HMM-based Mixed-language (Mandarin-English) Speech Synthesis

Yao Qian¹, Houwei Cao*¹,², Frank K. Soong¹

¹Microsoft Research Asia, Beijing {yaoqian, frankkps}@microsoft.com, hwcao@ee.cuhk.edu.hk
²The Chinese University of Hong Kong, Hong Kong

Abstract—English words or short phrases embedded in Mandarin utterances have become more common among bilingually educated people like college students in China. Similarly, it becomes highly desirable that TTS systems can synthesize mixed-language speech properly. Recently, we proposed an HMM-based bilingual TTS to synthesize a target language when only monolingual source language recording from a speaker is available. In this paper, we extend it to synthesize mixed-language sentences. A cross-language state mapping is first established between decision trees built from the English and Mandarin recordings of a bilingual speaker. Via the mapping, English words or phrases embedded in Mandarin sentences can then be synthesized. The bilingual state-mapping is extended to monolingual speaker to perform mixed-language synthesis. Perceptual test results show: (1) decent intelligibility, confirmed by an English word transcription accuracy of 86%; (2) good speech quality with an average MOS score of 3.2.

Keywords—Speech synthesis, HMM-based TTS, Mixed-language speech synthesis

I. INTRODUCTION

There are many studies on multilingual TTS with the same voice, i.e., polyglot speech synthesis [1-4]. A straightforward approach to bilingual HMM-based TTS is to use a multilingual speech database recorded by a multilingual speaker [1]. However, it is difficult to find a professional polyglot speaker who can speak more than three languages. Other approaches use phone mapping to find acoustically similar units from the source language spoken by a mono-lingual speaker to synthesize the language without any recording data (target language). The cross-language phone mappings where similar sounding phones are mapped to each other, is based upon a good similarity measure. The measure can compares two articulatory features and assign appropriate weights via subjective perception test, i.e., by asking native speakers to evaluate the intelligibility and naturalness of the synthesized speech via the mapping[3]. In [2], speech segments of speech database of one speaker (source language) are compared with the speech segments of another speaker (target speaker) and the closest one in cepstral similarity measure is chosen. The above polyglot systems are mostly based on a concatenation technique. Recently, HMM-based TTS has been successfully applied to TTS synthesis for many different languages [5]. In [4], an average voice based on HMMs 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 languages in the training set.

Mandarin and English belong to different language families and they are pretty unrelated both phonetically and grammatically. The phones in the two languages are quite different in terms of their IPA symbols. The phone mapping-based approaches to polyglot speech synthesis are not quite feasible for building a good bilingual (Mandarin-English) TTS. Recently we proposed a HMM-based lingual TTS system, in which cross-language state sharing and mapping are used to synthesize natural speech from a given bilingual text [6,7]. In this study, we focus on the mixed-language text speech synthesis and evaluation of the system.

II. AN HMM-BASED BILINGUAL TTS SYSTEM

The proposed HMM-based lingual TTS system is described in the following sections.

A. Context-dependent State Sharing across Languages

A straightforward approach to bilingual HMM-based TTS with same voice is to use a speech corpus of the two languages recorded by a bilingual speaker. To reduce the recording effort and to maximize the training efficiency, the smallest phone set that covers all phones in the two languages is used for training HMMs. Since Mandarin and English are highly unrelated phonetically, their phones are not very sharable in terms of the corresponding IPA symbols. However, the more refined sub-phonemic units of the two languages are more similar than their phonemic counterparts. In other words, sub-phonemic units form the basic components of phonemes and they tend to be more similar across different languages than the phonemic units. Allophones, which are highly context dependent, provide more chances for unit sharing between Mandarin and English. We use context-dependent state sharing across languages to build an HMM-based bilingual, Mandarin (Man.) and English (Eng.), TTS system. The schematic diagram of the system is illustrated in Fig. 1.

A training corpus of Mandarin and English speech sentences, recorded by a bilingual speaker, and a universal phone set of all English and Mandarin phones, are used to construct the bilingual HMM-based TTS system. Phone models of rich contexts, including short contexts like left or right phones or longer contexts like phone positions and Part Of Speech (POS), are used to capture the co-articulation effects in speech signals. However, constrained by limited training data,
we need to tie (cluster) models of rich contexts into generalized ones for predicting unseen contexts in test sentences. State tying via a clustered decision tree is implemented and states of different central phones of the two 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 a velar plosive, such as /g/ (Eng.), /k/ (Man.) or /k/ (Man.)?

b) language-dependent questions: e.g. E_unvoiced_fricative, Does the state belong to an English unvoiced fricative, such as /θ/ (Eng.), /s/(Eng.), /h/(Eng.), /ʃ/ (Eng.), or /θ/(Eng.)?

c) language-specific questions: e.g. E_Voiced_Stop, Does the state belong to a velar plosive, such as /b/(Eng.), /d/(Eng.), or /g/ (Eng.)? All attributes of the question only exist in English, i.e. there are no Mandarin voiced stops. Based on their manners and places of articulation, supra-segmental features, etc., we construct questions to jointly tie states of English and Mandarin phone models.

