CROSS-LINGUAL FRAME SELECTION METHOD FOR POLYGLOT SPEECH SYNTHESIS

Chia-Ping Chen,

National Sun Yat-Sen University
Dept. of Computer Science and Engineering
Kaohsiung, Taiwan

Yi-Chin Huang, Chung-Hsien Wu, and Kuan-De Lee

National Cheng-Kung University
Dept. of Computer Science and Information Engineering
Tainan, Taiwan

ABSTRACT

A novel approach is proposed to creating a polyglot speech synthesis system without the need of collecting speech data from a bilingual (or multilingual) speaker, which is often expensive or even infeasible. Given a target speaker with data in the first language (Mandarin in this study), the basic idea is to construct artificial utterances in the second language (English) via selection of speech sample frames of the given speaker in the first language. As the speaker needs not be polyglot, this method is generally applicable to any speaker and any languages. In the search for optimal frame sequence selection, the candidate set is constrained by a decision tree for phone segments in the speech data of both languages, and the cost function depends on the context-dependent articulatory and auditory features. Evaluation results show that good performance regarding similarity (speaker identity) and naturalness (speech quality) can be achieved with the proposed method.

Index Terms— polyglot speech synthesis, frame selection, articulatory features, auditory features

1. INTRODUCTION

By definition, a polyglot speech synthesis system [1] can output synthesized speech of multiple languages with the *same* voice. The polyglot synthesis is clearly more challenging than multilingual speech synthesis systems where the voices can be *different*. Several methods have been proposed for polyglot speech synthesis. In [2], a cross-language mapping relation between the Mandarin and English decision trees for tied states is learned based on bilingual data from a bilingual reference speaker, and is subsequently used in the model training in the source language and the synthesis in the target language for another speaker. These approaches require the collection of speech data from a polyglot speaker fluent in the set of languages of interest. As the number of languages increases, the data collection task will become quite difficult. If the phone sets of different languages could be accurately shared, it would be possible to apply the acoustic phone models trained in one language to the synthesis of speech of another language, achieving polyglot functionality [3]. However, the effectiveness of such an approach depends on the language pair. For languages from different families, e.g., English and Mandarin, such a mapping often becomes crude and not reliable.

Alternatively, model adaptation and text-independent voice conversion [4, 5] methods offer the possibility for creating artificial language-dependent utterances, which can be used in a polyglot system. In model adaptation, a language-dependent synthesis model is trained, and then adapted with the speech of the target speaker. In voice conversion, the output speech of a language-dependent synthesis model is converted to that of the target speaker. Our method is different from the above approaches in that we apply frame-level data selection to string frames together to be artificial utterances in a different language for the target speaker. The articulatory and auditory features are used for frame sequence selection.

The rest of this paper is organized as follows. The proposed method with details in Section 2. The evaluation results are presented with comments in Section 3. Finally, concluding remarks are stated in Section 4.

2. METHODOLOGY

2.1. Outline

In our system, the first language of the target speaker is Mandarin, and the second language is English. To create such a polyglot synthesis system through the proposed method, we need a speech corpus in Mandarin of the target speaker and a speech corpus in English. Let us denote the target speaker by s, the Mandarin corpus (of s) by $\mathcal{D}_m$, and the English corpus (of a different speaker) by $\mathcal{D}_e$. We carry out the following steps.

- Train a Mandarin synthesis model for s from $\mathcal{D}_m$; denote the trained model by $\mathcal{M}_m$;
- For each utterance $u_i \in \mathcal{D}_e$, construct an artificial utterance $\tilde{u}_i$ using frames in $\mathcal{D}_m$ and certain post-processing;
- Denote the set of $\tilde{u}_i$ by $\tilde{\mathcal{D}}_e$, which is parallel to $\mathcal{D}_e$ with the voice of s;
- Train an English synthesis model from $\tilde{\mathcal{D}}_e$; denote the trained model by $\tilde{\mathcal{M}}_e$;

The polyglot synthesis system uses the models $\mathcal{M}_m$ and $\tilde{\mathcal{M}}_e$. The main idea here is to build the corpus $\tilde{\mathcal{D}}_e$ parallel to $\mathcal{D}_e$ in a novel way. Given $u \in \mathcal{D}_e$, we need a plausible method of selecting frames from $\mathcal{D}_m$ for the construction of $\tilde{u} \in \tilde{\mathcal{D}}_e$. Fig. 1 shows the system framework, which consists of tree-based phone clustering, feature extraction, and unit-selection based alignment. They are explained in details in the following subsections.

