A COMPARISON OF SUPERVISED AND UNSUPERVISED CROSS-LINGUAL SPEAKER ADAPTATION APPROACHES FOR HMM-BASED SPEECH SYNTHESIS

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

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

ABSTRACT

The EMIME project aims to build a personalized speech-to-speech translator, such that spoken input of a user in one language is used to produce spoken output that still sounds like the user's voice however in another language. This distinctiveness makes unsupervised cross-lingual speaker adaptation one key to the project's success. So far, research has been conducted into unsupervised and cross-lingual cases separately by means of decision tree marginalization and HMM state mapping respectively. In this paper we combine the two techniques to perform unsupervised cross-lingual speaker adaptation. The performance of eight speaker adaptation systems (supervised vs. unsupervised, intra-lingual vs. cross-lingual) is compared using objective and subjective evaluations. Experimental results show the performance of unsupervised cross-lingual speaker adaptation is comparable to that of the supervised case in terms of spectrum adaptation in the EMIME scenario, even though automatically obtained transcriptions have a very high phoneme error rate.

***Index Terms*—** unsupervised cross-lingual speaker adaptation, decision tree marginalization, HMM state mapping

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). Cross-lingual speaker adaptation is thus one of the key goals of EMIME.

Such a speech-to-speech translator consists of speech recognition, machine translation and speech synthesis. EMIME focuses on speech recognition and synthesis. Bridging the gap between speech recognition and synthesis [1] is also an implicit goal. Thus, we hope to employ a unified modelling framework which applies to both recognition and synthesis. As speech recognition is typically HMM-based and we want to easily alter the voice identity of output speech, the HMM-based speech synthesis technology [2, 3] is the ideal choice. As a statistical parametric approach, the HMM-based framework provides a great deal of flexibility, especially with respect to its generality across languages and the ease of altering voice characteristics of models. Consequently, this paper investigates cross-lingual speaker adaptation based on unified HMM modelling.

We proposed a decision tree marginalization technique in [4] for unified HMM modelling, by which speech recognition can be performed with speech synthesis models. We found that this technique made it feasible to conduct unsupervised intra-lingual speaker adaptation in a unified modelling framework. As a result, employing the HMM state mapping technique [5] as well as decision tree marginalization should make unsupervised cross-lingual speaker adaptation viable in a unified modelling framework. We investigate the viability of the combination of these techniques in this paper.

In Section 2, decision tree marginalization and HMM state mapping are briefly reviewed. In Section 3, details on applying the two techniques simultaneously to unsupervised cross-lingual speaker adaptation are described. We then compare the performance of supervised and unsupervised cross-lingual speaker adaptation systems in the context of English and Mandarin Chinese in Section 4. Conclusions follow in Section 5.

2. COMPONENT TECHNIQUES

2.1. Decision Tree Marginalization

Decision tree marginalization [4] allows deriving speech recognition models from a full-context speech synthesis model set according to given triphone labels. Hence, the first stage is training a conventional HMM-based speech synthesis system from scratch, of which each HMM state emission distribution is typically composed of a single Gaussian PDF.

Conventionally, making a new synthesis model is carried out by traversing a synthesis decision tree according to the new full-context label and eventually assigning one leaf node to it. The basic idea of decision tree marginalization is fairly straightforward in the sense that it generates a triphone recognition model in almost the same manner. The only difference from making a new synthesis model is that both children of a decision tree intermediate node of the synthesis system are traversed when the question associated with the intermediate node is irrelevant to any triphone context. So finally a triphone label is associated with more than one leaf node, which form a state emission distribution of multiple Gaussian components. In other words, a triphone model for recognition constructed by decision tree marginalization can be viewed as a linear combination of full-context single Gaussian models for synthesis. No model parameters are changed during the whole process. See Figure 1 for an example.

The decision tree marginalization process described above is actually a special case. It can be extended such that an arbitrary context combination of full-context labels is marginalized out. For instance, we can create tonal monophone models by marginalizing out all the contexts that are unrelated to the base phone context and tone information.

