Cross-lingual speaker adaptation for HMM-based speech synthesis using joint-eigenvoices with a perceptual characteristic space

学籍番号 2 3 4 1 3 5 8 2 氏名 De Franca Oliveira Viviane
指導教員名 徳田 恵一

1 Introduction

Research on cross-lingual speaker adaptation (CLSA) [1, 2] has been performed to enable the output of a Speech-to-Speech Translation system to sound like the input speaker (target speaker). To realize such a system, a state mapping (SM)-based method has been proposed [2]. However, a bilingual database is required to sufficiently represent the relation between source and target languages. To improve the performance of CLSA systems without a bilingual database, we proposed a new approach, where a language-independent space of perceptual characteristics (PCs) is used as an intermediate space in the mapping between languages. During this work, two methods were proposed: The first one models the three spaces spanned by source and target languages and the PCs separately and uses speaker interpolation in the target language to generate data for the target speaker. On the other hand, the second method provides the modeling of all the three spaces into a unique framework, using the maximum likelihood estimation to update all the model parameters.

2 CLSA based on perceptual characteristics and speaker interpolation

Some characteristics of the voice that may be perceived by the human ear are independent of the spoken language, being mainly influenced by inherent characteristics of the speakers, such as gender and age. Therefore, they can be used for establishing the relation between the two languages in a CLSA system [3]. In this approach, a set of these characteristics was used to construct an intermediate space for the mapping between source and target languages (PC space). In [3], each speaker is represented by a super-vector, which is a data structure that contains all the mean vectors from the output probability distributions of a speaker dependent model. In this method, whenever a new target speaker enters the source language speaker space, it is firstly projected to the PC space and, once an appropriate representation for this speaker is found in that space, it is projected to the target language. Finally, speaker interpolation is performed in the target language to reconstruct the super-vector of the target speaker. An overview of the method is depicted in Fig. 1.

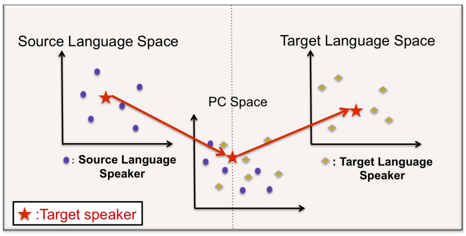

Fig. 1: Overview of the CLSA system using a space of perceptual characteristics and speaker interpolation

3 Joint-eigenvoices for CLSA

The joint-eigenvoice method is based on factor analysis (FA) [4]. To adopt the FA framework in the proposed method, we slightly modified the classical FA model to be able to represent consecutive data, considering that parts of the input data matrix are unobserved. A graphical representation of this input data is depicted in Fig. 2. In the Figure, $\boldsymbol{D}$ represents the matrix of observation data, where each column is a data vector, equivalent to $\boldsymbol{o}$. Moreover, $\boldsymbol{D}^{(S)}$ and $\boldsymbol{D}^{(T)}$ are matrices, where each column represents a data vector from a training speaker, which consists of a super-vector followed by a score representation in the PC space. The shaded areas represent the latent (hidden) data of the model.

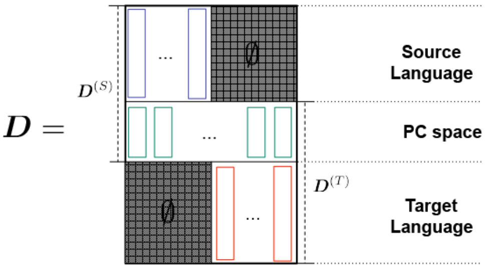

Fig. 2: Graphical representation of the input data for the joint-eigenvoice proposed framework

