Label Transform Based Cross-Language Speaker Adaptation in Bilingual (Mandarin-English) TTS

Yongjin So, Jia Jia, Yongxin Wang and Lianhong Cai
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology（TNList）
Department of Computer Science and Technology, Tsinghua University, Beijing, China
{xuyj03,jiajia,wangyongxin}@mails.tsinghua.edu.cn,clh-dcs@tsinghua.edu.cn

Abstract

This paper studies the cross-language speaker adaptation for HMM-based speech synthesis. To solve the problem when the adaptation data and the main corpus are not in the same language, we proposed a label transform based cross-language speaker adaptation approach. In order to transform the phone sequence between English and Chinese, a new Mandarin-English phonetic alphabet–HCSIPA is designed. Then, in addition to the traditional Kullback-Leibler Divergence, a phoneme similarity measure: AMD, which take articulation difference into account, is proposed to get the similarity between phonemes. Finally, a perception-based phoneme mapping strategy is implemented to increase the mapping accuracy between Mandarin and English phonemes. The perceptual tests verify the rationality of our approach. The adapted speeches have high natural quality, and are judged as similar to the target speaker.

1. Introduction

With the development of HMM-based text-to-speech synthesis, polyglot speech synthesis, in which one engine can synthesize multiple languages using the same voice, is often demanded, e.g. in the application of spoken language translation (SLT).

A straightforward approach is to use a polyglot corpus, in which a multilingual speech corpus is recorded by one multilingual speaker [1]. However, it is difficult to find such a person who can speak multiple languages with professional levels.

There are many studies on polyglot speech synthesis using multilingual speech corpus in which speech of different languages is recorded with different speakers. A common approach uses another bilingual corpus recorded by another speaker to build a mapping model between different languages. Various mapping approaches were proposed, e.g. phone mapping, state mapping [2-5], frame mapping [6] and Gaussian component mapping [7]. Such approach can be called as the method based on language adaptation. However, when we have a large corpus of speaker A in language X and a small corpus of speaker B in language Y, such language adaptation cannot synthesis speech of B in language X, because the corpus of target speaker is not enough to train a model on which language mapping can be applied.

To resolve this problem, label transform based speaker adaptation approach is proposed [8]. The text labels of language Y are transformed into text labels of language X using a certain cross-language phone mapping rule, and existing inner-language speaker adaptation is applied to adapt the model of speaker A to speaker B in language X. However in [8], Chinese initial/final is mapped to English phoneme sequence only based on phonetic knowledge, similarity between phonemes of different language is not measured.

There are also many researches on the similarity measure between phonemes of different language. In order to synthesis loan words and person names with language Y in a TTS of language X, the xenophone of language Y is transcribed by native speaker with language X [14-16], and the phoneme distance of language X and Y is measured by occurrence percentage. Such approach has a high dependence on testers, and it is difficult to find such several testers.

The KLD has been typically used to measure the similarity of phone with different languages. But considering it does not take any language-specific information into account, the similarity measure guided by phonological knowledge is proposed in [17]. However, it is not integrated with acoustic distance, and it is used in language adaptation. In [18], an acoustic-phonetic unit similarity is proposed. The phonemes are hierarchically classified by phonetic questions, and the distance is measured the distance in the hierarchical structure. This makes many different phoneme pairs would have the same distance, so it is hard to be used in choosing a nearest phoneme.

978-1-4673-0174-9/12/$31.00 ©2012 IEEE

In this paper, we focus on label transform based cross-lingual speaker adaptation. Firstly, a new Mandarin-English phonetic alphabet–HCSIPA is designed, which enables us to transform the phoneme labels across language more easily. Then, unlike [8], we use a similarity measure to find phoneme mapping. In addition to the traditional KLD distance, a phoneme similarity measure: AMD, which takes articulation difference into account, is proposed to get the similarity between phonemes. Finally, a perception-based phoneme mapping strategy is implemented to increase the mapping accuracy between Mandarin and English phonemes.

2. Phoneme Set Construction and Phoneme Similarity Measure

In order to map labels between different languages, a uniform phoneme set should be used for the both languages. A new Mandarin–English phonetic alphabet–HCSIPA is proposed in this paper.