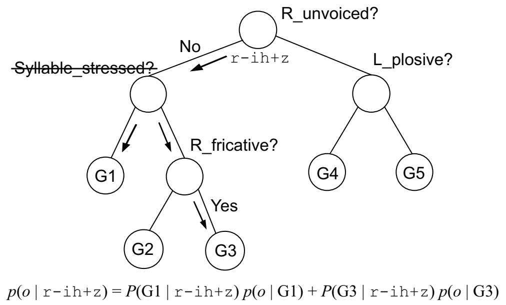
R_unvoiced?
No
r-ih+z
~~Syllable_stressed?~~
L_plosive?
R_fricative?
G1
G4
G5
Yes
G2
G3
$p(o \mid \texttt{r-ih+z}) = P(\text{G1} \mid \texttt{r-ih+z})\, p(o \mid \text{G1}) + P(\text{G3} \mid \texttt{r-ih+z})\, p(o \mid \text{G3})$

Fig. 1. An example of decision tree marginalization, showing how a new recognition model "r-ih+z" is derived from a decision tree of a speech synthesis system ("L_" / "R_": left/right phone; "G?": clustered state emission distribution PDFs)

2.2. HMM State Mapping

We consider the case in which we have adaptation data in an input language ($L1$) and an average voice model set for synthesis in an output language ($L2$). In theory, this prevents us from directly adapting the voice identity of the average voice model set into that of the adaptation data, because language mismatch eliminates all the correspondence between the data and the model set. Two possible solutions are (i) training a bilingual model set [6] and (ii) reconstructing the correspondence. HMM state mapping [5] is an effective method capable of reconstructing the correspondence for cross-lingual speaker adaptation when a bilingual model set is unavailable.

HMM state mapping requires two decent average voice model sets in $L1$ and $L2$, respectively. The two average voices are presumed to sound like a single person. Each state-cluster of $L1$ (or $L2$) is then associated with the most similar one of $L2$ (or $L1$) by matching state-cluster PDFs in the two model sets which have minimum (symmetric) Kullback-Leibler divergence between them. It is not guaranteed that every state-cluster of $L2$ (or $L1$) is touched. Untouched ones are ignored typically. Wu *et al.* [5] proposed two ways of applying state mapping rules to cross-lingual speaker adaptation:

Transform version is performed by first generating speaker dependent transforms by carrying out intra-lingual speaker adaptation using the acoustic model set trained for $L1$. Following this, voice characteristics of the acoustic model set in $L2$ are converted by applying these speaker-dependent transforms to state-clusters of $L2$'s acoustic models, according to prepared state mapping rules between the two acoustic model sets.

Data version is performed by first mapping state-clusters of the acoustic model set in $L1$ to those of $L2$'s acoustic models. Then adaptation data in $L1$ is associated with state-clusters of $L2$ through state-clusters of $L1$. Finally the adaptation data in $L1$ is treated as if it were in $L2$ and adaptation is performed using $L2$'s acoustic models in the "intra-lingual" sense.

3. COMBINING DECISION TREE MARGINALIZATION AND HMM STATE MAPPING

As discussed above, decision tree marginalization makes it feasible to perform unsupervised intra-lingual speaker adaptation and HMM state mapping makes it feasible to perform supervised cross-lingual speaker adaptation. We expected that their combination would enable unsupervised cross-lingual speaker adaptation.

First of all, we prepared HMM state mapping rules using two average voice synthesis model sets in $L1$ and $L2$, respectively, and performed speech recognition with the help of decision tree marginalization in order to obtain estimated triphone transcriptions of adaptation data uttered in $L1$.

Once estimated triphone transcriptions of adaptation data were available, either the transform version or the data version of HMM state mapping was used for "supervised" cross-lingual speaker adaptation. Note that estimated transcriptions were triphone sequences in $L1$. So rather than the synthesis model set in $L1$, it is the recognition models of $L1$ constructed by decision tree marginalization that were involved in the "supervised" cross-lingual speaker adaptation.

