Phonological Knowledge Guided HMM State Mapping for Cross-Lingual Speaker Adaptation

Hui Liang[1,2], *John Dines*[1]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
hliang@idiap.ch, dines@idiap.ch

Abstract

Within the HMM state mapping-based cross-lingual speaker adaptation framework, the minimum Kullback-Leibler divergence criterion has been typically employed to measure the similarity of two average voice state distributions from two respective languages for state mapping construction. Considering that this simple criterion doesn't take any language-specific information into account, we propose a data-driven, phonological knowledge guided approach to strengthen the mapping construction – state distributions from the two languages are clustered according to broad phonetic categories using decision trees and mapping rules are constructed only within each of the clusters. Objective evaluation of our proposed approach demonstrates reduction of mel-cepstral distortion and that mapping rules derived from a single training speaker generalize to other speakers, with subtle improvement being detected during subjective listening tests.

Index Terms: phonological knowledge, minimum generation error, cross-lingual speaker adaptation, HMM-based TTS

1. Introduction

The language barrier is an important hurdle to overcome in order to facilitate better communication between people across the globe. Real-time automated speech-to-speech translation is a technology that could provide means to bridge the gap between languages, thus it is an important research topic. One component technology of speech-to-speech translation is speaker adaptation for speech synthesis, which would enable translated speech to be produced with a user's input voice characteristics.

HMM-based speech synthesis lends itself particularly well to speech-to-speech translation since it includes a range of speaker adaptation algorithms that centre around the so-called *average voice* paradigm [1]. In the context of speech-to-speech translation, we generally use the term *cross-lingual speaker adaptation*, which essentially means adapting the voice identity of average voice models to that of given adaptation data in a different language to that of the average voice models.

State mapping for cross-lingual speaker adaptation is performed by taking average voice models trained in the input (adaptation) and target (synthesis) languages and finding the closest matching states between the two models. Since the HMM state mapping technique was introduced [2], the minimum Kullback-Leibler divergence (KLD) criterion has been typically employed to establish this mapping. This purely data-driven criterion, though working acceptably well for cross-lingual speaker adaptation [3, 4], may not always produce meaningful state mapping rules especially when the two languages are quite distinct in terms of phonology.

In this work we propose to introduce phonological knowledge into the above-mentioned state mapping method. Our key idea is classifying average voice state distributions from two languages into phonologically constrained clusters and then constructing mapping rules only within each of the clusters. We achieve this by decision tree based clustering [5]. Sub-optimal phonological constraints (i.e. questions for node splitting) are discovered using a small set of bilingual development data, on which resulting state distribution clusters maximally provide improvement to cross-lingual speaker adaptation in terms of mel-cepstral distortion. In this paper we evaluate the effectiveness of our proposed method as well as the generality of the optimal set of mapping rules found for a particular speaker.

2. Current Mapping Construction Method

We call the language a target speaker speaks in adaptation data "input language (L_{in})" and the language in which speech is synthesized "output language (L_{out})". The state of the art of cross-lingual speaker adaptation is presented in [3], where each average voice state distribution from some language and its closest match from another language in terms of minimum KLD constitute a mapping rule. It was shown that the mapping could be performed as *transform mapping* (from each state in L_{out} to a state in L_{in}) or *data mapping* (from each state in L_{in} to a state in L_{out}). In this paper we present state mapping from the data mapping perspective since our previous analysis [4, 6] has shown a preference for this approach, though it may equally generalise to transform mapping as well. We also concentrate on adaptation of spectral features where mel-cepstral distortion (MCD) is employed as the objective measure.

In data mapping, adaptation data in L_{in} is associated with average voice state distributions of L_{out}. Then cross-lingual speaker adaptation is carried out in the intra-lingual manner. Our previous work showed the role that phonological mismatch between languages played in cross-lingual adaptation performance, hence it is natural to question the optimality of the minimum KLD criterion for state mapping, since it doesn't take into account any language-specific knowledge. To test the optimality of the minimum KLD criterion, we repeated the data mapping experiments in [6] as a preliminary examination – adapting an English average voice model with 100 Mandarin adaptation utterances in speaker MMh's voice (see Section 4 for MMh), but using mapping rules defined by the k-th best match in L_{out}.