Figure 1. The schematic diagram of an HMM-based bilingual, Mandarin (Man.) and English (Eng.), TTS system

B. Context-dependent State Mapping

Section II.A is a straightforward way to build a bilingual, Mandarin and English, TTS system. However, it is not easy to find professional speakers who are fluent in both languages when we need to build an inventory of bilingual voice-fonts of multi-speakers. Also, it is an interesting research topic on how to synthesize target language speech when only monolingual source language recording of a speaker is available. We propose to establish a tied, context-dependent state mapping across different languages from a bilingual speaker first and then use it as a basis to synthesize other monolingual speaker’s voice in the target language.

First, we built two language-specific decision trees by using the bilingual data recorded by one speaker. Each leaf node in the Mandarin decision tree has a mapped leaf node, in the minimum K-L divergence sense, in the English one. The tied, context-dependent state mapping (from Mandarin to English) is shown in Fig. 2. The directional mapping from Mandarin to English can have more than one leaf nodes in the Mandarin tree mapped to one leaf node in the English tree. As shown in the figure, two nodes in the Mandarin tree are mapped into one node in the English tree. The mapping from English to Mandarin is similarly done but in a reverse direction.

Figure 2. The illustration of a tied, context-dependent state mapping from a Mandarin decision tree to an English decision tree.

In HMM-based speech synthesis, spectral and pitch features are separated into two streams and stream-dependent models are built to cluster two features into separated decision trees. Both English and Mandarin have trees of spectrum, pitch and duration. Leaf nodes of those trees are used to set a mapping between English and Mandarin.

To synthesize speech in a new language without pre-recorded data from the same speaker, we can utilize the mapping established with bilingual data and new monolingual data recorded by a different speaker. For example, a context-dependent state mapping trained from speech data of a bilingual (English-Mandarin) speaker A can be used to choose the appropriate states trained from speech data of a different, monolingual Mandarin speaker B to synthesize English sentences. The same structure of decision trees should be imposed on training HMM parameters of both speakers A and B.

III. EXPERIMENTS AND RESULTS

A. 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 Mandarin corpus recorded by another female. Table I lists the number of sentences in these two corpora. Testing data consists of 140 mixed language sentences (90 sentences with original recordings). Speech signals are sampled at 16 kHz, LPC analyzed with a Hamming window of 25-ms and shifted every 5-ms. LPC spectral features are transformed into 40th-order
LSPs and their dynamic counterparts. Five-state, left-to-right HMMs with single Gaussian, diagonal variance are adopted for phone model training.

<table>
<thead>
<tr>
<th>Corpus/Speaker</th>
<th>Bilingual</th>
<th>Monolingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>English</td>
<td>1,024</td>
<td>N/A</td>
</tr>
<tr>
<td>Mixed-language</td>
<td>600</td>
<td>N/A</td>
</tr>
</tbody>
</table>

We built two systems as follows:

1) **System I: State sharing across languages**

The sentences from the bilingual speaker A are used to train HMMs. Context-dependent state sharing across languages as discussed in Section IIA is applied. In synthesis, input text is converted into a sequence of contextual phone labels through a bilingual TTS text-analysis frontend (Microsoft Mulan) [8]. The corresponding parameters of contextual states in HMMs are retrieved via mixed-language decision trees. Then LSP, gain and F0 trajectories are generated in the maximum likelihood sense. Finally, speech waveforms are synthesized from the generated parameter trajectories.

We built two voices in this system. In one voice (Voice I), all sentences are used in training, while in another voice (Voice II), 600 mixed-language sentences are excluded in the training data. Since there are no mixed-language sentences in the training data of Voice II, the context of phones at a language switching boundary, e.g., the left phone or the right phone, is replaced with the nearest context in the language which the central phone belongs to. For example, the triphone /œɪ(ʌ)/-\(\overline{\nu}l(\overline{C})\) will be replaced with /\(\overline{\nu}l(C)\)-\(\overline{\nu}l(\overline{C})\)/, where the left context /\(\overline{\nu}l(\overline{C})\)/ is the nearest Mandarin substitute for /\(\overline{\nu}l(E)\)/ according to the KLD measure [6].

2) **System II: State mapping across languages**

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

**B. Evaluation Results**

1) **Subjective evaluation**

The 50 mixed-language sentences generated by the two voices of system I are evaluated subjectively in an AB preference test by three subjects. The results of preference test are shown in Table II. The preference score of Voice I (67%) is significantly higher than Voice II (8%). The main perceptually noticeable difference in the paired sentences synthesized by Voices I and II is at the language transitions between English and Chinese in those sentences.