As there are also phonemes that only exist in only one language, similar phonemes must be used in adaption process. A new articulation method distance is proposed in this paper in addition to the traditional KLD distance, which pays more attention to the articulation differences of phonemes, and is more consistent with human perception. For each Mandarin and English phoneme, the phonemic similarity is calculated using an integration of KLD and AMD: KLD+AMD.

2.1. A new Mandarin–English Phonetic Alphabet– HCSIPA

We built a new Mandarin–English phonetic alphabet–HCSIPA referring to X-SAMPA table [9] for English and SAMPA-SC [10] for Chinese.

The construction of HCSIPA is mainly based on IPA, while also paying attention to the phoneme characteristics of Mandarin and English. Unlike [5] in which phonemes are sub-divided by prosodic features, the phoneme in HCSIPA would only distinguish the place and manner of articulation, e.g. nasal, fricative, labial, etc. All of phonemes are expressed with two English letters, allowing it to be easily applied into phoneme labels.

Some phonemes in IPA, HCSIPA and alphabet of [5] are shown in Table1. 38 consonants and 29 vowels are used to denote Mandarin and English phonemes in HCSIPA. Nine consonants and eleven vowels are shared between the two languages, which are 29.85% of all phonemes.

2.2. Articulation Method Distance

In the label transform, for phonemes that only exist in one language, a most similar phoneme in the other language is used as a substitute. Usually, similarity between phonemes is measured by KLD [11] distance of their HMM models, as defined in Eqs. (1) and Eqs. (2).

$$\mathrm{D_{KL}}(\mathrm{A},\mathrm{B}) = \frac{1}{n}\sum_{i=1}^{\mathrm{n}} \mathrm{D_{KL}}(A_i, B_i) \tag{1}$$

$$\mathrm{D_{KL}}(A_i, B_i) = \frac{1}{2}\mathrm{tr}\{(\Sigma_{A_i}^{-1} + \Sigma_{B_i}^{-1})(\mu_{A_i} - \mu_{B_i})(\mu_{A_i} - \mu_{B_i})^{\mathrm{T}} + \Sigma_{A_i}\Sigma_{B_i}^{-1} + \Sigma_{B_i}\Sigma_{A_i}^{-1} - 2\mathrm{I}\} \tag{2}$$

where n is number of states, A_i and B_i are the i-th states of phoneme A and phoneme B, Σ and μ are the corresponding covariance matrix and mean vector of the GMM distribution of a state.

However, the KLD distance is not always consistent with human perception. For some phoneme, the KLD-nearest phoneme is perceptually far. These cases often appear in consonants, and sometimes a consonant can even be mapped to a vowel by KLD.

To solve this problem, we proposed a new distance measure based on the assumption that two phonemes perceived as similar also have similar place and manner of articulation, for example, the similar consonants hh (h in “house” from English) and hx (h in hong from Mandarin) are both back and unvoiced consonants with only a little difference in place of articulation. Based on this, we proposed an articulation method distance (AMD) representing the distance between two phonemes’ place and manner of articulation. Here the articulation method includes the place and manner of articulation.

For calculation of AMD, a binary-value property vector **a** is prepared for each phoneme. Each dimension of the vector represents one property (simple or complex) of the articulation method of the phoneme. Then AMD distance is the defined in Equation (3).

$$\mathrm{D_{AM}}(\mathrm{A},\mathrm{B}) = \mathrm{d}(\mathbf{a}_{\mathrm{A}}, \mathbf{a}_{\mathrm{B}}) = \frac{\mathrm{N}_{\mathbf{a_i}=\mathbf{b_i}}}{\mathrm{N_T}} \tag{3}$$

where $\mathbf{a}_{\mathrm{A}}$ and $\mathbf{a}_{\mathrm{B}}$ are the property vectors of phoneme A and phoneme B, $\mathrm{N_T}$ is the dimension of the property vector, $\mathrm{N}_{\mathbf{a_i}=\mathbf{b_i}}$ is the number of dimensions where $\mathbf{a}_{\mathrm{A}}$ and $\mathbf{a}_{\mathrm{B}}$ have the same value.

A total of 67 properties are used to construct the property vector in this paper. Some of the properties are simple property, which only relate to one aspect of articulation (e.g. consonant, vowel, nasal, etc.), while others are complex properties that relate to multi aspects of articulation (e.g. central vowel, unvoiced fricative, etc.).