2.2. Decision Tree for Phone Segments

The phonemes in Mandarin and English are clustered via a decision tree. The question set consists of language-independent questions such as place/manner of articulation, simple vowel or diphthong, and vowel position. Such questions help to cluster phones which are similar in their articulatory features, independent of the language. The phone segments in $\mathcal{D}_e$ and $\mathcal{D}_m$ are used as data in growing the decision tree by maximizing the reduction of entropy. Furthermore, it is mandatory that each leaf node (class) must contain at least one Mandarin segment and one English segment.

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

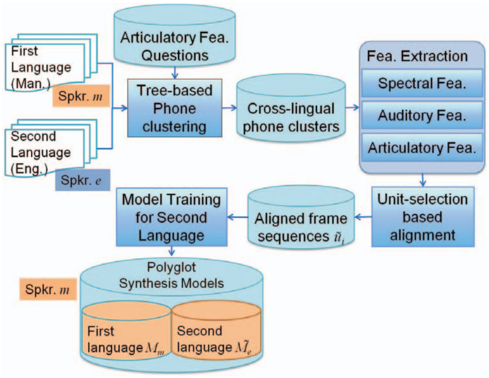

Articulatory Fea. Questions
First Language (Man.)
Spkr. m
Tree-based Phone clustering
Cross-lingual phone clusters
Fea. Extraction
Spectral Fea.
Auditory Fea.
Articulatory Fea.
Second Language (Eng.)
Spkr. e
Model Training for Second Language
Aligned frame sequences $\tilde{u}_i$
Unit-selection based alignment
Polyglot Synthesis Models
Spkr. m
First language M_m
Second language $\tilde{M}_e$

Fig. 1. Framework of the proposed polyglot synthesis system.

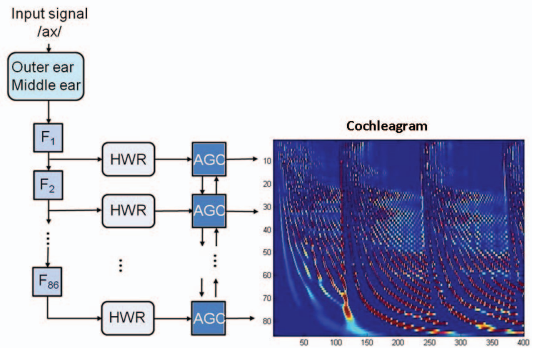


Fig. 2. An example of the auditory feature extraction.

2.3. Speech Features

2.3.1. *Mel-Generalized Cepstral Coefficients*

The mel-generalized cepstral (MGC) features are the input to vocoder for waveform generation. 25 MGC coefficients are extracted based on the mel-generalized cepstral analysis. The pitch of each frame is estimated to adapt the length of analysis window [6] to minimize the effects of non-ideal periodic extension. Furthermore, vocal tract length normalization (VTLN) is applied to reduce the spectral difference among different speakers. VTLN is implemented through a linear transformation [7] of the MGC coefficients.

2.3.2. *Auditory Features*

The auditory features are used in computing substitution cost in frame selection. We use the auditory features based on Lyon's auditory model [8], which models the characteristics of the human sound perception, such as non-linear frequency scale, equal-loudness contour, and masking effects. Fig. 2 shows an example of the auditory feature extraction steps and output feature vector of English phone */ax/*. First, the pre-emphasized signal is filtered by 86 cascaded filters. Then, using a series of half-wave rectification (HWR) and related automatic gain controls (AGC), the resultant cochleagram is used as the auditory feature. Our goal is to realize the synthesized speech which is perceived as uttered by the target speaker. Thus, the auditory features should be more suitable than the traditional MGC features.

