A CROSS-LANGUAGE STATE MAPPING APPROACH TO BILINGUAL (MANDARIN-ENGLISH) TTS

Hui Liang[*] [1,2], *Yao Qian*[1], *Frank K. Soong*[1], *Gongshen Liu*[2]

[1]Microsoft Research Asia, Beijing, P. R. China
[2]School of Information Security Engineering, Shanghai Jiao Tong University, P. R. China
hui.ts.liang@gmail.com, {yaoqian,frankkps}@microsoft.com, lgshen@sjtu.edu.cn

ABSTRACT

We propose a cross-language state mapping approach to HMM-based bilingual TTS. Two language-dependent decision trees are built first with a bilingual speech database recorded by a single speaker. A state mapping for every leaf node in the decision tree of a target language is created by finding the nearest leaf node in the tree of a source language. Kullback-Leibler divergence between two distributions is used to find the nearest leaf node. To synthesize target language speech by a monolingual, (source language) speaker's voice, we find HMM parameters trained by the monolingual (source language) speaker in the mapped leaf nodes. Similar mappings can be constructed by reversing the source and target languages. With these bi-directional cross-lingual mappings, we can synthesize bilingual or mixed-code speech by HMMs trained by any monolingual speaker. High voice (speaker) similarity is preserved in synthesized speech of the target language. Two perceptual tests on synthesized Mandarin speech confirms high intelligibility with a Chinese character transcription accuracy of 92.1% and an MOS score of 3.08.

Index Terms — Bilingual, state mapping, new language synthesis, HMM-based TTS

1. INTRODUCTION

With globalization of today's world, many telecommunication applications, e.g. information inquiry, reservation and ordering, and reading emails by TTS, demand a multilingual TTS system, in which one engine can synthesize multiple or even mixed-languages by the same voice. In foreign language learning, a multilingual TTS system can be a useful learning aid for foreign language learners. The aid can be even more attractive if sentences in a foreign language can be synthesized in a learner's own voice.

There are many studies on multilingual TTS systems [1-8]. In [1], a universal algorithm was proposed to synthesize multiple languages. Most multilingual TTS systems, which are based on a concatenation technique, use a pre-recorded multilingual corpus uttered by the same speaker and share a common unit selection module across different languages, while language-specific processing, e.g., phone and text analysis, is not shared. In [7], a synthesizer can synthesize mixed phonetic transcriptions in different languages by monolingual voices. It uses a phoneme mapping algorithm, which is based upon similarity of phonetic-articulation features between phonemes specified by IPA. Recently, HMM-based TTS has been successfully applied to TTS synthesis of many different languages [9]. An HMM is a statistically trained parameterized model. Spectral envelopes, fundamental frequencies and state durations are modeled simultaneously by corresponding HMMs. For a given text sequence, speech parameter trajectories and corresponding signals are generated from trained HMMs in the Maximum Likelihood (ML) sense.

Acoustic similarity between two phones can be measured in Kullback-Leibler divergence (KLD) between corresponding HMMs. KLD provides a useful measure to facilitate parameter sharing and mapping among HMMs. In [8], an average voice is first trained by using speech data of several speakers in different languages. The average voice model is then adapted to a specific speaker. As a result, the specific speaker can then be trained to speak all the languages in the training set.

Recently we proposed to build an HMM-based, Mandarin and English, bilingual TTS system [10]. In this system, we use a bilingual corpus recorded by a single speaker and construct a new, mixed-language TTS with both language-specific and language-independent questions to facilitate phone sharing across the two languages. However, to build a large inventory of bilingual voice-fonts of multi-speakers is not trivial when we need to find speakers who are fluent in both languages to record their bilingual voices.