The construction of property vectors makes some important simple properties appear repeatedly, implicitly increasing the weight of each simple question. In this property set, 64.18% of the properties are consonant-related, which increases similarity measure accuracy for consonants.

Table1. Comparison of HCSIPA and Alphabet in [5]

IPA	HCSIPA	Alphabet in [5]
/ r /	rc	/ r /
/ ɹ /	rr	
/ ɛ /	ee	/ɛ˥/ /ɛ˧/ /ɛ˦/
/ a /	ai	/a˥/ /a˧/ /a˦/
/ æ /	ae	

2.3. Phoneme Distance Measure based on KLD+AMD

AMD makes the similarity measure closer to perception by utilizing articulation method as a similarity measure. However, AMD cannot reflect the acoustic feature difference brought by different speakers or different speaking styles. To include both acoustic and articulation differences into the distance measure, the KLD and AMD are integrated using Eqs. (4).

The D_{KL}-nearest, D_{AM}-nearest and D_{KLAM}-nearest top 3 results for some phonemes are shown as Table2. The bold phonemes are the ones that are perceptually similar to the phonemes in the first column. The result shows that KLD+AMD based similarity measure can represent the phonemic distance of different language more accurately.

$$D_{KLAM}(A,B) = D_{KL}(A,B) \times D_{AM}(A,B) \qquad (4)$$

Table2. Top 3 nearest phonemes in various similarity measures. The bold phonemes are perceptually similar to the phoneme in the first column.

HCSIPA /IPA/	KLD	AMD	KLD+AMD
dd / d /	**tt / t/**	**tt / t/**	**tt / t/**
	bb / b/	**dh / ð/**	**th / tʰ /**
	th / tʰ /	ll / l/	**dh / ð/**
va / ʌ/	ea / ə/	**ai / a /**	**ai / a /**
	aa /ɑ/	oo / ɔ/	ea / ə/
	eo / ɤ/	rc / r /	**aa /ɑ/**
sh / ʃ/	**sc / ʂ/**	ch / tʃ /	**sc / ʂ/**
	zs / ʒ/	**sc / ʂ/**	qh / tʂʰ/
	qh / tʂʰ/	qh / tʂʰ/	ch / tʃ /

3. Cross-Lingual Speaker Adaptation Based on Label Transform

3.1. Framework of Approach

Here we propose a method to solve the following cross-language speaker adaptation problem: when there are large corpuses of language X recorded by speaker A, and a small corpus of language Y recorded by speaker B, how to synthesis speech in language X in the voice of B?

To solve this problem, the labels in language Y are transformed into labels in language X, and speaker adaptation is done base on the large corpus and the small corpus with transformed labels to get the model for B in language X.

For shared phonemes in Mandarin and English, its text labels do not needed to be transformed. But for phonemes which only exist in one language, a similar phoneme should be found in the other language to be used in label transform. To increase the mapping accuracy between Mandarin and English phonemes, a perception-based phoneme mapping strategy is applied.

3.2. Phoneme Mapping Strategy Based on Perception Classification

The similar phoneme is found using our proposed phoneme distance measure as integration of KLD and AMD. However, there are still some problems to be solved.

One problem is that some phonemes in one language do not have a similar phoneme in the other. Even the most similar one would be too different to be treated as a substitute. For example, no similar phoneme exists in Mandarin for the English phonemes ww (/ w/), je (/ j/), vi (/ v/). Another problem is that the nearest phoneme is not always the most similar in perception.

To solve the above problems, perceptually similar phoneme groups are constructed. The phonemes in HCSIPA are divided into 27 groups, while phonemes in each group can be perceived as similar. Among these groups, 8 of them contain only one phoneme. Among the 8 single phoneme groups, 5 (ff(/ f /), ll(/ l/), mm(/ m /), nn(/ n/) and ng(/ ŋ/)) are shared phonemes that do not need transform and 3 (ww(/ w/), je(/ j/) and vi(/ v/)) are phonemes that similar phonemes cannot be found.

Thus, phoneme mapping strategy of phonemes that only exist in one language would be to find a nearest phoneme in a perceptually similar phoneme group. If the group contains only one phoneme and the phoneme is not a common phoneme, it would not participate in speaker adaptation.