The modeling of source and target languages are done separately and the data vector of each speaker ($\boldsymbol{D}_i^{(S)}$ and $\boldsymbol{D}_k^{(T)}$) can be expressed as:

$$\begin{aligned}\boldsymbol{D}_i^{(S)} &= \boldsymbol{W}^{(S)}\boldsymbol{a}_i^{(S)} + \boldsymbol{\mu}^{(S)}\\ \boldsymbol{D}_k^{(T)} &= \boldsymbol{W}^{(T)}\boldsymbol{a}_k^{(T)} + \boldsymbol{\mu}^{(T)},\end{aligned} \tag{1}$$

where i, k are indices that represent each training speaker in the source and target language sets, respectively, $\boldsymbol{W}^{(S)}$, $\boldsymbol{W}^{(T)}$ are the factor loading matrices and $\boldsymbol{a}_i^{(S)}$, $\boldsymbol{a}_k^{(T)}$ are the factors. Moreover, $\boldsymbol{\mu}^{(S)}$ and $\boldsymbol{\mu}^{(T)}$ represent mean vectors, which are distributed by $\mathcal{N}(0, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{a}_i}^{(S)})$ and $\mathcal{N}(0, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{a}_k}^{(T)})$, respectively. Because a part of the components of $\boldsymbol{W}^{(S)}$ and $\boldsymbol{W}^{(T)}$ (i.e. the components that represent the PC space), are shared, these matrices are estimated simultaneously. According to this model, the output probabilities of the data vectors can be expressed as:

$$\begin{aligned}P(\boldsymbol{D}_i^{(S)}) &= \int P(\boldsymbol{D}_i^{(S)}|\boldsymbol{a}_i^{(S)})P(\boldsymbol{a}_i^{(S)})d\boldsymbol{a}_i^{(S)}\\ &= \mathcal{N}(\boldsymbol{\mu}^{(S)}, \boldsymbol{W}^{(S)}\boldsymbol{W}^{(S)T} + \Sigma^{(S)}),\\ P(\boldsymbol{D}_k^{(T)}) &= \int P(\boldsymbol{D}_k^{(T)}|\boldsymbol{a}_k^{(T)})P(\boldsymbol{a}_k^{(T)})d\boldsymbol{a}_k^{(T)}\\ &= \mathcal{N}(\boldsymbol{\mu}^{(T)}, \boldsymbol{W}^{(T)}\boldsymbol{W}^{(T)T} + \Sigma^{(T)}).\end{aligned} \tag{2}$$

Unlike the interpolation-based method [3], this method provides the modeling of all the three spaces into a unique scheme, where the parameters can be updated using the maximum likelihood (ML) estimation. Moreover, the eigenvoices can be properly estimated from all the input data.

For the synthesis, whenever a new speaker enters the source language space, a set of factors is firstly estimated for this target speaker and then used directly for the estimation of the super-vector in the target language, as follows:

$$\begin{aligned}\hat{\boldsymbol{D}}_{T,i}^{(S)} &= \arg\max_{\boldsymbol{D}_{T,i}^{(S)}} P(\boldsymbol{D}_{T,i}^{(S)}|\boldsymbol{a}_i^{(S)},\Lambda)\\ &= \boldsymbol{W}_T\hat{\boldsymbol{\mu}}_{\boldsymbol{a}_i}^{(S)} + \boldsymbol{\mu}_T,\end{aligned} \qquad (3)$$

where $\boldsymbol{D}_{T,i}^{(S)}$ is the super-vector in the target language space which corresponds to a given target speaker from the source language, $\boldsymbol{W}_T$ and $\boldsymbol{\mu}_T$ represent the factor loading matrix and mean vector, respectively, for the target language without the PC space values, and $\hat{\boldsymbol{\mu}}_{\boldsymbol{a}_i}^{(S)}$ is the value of $\boldsymbol{a}_i^{(S)}$ that maximises $P(\boldsymbol{a}_i^{(S)}|\boldsymbol{D}_{S,i}^{(S)},\Lambda)$.

4 Experiments

4.1 Subjective listening evaluations

To evaluate the performance of the proposed approach, CLSA experiments were conducted using the WSJ0 database (English) and the JNAS database (Japanese). The procedure used to construct the PC space and the experimental conditions for preparing the models were the same as described in [3].

