An Analysis of Language Mismatch in HMM State Mapping-Based Cross-Lingual Speaker Adaptation

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

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

Abstract

This paper provides an in-depth analysis of the impacts of language mismatch on the performance of cross-lingual speaker adaptation. Our work confirms the influence of language mismatch between average voice distributions for synthesis and for transform estimation and the necessity of eliminating this mismatch in order to effectively utilize multiple transforms for cross-lingual speaker adaptation. Specifically, we show that language mismatch introduces unwanted language-specific information when estimating multiple transforms, thus making these transforms detrimental to adaptation performance. Our analysis demonstrates speaker characteristics should be separated from language characteristics in order to improve cross-lingual adaptation performance.

Index Terms: HMM-based TTS, cross-lingual speaker adaptation, HMM state mapping, language mismatch

1. Introduction

The language barrier is an important hurdle to overcome in order to facilitate better communication between people across the globe. It would be exciting and extremely helpful if we had a real-time automated speech-to-speech translator, especially when the translator could reproduce a user's input voice characteristics in its output speech. This is exactly the principal goal of the EMIME project (Effective Multilingual Interaction in Mobile Environments [1]). Cross-lingual speaker adaptation is thus one of the key techniques that EMIME demands.

Cross-lingual speaker adaptation basically means altering the voice identity of average voice models given adaptation data in a different language. Unlike intra-lingual speaker adaptation, all the correspondence between the average voice models and the adaptation data is lost. By constructing mapping rules between model distributions across the two languages, the recently developed HMM state mapping technique [2] succeeds in relating those average voice models to the given adaptation utterances, and thus points out a promising research direction for cross-lingual speaker adaptation.

If we compare the problem of cross-lingual speaker adaptation with conventional problems in intra-lingual adaptation, there is an inherent challenge aside from the obvious lack of correspondence between adaptation data and average voice models. This challenge lies in the fact that we would like to apply adaptation algorithms such as maximum likelihood linear transformation [3], so that maximizing the likelihood on given adaptation data in an input language should also generalize to an increase of the likelihood (and objective/subjective synthesis quality) on unseen data in an output language. The adaptation algorithms employed to date make no such guarantee of generalization, but in practice have been found to work acceptably well [2, 4]. Nonetheless, in performing such cross-lingual adaptation, it is evident that a language mismatch factor is introduced between underlying average voice models and adaptation data.

Alleviating the influence of language mismatch should improve the performance of HMM state mapping-based cross-lingual speaker adaptation and eventually make it comparable to that of intra-lingual adaptation. However, it is first necessary to clarify how this language mismatch can impact cross-lingual adaptation. In this paper, we detail an investigation of the effects of language mismatch on cross-lingual speaker adaptation in order to fully understand the underlying mechanism and to discover potential directions for further improvements.

In this paper, we firstly summarize the basic idea of HMM state mapping and all of its four possible implementations. Secondly, we decompose language mismatch on the surface into four sources for the sake of clarification and investigation. Experimental results of intra-lingual and cross-lingual speaker adaptation are then presented respectively and this is followed by detailed analysis of the influence of language mismatch. The last section carries our conclusions.

2. HMM State Mapping

First of all, we define the language in which speech is synthesized as the *output language* and the language of given adaptation utterances from a target speaker as the *input language*.

2.1. Basic Idea and Implementations

An effective technique for cross-lingual speaker adaptation, HMM state mapping, has been proposed by Wu *et al.* [2]. The technique requires two monolingual average voice model sets in input and output languages, respectively, as a prerequisite. By establishing mapping rules between those average voice model distributions across the input and output languages, HMM state mapping is capable of relating the two different languages such that adaptation data in the input language can be utilized to adapt average voice models of the output language. Kullback-Leibler divergence (KLD) is employed as a measure of state distribution similarity. Specifically, for each state distribution of one language, a state distribution which has the minimum KLD value within the average voice model set of the other language is found, and then the two state distributions form a mapping rule.

