ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2008)
Kunming, China
December 16-19, 2008

CROSS-LINGUAL SPEAKER ADAPTATION FOR HMM-BASED SPEECH SYNTHESIS

Yi-Jian Wu[†], *Simon King*[‡], *Keiichi Tokuda*[†]

[†] Nagoya Institute of Technology, Japan
[‡] Centre for Speech Technology Research, University of Edinburgh, UK
yjwu@sp.nitech.ac.jp, Simon.King@ed.ac.uk, tokuda@nitech.ac.jp

ABSTRACT

This paper explores a cross-lingual speaker adaptation technique for HMM-based speech synthesis, where a source voice model for English is transformed into a target speaker model using Mandarin Chinese speech data from the target speaker. A phone mapping-based method is adopted to map Chinese Initial/Finals into English phonemes and two types of mapping rules, including one-to-one and one-to-sequence mappings, are compared. In order to avoid having to map prosodic features between languages, the adaptation procedure uses regression classes and transforms that are constructed for triphone models, then used to adapt the phonetic-and-prosodic-context-dependent models. From the experimental results, we found that a one-to-sequence phone mapping is better than a one-to-one mapping, and that the similarity between adapted English speech and target Chinese speaker is reasonable.

Index Terms— Speaker adaptation, cross-lingual, HMM-based speech synthesis

1. INTRODUCTION

Spoken language translation (SLT) systems have been under development for many years. The aim of a spoken language translation system is to recognize speech from a speaker in a source language, translate it to a target language and then produce corresponding speech using a text-to-speech technique. In a recently started European FP7 project – Effective Multilingual Interaction in Mobile Environments (EMIME) [1] – we are developing methods to personalize such SLT systems. In particular, the synthesized speech in the target language should sound like the input speaker, even though that speaker can not speak the target language. This problem has been previously explored by others in the TC-Star project [2], using cross-lingual voice conversion techniques [3], and the related problems had also been investigated in multi-lingual speech synthesis [4].

Our method uses the unique capabilities of HMM-based speech synthesis [5, 6]. One of these is the ability to adapt the models in order to modify the characteristics of the synthesized speech, including speaker identity, speaking style, and so on. This is achieved by modifying the HMM parameters using model adaptation technique. Several model adaptation algorithms, which were originally proposed for speech recognition, including Maximum a Posteriori (MAP), Maximum Likelihood Linear Regression (MLLR) [7], Constrained MLLR (CMLLR) [8], and so on, have been applied to HMM-based speech synthesis [9, 10]. It has been demonstrated that speaker adaptation of an "Average Voice" model [11] is superior to speaker adaptation of a speaker-dependent model. In our framework, an Average Voice model trained from the speech data of several speakers is used as the initial model for speaker adaptation.

In this study, we explore a cross-lingual speaker adaptation technique for HMM-based speech synthesis, where a source Average Voice model for one language (English) is transformed into a speaker-specific model using adaption data from the target speaker in another language (Mandarin Chinese). The adapted model can be used to synthesize English, with the speaker characteristics of the target speaker. Note that only Mandarin speech data are required for the target speaker.

To realize such cross-lingual speaker adaption, one simple approach would be to regard the Chinese adaptation data as English data. The labels for this data would be obtained by mapping the Chinese labels to appropriate English labels, and then apply model adaptation in the usual way. However, since "full context" labels, which include both phonetic and prosodic information, are used in HMM-based speech synthesis, we would need to map not only phonetic categories, but also prosodic labels from Chinese to English. For the phonetic mapping, i.e., mapping Chinese Initials/Finals to English phonemes, we designed two types of mapping rules – one-to-one and one-to-sequence mappings – by considering the phonetic definition of these units in the IPA [12] and their acoustic realizations. For the prosodic label mapping, it is extremely hard to design a mapping between Chinese and English, since some of the prosodic features in Chinese and English are quite different. In order to avoid using a prosodic feature mapping, we use an ingenious adaptation procedure in which the regression classes and transform matrices are built for triphone models, and then applied to full context models [15]. These are models of phonemes in a a particular phonetic context (two preceding and two following phones) and a particular prosodic context (various features are used, e.g., stress of current and surrounding syllables, position in the utterance, etc.).

