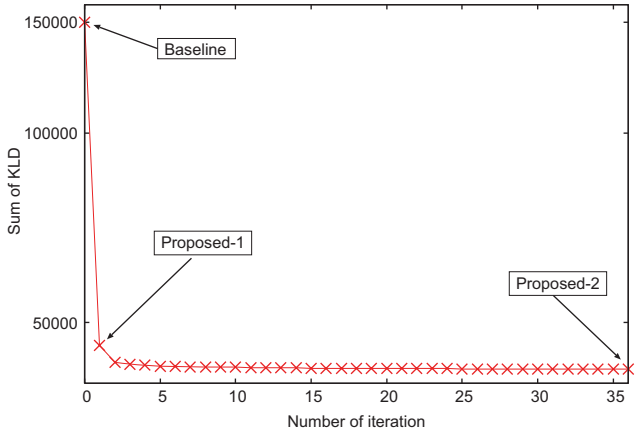

Fig. 4. The sum of KLD in iterative updates of global transform and state mapping.

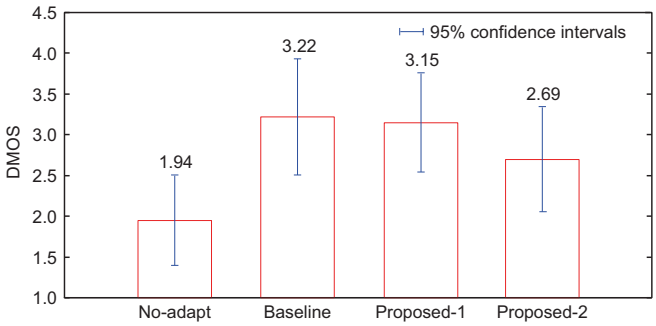

Fig. 5. DMOS of synthesized speech using Japanese average voice trained with 43 male and 43 female speaker's data.

to represent the differences in the speaker characteristic only. Consequently, we guess this degrades the phonological information of the average voice models in the output language. This coincides with the result of the second experiment. From the second experimental result, it can be considered that larger improvements in speaker characteristics was obtained by the proposed method than the first experiment, because the average voice models extremely differ from each other. However, while improving speaker characteristics, the phonological structure was also collapsed as Japanese language. This is the reason that the proposed method could not achieved a better result than the original state mapping based method. Accordingly, the phonological differences based on language need to be separated from the speaker characteristic in average voice models before estimating the transform.

V. Conclusions

In this paper, we proposed an approach to reduce the differences between language-dependent average voices by applying a global linear transform to average voice model in cross-lingual speaker adaptation. From the experimental results, we cannot validate the effectiveness of our approach compared with the original state mapping based method. This is because the transform affects not only speaker characteristics but also language identity in acoustic features. Therefore, if we can separate the language identity from the speaker characteristic in average voice models, the speaker characteristic can be

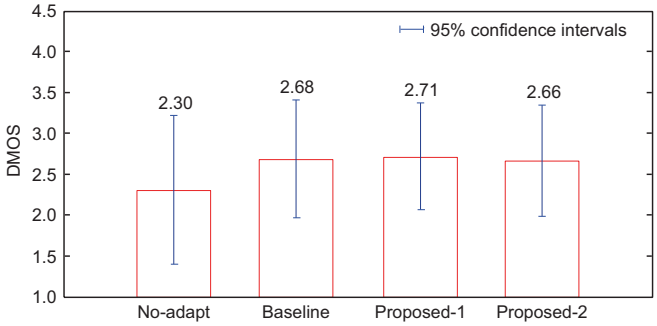

Fig. 6. DMOS of synthesized speech using Japanese average voice trained with only 43 male speaker's data.

adapted without influences of the phonological differences which are dependent on languages. Future work is to investigate a method to separate speaker and language identities. We will also introduce a cross-lingual speaker adaptation technique using bilingual speech databases.

Acknowledgments

This research was partly funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project), and the Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communication, Japan.

References

[1] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," Proc. Interspeech 2009, pp. 528–531, 2009.

[2] Y. Qian, H. Lang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin - English) TTS," IEEE TASLP, vol. 17, no. 6, pp. 1231–1239, 2009.

[3] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Unsupervised speaker adaptation for speech-to-speech translation system," IEICE Technical Report, vol. 109, no. 356, pp. 13–18, 2009.

[4] H. Liang, J. Dines, and L. Saheer,"A comparison of supervised and unsupervised cross-lingual speaker adaptation approaches for HMM-based speech synthesis," Proc. ICASSP 2010, pp. 4598–4601, 2010.

[5] M. Gibson, T. Hirsimaki, R. Karhila, M. Kurimo, and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," Proc. ICASSP 2010, pp. 4642–4645, 2010.

[6] Y. J. Wu, S. King, and K. Tokuda, "Cross-Lingual speaker adaptation for HMM-based speech synthesis," Proc. ISCSLP, pp. 9–12, 2008.

[7] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis from HMMs using dynamic features," Proc. ICASSP, pp. 389–392, 1996.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. Eurospeech, pp. 2347–2350, 1999.

[9] 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.

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

[11] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE TASLP, vol. 17, no. 6, pp. 1208–1230, 2009.

[12] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sound," Speech Communication, vol. 27, pp. 187–207, 1999.