In [2], two ways of applying such state mapping rules to cross-lingual speaker adaptation were proposed:

Data transfer 1. For the sets ($\mathbb{S}_{\text{in}}$ and $\mathbb{S}_{\text{out}}$) of state distributions of input and output languages, establish a set of mapping rules $\mathcal{M}_{\text{d}}$: $\mathcal{M}_{\text{d}}(\mathbb{S}_{\text{in}}) = \mathbb{S}_{\text{out}}$. This mapping direction is aimed at guaranteeing all the adaptation data will be used.

2. Transfer all the adaptation data in the input language from $\mathbb{S}_{\text{in}}$ to $\mathbb{S}_{\text{out}}$ according to $\mathcal{M}_{\text{d}}$.
3. Perform “intra-lingual” speaker adaptation on the side of the output language.

Transform transfer 1. For the sets ($\mathbb{S}_{\text{in}}$ and $\mathbb{S}_{\text{out}}$) of state distributions of input and output languages, establish a set of mapping rules $\mathcal{M}_{\text{t}}$: $\mathcal{M}_{\text{t}}(\mathbb{S}_{\text{out}}) = \mathbb{S}_{\text{in}}$. This mapping direction is aimed at guaranteeing each state distribution in $\mathbb{S}_{\text{out}}$ will be assigned a transform.

2. Perform intra-lingual speaker adaptation on the side of the input language.
3. Transfer a resulting transform from $\mathbb{S}_{\text{in}}$ to $\mathbb{S}_{\text{out}}$ for each state distribution of the output language according to $\mathcal{M}_{\text{t}}$.

2.2. Other Possible Implementations

In order to obtain a full picture of the influence of language mismatch, we propose another two ways of applying HMM state mapping rules:

Regression tree transfer 1. According to the mapping rules $\mathcal{M}_{\text{t}}(\mathbb{S}_{\text{out}}) = \mathbb{S}_{\text{in}}$, add each state distribution S_{out} of an output language into the regression class which the state distribution of an input language, $\mathcal{M}_{\text{t}}(S_{\text{out}})$, belongs to.

2. Remove state distributions of the input language from regression classes of the input language, and then remove empty regression tree leaf nodes of the input language.
3. Like the data transfer implementation, associate adaptation data in the input language with state distributions of the output language.
4. Estimate transforms over average voice distributions of the output language and regression tree structure of the input language. Each distribution of the output language obtains a transform as a result.

Conceptually, this is equivalent to transferring regression tree structure of the input language to the output language side.

Distribution transfer 1. According to the mapping rules $\mathcal{M}_{\text{d}}(\mathbb{S}_{\text{in}}) = \mathbb{S}_{\text{out}}$, add each state distribution S_{in} of an input language into the regression class which the state distribution of an output language, $\mathcal{M}_{\text{d}}(S_{\text{in}})$, belongs to.

2. Remove state distributions of the output language from regression classes of the output language, and then remove empty regression tree leaf nodes of the output language.
3. Estimate transforms over average voice distributions of the input language and regression tree structure of the output language.
4. As transforms are assigned to regression classes rather than state distributions, average voice distributions of the output language are assigned transforms indirectly.

Conceptually, this is equivalent to transferring state distributions of the input language to the output language side.

3. Decomposition of Language Mismatch

On the surface, language mismatch in the cross-lingual speaker adaptation context refers to the mismatch between the language identity of adaptation data (L_{data}) and that of average voice state emission pdfs for synthesis ($L^{\text{syn}}_{\text{pdf}}$). In practice, language mismatch may occur during estimation of adaptation transforms and during synthesis. Hence we are also concerned with the language identities of average voice state emission pdfs and regression tree structure involved during transform estimation ($L^{\text{adapt}}_{\text{pdf}}$ and $L^{\text{adapt}}_{\text{reg}}$, respectively). It is apparent that language mismatch can occur in the four possible ways below (as illustrated in Table 1):

1. between L_{data} and $L^{\text{adapt}}_{\text{pdf}}$ during transform estimation
2. between L_{data} and $L^{\text{adapt}}_{\text{reg}}$ during transform estimation
3. between $L^{\text{syn}}_{\text{pdf}}$ and $L^{\text{adapt}}_{\text{pdf}}$ during synthesis
4. between $L^{\text{syn}}_{\text{pdf}}$ and $L^{\text{adapt}}_{\text{reg}}$ during synthesis