2. AN HMM-BASED BILINGUAL TTS SYSTEM

A conventional, rather straightforward approach to bilingual HMM-based TTS is to build two monolingual HMM systems by sharing common phones between two languages. To maintain a universal voice quality, usually a corpus of two covered languages is recorded by one bilingual speaker. To reduce a recording effort and to maximize training efficiency of a speech database, the smallest possible phone set which covers all the phones of two languages is used for training HMMs. For a case of the study of Mandarin-English bilingual TTS, we realize that Mandarin is a tonal language in the Sino-Tibetan family, while English is a stress-timed language in the Indo-European family. In terms of phonemes (IPA symbols) of the two languages, only a small percentage of phonemes, i.e., eight consonants and two vowels can be shared [10]. To improve possible sharing, we argue that despite the difference between phonemic structures of the two languages, it may still be possible to find more common acoustic attributes at a granular, sub-phonemic level. Numerous allophones, which are used in specific phonetic contexts, provide more chances of sharing HMM states between Mandarin and English. We have

[*] An intern in the Speech Group, Microsoft Research Asia

1-4244-1484-9/08/$25.00 ©2008 IEEE

proposed to share context-dependent HMM states for constructing a bilingual (Mandarin-English) TTS system [10] as illustrated in Figure 1.

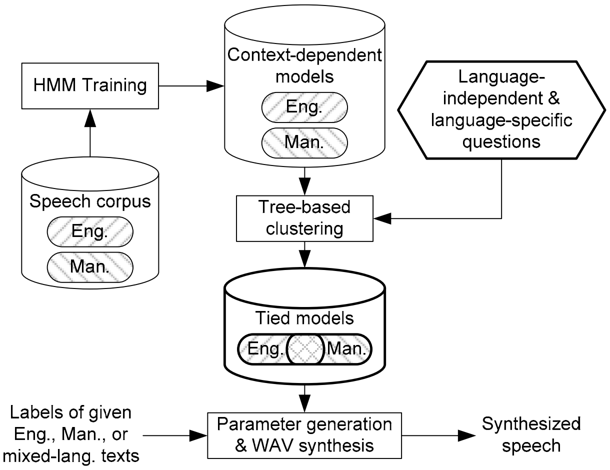
HMM Training
Context-dependent models
Eng.
Man.
Language-independent & language-specific questions
Speech corpus
Eng.
Man.
Tree-based clustering
Tied models
Eng.
Man.
Labels of given Eng., Man., or mixed-lang. texts
Parameter generation & WAV synthesis
Synthesized speech

Fig. 1. The schematic diagram of an HMM-based bilingual (Mandarin-English) TTS system

In this system, the training corpus consists of Mandarin and English sentences recorded by a bilingual speaker and a universal phone set, or a union of all the phones in English and Mandarin is used. Phone models of rich contexts, e.g. tri-phone, or models with even longer contexts like phone positions and POS, are used to capture acoustic co-articulation effects in HMM-based TTS. However, limited by the available training data, we have to tie the models of rich contexts into generalized ones so as to predict unseen contexts in testing. State tying via a clustered decision tree is used for this purpose. To train bilingual HMM TTS systems, states from different central phones across different languages are allowed to be tied together. Questions used in growing decision trees include:

a) language-independent questions: e.g. *Velar_Plosive*, Does the state belong to velar plosive phones, which contain /g/ (Eng.), /k^h/ (Eng.), /k/ (Man.) or /k^h/ (Man.)?
b) language-specific questions: e.g. *E_Voiced_Stop*, Does the state belong to English voiced stop phones, which contain /b/, /d/ and /g/?

According to the manners and places of articulations, supra-segmental features, etc., we construct the questions to tie states of English and Mandarin phone models together.

The new bilingual TTS system with context-dependent HMM state sharing across languages outperforms a simple baseline system where two individual, language-dependent HMMs are trained separately [10]. The new system has a smaller, about 40% less, footprint than the baseline system. Quality wise, the new system is either the same for non-mixed, Mandarin or English synthesis as the baseline or much better for mixed-language synthesis. The higher quality of mixed-language synthesis is confirmed by favorable subjective preference test results, 60.2% vs 39.8% (α = 0.001, CI = [0.1085, 0.3004]) [10].