Table 1. The articulatory features.

vowel	fricative	nasal	stop	approximate	high
labial	glottal	dental	round	retroflex	mid
voiced	anterior	velar	back	continuant	low
tense	coronal	vocalic	silence		

2.3.3. *Articulatory Features*

The articulatory features (AF) are used in computing concatenation cost in frame selection. As AF is cross-lingual and speaker-independent [9], using AF in distance estimation between neighboring frames is a feasible choice. The AF features are the frame-level posterior probabilities of 22 speech articulatory features listed in Table 1. They are estimated by an artificial neural network. The input layer consists of nodes for the static and dynamic MFCC features, while the output-layer nodes correspond to articulatory features. The hidden layer consists of 100 nodes.

2.3.4. *Feature Representation of a Frame*

Each frame is represented by the MGC features or the auditory features of the current frame and the *enhanced* AF vector covering three consecutive frames. That is,

$$f^T = (v^T, u^T) = (v^T, u_p^T, u_c^T, u_n^T), \tag{1}$$

where v is the spectral subvector, u is the enhanced AF subvector with u_p, u_c, u_n being respectively the basic AF vectors of the previous frame, current frame, and next frame.

2.4. Cross-Lingual Frame Selection

For each utterance in English, we select the frames in the Mandarin speech corpus provided by the target speaker to construct an artificial utterance via the following steps.

- Align the input utterance to a string of phone segments;
- For an English phone segment, say p, in the utterance, find the leaf node it belongs to in the phone-level decision tree, say $\mathcal{T}$;
- For each frame in p, find Mandarin candidate frames from the segments in the same leaf node of $\mathcal{T}$, based on a distance measure between speech feature vectors;
- Use the dynamic programming algorithm to find the best sequence of candidate frames for the entire utterance;
- Apply pitch transformation to the artificial utterance.

In the following subsections we explain the details of each step.

2.4.1. *Candidate Set and Substitution Cost*

For the substitution cost, i.e., the cost for substituting an English frame by a Mandarin frame, Euclidean distance is used between the auditory feature vectors

$$\delta_s(v_e, v_m) = |v_e - v_m|, \tag{2}$$

where v_e (v_m) is the auditory feature vector of an English (Mandarin) frame. Based on Eq. (2), a candidate set for an English frame consisting of n Mandarin frames in the same leaf node can be formed. Using a candidate set has the benefit of reducing the concatenation cost at the expense of increasing the substitution cost, potentially leading to a reduced total cost.

2.4.2. *Concatenation Cost*

The concatenation cost is

$$\delta_c(u, w) = \frac{1}{2}(|u_c - w_n| + |w_c - u_p|), \quad (3)$$

where u and w are the enhanced AF vectors of the current frame and the previous frame. In Eq. (3), one term is the Euclidean distance between the current AF vector of the current candidate (u_c) and the next AF vector of the previous candidate (w_n), and the other term is the distance between the previous AF vector of the current candidate (u_p) and the current AF vector of the previous candidate (w_c). By adding these two measurements, the contextual information of the frame sequence is taken into account to reduce discontinuity in articulatory features.

2.4.3. *Dynamic Programming*

Given an English utterance of I frames, denoted by

$$\mathbf{v}_e = (v_{e1}, \ldots, v_{eI}),$$

the unit selection for the construction of an artificial target-speaker utterance can be formulated as

$$\mathbf{f}^* = (f_1, \ldots, f_I)^* = \underset{f_1,\ldots,f_I}{\arg\min} \quad \text{cost}(\mathbf{f}, \mathbf{v}_e), \quad (4)$$

where

$$\text{cost}(\mathbf{f}, \mathbf{v}_e) = \sum_{i=1}^{I} \delta_s(v_{ei}, v_i) + \sum_{i=2}^{I} \delta_c(u_i, u_{i-1}). \quad (5)$$

Note that $f_i^T = (v_i^T, u_i^T)$ as defined in Eq. (1).

With n candidate frames for each frame, the search state space is essentially $O(n^T)$. Using the dynamic programming (DP) algorithm, the search time complexity can be reduced to $O(In^2)$.

2.4.4. *Pitch Transformation*

Mandarin and English have different intonation patterns, so it is important to transform the pitch contour of the artificial utterance to sound more like the second language. A Gaussian mixture model (GMM) is trained with the pitch data extracted from $\mathcal{D}_e$ and $\tilde{\mathcal{D}}_e$ and used in the transformation [10]. In other words, the transformation is based on means and variances of the pitch data.