4. Experiments

4.1 Experiment Setups

A corpus containing 1,400 English sentences and 1,400 Chinese sentences recorded by one female bilingual speaker (C) were used to calculate the similarity between phonemes. A corpus containing 5,400 Chinese (language X) sentences of a female speaker (A) were used to train the Mandarin model, and 20, 100 and 1,000 English (language Y) sentences of another female speaker (B) were used for adaptation. Ten Chinese sentences which are not in the training set are used for intelligibility, quality and similarity tests.

All speech waveforms were sampled at 16 KHz. The tools and scripts from HTS-2.1.1 were used for model training [12]. TTS feature vectors are comprised of 135 dimensions: 39-dimension STRAIGHT mel-Cepstral coefficients, log F0, 5 band-filtered aperiodicity measures, and their dynamic and acceleration coefficients. The CSMAPLR algorithm was adopted for speaker adaptation [13]. For synthesis, STRAIGHT synthesis filter was used to generate the speech waveform.

4.2 Experimental Results

Here we compared the speech quality of synthesized speech (MOS) and similarity between synthesized speech and original speaker (B) (DMOS) using different phoneme similarity measure and different amount of adaptation data. And we also used the transcribed character accuracy to show that our cross-lingual adaptation does not bring apparent decline to intelligibility in synthesized speech. Finally the effect of phoneme mapping strategy is estimated.

Ten native Mandarin speakers were asked to give their scores on the speech quality and similarity with target speaker in a five-point scale: 5=excellent, 4=good, 3=fair, 2=poor, 1=bad. They were also asked to transcribe 10 synthesized sentences for intelligibility test, and the result is represented by Chinese character accuracy rate.

4.2.1. Different similarity measure. Table3 and Fig.1 show the scores of intelligibility, speech quality and similarity of adapted speech when KLD, AMD and KLD+AMD are used as phonemic similarity measure.

From the result, we can see that AMD can greatly improve the overall performance of cross-language speaker adaptation. When KLD+AMD is used, the performance can be further improved.

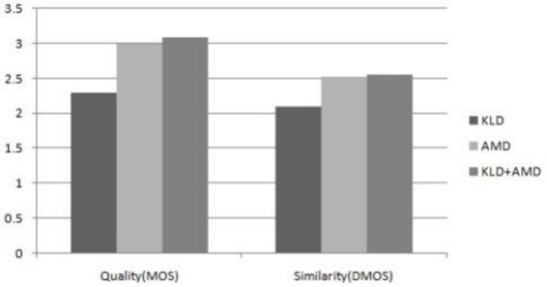

Fig.1 speech quality and similarity score with different phoneme similarity measures

Table3. Intelligibility with different similarity measures

	KLD	AMD	KLD+AMD
Character Accuracy	60.61%	88.49%	90.3%

4.2.2. Different amount of adaptation data. The intelligibility, speech quality and similarity achieved with different amount of adaptation data are shown as Table4 and Fig.2. The results show that decline in overall performance is not very significant with the reduction in the amount of adaptation data. It means that our cross-language speaker adaptation approach can be used when there is only a small amount of target speaker's speech.

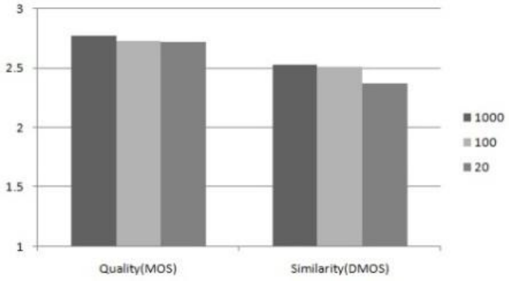

Fig.2 speech quality and similarity score with different amount of adaptation data

Table4. Intelligibility with different amount of adaptation data

Amount of adaptation data	1000	100	20
Character Accuracy	92.12%	89.09%	86.06%

4.2.3. Effect of phoneme mapping strategy. Here we compared the performance with and without similar phoneme groups. Performance improvement with our mapping strategy can be seen from Table5 and Fig.3.

The results show that our phoneme mapping strategy can improve the adaptation performance. In particular, it improves the speech quality more obvious.

