STATE MAPPING FOR CROSS-LANGUAGE SPEAKER ADAPTATION IN TTS

Yi-Ning CHEN[1], Yang Jiao[1, 2+], Yao Qian[1], and Frank K. Soong[1]

[1]Microsoft Research Asia, Beijing, China
[2]Beijing University of Posts and Telecommunications, Beijing, China
[1]{ynchen, yaoqian, frankkps}@microsoft.com, [2]jiao.yang.bupt@gmail.com

ABSTRACT

Cross-language speaker adaptation has many interesting applications, e.g. speech-to-speech translation. However, in cross-language speaker adaptation, a common phoneme set, assumed to be used by different speakers of the same language, does not exist any longer. Instead, a nearest neighbor based phoneme mapping from one language to the other has been adopted. In this study, we used our recently proposed sub-phonemic HMM state mapping for cross-language adaptations. The sub-phonemic HMM states, due to their phonetic segment nature, tend to be more sharable across different languages than phonemes. Kullback-Leibler divergence, an information-theoretic measure, is chosen here to measure the similarity between given states in different languages. Experimental results show that new state mapping outperforms the phoneme mapping baseline system in terms of three objective measures: log spectral distance, F0 adaptation error and F0 correlations. In comparing with intra-language adaptation, the cross-language result of the new algorithm is also fairly decent.

Index Terms— HMM-based TTS, Speaker adaptation, Cross language, Kullback-Leibler divergence.

1. INTRODUCTION

Speaker adaptation is useful for creating customized voice fonts with limited adaptation data. In the speech-to-speech translation, the source speaker's speech is first converted (recognized) into a text sequence in its source language, translated into the target language, finally synthesized in the target language with a voice similar to the source speaker. In most cases, the source speaker cannot speak the target language. This is where cross-language speaker adaptation is needed. In a standard flowchart of speech-to-speech translation, the voice conversion is adopted as shown in the left hand side of Fig. 1 [1]. In the figure, *V* refers to voice and *L* refers to language; *T* refers to target and *S* refers to source. For example, V_TL_S denotes the voice of the target speaker with the source language.

However, the voice conversion part can be replaced by speaker adaptation. The new flowchart is shown in the right hand side of the figure. The standard voice conversion method does not consider the text of speech, which is available in speech-to-speech translation. Hence, adopting speaker adaptation in this scenario can be beneficial. If speech synthesis is carried out in a Hidden Markov Model (HMM) [2], the adaptation process is more straightforward [3].

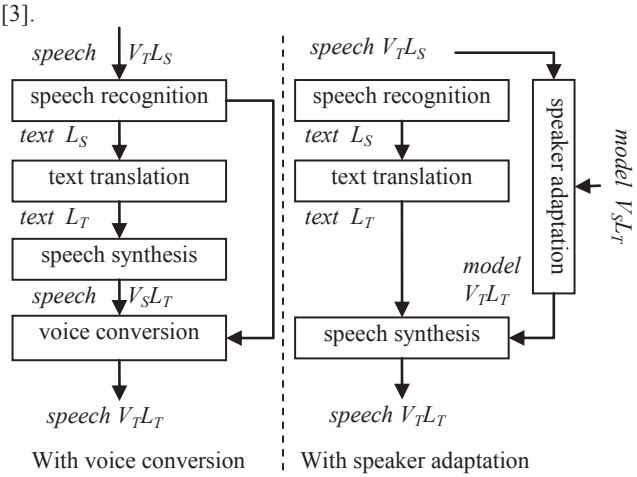

Figure 1. Flowchart of a speech-to-speech translation. (The voice conversion part is in [1])

There are many research results related to this topic; e.g. voice conversion with a Gaussian Mixture Model (GMM) mapping [4], multilingual speech recognition [5-7] and synthesis [8-10].

In multilingual speech recognition and synthesis, we need to create a mapping between two languages. Here the mapping can be a word, syllable, phoneme, or other. Then two questions must be answered; 1) what basic units for mapping? 2) How to measure the similarity between two given units?