3. EVALUATION

3.1. Data

For the Mandarin corpus $\mathcal{D}_m$, it is desirable to have enough data to cover the contextual variation and to provide a large inventory for the approximation of English. The read speech of a male speaker (speaker MR00) of the Tsing Hua Corpus of Speech Synthesis (TH-CoSS), which has been used in a HMM-based Mandarin synthesis system [11], is used. This subset consists of $5,406$ utterances, with $98,749$ syllables. For the English corpus $\mathcal{D}_e$, the speech data of speaker bdl of the CMU ARCTIC database is used, which is a phonetically-balanced subset with $1,131$ utterances.

A few important figures are summarized as follows. 25 MGC or 86 Lyon's model coefficients are extracted for computing the substitution costs, and 22 articulatory features are extracted via artificial neural networks for computing the concatenation costs. The speech sampling rate is $16,000$ Hz. The feature-extraction frame size is 40 ms with a shift of 5 ms. In total, $1,131$ artificial English utterances are constructed.

3.2. Evaluation Results

3.2.1. *Feature Evaluation*

The innovation of our method is to use the auditory features and the articulatory features in the process of frame selection. In order to decide whether this is beneficial, we compare the proposed method with the traditional method, in which the MGC features are used in the computation of the substitution and concatenation costs.

In the evaluation of substitution cost, we compare using MGC and auditory features in Eq. (2). Subjective tests for naturalness with 10 native Mandarin subjects are conducted. There are 20 artificial English utterances generated by the proposed method and texts are randomly selected from the English corpus. We compare the speech quality of the utterance using Mean Opinion Score (MOS) test. The average MOS of using auditory features is 3.0, which is better than 2.26 using MGC features.

In the evaluation of concatenation cost computation, we compare using the articulatory features with using MGC features in Eq. (3). In the ABX tests for similarity with 10 subjects, in 62% of the cases the synthesized speech using the articulatory features is considered better than using the MGC features.

3.2.2. *Comparison with State-Mapping Based Method*

The proposed method is compared with the state mapping based method proposed in [12]. A cross-lingual state mapping between English and Mandarin synthesis model sets is established using the Kullback-Leibler divergence (KLD) criterion.

In the state mapping method, the Mandarin and English model sets are trained using the same corpora as those used in the proposed method. For model adaptation from bdl to the target speaker MR00, a phonetically-balanced set of 100 artificial English utterances is selected manually from $\tilde{\mathcal{D}}_e$, generated by the proposed method. This set is also used as the parallel data for GMM training for pitch transformation.

10 prompts randomly chosen from English news paper are synthesized by the proposed method and the state mapping based method (henceforth referred to as the baseline). The subjects are asked to evaluate the *naturalness* (speech quality), *similarity* (speaker identity), and *intelligibility* of the synthesized speech.

Regarding similarity, the proposed method is preferred over the baseline in 87% of the ABX test, as shown in Fig. 3(a). In the speech quality test, the average MOS of our method is 3.04, which is better than the baseline (2.71). Regarding intelligibility, the average MOS scores are 4.2 (reference), 3.3 (our method), and 2.5 (baseline), where the reference is the English synthesis model trained with $\mathcal{D}_e$. These results are summarized in Fig. 3(b).

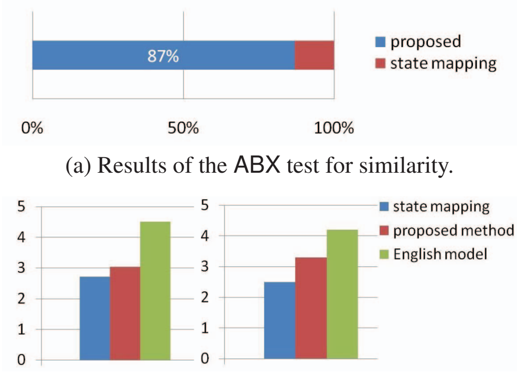
We also compare the substitution costs between the synthesized utterances and the reference ones. The histograms of the substitution costs are shown in Fig. 4. The mean of the substitution cost is 0.074 for the baseline method, and 0.048 for our method. The standard deviation is respectively 0.023 and 0.013. Therefore, our method outperforms the state mapping method in both subjective and objective measures.