This strategy has another important advantage when used in an unsupervised fashion: it only requires phonetic labels to be automatically recognized for the adaptation data. It would be considerably more difficult to automatically recognize the prosodic labels with sufficient accuracy.

The rest of this paper is organized as follows. In section 2, we first briefly review speaker adaptation within one language and then present the details of our phone mapping based method for cross-lingual speaker adaptation. In section 3, we describe the experiments used to evaluate the performance of the proposed cross-lingual speaker adaptation method and present the results. Finally, our conclusions and suggestions for future work are given in section 4.

2. FROM INTRA-LINGUAL TO CROSS-LINGUAL SPEAKER ADAPTATION

2.1. Intra-lingual speaker adaptation

Intra-lingual speaker adaptation (usually just called "speaker adaptation"), transforms a source model to a target speaker using a limited amount of speech data from the target speaker. Initially developed for use in HMM-based speech recognition, many model adaptation

978-1-4244-2942-4/08/$25.00 (c) 2008 IEEE

Chinese Initial/Final	One-to-one mapping	One-to-sequence mapping
/f/	/f/	/f/
/sh/	/s/	/s r/
/ai/	/ay/	/ay/
/iao/	/aw/	/ih aw/
/iong/	/oo/	/ih oo ng/

Table 1. Examples of the mappings from Chinese Initial/Finals to English phonemes

algorithms, including MAP, MLLR, CMLLR, etc., have been proposed. The purpose of speaker adaptation for speech recognition is to reduce the mismatch between source model and target speaker, and thus improve the recognition accuracy for the target speaker. Adaptation can be supervised (i.e., the correct labels are available for the adaptation data) or unsupervised (where the labels for the adaptation data must be obtained automatically, by using the unadapted models to perform ASR, for example).

In HMM-based speech synthesis, speaker adaptation techniques can be used to adapt the source model using speech data from target speaker, and thus make the speech synthesized from the adapted model sound like the target speaker. Several adaptation algorithms have been borrowed from speech recognition and further developed [10] for HMM-based speech synthesis. Since the purpose of speaker adaptation for speech synthesis is different from that for speech recognition, a speech synthesis-specific adaptation algorithm, called Minimum Generation Error Linear Regression (MGELR), has also been proposed [13]. The use of speaker adaptive training (SAT) to construct the Average Voice model has also been found to improve performance [14].

2.2. Cross-lingual speaker adaptation

In this paper, we adopt a phone mapping based method for cross-lingual (English to Chinese) speaker adaptation. In this method, we first map the Chinese context labels into English context labels, and then apply the model adaption technique in a similar way to intra-lingual speaker adaptation.

2.2.1. Phonetic label mapping

The phonetic label mapping between Chinese and English is achieved by mapping Chinese Initials/Finals to English phonemes. There are two basic ways to obtain such a phone mapping. The first way is to manually design the mapping rules by considering the phonetic definition of the units in a universal phoneme set (such as IPA [12]) and their acoustic realization. The other way is to calculate the distance between the phonetic units using statistics from speech data. The latter method usually requires a bilingual speech corpus uttered by the same speaker. However, such a corpus is not available here. Therefore, we chose the first method and manually designed two sets of mapping rules:

- One-to-one mapping: map one Chinese Initial/Final to one English phone.
- One-to-sequence mapping: map one Chinese Initial/Final to a sequence of English phones.

Some examples of the mappings are shown in Table 1. Chinese Initials can mostly be directly related to one English consonant. Therefore, the one-to-one and one-to-sequence mappings are the same for most Chinese Initials except /zh/, /ch/, /sh/. For Chinese Finals, the mapping is more difficult and complicated, since one Chinese Final usually consist of several vowels/nasals. In the one-to-one mapping, we map one Chinese Final to one English phone by considering the main or central part of the Chinese Final. In the one-to-sequence mapping, we decompose the Chinese Final into several vowels/nasals, and map each of them to an English phoneme.

Both sets of mappings rules have their own advantages and disadvantages. Although the one-to-one mapping is not accurate enough for some Chinese Finals, the number of states in the resulting HMM sequence after mapping is appropriate, since one 5-state HMM is usually used for a single Chinese Final or a single English phoneme in HMM-based speech synthesis. The one-to-sequence mapping is more phonetically accurate, but may result in an inappropriate number of states in the HMM sequence after mapping. For example, the Chinese Final /iong/ can be mapped to the English phone sequence /ih oo ng/. This results in a 15-state model for /iong/ which was originally modeled by a 5-state model.