We evaluated for ten values of k in turn and calculated corresponding MCD measurements. Results in Table 1 show that while KLD does generally increase with increasing k, this is only apparent for k>5. This suggests that while KLD is an effective measure, there may also exist additional latent factors

k	MCD (dB)	k	MCD (dB)
1	7.67	10	7.76
2	7.64	20	7.98
3	7.64	30	8.16
4	7.64	40	8.38
5	7.80	50	8.48

Table 1: *Results under the k-th best minimum KLD criterion for data mapping cross-lingual speaker adaptation*

that may be combined with KLD to achieve more effective mapping. In particular, the introduction of additional knowledge based on our understanding of the two languages' phonology may be used to guide the mapping.

3. Phonological Knowledge Guided State Mapping Construction

3.1. Basic idea

The minimum KLD criterion is used to construct mapping rules between average voice state distributions of context-dependent phones in L_{in} and L_{out}, but without taking into account our knowledge of their underlying phone categories. It can be seen that this approach could potentially lead to mapping rules that make little sense at the phone level (for instance, an L_{in} vowel state mapped to an L_{out} plosive state). Therefore we propose to introduce phonological knowledge in order to avoid such mappings from occurring. Specifically, we propose to classify average voice state distributions from L_{in} and L_{out} into phonologically constrained clusters such that mapping rules are constructed under the minimum KLD criterion, but only within each of these phonologically constrained clusters. Hence a state gets mapped to its phonologically similar states only.

3.2. Data-driven fashion for state classification

The challenge is to derive phonologically constrained clusters in a data-driven manner since it has been previously observed that purely knowledge-based approaches are not effective [7]. As a result, we employ decision tree-based state clustering in a similar fashion to cluster well-trained state distributions of L_{in} and L_{out} average voice models. Each leaf node of the decision trees is a phonologically constrained cluster.

3.2.1. Question design

Out of hundreds of phonetic and prosodic contexts used in HMM-base speech synthesis, the most important ones for spectrum are generally considered to be the triphone part – left phoneme ("L-"), central phoneme ("C-") and right phoneme ("R-"). Consequently, we consider the triphone contexts as the essential factors for clustering of average voice state distributions of L_{in} and L_{out} and create seven phoneme classes based on articulation manners that are commonly shared across our L_{in} and L_{out} – silence, vowel, plosive, fricative, affricate, approximant and nasal. Thus, we have a total of 21 questions for decision tree-based state clustering. A state distribution is considered to be a member of a phoneme class if any context-dependent phone to which it is tied belongs to this class, consequently, a state may have membership to multiple classes.

3.2.2. Question selection criterion for each node

Utterances from one or more speakers are selected as development data, which has no intersection with training data of average voice models, adaptation data or test data. Minimum generation error (MGE) [8] is used as the question selection criterion for each node. In order to find the best split for a node X, average voice state distributions belonging to X are clustered according to each question and the improvement is found by: (i) recalculating mapping rules between L_{in} and L_{out} based on each of the possible node splits; (ii) performing cross-lingual speaker adaptation in the data mapping fashion [4] with these newly formed mapping rules in X and all the existent ones in the rest untouched leaf nodes; and (iii) calculating the MCD change on held-out development data. The question producing the best improvement is selected for splitting node X eventually. The overall procedure is summarised below:

1. Form N root nodes by pooling all average voice state distributions from L_{in} and L_{out} for each of the N states, where N is the number of emitting states per HMM.
2. Find the next non-terminal leaf node across the N trees in the manner of breadth-first search.
3. Find the best split for this leaf node under the MGE criterion. If either of the following conditions is true it is considered a terminal leaf node, otherwise the node is split according to the selected question:
   (a) One or both children contain state distributions from only one language;
   (b) The best split produces an MCD reduction less than threshold $\varepsilon_{\Delta\text{MCD}}$ ($\varepsilon_{\Delta\text{MCD}}>0$).
4. Go back to Step 2 or stop growing the decision trees when all leaf nodes are terminal leaves.