We conducted formal listening tests (MOS and DMOS) to measure the naturalness and speaker similarity of speech generated by the interpolation-based method. Moreover, an informal subjective test was carried out to evaluate the performance of the joint-eigenvoice method. We confirmed that the speaker characteristics generated by the proposed approach in both methods were similar to those of the target speakers. We also plan on conducting formal listening tests to evaluate the joint-eigenvoice method in the near future.

In the MOS and DMOS tests, the following methods were compared: No-adaptation (average voice model), SM (State Mapping), PC-1 (SD model of the closest speaker in the PC space) and PC-2 (interpolation of the two closest speakers in the PC space). The results are depicted in Figs. 3 and 4. From the MOS test results, we verify that the proposed approach improves the naturalness of synthesized speech when compared to the SM method. On the other hand, little improvement was verified for speaker similarity in the DMOS test, which can be attributed to the sparsity of speakers in the training sets, once the number of speakers needed to cover the PC space is unknown. Nevertheless, the proposed method performs better than No-adaptation for speaker similarity and, considering the advantage obtained in the MOS test, we conclude that the proposed framework is efficient in implementing the CLSA system.

4.2 Joint-eigenvoice method simulation

We also performed a simulation experiment to evaluate the parameter estimation performance of the joint-eigenvoice method. For the simulation, we used a matrix of randomly generated values as input data and calculated the mean square error (MSE) between the estimated super-vector for the synthesis and the original generated super-vector. Fig. 5 shows the variation of this quantity over the iterations of the re-estimation algorithm. From this, we conclude that the method can efficiently estimate the super-vector in the target language.

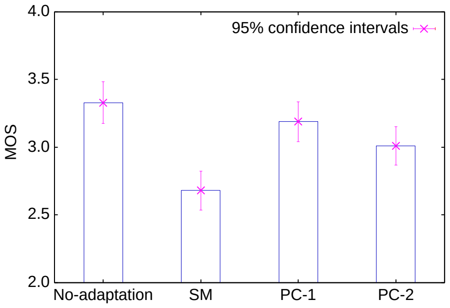

Fig. 3: Results from the MOS subjective listening test for naturalness

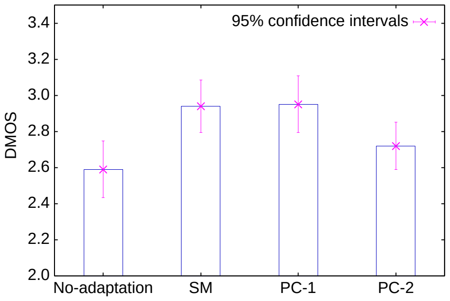

Fig. 4: Results from the DMOS subjective listening test for speaker similarity

5 Conclusion

In this work, we proposed a novel approach for CLSA, where a space of language-independent perceptual characteristics of speech was used to establish the relation between source and target languages. Two methods were proposed: The first one used three separate spaces for the modeling and performed speaker interpolation in the target language, while the second one provided unified modeling of the three spaces spanned by source and target languages and the PCs, using the ML estimation to update the model parameters. The results of the listening tests suggest that the proposed framework produces speech with voice characteristics similar to the target speaker, and obtains better speech quality than the state mapping approach. As future work, we plan on constructing the PC space with a greater variety of speakers and conducting more formal listening tests.

References

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

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

[3] V. F. Oliveira, S. Shiota, Y. Nankaku, K. Tokuda, "Cross-lingual Speaker Adaptation for HMM-based Speech Synthesis based on Perceptual Characteristics and Speaker Interpolation," in *Proc. Interspeech 2012*, 2012.

[4] C. Bishop, "Pattern Recognition and Machine Learning," Springer, 2006.

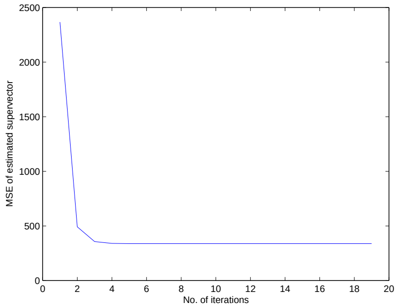

Fig. 5: MSE between the super-vector estimated by the joint-eigenvoice method and the original generated data vector.