It should be noted that not all English phonemes occur in the mapping rules (e.g., /th/) which means there will be no adaptation data directly related to the models for this phoneme. However, these models can still be adapted using MLLR-based adaptation frameworks, because adaptation data related to other models that are in the same regression class (or parent class) as models for /th/ will be used. The Chinese data, after mapping the phone labels to English, will have a different distribution of phonemes than we would find in actual English data. This may adversely affect the adaptation performance.

2.2.2. Prosodic label mapping

Chinese is a tonal language and English is an accent language. It is therefore difficult to map Chinese tones to corresponding appropriate English prosodic labels. Furthermore, the syllable structure of Chinese is quite different from that of English. One Chinese syllable consists of an Initial and a Final, which may includes several vowels. In order to avoid the need to construct a prosodic mapping between the two languages, we used a method similar to that in [15], where an unsupervised adaptation for HMM-based speech synthesis was conducted without recognizing prosodic labels.

2.2.3. Adapting full context models

For each full context dependent model, we can obtain the corresponding triphone model by ignoring the prosodic contextual factors and dropping some phonetic contextual factors. During training, we construct a set of regression classes and a regression tree for triphone models. We then train the transform matrices using Chinese speech data with English phonetic labels obtained by one of the two mapping methods described earlier. These trained transform matrices cannot be applied directly to the full context models with tied parameters, since the tying structure of those models may be incompatible with the regression classes for triphone models. For example, two full context model parameters may be tied, but be in different regression classes. The solution is to simply to untie the clustered models, either completely or at least enough to ensure no tied group of parameters contains members in more than one regression class.

2.2.4. Adaptation procedure

In summary, the cross-lingual adaptation procedure is:

1. Train a set of English Average Voice full context models.
2. Create a set of English triphone models by untying, reclustering and retraining these English full context models.

3. Construct the regression classes and regression tree for these triphone models.
4. Partially untie the English full context models, so that they become compatible with these regression classes.
5. Map the Chinese phonetic labels for the Chinese adaptation data to English phonetic labels, and thus obtain English triphone labels for the Chinese adaptation data.
6. Train the transform matrices for the English triphone models using these data.
7. Adapt the English full context models using these transforms.

3. EXPERIMENTS

3.1. Experimental setups

Data taken from the CMU-ARCTIC English database [16] – about 1 hour of speech data from each of 4 males (awb, bdl, rms, jmk) and 1 female (clb) – was used to train the English Average Voice model The Chinese speech database from the Blizzard Challenge 2008 [17] was used as the target speaker data. All speech waveforms were sampled at a rate of 16KHz. The acoustic features, including F0 and mel-cepstral coefficients, were extracted with a 5ms shift. The feature vector consists of static features, including 25-th order mel-cepstral coefficients, log F0, their delta and delta-delta coefficients. A 5-state left-to-right no-skip HMM was used to model each English phoneme (or Chinese Initial/Final), and MSD-HMMs [18] were used for F0 modeling. The tools and scripts from HTS-2.1 [19] were used for model training and adaptation. The CMLLR-based method was adopted for model adaptation in the experiment. For synthesis, a Mel Log Spectrum Approximation (MLSA) filter [20] was used to generate the speech waveform. We investigated several different configurations for adaptation:

1. Different amounts of adaptation data: 10, 100 or 1000 Chinese utterances for adaptation.
2. Adaptation of only some features: duration, F0 or Mel-cepstral coefficients.
3. Different phone mapping rules: one-to-one mapping or one-to-sequence mapping.

3.2. Experimental Results

3.2.1. Different amount of adaption data

In an initial informal listening test, the quality of synthesized speech after cross-lingual speaker adaptation appeared to be reasonable even when using only 10 Chinese utterances for adaptation. Increasing the amount of adaptation data from 10 utterances to 1000 utterances causes the synthesized speech became more stable and clearer. However, the quality of synthesized speech after cross-lingual speaker adaptation is still worse than the synthesized speech from the source average-voice model. In the remaining experiments, 100 Chinese utterances were used for cross-lingual speaker adaptation.