MGE is a time-consuming optimization criterion [8], nonetheless, there are merely 21 questions in all, thus, the computational cost is still manageable. As a post-process, the proposed method degenerates into the purely minimum KLD criterion-based approach if it ends up with no node being split (i.e. no phonologically constrained clusters created).

4. Experiments and Analysis

We trained two average voice, single Gaussian-per-state synthesis models on the corpora Speecon (12.3 hours in Mandarin as L_{in}) and WSJ-SI84 (15.0 hours in English as L_{out}), respectively, in the HTS-2007 framework [9]. The HMM topology used was five-state (i.e. N=5) and left-to-right with no skip. Speech features were 39th-order STRAIGHT [10] mel-cepstra, $\log F_0$, five-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz recordings with a window shift of 5ms. All the following cross-lingual adaptation experiments were performed on these two average voice models, using the CSMAPLR [11] algorithm for speaker adaptation and global variances calculated on adaptation data for synthesis.

4.1. Speakers and speech data

Two male (MM3 and MM6) and two female (MF2 and MF7) speakers were selected from a bilingual corpus recorded in an anechoic chamber [12]. One more male speaker, MMh, whose voice was recorded in the same chamber, was also involved. The five speakers read exactly the same prompts in both Mandarin and English. MF2 was a truly bilingual speaker of Man-

darin and English, and the remaining four were native Mandarin speakers. MMh, MF7 and MM3 had reasonably natural English accents but MM6's English was strongly Mandarin-accented. Therefore, only MF2, MMh, MF7 and MM3 were considered as training speakers for our proposed approach.

Adaptation data of each of the five speakers consisted of 100 Mandarin utterances (files 0026∼0125). Development data of each of the four training speakers consisted of 100 English utterances (files 0026∼0125). Test data of each of the five speakers consisted of 25 English utterances (files 0001∼0025).

4.2. Cross-lingual adaptation approaches

We conducted four groups of experiments (L_{in} = Mandarin, L_{out} = English). Within each group, mapping rules of classified states for mel-cepstra were derived from one of the four training speakers by means of our proposed method while those for $\log F_0$, band aperiodicity and duration were still constructed purely by the minimum KLD criterion. Then these mapping rules were used for cross-lingual adaptation (all the four kinds of parameters) of the English average voice for each of the four remaining speakers. $\varepsilon_{\Delta\text{MCD}}$ was set to 0.0005dB. Our baseline system merely involved the minimum KLD criterion in construction of mapping rules for all kinds of features.

In this study we only investigated global transform based adaptation due to present computational demands of the MGE-based decision tree construction. In addition, our previous study [6] demonstrates that using regression class tree based adaptation is detrimental to cross-lingual speaker adaptation. Hence, we consider this as a topic for future work.

4.3. Objective evaluation

Original recordings of test data of the five speakers were force-aligned using the English average voice models and speech samples for objective evaluation were synthesized as per the resulting alignments. Results of objective evaluation of the four groups of cross-lingual adaptation experiments are presented in Figure 1 and Table 2. These MCD measurements were calculated on the entire test data set of the five speakers respectively.

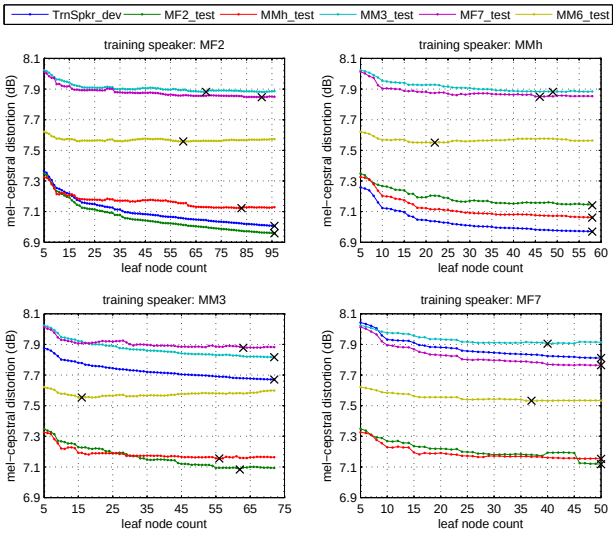

Figure 1: *Plot of MCD versus leaf node count during decision tree construction (Crosses indicate minimums on the curves. "TrnSpkr_dev" refers to the development data of respective training speakers. "_test" refers to test data. The six points on the vertical axis in each sub-figure come from the baseline.)*