<table>
<thead>
<tr>
<th>Similar</th>
<th>Different</th>
<th>Voice I better</th>
<th>Voice II better</th>
</tr>
</thead>
<tbody>
<tr>
<td>13%</td>
<td>12%</td>
<td>67%</td>
<td>8%</td>
</tr>
</tbody>
</table>

2) **Objective evaluation**

The synthesized sentences are also evaluated by objective measures to compare the difference between Voices I and II. Synthesis quality is measured objectively in terms of distortions between original speech and speech synthesized by System I. Since the predicted HMM state durations of generated utterances are in general not the same as those of original speech, we first measure the root mean squared error (RMSE) of phone durations of synthesized speech. Spectra and pitch distortions are then measured between original speech and synthesized speech where the state durations of the original speech (obtained by forced alignment) are used for speech generation. In this way, both spectrum and pitch are compared on a frame synchronous basis between the original and synthesized utterances.

Since perceptually noticeable differences are mainly at the language transition in the mixed-language sentences, we use objective measures to evaluate those parts. Here, language transition part is defined as one syllable for Mandarin or one word for English before/after language switch point. Table III shows the averaged log spectrum distance, RMSE of F0 and phone durations evaluated in the transition parts of 90 mixed-language test sentences generated by the two voices in system I. It indicates that the distortion difference between Voices I and II at language transitions is rather large, especially on the prosody (F0, Gain and duration) distortion. Adding mixed-language sentences into training data in Voice I helps to alleviate the problem of segmental and supra-segmental discontinuities between Mandarin and English language transitions. Since all training sentences are either exclusively Chinese or English in Voice II, there is no training data to train such language-switching phenomena. It is therefore more prone to the synthesis artifacts at the language switching points between English and Chinese words.

<table>
<thead>
<tr>
<th>Voice</th>
<th>RMSE of Duration (s)</th>
<th>RMSE of Gain (dB)</th>
<th>RMSE of F0 (Hz)</th>
<th>Log spectrum distance (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man.</td>
<td>0.038</td>
<td>5.12</td>
<td>20.39</td>
<td>3.98</td>
</tr>
<tr>
<td>Eng.</td>
<td>0.040</td>
<td>7.61</td>
<td>27.08</td>
<td>4.79</td>
</tr>
<tr>
<td>Man.</td>
<td>0.040</td>
<td>5.91</td>
<td>26.66</td>
<td>4.04</td>
</tr>
<tr>
<td>Eng.</td>
<td>0.040</td>
<td>7.61</td>
<td>34.62</td>
<td>5.16</td>
</tr>
</tbody>
</table>

The 50 mixed-language sentences are also synthesized by Mandarin HMMs in System II. The embedded English words or phrases are synthesized by using the mapping built from the leaf nodes of the monolingual English tree to those of the monolingual Mandarin tree via the recordings of a bilingual speaker. Since the mapping is built by two monolingual decision trees, the synthesis artifacts at the switches of English and Chinese words occur in the voice of System II. This phenomenon is the same as that in the voice II of System I. Table III indicates the artifacts in prosody are more obvious than the spectral ones, therefore we change the F0 mean of the Mandarin tree in the voice I of system I to that of the
monolingual speaker, then use the prosody generated from it for System II.

Two perceptual tests were conducted to evaluate the quality of mixed-language speech synthesized with Voice I in System I and the voice in System II. The first is an intelligibility test where five graduate students were asked to transcribe the embedded English word or phrases for 20 sentences, which are randomly selected from a set of 50 synthesized sentences. The second is a quality test where the same five subjects were asked to give their opinions on the synthesized speech quality in a five-point scale MOS: 5=excellent, 4=good, 3=fair, 2=poor, 1=bad. The intelligibility and quality test results are shown in Table IV. An 86% of English word transcription accuracy and an average MOS score of 3.2 are obtained by System II. Compared with Voice I of System I, trained with the recordings of a bilingual speaker, the performance of System II is quite acceptable.

<table>
<thead>
<tr>
<th>Intelligibility</th>
<th>Voice I in System I</th>
<th>Voice in System II</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOS</td>
<td>3.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

IV. CONCLUSIONS

In this paper, context-dependent state sharing is first used to build a mixed-language TTS through the bilingual recording of such a speaker. The discontinuities at segmental and suprasegmental between Mandarin and English language transitions in mixed-language synthesis are largely alleviated by adding mixed-language sentences into training data. We also propose to use context-dependent state mapping to synthesize embedded English words or phrases of mixed-language sentences by the recording of a mono-lingual, Mandarin speaker. The perception test results of synthesized mixed-language speech show that our approaches are viable.

REFERENCES