3.2.2. Adaptation of different acoustic features

There is no doubt that the adaptation of the spectrum is very effective for changing the speaker characteristics to the target speaker. When we only adapted the model parameters for spectral features and simply scaled the pitch range to that of the target speaker, the synthesized speech became similar to the target speaker.

The adaptation of F0 models introduced some tonal effect in the synthesized English speech. Whether such an effect is positive or not depends on the requirements of the application. Although it makes the prosody a little unnatural, compared to a native speaker of English, it does give the impression of English being spoken by a native speaker of Chinese.

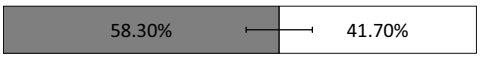

Fig. 1. Preference scores for different phone mapping

The adaptation of duration models reduced the quality of the synthesized speech, especially when using the one-to-sequence phone mapping. Prosodic structure is very different in Chinese and English, so we conclude that adaptation of the duration model does not make much sense in the proposed cross-lingual speaker adaptation scheme.

3.2.3. Effect of GV

The Global Variance (GV) parameter generation algorithm [21] had been shown to be effective in improving the quality of synthesized speech in HMM-based speech synthesis. We evaluated the effect of this technique for parameter generation using the adapted models. In our experiments, we found that using GV for Mel-ceptral parameter generation dramatically improves the quality of synthesized speech from the adapted model; this is consistent with the result in [21]. However, using GV for F0 parameter generation was not effective: the prosody became unnatural. One of the reasons for this may be that the global variance of F0 for one speaker varies, depending on the language being spoken. Typically, the GV of F0 in Chinese speech data is larger than that in English speech data from the same speaker, because of the dynamic range required to express tonal structure in Chinese. Therefore, it may not be appropriate use a GV model from Chinese F0 data for F0 parameter generation in English.

3.2.4. Different phone mapping

In order to compare the adaptation performances using different phone mapping rules, a formal listening preference test was conducted. 40 sentences, which were not included in the training data, were synthesized from the adapted models using each of the two phone mapping methods. Only the spectral parameters of the models were adapted, and the generated F0 trajectory was scaled to have the pitch range of the target speaker. GV was only applied to the spectral parameters. Eight listeners were presented pairs of synthesized speech, and asked which one sounded best.

Fig. 1 shows the preference score, with the horizontal line indicating the 95% confidence interval. The one-to-sequence phone mapping resulted in slightly better performance than the one-to-one mapping. Listening to the synthesized speech samples, we found that the speech synthesized using the one-to-sequence mapping was clearer but sometimes became unstable. As we mentioned in Sec. 2.2.1, the one-to-one mapping is not accurate for mapping some Chinese Finals to one English phoneme. For example, the Chinese Final /iong/ is mapped to the English vowel /oo/ in the one-to-one mapping, which means the speech data for /iong/ is used for adaptation of models of /oo/. This inaccurate mapping means that inappropriate speech data are used for model adaption, resulting in "muffled" synthetic speech. The problem of an inappropriate number of states resulting from the one-to-sequence mapping may introduce instability, because a lengthy sequence of states must be aligned with a relatively short region of adaptation data.

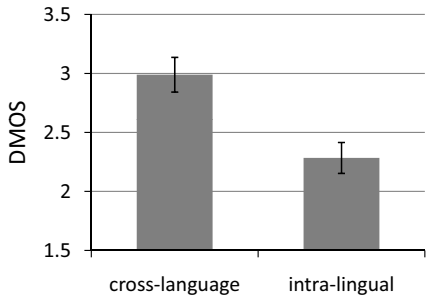

Fig. 2. DMOS scores after speaker adaptation

3.2.5. Speaker similarity

Finally, we conducted formal listening tests to evaluate the similarity in speaker identity between the synthesized speech after cross-lingual speaker adaptation and the speech of the target speaker. In order to remove the influence of the vocoder (i.e., parameterization followed by reconstruction of the speech using the MLSA filter), we used synthetic speech from a speaker-dependent model, rather than natural speech from the target speaker. Since no English speech data for the target speaker were available, we trained a model for Chinese. Therefore, Chinese utterances were compared to English utterances in the listening test.