TrnSpkr	Data set	ΔMCD	Data set	ΔMCD
MF2	MF2_dev	0.36	MF2_test	0.39
	MMh_test	0.20	MM3_test	0.14
	MF7_test	0.16	MM6_test	0.05
MMh	MMh_dev	0.29	MF2_test	0.21
	MMh_test	0.26	MM3_test	0.14
	MF7_test	0.16	MM6_test	0.06
MM3	MM3_dev	0.21	MF2_test	0.26
	MMh_test	0.16	MM3_test	0.21
	MF7_test	0.13	MM6_test	0.02
MF7	MF7_dev	0.23	MF2_test	0.23
	MMh_test	0.17	MM3_test	0.11
	MF7_test	0.25	MM6_test	0.09

Table 2: *MCD reduction (ΔMCD) in dB due to the proposed method, i.e., the difference of the leftmost and rightmost values on each curve in Figure 1*

It can be seen from Figure 1 that mapping rules optimized on the development data of a bilingual speaker consistently provided improvement on their own test data. When applying such mapping rules to other target speakers, it is observed that the MCD curves of these target speakers still had a nearly monotonically decreasing tendency. In other words, speaker-dependently constructed mapping rules still maintained a degree of speaker-independence. The exception was MM6, who received the least MCD reduction among all the speakers. This result may come from the fact that MM6 has the most pronounced accent when speaking English, thus resulting in clustered mapping rules that do not generalise to his speech.

4.4. Impact of phonological knowledge on mapping rules

A total of 2975 mapping rules were constructed, one for each of the 2975 states in the Mandarin average voice model. Figure 2 shows how k varied under the data-driven use of phonological constraints.

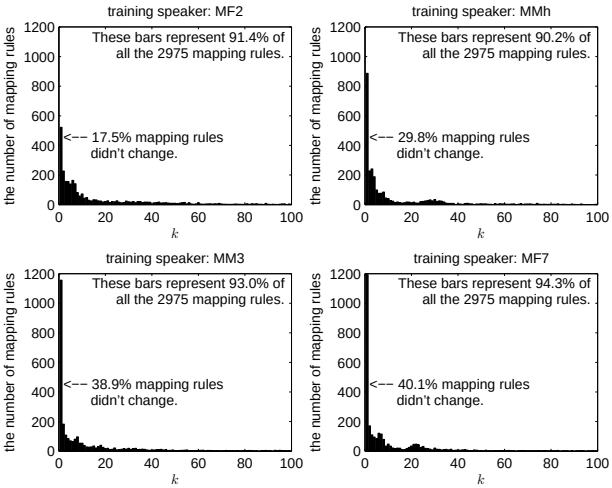

Figure 2: *Histogram of KLD rank (k) using the proposed approach*

We observe two common traits in the four sub-figures of Figure 2. Firstly, the bars corresponding to k=1 are significantly taller than any others and tall bars concentrate in the range of k<20. Thus, the minimum KLD criterion continues to play a dominant role and KLD remains a good measure of phonological similarity of context-dependent models from two different languages. Secondly, a significant proportion (minimum of

59.9%) of mapping rules changed under our proposed approach. Thus, it is also evident that the minimum KLD criterion on its own may not be sufficient, as suggested by our initial analysis. It is also interesting to note from both Table 2 and Figure 2 that our approach has the most impact on the truly bilingual speaker MF2, in terms of the number of changed mapping rules, MCD improvement and providing the best generalisation to the other speakers (except MM6, as was discussed previously).

4.5. Questions used for root node splitting

One means to analyse the generalisation of the proposed approach is to consider the questions that have yielded the greatest MCD improvement. We show in Table 3 the questions in the root node of each decision tree (which also gave the greatest MCD improvement) for each of the training speakers.