Phonemes are most widely used [5, 6, 9] as the basic units for mapping. However, if phoneme sets of the two languages are significantly different, like Chinese and English, the mapping will be unsatisfactory [8]. Allophones, e.g. right and left context-dependent tri-phone, is another choice and it tends to be more flexible [7]. In this paper, we select HMM states as the basic unit, a unit even shorter than phonemes. In HMM, a state represents a distinctive, acoustic-phonetic event. Even for languages of fairly different phoneme set are different, the acoustic-phonetic events are still close enough [9]. In this paper, we use the state mapping for our cross language adaptation task.

For the similarity measure, the International Phonetic Alphabet (IPA) is used to measure the similarity in the terms of place of articulation and manner of articulation [11]. The confusion matrix is adopted in the speech recognition task [5, 7]. However, a more direct way is to measure the acoustic distribution between two

[+] The work is done when the second author worked in MSRA as an intern.

states. The Kullback-Leibler divergence (KLD) is a well established measure for that [12], it is used a lot in speech synthesis [13-15]. In this paper, we also use KLD as our similarity measure.

The structure of this paper is as follows. In Section 2, we introduce the method of state mapping. In Section 3, the method of phone mapping is introduced as a reference method. Section 4 gives a detailed introduction about KLD and its forms in HMM-based speech synthesis. Section 5 shows the results and conclusions are drawn in Section 6.

2. STATE MAPPING

In HMM-based speech synthesis, states are the smallest component. After the distance is defined, for each state in one language, we can find the corresponding one in the other language by minimizing the distance:

$$\hat{S}^X = \arg\min_{S^X} D\left(S^X, S_j^Y\right) \tag{1}$$

where, S_j^Y is a state in language Y. D is the distance between two states.

Details of the state mapping can be found in Fig. 2. We want to create a state mapping from Speaker *A*, language *X* to Speaker *C*, Language *Y*. To make it simple, we will use V_AL_X to represent Speaker *A*, Language *X*. We will not directly create the state mapping with Equation (1) since both the speaker and language are different. We introduce an auxiliary speaker who can speak in both Language *X* and Language *Y*. We will call it Speaker B. With this new speaker, the mapping between V_AL_X and V_CL_Y is changed to three parts, the mapping between V_AL_X and V_BL_X, the mapping between V_BL_X and V_BL_Y, and the mapping between V_BL_Y and V_CL_Y. These three mappings are shown in Fig. 2 as three bold dark arrows. In them, the mapping between V_BL_X and V_BL_Y are speaker irrelevant, the other two are language irrelevant.

Since speaker *A* and speaker *B* are in the same language, we use the same decision tree for them. There is no guarantee that all the leaf nodes in speaker *B* can be seen in speaker A. Hence, the decision tree of V_AL_X will be a subset of V_BL_X. Then, for one state in V_BL_X, there will be one corresponding state in V_AL_X with the same context.

For each state in V_BL_Y, we can find the closest state in V_BL_X with (1). Then the mapping between V_BL_X and V_BL_Y is created.

For the mapping between V_BL_Y and V_CL_Y, we can assume they have the same decision tree structure like any another intra-language adaptation.

With the steps above, a state mapping between V_CL_Y and V_AL_X is created. The adaptation can be implemented as the standard intra-language adaptation.

This auxiliary speaker can be a bilingual speaker. Or in a more general case, it can be an average voice model [16] of two languages. In both cases, the two models V_BL_Y and V_BL_X are speaker irrelevant. In this paper, since we only have 1-2 speakers for each language, we create the mapping with a bilingual speaker.

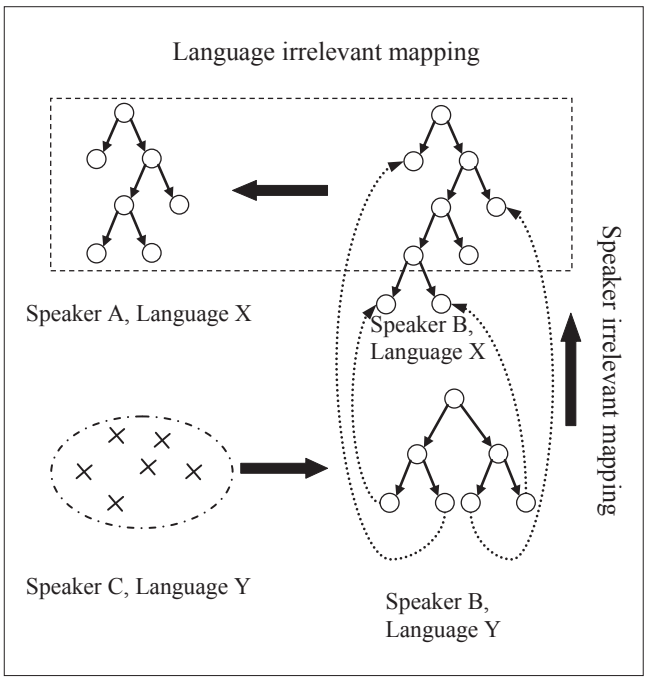

Figure 2. State mapping creating.