4. EXPERIMENTS

4.1. Experimental Setup

We trained two average voice, single Gaussian synthesis model sets on the corpora SpeeCon (Mandarin) and WSJ SI84 (English), respectively, and derived HMM state mapping rules and eight synthesis systems from them. Half of the eight systems were supervised and the rest were unsupervised. We collected bilingual adaptation data from two Chinese students (H and Z) who also spoke English well. The Mandarin and English prompts, which were not included in our training data, were also selected from SpeeCon and WSJ, respectively. Mandarin and English were defined as input ($L1$) and output ($L2$) languages, respectively, throughout our experiments.

System name format: (**S**/**U**) (**1**/**2**) - (**D**/**T**/**M**)

S/U	supervised / unsupervised
1/2	cross-lingual / intra-lingual
D/T	data/transform version of HMM state mapping
M	Decision tree marginalization was used instead of HMM state mapping. The average voice model set of Mandarin ($L1$) was therefore unnecessary.

Following this naming rule, the eight synthesis systems were S2, S1-M, S1-T, S1-D, U2, U1-M, U1-T and U1-D:

S2 purely built on the English side

S1-M We marginalized out all the English-specific contexts first. As a result, a Mandarin full-context label was associated with more than one English state-cluster. Then Mandarin adaptation data could be treated as English data for "intra-lingual" speaker adaptation.

S1-T & S1-D as described in Section 2.2

U2 purely built on the English side; as described in Section 2.1

U1-M We marginalized out all the non-triphone contexts and then recognized Mandarin adaptation data with English models. Mandarin adaptation data was thus associated with the English average voice model set.

U1-T & U1-D as described in Section 3

As decision tree marginalization was engaged in all the four unsupervised systems and S1-M, their transforms were estimated over multiple Gaussian component models instead of single Gaussian ones.

Speech features were 39th-order mel-cepstra, $\log F_0$, five dimensional band aperiodicity, and their delta and delta-delta coefficients. The CSMAPLR [7] algorithm and 40 adaptation utterances were used. Global variances were calculated on adaptation data. A simple phoneme loop was adopted as a language model for recognition. The average phoneme error rate was around 75%.

4.2. System Evaluation

We calculated RMSE of mel-cepstrum (MCEP) and F_0, as well as correlation coefficients and voicing error rates of F_0, for objective evaluation. See Table 1 ("AV" means "average voice").

	MCEP		F_0			
	RMSE (/frm)		RMSE (Hz/frm)		CorrCoef	
	H	Z	H	Z	H	Z
AV	1.39	1.43	26.0	35.9	0.46	0.49
S2	1.04	1.04	11.8	9.6	0.46	0.56
U2	1.06	1.08	13.0	14.0	0.47	0.54
S1-T	1.23	1.22	20.0	12.6	0.47	0.51
U1-T	1.24	1.26	21.1	16.5	0.48	0.53
S1-D	1.13	1.14	19.5	12.6	0.47	0.51
U1-D	1.13	1.13	22.7	17.3	0.48	0.55
S1-M	1.10	1.11	25.9	22.3	0.48	0.54
U1-M	1.10	1.11	25.1	21.0	0.48	0.53

Table 1. Objective evaluation results

The proposed method was mainly designed for spectrum adaptation. Table 1 confirms that the performance of unsupervised adaptation is comparable to that of supervised adaptation no matter which approach was applied, especially in terms of spectrum. According to Table 1:

(1) Intra-lingual systems provided the best performance in terms of spectrum adaptation, which makes sense as there was no language mismatch.

(2) It is not surprising that S1-T and U1-T provided worse performing spectrum adaptation, because the transforms were estimated on the Mandarin side but used to adjust the English average voice models; there was an obvious language mismatch.

(3) In contrast, mapping rules were applied to the Mandarin adaptation data before transform estimation when the data version of HMM state mapping was used. Since transforms were directly estimated on the Mandarin data and the English average voice models, the language mismatch in S1-D and U1-D could be partly alleviated by the maximum likelihood linear transformation (MLLT) based adaptation algorithm. RMSE of MCEP thus decreased.

(4) In S1-M and U1-M, without any explicit mapping rules, the Mandarin adaptation data was directly associated with PDFs of the English average voice models by prior phonetic knowledge and in an ML-based data-driven manner, respectively. This could be regarded as an automatic, more precise, mapping process. So S1-M and U1-M could be slightly better than S1-D and U1-D in terms of spectrum.