	MF2	MMh	MM3	MF7
2	L-nasal	L-nasal	L-nasal	L-nasal
3	C-nasal	C-nasal	C-vowel	C-nasal
4	C-nasal	C-nasal	C-affricate	C-affricate
5	R-fricative	C-affricate	C-nasal	C-affricate
6	L-silence	L-plosive	L-plosive	L-silence

Table 3: *Root node questions for emitting states at each of the five positions (2∼6) in an HMM*

It is interesting to see that most questions chosen by our proposed method were shared across speakers. The occurrence confirms that phonological constraints played a remarkably speaker-independent role in optimizing mapping rule construction.

4.6. Subjective evaluation

Subjective evaluation was performed in the form of AB and ABX listening tests for naturalness and speaker similarity, respectively. All of the speech samples were selected from the experiment group corresponding to the top-left sub-figure in Figure 1, since MF2 seems to provide the best generalisation to other speakers. We synthesised five sentences from the 25 used in the objective evaluation for each of the five speakers using the baseline and proposed approaches. Note that we used unadapted duration from the English average voice models. The evaluation comprised a total of 50 AB/ABX comparisons. Subjective evaluation results are shown in Figure 3.

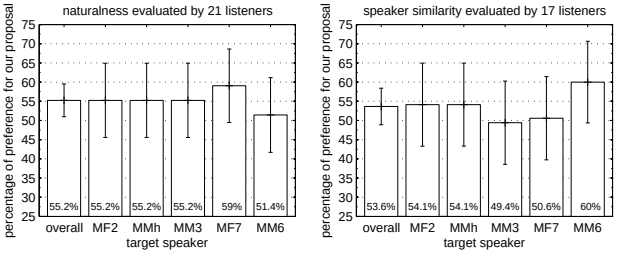

Figure 3: *Subjective evaluation results (Whiskers indicate 95% confidence intervals.)*

From informal listening, we noted that speaker similarity was not greatly impacted by the proposed approach, but naturalness was marginally improved (speech was produced with less 'muffled' characteristics than the baseline). Our perception is reflected in Figure 3. The lack of improvement in speaker similarity may in part come from limitations of the global transform that has been used in these experiments. A few speech samples can be found at http://www.idiap.ch/~hliang/demos/IS2011/.

5. Conclusions

The effectiveness and generality across speakers of phonological knowledge guided state mapping construction have been demonstrated in this paper. Though the consequent improvement that has been achieved so far is subtle, this method provides us with a promising future direction to improve cross-lingual speaker adaptation. We expect that optimizing state mapping rules on speech data of multiple bilingual speakers would result in a more robust set of mapping rules. The question set design is also worthy of further investigation. Lastly, we plan to investigate applying this method to regression class based adaptation.

6. References

[1] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training", *IEICE Trans. on Information and Systems*, vol. E90-D, no. 2, pp. 533–543, Feb. 2007.

[2] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1231–1239, Aug. 2009.

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

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

[5] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling", in *Proc. of the Workshop on Human Language Technology*, 1994, pp. 307–312.

[6] H. Liang and J. Dines, "An analysis of language mismatch in HMM state mapping-based cross-lingual speaker adaptation", in *Proc. of Interspeech*, Sep. 2010, pp. 622–625.

[7] Y.-J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis", in *Proc. of ISCSLP*, Dec. 2008, pp. 1–4.

[8] Y.-J. Wu, W. Guo, and R.-H. Wang, "Minimum generation error criterion for tree-based clustering of context-dependent HMMs", in *Proc. of Interspeech*, Sep. 2006, pp. 2046–2049.

[9] 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 Trans. on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[10] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds", *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, Apr. 1999.

[11] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[12] M. Wester and H. Liang, "The EMIME Mandarin Bilingual Database", University of Edinburgh, Tech. Rep. EDI-INF-RR-1396, Feb. 2011.

ACKNOWLEDGEMENT: Research leading to the results in this paper was funded from the Seventh Framework Programme (FP7/2007-2013) of the European Union under the grant agreement 213845 (the EMIME project).