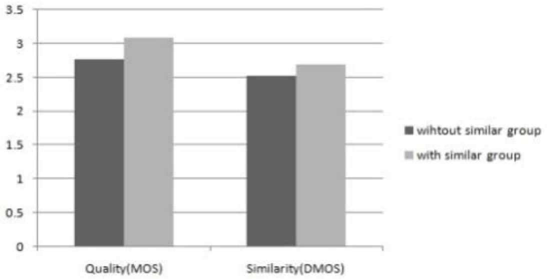

Fig.3 speech quality and similarity score with and without similar phoneme groups

Table5. Intelligibility with and without similar phoneme groups

	without	with
Character Accuracy	90.3%	92.12%

5. Conclusion

We propose a label transform based cross-language speaker adaptation approach. Specifically, a Mandarin-English phonetic alphabet-HCSIPA, a phoneme similarity measure KLD+AMD, and a perception-based phoneme mapping strategy are proposed for label transform between Mandarin and English.

The perceptual tests show the effectiveness of our approach in small amount of adaptation data. The similarity measure based on KLD+AMD and the phoneme mapping strategy can greatly improve the overall performance of cross-language speaker adaptation. Through our approach, a MOS score of 3.09 and a DMOS score of 2.69 could be obtained, and the intelligibility of synthesized speech is maintained at 92.1%,

6. Acknowledgement

This work is supported by the National Natural, and Science Foundation of China (61003094, 90820304, 90920302). The paper is also funded by Tsinghua National Laboratory for Information Science and Technology（TNList）Cross-discipline Foundation, and supported by the Innovation Fund of Tsinghua-Tencent Joint Laboratory.

References

[1] C. Traber *et al.*, “From multilingual to polyglot speech synthesis,” in *Proc. Eurospeech*, pp. 835-838, 1999.

[2] Y. Qian *et al.*, “HMM-based mixed-language (Mandarin-English) speech synthesis,” in *Proc. of ISCSLP*, pp. 1-4, 2008.

[3] H. Liang, Y. Qian and F.K. Soong, “An HMM-based bilingual (Mandarin-English) TTS,” in *Proc.6th ISCA Speech Synth. Workshop*, pp. 137-142, 2007.

[4] H. Liang et al., “A cross-language state mapping approach to bilingual (Mandarin-English) TTS,” in *Proc. of ICASSP*, pp. 4641-4644, 2008.

[5] Y. Qian, H. Liang, and F.K. Soong, “A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS,” in *IEEE TASLP*, vol. 17, no. 6, pp. 137-142, 2009.

[6] Y. Qian, J. Xu, and F.K. Soong, “A frame mapping based HMM approach to cross-lingual voice transformation,” in *Proc. of ICASSP*, pp. 5120-5123, 2011.

[7] H. Cao, T. Lee, and P.C. Ching, “Cross-lingual speaker adaptation via Gaussian Component mapping,” in *Proc. Interspeech*, pp. 869-872, 2010.

[8] Y.J. Wu *et al.*, “Cross-lingual speaker adaptation for HMM- based speech synthesis,” in *Proc. of ISCSLP*, pp. 9-12, 2008.

[9] http://en.wikipedia.org/wiki/Extended_Speech_Assessment_Methods_Phonetic_Alphabet

[10] J.L. Zhang, “A machine-readable alphabet in Mandarin: SAMPA-SC,” in *ACTA ACUSTICA, Chinese*, pp. 81-87, 2009.

[11] Y. Zhao *et al.*, “Measuring attribute dissimilarity with HMM KL-divergence for speech synthesis,” in *Proc.6th ISCA Speech Synth. Workshop*, pp. 206-210, 2007.

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

[13] J. Yamagish et al., “Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm,” in *IEEE TASLP*, vol. 17, no. 1, pp. 66-83, 2009.

[14] J. Gustafson., “Transcribing names with foreign origin in the ONOMASTICA project,” in *Proc. of ICPHS*, 1995.

[15] R. Eklund and A. Lindstrom., “How to handle ‘Foreign’ sounds in Swedish Text-to-Speech conversion: approaching the ‘Xenophone’ problem,” in *Proc. of CSLP*, 1998.

[16] J. Steigner and M. Schroder., “Cross-language phonemisation in German Text-to-Speech synthesis,” in *Proc. InterSpeech*, 2007.

[17] H. Liang and J. Dines., “Phonological knowledge guided HMM state mapping for cross-lingual speaker adaptation,” in *Proc. InterSpeech*, 2011.

[18] V.B. Le, L. Besacier and T. Schultz., “Acoustic-phonetic unit similarities for context dependent acoustic model portability,” in *Proc. of ICASSP*, pp. I1101-I1104, 2006