What To Transfer	L_{data} (in. lang.)		$L^{\text{syn}}_{\text{pdf}}$ (out. lang.)	
	$L^{\text{adapt}}_{\text{pdf}}$	$L^{\text{adapt}}_{\text{reg}}$	$L^{\text{adapt}}_{\text{pdf}}$	$L^{\text{adapt}}_{\text{reg}}$
transform	○	○	×	×
distribution	○	×	×	○
data	×	×	○	○
regression tree	×	○	○	×
intra-lingual	○	○	○	○
pseudo-intra-lingual	○	×	○	×

Table 1: *Language mismatch overview (“×”: mismatched; “○”: matched; see Section 4.1 for the pseudo-intra-lingual case)*

4. Experiments and Analysis

Throughout the following experiments, we used Mandarin Chinese as the input language and English as the output language. We trained two average voice, single Gaussian synthesis model sets on the corpora SpeeCon (Mandarin) and WSJ SI84 (English), respectively, in the HTS-2007 framework [5]. The HMM topology was five-state and left-to-right with no skip. Speech features were 39th-order mel-cepstra, $\log F_0$, five-dimensional band aperiodicity, and their delta and delta-delta coefficients, extracted from 16kHz WAV files with a window shift of 5ms. Speech data for adaptation and evaluation was sourced from a small bilingual corpus recorded in an anechoic studio and uttered by a male native Mandarin speaker who had a reasonably natural English accent. The CSMAPLR [6] algorithm and 100 adaptation utterances in Mandarin were used. Global variances for synthesis were calculated on adaptation data. We mainly focus on cross-lingual adaptation of mel-cepstrum and employ mel-cepstrum distortion (MCD) as an objective measure of adaptation performance.

4.1. Experiments on Intra-Lingual Speaker Adaptation

In the context of intra-lingual speaker adaptation, there is no language mismatch (see the fifth line of Table 1). Consequently, adaptation should behave in a “normal” fashion: it should reduce mel-cepstrum distortion with respect to reference speech and provide further improvements as more adaptation data becomes available (and more transforms can be estimated). We estimated several sets of transforms for confirmation and subsequent comparison. The description of experiments in the intra-lingual context is as follows:

1. Each stream was assigned a single global transform. So there was only one global transform for mel-cepstrum adaptation.
2. Each state of each stream was assigned a single global transform. So there were five global transforms in all for mel-cepstrum adaptation.
3. Various amounts of transforms were generated by setting different thresholds of generating a transform (i.e., `HADAPT:SPLITTHRESH`) in HTS [7].

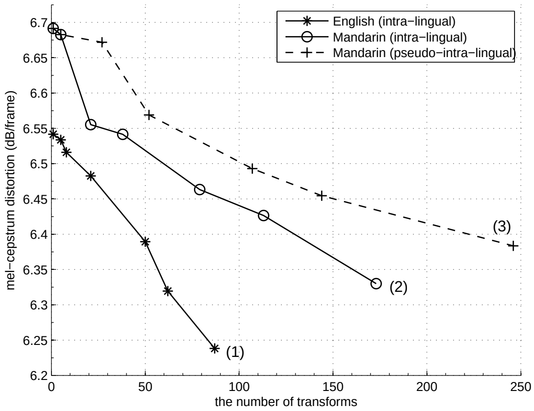

Figure 1: *MCD comparison (intra-lingual speaker adaptation with 100 adaptation utterances)*

It can be confirmed from the two solid lines in Figure 1 that a great number of transforms can better characterize the voice of a target speaker in the intra-lingual context. As transforms generated by distribution transfer (see Section 4.2) were estimated over average voice models in Mandarin, we also synthesized Mandarin speech with them for further analysis. This is the pseudo-intra-lingual case, for its $L_{\text{reg}}^{\text{adapt}}$ is English.

4.2. Experiments on Cross-Lingual Speaker Adaptation

We carried out cross-lingual speaker adaptation with each of the four HMM state mapping-based implementations detailed in Section 2. In each case we performed adaptation with a different number of transforms as we previously did in the intra-lingual adaptation. Objective evaluation results of cross-lingual adaptation experiments are presented in Figure 2.