3.2.3. *Evaluation of Unseen Phones*

The "unseen phones" are the phones in one language which do not exist in the other language. Therefore, the instances of these phones in $\tilde{\mathcal{D}}_e$ are arguably the most artificial parts. In the test of naturalness of the unseen phones, we use the following carrier sentence: "w_ϕ is good w_ϕ for the w_ϕ", where w_ϕ is the word containing the unseen

Table 2. Unseen phones: English phones not seen in Mandarin.

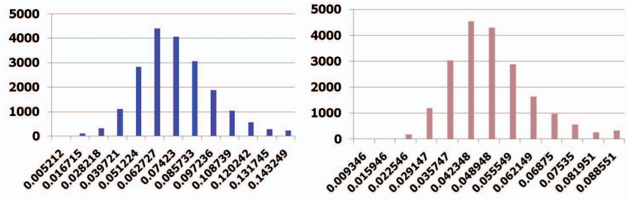
/ae/	at	/sh/	show	/ah/	but	/th/	thank
/ch/	chair	/uh/	should	/dh/	that	/v/	very
/ih/	big	/w/	way	/jh/	just	/y/	yes
/ng/	sing	/z/	zoo	/oy/	boy	/zh/	measure



(a) Results of the ABX test for similarity.

(b) Results of MOS for naturalness (left) and intelligibility (right)

Fig. 3. Comparison with state-mapping based method.



(a) State mapping

(b) Proposed method

Fig. 4. Histograms of the substitution costs.

phone ϕ, for subjective listening tests. The unseen phones ϕ and w_ϕ are listed in Table 2. The averaged MOS score over 10 subjects and 16 prompts containing unseen phones using the proposed method is 3.8, which is better than the baseline (2.6).

3.3. Discussion

The proposed frame selection method and the baseline state mapping method are both extensions of the idea of phone mapping. Since the proposed method performs frame-level selection, it is in principle more refined than state mapping based on state level operations. Our evaluation results indeed support this argument, by showing that our method outperforms the baseline method.

Since the frames in an utterance of $\tilde{\mathcal{D}}_e$ can come from many different utterances (phone segments), it is likely to hurt the intelligibility. This is indeed the case, as the performance level is not as good as the reference model trained by $\mathcal{D}_e$.

4. CONCLUSION

In this paper, a frame level unit-selection based method for polyglot speech synthesis is stated, implemented, and evaluated. The articulatory features and auditory features are employed in the selection process to achieve high-quality synthesis output. Experiment results show that the implemented system can outperform systems based on state mapping method.

On a practical side, our approach to polyglot synthesis renders corpus preparation task straightforward since the speech data of each language can come from different speakers, exempting the need for a polyglot source speaker. It can be easily generalized to any speaker and any number of languages.

5. REFERENCES

[1] Christof Traber, Karl Huber, Karim Nedir, Beat Pfister, Eric Keller, and Brigitte Zellner, "From multilingual to polyglot speech synthesis," in *Proc. Eurospeech*, 1999, pp. 835–838.

[2] Yao Qian, Hui Liang, and Frank K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin-English) TTS," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, pp. 1231–1239, August 2009.

[3] Leonardo Badino, Claudia Barolo, and Silvia Quazza, "Language independent phoneme mapping for foreign TTS," in *Proc. 5th ISCA SSW*, 2004.

[4] David Sündermann, Harald Höge1, Antonio Bonafonte, Hermann Ney, and Julia Hirschberg, "Text-independent cross-language voice conversion," in *Proc. Interspeech*, 2006.

[5] Chung-Hsien Wu, Chi-Chun Hsia, Te-Hsien Liu, and Jhing-Fa Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1109–1116, July 2006.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[7] Michael Pitz and Hermann Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[8] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. ICASSP-82*, 1982, pp. 1282–1285.

[9] Sebastian Stüker, *Multilingual Articulatory Features*, Ph.D. thesis, Carnegie Mellon University, 2003.

[10] Chung-Hsien Wu, Chi-Chun Hsia, Chung-Han Lee, and Mai-Chun Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1394–1405, August 2010.

[11] Chi-Chun Hsia, Chung-Hsien Wu, and Jung-Yun Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis,," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1994–2003, November 2010.

[12] Yi-Jian Wu, Yoshihiko Nankaku, and Keiichi Tokuda, "State mapping based method for cross-lingual speaker adaptation in hmm-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 528–531.