3. PHONE MAPPING

Phone mapping between different languages is used as a reference algorithm. In this paper, we will use the similar method as in Section 2 to create the phone mapping. The formula of finding the state mapping is:

$$\hat{H}^X = \arg\min_{H^X} D\left(H^X, H_j^Y\right) \tag{1}$$

where, H^X is a phoneme in language *X*, H_j^Y is a phoneme in language *Y*, and *D* is the distance between two phonemes. In this paper, we use the context independent model to present each phoneme.

Similar with Section 2, an auxiliary speaker is introduced to create the mapping. After the mapping is created, the script of one language can be converted to the script of another language. Then, the cross-language adaptation becomes intra-language adaptation.

4. KULLBACK-LEIBLER DIVERGENCE

The distance of two probability distributions is quite useful and has been studied for a long time. The name of the Kullback-Leibler divergence is first discussed by Hastie [17]. Although Kullback himself preferred other names [18], the KLD is a standard form today. In addition, we must point out that this distance was first introduced by Jeffreys in 1946 [19].

The formula or KLD between two distributions *p* and *q* can be defined as,

$$D_{KL}\left(p \,\|\, q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{3}$$

In this paper, we will use the notation of Equation (3) for the asymmetry version of KLD, and we will use the notation of Equation (4) for the symmetry version of KLD. Hence, we have,

$$J\left(p,q\right) = D_{KL}\left(p \,\|\, q\right) + D_{KL}\left(q \,\|\, p\right) \tag{4}$$

4.1. Multi-space probability distribution

Multi-space probability distribution (MSD) is proposed for F0 pattern modeling by Tokuda [2]. In MSD, the whole sample space Ω can be divided by G subspaces with index g.

$$\Omega = \bigcup_{g=1}^{G} \Omega_g \tag{5}$$

Each space Ω_g has its probability ω_g, where

$$\sum_{g=1}^{G} \omega_g = 1 \tag{6}$$

Hence, the probability density function of MSD can be written as,

$$p(x) = \sum_{g=1}^{G} p_{\Omega_g}(x) = \sum_{g=1}^{G} \omega_g M_g(x) \tag{7}$$

In it,

$$\int_{\Omega_g} M_g(x)\,dx = 1 \tag{8}$$

From (7), (6), and (8), it looks very similar to the multiple mixtures; however, they are not the same. In the mixture condition, distributions of components are overlapped. However, in MSD, if all the subspaces are different, the distributions of them are not overlapped. Hence, in MSD, we will have,

$$M_g(x) \equiv 0 \quad \forall x \notin \Omega_g \tag{9}$$

This property will help us in calculating the distance between two distributions. In the next subsection, we will introduce one of these distances, the Kullback-Leibler Divergence.

4.2. Kullback-Leibler divergence for multi-space probability distribution

Putting (7) and (9) into (3), the KLD of MSD can be found using Equation (10).

$$\begin{aligned} D_{KL}(p \| q) &= \int_{\Omega} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \\ &= \sum_{g=1}^{G} \left\{ \int_{\Omega_g} \omega_g^p M_g^p(x) \log\left(\frac{\omega_g^p M_g^p(x)}{\omega_g^q M_g^q(x)}\right) dx \right\} \\ &= \sum_{g=1}^{G} \left\{ \omega_g^p \log\left(\frac{\omega_g^p}{\omega_g^q}\right) + \omega_g^p \int_{\Omega_g} M_g^p(x) \log\left(\frac{M_g^p(x)}{M_g^q(x)}\right) dx \right\} \\ &= \sum_{g=1}^{G} \left\{ \omega_g^p D_{KL}\left(M_g^p \| M_g^q\right) \right\} + \sum_{g=1}^{G} \left\{ \omega_g^p \log\left(\frac{\omega_g^p}{\omega_g^q}\right) \right\} \end{aligned} \tag{10}$$

From this equation, we can see that the KLD of MSD has two terms; one is the weighted sum of KLD of each subspace; the other is the KLD of the weight distribution.