4.3. Analysis of the Influence of Language Mismatch

4.3.1. Overall impact

Taking a look at Figures 1 and 2, we find that the seven polylines can be divided into three groups:

(a) All the polylines in Figure 1: all the cases of intra-lingual adaptation show similar behavior, though the misuse of English regression tree structure in the pseudo-intra-lingual case introduces the mismatches between L_{data} and $L_{\text{reg}}^{\text{adapt}}$ and between $L_{\text{pdf}}^{\text{syn}}$ and $L_{\text{reg}}^{\text{adapt}}$ that result in worse adaptation performance.

(b) Polylines 1 and 2 in Figure 2: these results pertain to cross-lingual adaptation using the state emission pdfs mapped from English average voice models. Both implementations give the lowest MCD values and do not appear to be impacted by regression tree structure.

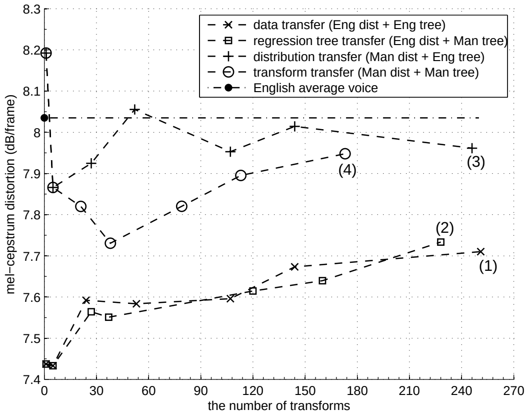

Figure 2: *MCD comparison (cross-lingual speaker adaptation with 100 adaptation utterances)*

(c) Polylines 3 and 4 in Figure 2: these systems use adaptation transforms estimated over the state emission pdfs of Mandarin average voice models. The worst performance is achieved with the distribution transfer implementation, which involves language mismatches during both transform estimation and synthesis.

It is apparent that the different sources of language mismatch can have a significant impact on cross-lingual speaker adaptation. The most severe mismatch appears to be that between the distributions used to estimate adaptation transforms and the distributions to which the transforms are applied during synthesis (i.e., $L_{\text{pdf}}^{\text{adapt}}$ and $L_{\text{pdf}}^{\text{syn}}$). The language mismatch related to regression tree structure appears to be less severe and less predictable in their severity.

4.3.2. Considering the number of transforms

Polyline 4 in Figure 2 and Polyline 2 in Figure 1 actually correspond to the same set of transforms, applied to English (cross-lingual adaptation) and Mandarin (intra-lingual adaptation) synthesis respectively. The monotonically decreasing Polyline 2 in Figure 1 is what we would expect (and desire) from using an increasing number of transforms. However, when applied to cross-lingual speaker adaptation, we note quite different behavior – the performance first improves and then degrades after a certain number of transforms have been estimated (see Polyline 4 in Figure 2). Likewise, the performance of data and regression tree transfer, corresponding to Polylines 1 and 2, degrades immediately after more than one transform per state have been estimated. We can explain this behavior in terms of overfitting.

When adapting average voice models, the resulting combined models and transforms should match adaptation data. In a speaker adaptation scenario, the transforms would ideally be learning only speaker-dependent characteristics to transform the average voice models to speaker-dependent models, but in practice, language-dependent characteristics are also captured. In the case of transform transfer, whereby transforms are estimated over input language average voice models, speaker-only characteristics are better captured by the transforms since there is no language mismatch during transform estimation. As a result, using multiple transforms can be beneficial up to a certain point, after which the transforms become more and more language-specific and performance degrades. In the case of data and regression tree transfer, there is an inherent language mismatch between average voice distributions for transform estimation and adaptation data. Hence, transforms immediately begin to be strongly influenced by this mismatch and using multiple transforms is immediately detrimental.

Despite the apparent advantage of transform transfer to better take advantage of multiple transforms, it still performs worse than data and regression tree transfer. It would appear that transform transfer, while modeling less of input language characteristics, is less suitable for adapting models in the output language. Thus, data transfer and regression tree transfer seem to provide the best way forward, but the challenge will be to develop techniques that are able to take advantage of larger quantities of adaptation data by using multiple transforms. Primarily, this will require a means to separate the effects of language and speaker mismatches that are both being captured presently.