3. STATE MAPPING FOR NEW LANGUAGE SYNTHESIS

To build a bilingual TTS system by recording a bilingual, single speaker database is no longer feasible if such a speaker is not available. Also, it is academically interesting to investigate how to synthesize a target language when only monolingual(source language) recordings from a desired speaker is available.

Various speaker adaptation techniques have been successfully applied to HMM-based speech recognition. For synthesis, we can adapt bilingual HMMs to any specified monolingual speaker by means of supervised Maximum Likelihood Linear Regression (MLLR) adaptation [11]. For a small amount of adaptation data, a global transform can be generated and applied to every Gaussian component in a model set. As more adaptation data becomes available, the Gaussian components can be grouped into broad phonetic classes such as silence, glides, stops, nasals, etc. Then class specific transforms can be constructed. However, it is difficult to do speaker adaptation across different languages, especially when two languages are phonetically distant and very few phones can be shared.

We propose a tied, context-dependent state mapping between two languages. The mapping is established first from a bilingual speaker and then used to synthesize target language speech from another monolingual (source language) speaker's voice. Details are given in the following subsections.

3.1. Establishing State Mapping across Languages

A tied, context-dependent state mapping between two languages needs to be established with a bilingual speech database recorded by a single speaker. Two language-specific decision trees are created separately with English and Mandarin data in the bilingual database. Each leaf node in the Mandarin decision tree has a mapped leaf node, in the minimum Kullback-Leibler divergence (KLD) sense, in the English tree. The state mapping (from Mandarin to English) is shown in Figure 2. This directional mapping can have more than one leaf nodes in the Mandarin (target) tree mapped to the same leaf node in the English (source) tree as shown in the figure. Mappings from English to Mandarin can be similarly done in a reverse direction.

KLD is an information-theoretic measure of (dis)similarity between two probability distributions. When the temporal structure of HMMs is aligned by dynamic programming, KLD can be further modified to measure the difference between HMMs of two evolving speech sounds [12,13]. For two given distributions P and Q of continuous random variables, a symmetric form of KLD between P and Q is:

$$D_{KL}(P,Q) = \int p(x)\log\frac{p(x)}{q(x)}dx + \int q(x)\log\frac{q(x)}{p(x)}dx \qquad (1)$$

where p and q denote the densities of P and Q. For two multivariate Gaussian distributions, Eq. (1) has a closed form:

$$\begin{aligned} D_{KL}(P,Q) = \frac{1}{2}\mathrm{tr}\{(\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Sigma}_q^{-1})(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^{\mathrm{T}} \\ + \boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_q^{-1} + \boldsymbol{\Sigma}_q\boldsymbol{\Sigma}_p^{-1} - 2\boldsymbol{I}\} \end{aligned} \qquad (2)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are corresponding mean vectors and covariance matrices, respectively.

In HMM-based speech synthesis, spectrum, pitch and duration are separated into three streams and stream-dependent models are built to cluster those three features into separate decision trees. Spectrum and duration features are modeled by

HMMs. We apply Eq. (2) to measure the similarity between two tied states of HMMs. Pitch features are modeled by MSD-HMM, which was proposed to model two, discrete and continuous, probability spaces, discrete for unvoiced segments and continuous for voiced F0 contours [14]. The upper bound of KLD between two states of MSD-HMMs is written as:

$$\begin{aligned} D_{KL}(P,Q) \le\ & (w_0^p - w_0^q)\log\frac{w_0^p}{w_0^q} + (w_1^p - w_1^q)\log\frac{w_1^p}{w_1^q} \\ & + \frac{1}{2}\mathrm{tr}\{(w_1^p \boldsymbol{\Sigma}_p^{-1} + w_1^q \boldsymbol{\Sigma}_q^{-1})(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^{\mathrm{T}} \\ & + w_1^p(\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_q^{-1} - \boldsymbol{I}) + w_1^q(\boldsymbol{\Sigma}_q\boldsymbol{\Sigma}_p^{-1} - \boldsymbol{I})\} \\ & + \frac{1}{2}(w_1^q - w_1^p)\log\left|\boldsymbol{\Sigma}_p\boldsymbol{\Sigma}_q^{-1}\right| \end{aligned} \tag{3}$$

where w_0 and w_1 are the prior probabilities of unvoiced and voiced subspaces, respectively. Both English and Mandarin have trees of spectrum, pitch and duration. Each leaf node of those trees is used to establish a mapping between English and Mandarin.

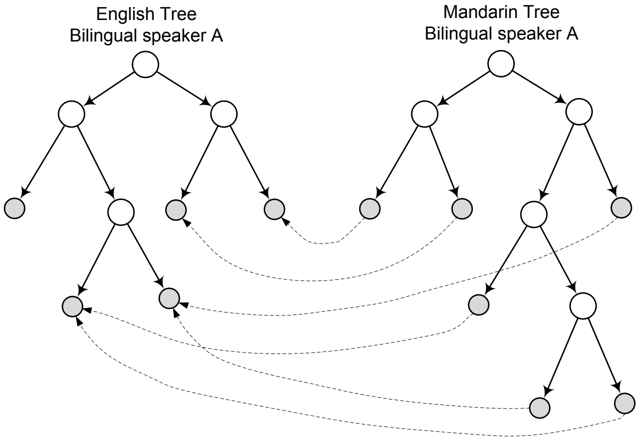

Fig. 2. Cross-language state mapping from a Mandarin decision tree to an English one.

3.2. Using State Mapping for a Monolingual Voice

To synthesize speech in a target language without pre-recorded data from a desired speaker who is monolingual in a source language, we have to utilize the state mapping established previously with data of a bilingual speaker. A context-dependent, state mapping trained from speech data of a bilingual (English-Mandarin) speaker A is used to choose appropriate states trained from speech data of a different, monolingual English (source language) speaker B to synthesize Mandarin (target language) sentences as illustrated in Figure 3.

In the training phase, the same decision tree structure used for the bilingual speaker A (on his English data) is shared in training models for the monolingual, English speaker B. In other words, the data of speaker B traverse through the English decision tree of speaker A by following the same contextual questions at each tree-node splitting to update the mean and variance accordingly. In synthesizing a Mandarin sentence with the monolingual English speaker's model, appropriate contextual labels of the input Mandarin sentence are firstly retrieved by traversing the Mandarin decision tree trained by speaker A with the corresponding leaf nodes. Via state mappings between the two decision trees, HMM parameters at the leaf nodes of the English tree of speaker B can be retrieved. Similarly, the state mappings from English to Mandarin can be used for synthesizing English sentences with a monolingual, Mandarin speaker C's voice.

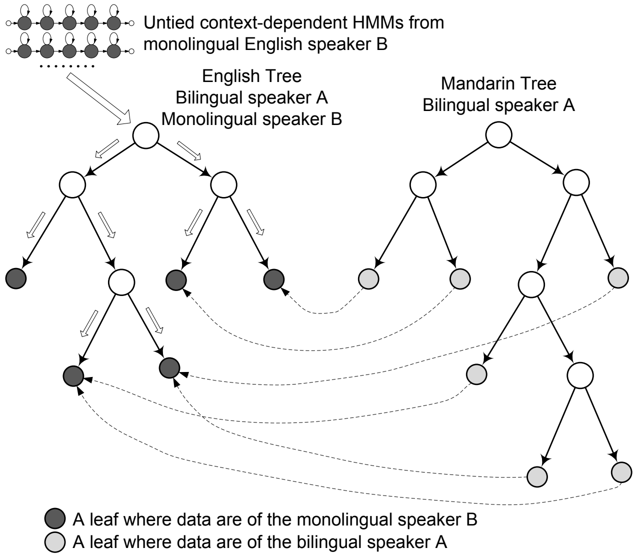
Untied context-dependent HMMs from
monolingual English speaker B
English Tree
Bilingual speaker A
Monolingual speaker B
Mandarin Tree
Bilingual speaker A
A leaf where data are of the monolingual speaker B
A leaf where data are of the bilingual speaker A

Fig. 3. Synthesizing Mandarin sentences by using a monolingual English speaker B's voice via cross-language state mappings.