4.3. Kullback-Leibler divergence for HMM

Given two HMMs, their KLD is defined as,

$$D_{KL}(p \| q) = \int p(\mathbf{o}^{1:t}) \log \frac{p(\mathbf{o}^{1:t})}{q(\mathbf{o}^{1:t})} d\mathbf{o}^{1:t} \tag{11}$$

where $\mathbf{o}^{1:t}$ is the observation sequence runs from time 1 to *t*.

Only an upper bound of KLD can be calculated with close form. When the state numbers of two HMMs are not the same, the computation is more complex.

Since phone mapping is only a reference method here, we will not discuss details in this paper. A detailed version can be found in Do [20] and Liu's [21] work for readers' interests.

5. EXPERIMENTS AND RESULTS

5.1. Target of experiment

This paper focuses on cross-language adaptation. Hence, our goal is to make the cross-language adaptation as good as intra-language adaptation. In this paper, we use a bilingual (English, Mandarin) speaker to do the test. We adapt the source model separately for the speaker's English and Mandarin sentences. Then we will compare the sentences synthesized by these two adapted models. In this paper, we will measure the objective measures between the two versions. Details of the experiment setting can be found in the next subsection.

5.2. Experiment settings

The corpuses of two speakers are used in this experiment. An English male speaker, Tom, is used as the source speaker, and a bilingual female speaker, ZT, as the target speaker. If we could have found another bilingual speaker, we would have used that person as the auxiliary speaker. In this paper, we use ZT as the auxiliary speaker for an oracle experiment. The number of sentences used for training, mapping, adapting, and testing, are listed in Table 1.

Table 1. Sentence numbers for each speaker.

	English	Mandarin
Tom	10000	NA
ZT (Mapping)	1000	1000
ZT (Adaptation)	50	50
Testing	1000	1000

The system is built with HTS2.1 [22]. The feature is a 40-dimension Linear Prediction Cepstrum Coefficient (LPCC). Global variance is not deployed. MLLR is adopted for adaptation.

5.3. Objective measures

In this paper, we use three objective measures. They are the log-spectrum distance (S_D), the root mean square error of F0 (d_{f0}), and the correlation coefficient of F0.

In this paper, the duration model of the target speaker is used. Then, the sentences synthesized by the adaptation model and which synthesized by the model of target speaker are compared, and the three objective measures are calculated.

5.4. Results and discussion

5.4.1. Phone mapping and state mapping

The first experiment is to compare speaker adaptation performance difference between phone mapping and state mapping. The results are shown in Table 2.

Table 2. Phone mapping VS. state mapping.

	S_D (dB)	d_{f0} (Hz)	Correlation
Phone mapping	3.97	23.98	0.295
State mapping	3.55	20.40	0.312

From the results, the state mapping has 0.42 dB improvement in the log spectrum distance and about 3.58 Hz in F0.

5.4.2. Intra-language and cross-language adaptation

The second experiment is to compare the speaker adaptation performance difference between cross-language adaptation and intra-language adaptation. As a reference, the objective measures without adaptation are also listed in the same table.

Table 3. Cross-language VS. intra-language.

	S_D (dB)	d_{f0} (Hz)	Correlation
Intra-language	2.56	18.91	0.408
Cross-language	3.55	20.40	0.312
No adaptation	7.12	72.61	0.500

From this result, the log spectrum distance has 0.99 dB difference between intra-language adaptation and cross-language adaptation. Consider the 7.12 dB difference between the source and the target speaker, and the difference when people are speaking different languages, we would say the decrease from intra-language to cross-language is fair. However, if we compare the correlation coefficient, we will find it is even worse than that without adaptation. That means only the method of F0 mean shifting can achieve a better performance than adaptation even with intra-languages.

6. CONCLUSIONS

Cross-language adaptation is useful in scenarios such as speech-to-speech translation. However, without a common phoneme set like intra-language adaptation, a mapping between two languages is necessary. In this paper, a novel algorithm of cross-lingual adaptation is introduced. The HMM state, a sub-phonemic unit which represents acoustic-phonetic event, is used as the basic unit for mapping. Kullback-Leibler divergence is used as the measure for the similarity of two states.