(5) Unfortunately, the great prosody distinction between English and Mandarin meant F_0 adaptation was not nearly as effective.

Speaker	Language	Mean	StD	Min	Max
H	Mandarin	137.9	25.2	72.9	236.3
H	English	128.7	11.8	64.1	222.6
Z	Mandarin	117.9	15.4	58.1	182.1
Z	English	112.0	10.3	59.3	186.1

Table 2. F_0 statistics (Unit: Hz)

Initially we synthesized speech with adapted pitch contours, but unnatural pitch patterns resulting from unsupervised cross-lingual speaker adaptation were perceived during informal listening evaluation. In addition, Table 2 confirms that the prosody of English (i.e. stress-timed & atonal) is distinct from that of Mandarin (i.e. syllable-timed & tonal). Hence, pitch and duration of utterances to be subjectively evaluated were synthesized by the English average voice model set. We then shifted the mean F_0 value of each synthesized pitch contour to that of speech data of the corresponding bilingual speaker (H or Z). So our formal listening test merely focused on the performance of spectrum adaptation.

Our formal listening test consisted of two sections: naturalness and speaker similarity. In the naturalness section, a listener was requested to listen to a natural utterance first and then utterances synthesized by the eight systems each as well as vocoded speech in a random order. Having listened to each synthesized utterance, the listener was requested to score what he/she heard on a 5-point scale of 1 through 5, where 1 meant "completely unnatural" and 5 meant "completely natural". The speaker similarity section was designed in the same fashion, except that a listener was requested to listen to one more utterance which was synthesized directly by the average voice models and the 5-point scale was such that 1 meant "sounds like a totally different person" and 5 meant "sounds like exactly the same person".

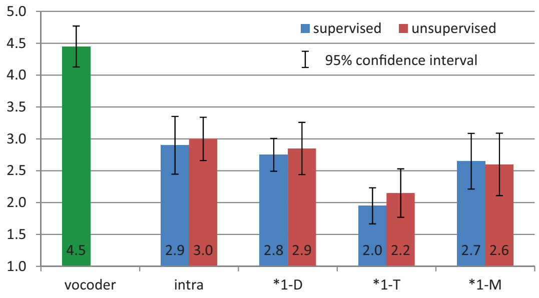

Fig. 2. Naturalness score (speaker H)

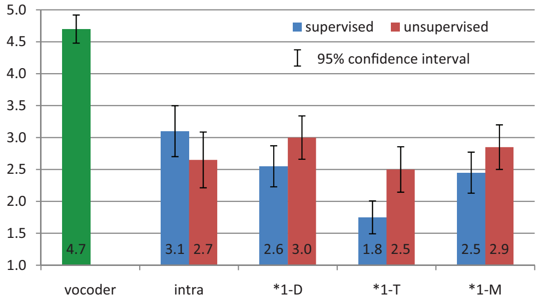

Fig. 3. Naturalness score (speaker Z)

Twenty listeners participated in our listening test. Because of the anonymity of our listening test, only two native English speakers can be confirmed. The results in Figure 2 and Figure 3 suggest that unsupervised cross-lingual speaker adaptation is comparable to or sometimes better than the supervised case in terms of naturalness. We noted that in the case of intra-lingual speaker adaptation with speaker Z's speech adaptation data, the supervised system S2 outperformed the unsupervised one U2. This is probably because speaker Z speaks Mandarin accented English while speaker H has a more natural English accent. In order to avoid the potential effect of non-standard English accents, only speaker H was involved in the speaker similarity evaluation.

It is observed from both objective and subjective evaluation results that for speaker H, *1-D and *1-M followed the intra-lingual adaptation systems closely while *1-T evidently underperformed. Reviewing the analysis of Table 1, we noted the state emission PDFs of *1-D, *1-M and intra-lingual systems for transform estimation were all in English, which was the output language, and that the difference was just language identities of their adaptation data. By contrast, both the emission PDFs and adaptation data of *1-T for transform estimation were in Mandarin, which was not the output language. Hence, it would appear that it is necessary to make sure we use output language distributions for estimation of cross-lingual speaker transforms. The language identity of adaptation data is less important than that of a model set to be adapted.