4. EXPERIMENTS AND EVALUATIONS

4.1. Experimental Setup

Two corpora, both phonetically and prosodically rich, in broadcast news style are used in our experiments. One is a bilingual (Mandarin and English) corpus, recorded by a female speaker. The other one is a monolingual English corpus recorded by a female speaker of American English. Table 1 lists the number of sentences in these two corpora. Testing data consist of 70 Mandarin sentences. Speech signals are sampled at 16 kHz, windowed by a 25-ms window with a 5-ms shift, and LPC spectral features are transformed into 40th-order LSPs and their dynamic counterparts. Five-state, left-to-right HMMs with single, diagonal variance Gaussian distributions are adopted for training phone models.

Table 1. The number of sentences in the two corpora

Corpus/Speaker	bilingual	monolingual
Mandarin	1,000	N/A
English	1,024	1,503

We use two different methods to synthesize Mandarin speech with the voice of the monolingual English speaker.

Method I: Speaker adaptation for bilingual HMMs

The bilingual corpus is used to build an HMM-based bilingual (Mandarin-English) TTS by means of context-dependent state sharing presented in Section 2. We use monolingual English sentences to adapt the bilingual HMMs in the MLLR sense for Mandarin synthesis.

Method II: Across-language state mapping
With two separate, language-specific decision trees built for Mandarin and English data in the bilingual corpus, mappings from leaf nodes of the Mandarin tree to those of the English tree are established as described in Section 3.1. Mandarin LSP trajectories are generated with parameters of the monolingual English voice in mapped states.

4.2. Evaluation Results and Analysis

In Method I we adapted the bilingual speaker's HMMs toward the monolingual (English) speaker's voice. However, in cross-language speaker adaptation there may be only limited phonetic units shared between the two languages. Consequently, even with ample adaptation data from the monolingual (source language) speaker, HMMs in the target language may still not be adequately adapted. With 500 sentences as adaptation data from the monolingual (source language) speaker to construct 32 transformation MLLR matrices, synthesized Mandarin speech still sounds similar to the original bilingual speaker. Since the adaptation based Method I is not successful, we switch to Method II of cross-language state mapping based target language synthesis.

In an oracle experiment with Method II where we use the English model of the bilingual speaker to synthesize Mandarin sentences, we notice that the dynamic range of F0 contours predicted by cross-language mapped HMMs is smaller than that by monolingual Mandarin HMMs. After analyzing the English and Mandarin training data, we find that the dynamic range of F0 in Mandarin sentences is much larger than that in English. F0 means and standard deviations of the two corpora are shown in Table 2. The larger variance of Mandarin F0 is partially due to the lexical tones of Mandarin where the intrinsic variations in four (or five) lexical tones increase its F0 dynamic range. Therefore, to synthesize Mandarin speech in Method II with a monolingual English voice, we change the F0 mean of the Mandarin tree in the bilingual model to that of the monolingual speaker and keep the same F0 variance of Mandarin speech. Resultant synthesized Mandarin speech sounds very similar to the original monolingual English speaker.

Two perceptual tests were conducted to evaluate the quality of Mandarin speech synthesized with Method II. Intelligibility Test: five native Mandarin speakers were asked to transcribe 20 sentences randomly selected from a set of 70 synthesized sentences. A 92.1% of Chinese character accuracy of transcription is obtained. Quality Test: the same five subjects were asked to give their opinions on the synthesized speech quality in a five-point scale MOS [15]: 5=excellent, 4=good, 3=fair, 2=poor, 1=bad. An average MOS score of 3.08 is obtained.