Three objective measures: log spectral distance, F0 adaptation error, and F0 correlations, are implemented in this paper to compare the adaptation performance difference. Compared with the phone mapping, this state mapping method is 0.42 dB better in the log spectrum distance, and 3.58 Hz better in the root mean square error of F0. Compared with the intra-language adaptation, the decrease in log spectrum distance is 0.99dB, which is fair compare with the 7.12 dB difference between source and target speaker.

7. ACKNOWLEDGEMENTS

The author would like to thank Houwei Cao for providing the state and phoneme mapping table between Mandarin and English.

8. REFERENCES

[1] D. Suendermann, H. Hoege, A. Bonafonte, H. Ney, and J. Hirschberg, "TC-Star: Cross-Language Voice Conversion Revisited," in Proc. of the TC-Star Workshop, 2006.

[2] K. Tokuda, T. Kobayashi, etc., "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," in Proc. of ICASSP, pp. 1315-1318, 2000.

[3] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, "Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR," in Proc. of SSW3, pp. 273-276, 1998.

[4] M. Mashimo, T. Toda, K. Shikano, N. Campbell, "Evaluation of cross-language voice conversion based on GMM and straight," in Proc. of EUROSPEECH, pp. 361-364, 2001.

[5] O. Andersen, P. Dalsgaard, and W. Barry, "Data-driven Identification of Poly- and Mono-phonemes for Four European Languages," in Proc. of EUROSPEECH, pp. 759-762, 1993.

[6] J. Kohler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," in Proc. of ICSLP, pp. 2195-2198, 1996.

[7] A. Zgank, B. Imperl, F. T. Johansen, Z. Kacic, B. Horvat, "Cross-lingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering," in Proc. of EUROSPEECH, pp. 2725-2729, 2001.

[8] H. Liang, Y. Qian, F. K. Soong and G. Liu, "A Cross-Language State Mapping Approach to Bilingual (Mandarin-English) TTS", in Proc. of ICASSP, pp. 4641-4644, 2008.

[9] A.W. Black, K.A. Lenzo, "Multilingual Text-to-Speech Synthesis," in Proc. of ICASSP, pp 761-764, 2004.

[10] M. Chu, H. Peng, et al., "Microsoft Mulan - a Bilingual TTS System," in Proc. of ICASSP, pp. 264-267, 2003.

[11] International Phonetic Association, "Report on the 1989 Kiel convention," *Journal of the International Phonetic Association*, vol. 19(2), pp. 67-82, 1989.

[12] S. Kullback and R.A. Leibler, "On Information and Sufficiency," Ann. Math. Stat., Vol. 22, pp. 79-86, 1951.

[13] Y. Zhao, C.S. Zhang, etc. "Measuring Attribute Dissimilarity with HMM KL-Divergence for Speech Synthesis," in Proc. of SSW6, pp. 206-210, 2007.

[14] Z.H. Ling, R.H. Wang, "HMM-Based Hierarchical Unit Selection Combining Kullback-Leibler Divergence with Likelihood Criterion," in Proc. of ICASSP, pp. 1245-1248, 2007.

[15] Y. Zhao, P. Liu, Y. Li, Y.N. Chen, and M. Chu. "Measuring Target Cost in Unit Selection with KL-Divergence between Context-Dependent HMMs," in Proc. of ICASSP, pp. 725-728, 2006.

[16] J. Yamagishi, and T. Kobayashi, "Average-Voice-based Speech Synthesis using HSMM-based Speaker Adaptation and Adaptive Training," IEICE Trans. Information and Systems E90-D, no.2, pp.533-543, Feb. 2007.

[17] T. Hastie, "A Closer Look at the Deviance," The American Statistician, 41, pp. 16-20, 1987.

[18] S. Kullback, "The Kullback-Leibler distance," The American Statistician, 41, pp. 340-341, 1987.

[19] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problem," in Proc. of Roy. Soc. A., vol. 186, pp. 453-461, 1946.

[20] M.N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," IEEE Signal Proc. Letters, vol. 10, No. 4, pp. 115-118, 2003.

[21] 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.

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