4.4. The Issue of the Amount of Adaptation Data

Since data transfer using global transforms provides the best adaptation performance amongst all the systems, it is still worth investigating the effect of the quantity of adaptation data. Experiments were carried out with different quantities of adaptation utterances and objective evaluation results are presented in Figure 3. Due to the size of our own bilingual corpus, we couldn't use more than 100 adaptation utterances.

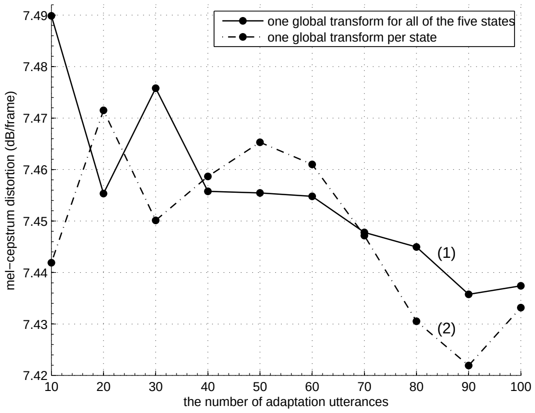

Figure 3: *MCD comparison (data transfer with the number of transforms fixed)*

Figure 3 shows a rough trend that more adaptation data helps to improve cross-lingual adaptation performance. Unfortunately, the use of global transforms limits the benefits of using more adaptation data, which can be seen in the almost negligible improvements observed in this experiment. This result further justifies the need for developing new techniques which can take advantage of a large quantity of adaptation data and multiple transforms.

4.5. Subjective Perception

We are mainly interested in objective measures, as they relate to the adaptation criterion most closely and thus should be a more sensitive reflection of the impacts of language mismatch. Nonetheless, objective measures don't always correlate with human perception [8]. We performed informal listening tests for confirmation.

We note, for the case of intra-lingual adaptation, that voice quality is always good and with more transforms speaker similarity improves. The fact that the target speaker doesn't have an American accent (to match the average voice models) makes the use of multiple transforms particularly important. In all cases of cross-lingual adaptation, speaker similarity is noticeably worse than the intra-lingual adaptation. For transform transfer, voice quality is maintained, but speaker similarity is poor. For data transfer and regression tree transfer, speaker similarity is better, but voice quality is degraded (a "muddy" quality that reflects the adaptation towards Mandarin). Furthermore, speech quality becomes distorted as more transforms are estimated – confirming the results obtained from our objective evaluations.

5. Conclusions

In this paper we have investigated how language mismatch degrades HMM state mapping-based cross-lingual adaptation. We have demonstrated the different sources of language mismatch and how these impact the different adaptation implementations. From our results we can conclude that though HMM state mapping is an effective method to relate two different languages it remains sensitive to the negative impacts of language mismatch. Reducing this mismatch is thus a key goal of future research.

Moreover, we have investigated in detail the impact of the number of transforms and quantity of adaptation data on cross-lingual adaptation. From the results of this study it becomes clear that current approaches are largely unable to take advantage of large quantities of adaptation data. In order to better reduce language mismatch and in so doing enable the effective use of multiple transforms, it will be necessary to introduce techniques that model the inherent differences between languages.

6. Acknowledgements

The research leading to these results was funded from the 7th Framework Programme (FP7/2007-2013) of the European Union under Grant Agreement 213845 (the EMIME project).

7. References

[1] Effective Multilingual Interaction in Mobile Environments, EMIME. [Online]. Available: http://www.emime.org

[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. of Interspeech*, Sep. 2009, pp. 528–531.

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

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

[6] 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 Transactions on Audio, Speech and Language Processing*, vol. 17, no. 1, pp. 66–83, Jan. 2009.

[7] The HMM-based speech synthesis toolkit, HTS. [Online]. Available: http://hts.sp.nitech.ac.jp

[8] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. of Interspeech*, Sep. 2009, pp. 1391–1394.