For comparison, a simple method for adapting the English Average Voice model to the target Chinese speaker was used. 100 utterances of speech data from an English speaker ('slt' from the ARCTIC database) whose voice characteristics are similar to the target Chinese speaker were used to adapt the Average Voice model.

We used the same 40 sentences as in the previous listening test, and synthesized speech from both intra-lingual (using the English speech of 'slt') and cross-lingual (using the Chinese speech of the target speaker) adapted models. Eight listeners were presented with pairs of synthesized speech samples (firstly one utterance from the speaker-dependent Chinese model and then one utterance from the adapted English model) and asked to give a DMOS score to each English speech sample. Other conditions of the listening test were the same as in the previous test. The results are shown in Figure 2. The similarity of the synthetic speech to the target speaker after cross-lingual speaker adaptation is better than that obtained by intra-lingual speaker adaptation (which uses the wrong target speaker). However, the quality of synthetic speech generated by cross-lingual speaker adaptation is worse than that from intra-lingual speaker adaptation.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have described some initial experiments in cross-lingual (English-Chinese) speaker adaptation technique for HMM-based speech synthesis. A phone mapping based method is introduced, where two sets of phonetic label mapping rules including one-to-one and one-to-sequence mapping are designed, and an ingenious adaptation procedure is adopted to avoid prosodic label mapping. From the experimental results, the one-to-sequence phone mapping is better than the one-to-one mapping, and the similarity between the adapted English speech and the target Chinese speaker is reasonable. Future work is to apply a state mapping instead of a phone mapping for cross-lingual speaker adaptation.

5. ACKNOWLEDGEMENTS

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

6. REFERENCES

[1] EMIME project: http://www.emime.org

[2] TC-Star project: http://www.tc-star.org

[3] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney and J. Hirschberg, "TC-Star: Cross-language voice conversion revisited," in *Proc. of the TC-Star Workshop 2006*, Spain, 2006.

[4] J. Latorre, K. Iwano and S. Furui, "New approach to polyglot synthesis: how to speak any language with anyone's voice," in *Proc. of Multilingual Speech and Language Processing*, 2006.

[5] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. of ICASSP*, pp. 389-392, 1996.

[6] 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 ICASSP*, vol. 5, pp. 2347-2350, 1999.

[7] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," in *Computer Speech and Language*, vol.9, no.2, pp. 171-185, 1995.

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

[9] T. Masuko, K. Tokuda, T. Kobayashi and S. Imai, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 273-276, 1998.

[10] J. Yamagishi, K. Ogata, Y. Nakano, J. Isogai and T. Kobayashi, "HSMM-based model adaptation algorithms for average-voice-based speech synthesis," in *Proc. of ICASSP*, pp. 77-80, May 2006.

[11] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," in *IEICE Trans. of Fundamentals*, vol. E86-A, no. 8, pp. 1956-1963, 2003.

[12] http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[13] L. Qin, Y.-J. Wu, Z.-H. Ling, R.-H. Wang and L.-R. Dai, "Minimum generation error lineal regression based model adaptation for HMM-based speech synthesis," in *Proc. of ICASSP*, pp. 3953-3956, Mar. 2008.

[14] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "A Training Method for Average Voice Model Based on Shared Decision Tree Context Clustering and Speaker Adaptive Training," in *Proc. ICASSP 2003*, vol. 1, pp.716-719, 2003.

[15] S. King, K. Tokuda, H. Zen and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. of Interspeech* (accepted), 2008.

[16] J. Kominek and A. Black, "The CMU ARCTIC speech databases for speech synthesis research," Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177, http://festvox.org/cmu_arctic/, 2003.

[17] http://www.synsig.org/index.php/Blizzard_Challenge_2008

[18] K. Tokuda, T. Masuko, N. Miyazaki and T. Kobayashi, "Hidden markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, pp. 229-232, 1999.

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

[20] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. of ICASSP*, pp. 93-96, 1983.

[21] T. Toda and K. Tokuda, "Speech parameter generation algorithm considering global variance for HMM-based speech synthesis," in *Proc. of Interspeech*, pp. 2801-2804, 2005.