Table 2. F0 means and standard deviations of the training data.

Speaker	bilingual		monolingual
Language	Mandarin	English	English
mean (Hz)	198.5	198.3	180.7
std (Hz)	49.62	37.39	34.55

5. CONCLUSIONS

We propose a cross-language, HMM-state mapping between two language-specific decision trees for synthesizing a target language using a monolingual (source language) voice. State mappings are created in the minimum KLD sense between leaf nodes of the bilingual trees. A sentence transcription based perceptual test confirms that the synthesized Mandarin speech is highly intelligible and a Chinese character accuracy rate of 92.1% is obtained. Although cross-language mapping in this study was tested only on a bilingual corpus recorded by a single speaker, the approach can be easily expanded to multiple (>2) languages.

6. REFERENCES

[1] Sproat, R. (Editor), *Multilingual Text-to-Speech Synthesis: the Bell Labs Approach*, Kluwer Academic Publisher, 1998.

[2] S. Quazza, L. Donetti, L. Moisa, and P. L. Salza, "Actor®: A Multilingual Unit-selection Speech Synthesis System", *Proc. of 4th ISCA Speech Synthesis Workshop*, 2001.

[3] A. W. Black, and K. A. Lenzo, "Multilingual Text-to-Speech Synthesis", *Proc. of ICASSP*, vol.3, pp.761-764, May 2004.

[4] M. Chu, H. Peng, Y. Zhao, Z. Y. Niu, and E. Chang, "Microsoft Mulan - A Bilingual TTS System", *Proc. of ICASSP*, vol.1, pp.264-267, April 2003.

[5] F. Deprez, J. Odijk, and J. D. Moortel, "Introduction to Multilingual Corpus-based Concatenative Speech Synthesis", *Proc. of Interspeech*, pp.2129-2132, August 2007.

[6] N. Campbell, "Talking Foreign - Concatenative Speech Synthesis and the Language Barrier", *Proc. of Eurospeech*, pp.337-340, September 2001.

[7] L. Badino, C. Barolo, and S. Quazza, "Language Independent Phoneme Mapping for Foreign TTS", *Proc. of the 5th ISCA Speech Synthesis Workshop*, pp.217-218, 2004.

[8] J. Latorre, K. Iwano, and S. Furui, "Polyglot Synthesis Using A Mixture of Monolingual Corpora", *Proc. of ICASSP*, vol.1, pp.1-4, March 2005.

[9] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis", *Proc. of ICASSP*, vol.3, pp.1315-1318, June 2000.

[10] H. Liang, Y. Qian, and F. K. Soong, "An HMM-based Bilingual (Mandarin-English) TTS", *Proc. of the 6th ISCA Speech Synthesis Workshop*, pp.137-142, August 2007.

[11] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker Adaptation for HMM-based Speech Synthesis System Using MLLR", *Proc. of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp.273-276, November 1998.

[12] T. A. Myrvoll, and F. K. Soong, "Optimal Clustering of Multivariate Normal Distributions Using Divergence and Its Application to HMM Adaptation", *Proc. of ICASSP*, vol.1, pp.552-555, April 2003.

[13] Y. Zhao, C. Zhang, F. K. Soong, M. Chu, and X. Xiao, "Measuring Attribute Dissimilarity with HMM KL-Divergence for Speech Synthesis", *Proc. of the 6th ISCA Speech Synthesis Workshop*, pp.206-210, August 2007.

[14] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-Space Probability Distribution HMM", *IEICE Trans. Inf. & Syst.*, vol.E85-D, no.3, pp.455-464, March 2002.

[15] *Methods for subjective determination of transmission quality*, Recommendation P.800, ITU-T Std., August 1996.