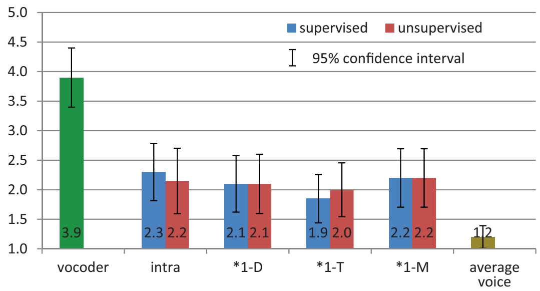

Fig. 4. Similarity score (Mandarin reference uttered by speaker H)

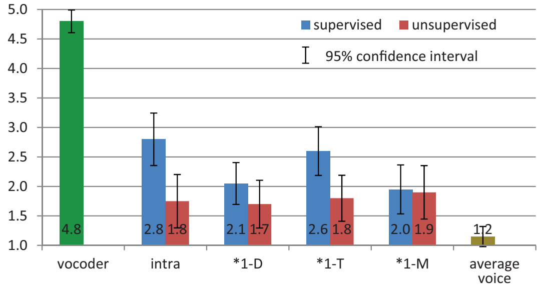

Fig. 5. Similarity score (English reference uttered by speaker H)

The results in Figure 4 were obtained in the EMIME scenario – speaker similarity has to be compared between natural speech in $L1$ and synthesized speech in $L2$. This figure shows unsupervised speaker adaptation is comparable to the supervised case in terms of speaker similarity. However, Figure 5, where both natural and synthesized speech were in English, shows an interesting contrast in that supervised adaptation outperformed the unsupervised case. We attribute this phenomenon to human perception being affected by language mismatch. Namely, because the prompt of a natural English utterance was the same as that of synthesized ones, and thus they were uttered with close prosody, the listeners could more easily perceive how similar/dissimilar a synthesized utterance was to a natural one, and tended to grade supervised adaptation with higher scores. In the case shown by Figure 4, the language mismatch made it more difficult for the listeners to compare a synthesized utterance with a natural one. The listeners didn't think either synthesized utterance (adapted supervisedly or unsupervisedly) sounded more similar/dissimilar to the natural one. This explanation needs to be confirmed by further experiments and analysis.

Comparing with the cross-lingual systems *1-D and *1-M, we didn't observe significantly better performance of the intra-lingual systems. This suggests the MLLT-based speaker adaptation technique is able to compensate for language mismatch between adaptation data and an average voice model set fairly well.

5. CONCLUSION

We implemented unsupervised cross-lingual speaker adaptation by combining recently developed decision tree marginalization and HMM state mapping techniques. It was observed that unsupervised cross-lingual speaker adaptation was comparable to the supervised case in terms of spectrum adaptation in the EMIME scenario. We have observed language mismatch is the main problem for cross-lingual speaker adaptation, so introducing some extra techniques to alleviate the mismatch before speaker adaptation would be helpful. Since prosody plays an important role in voice characteristics as well, we may need to pay more attention to improving prosody adaptation in order to deal with two dissimilar languages.

6. ACKNOWLEDGEMENTS

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

The authors would like to thank Dr. Junichi Yamagishi for providing HTS training scripts, Dr. Mirjam Wester for useful discussion concerning evaluation of cross-lingual speaker adaptation, Mr. Saheer Babu for making web pages for subjective listening tests, and Mr. Philip Garner for proofreading this manuscript.

7. REFERENCES

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

[2] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in *Proc. of Eurospeech*, Sept. 1999, pp. 2347–2350.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", in *Proc. of ICASSP*, June 2000, pp. 1315–1318.

[4] J. Dines, L. Saheer, and H. Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves", in *Proc. of Interspeech*, Sept. 2009, pp. 1395–1398.

[5] 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*, Sept. 2009, pp. 528–531.

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

[7] 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.