English
Research Express@NCKU Volume 30 Issue 10 - October 7, 2016
[ http://research.ncku.edu.tw/re/articles/e/20161007/4.html ]

# Polyglot Speech Synthesis Based on Cross-Lingual Frame Selection Using Auditory and Articulatory Features

Chia-Ping Chen2, Yi-Chin Huang1, Chung-Hsien Wu[1,*], and Kuan-De Lee1

[1] Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 701, Taiwan
[2] Department of Computer Science and Engineering, National Sun Yat-Sen University, Kaohsiung 800, Taiwan

chunghsienwu@gmail.com

IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 22, No. 10, October 2014, pp.1558-1570.

This study proposes an approach for polyglot speech synthesis based on cross-lingual frame selection using only mono-lingual speaker corpora. The basic idea is to generate artificial parallel utterances based on optimal frame selection, and use the generated utterances to adapt a synthesis model in a language foreign to the target speaker. As a result, a polyglot synthesis system can be developed based on mono-lingual speech databases of speakers speaking different languages.

Through frame selection, we attempt to mix-and-match frames of a target speaker in the primary language to create artificial utterances parallel to those of a reference speaker in the second language. Compared to longer units such as state segments or phone segments, frame-level manipulation is more fine-grained and more precise. The search cost for an optimal frame sequence would be prohibitive if full search were used. The main idea of the proposed method is to constrain the search space for a frame within a decision-tree leaf, which corresponds to a cluster of frames, to reduce the search cost. Furthermore, it is designed in the proposed method that the selection of adjacent frames favors similar articulatory attributes, so the frame-level similarity is enhanced by contextual articulatory features.

One important aspect of the proposed method is the representation of speech frames. Considering the fact that spectral features often faithfully convey the information of speaker identity, we include articulatory attributes as features since they are relatively language- and speaker-independent. In addition, the auditory features called *cochleagram* are included to better quantify perceptual similarities.

Key steps of the proposed method are:

1. Acquire speech data of the target speaker in the primary language (Mandarin) and speech data of the reference speaker in the second language (English);
2. Construct an articulatory attribute detector for the extraction of articulatory features;
3. Cluster segments of frames so that subsequently the candidate frames of a frame are chosen within the same cluster;
4. For each utterance in the English corpus, find the optimal Mandarin frame sequence;
5. Use the artificial utterances in English to adapt a speaker dependent English synthesis model to the target speaker.

An illustration of the proposed ideas is shown in Fig. 1. Artificial English utterances based on the voice of the target Mandarin speaker are obtained via frame selection, and used to adapt an English synthesis model. Combining the adapted English synthesis model and a Mandarin synthesis model trained on natural utterances, we achieve a polyglot synthesis system for the target speaker. Fig. 2 illustrates the framework of the overall system.

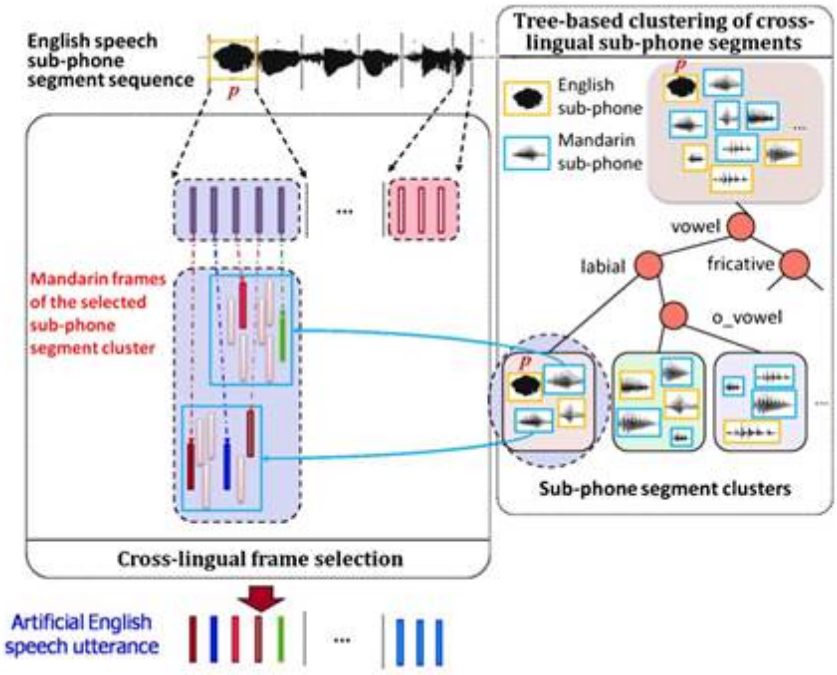


Fig. 1. Illustration of the proposed ideas.

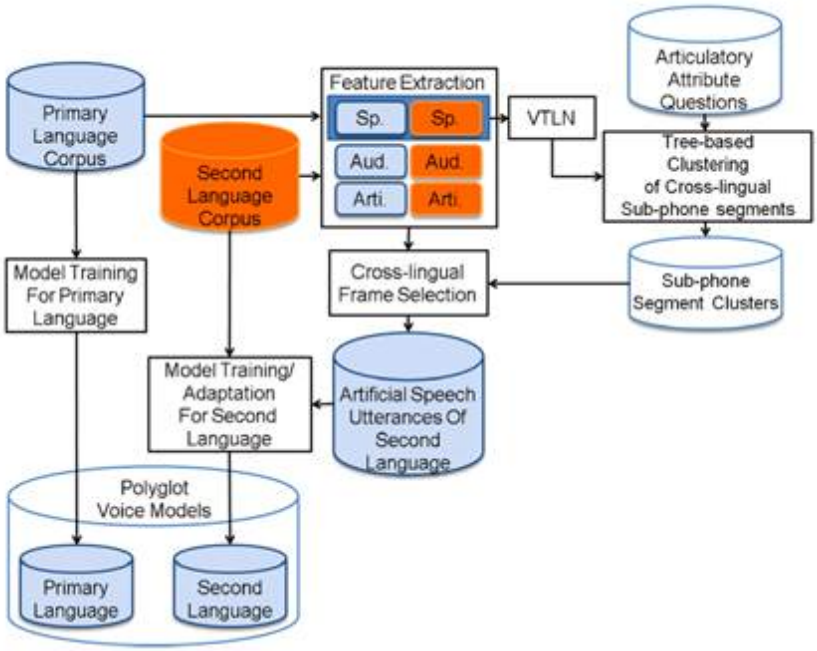


Fig. 2. Framework of the proposed polyglot synthesis system

For evaluation, a Mandarin-English polyglot system is implemented where the target speaker only speaks Mandarin. The results show that decent performance regarding voice identity and speech quality can be achieved with the proposed method.

References:

[1] A. Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” in Proc. *IEEE Int. Conf. Acoust., Speech, Signal Process*., May 1998, pp. 285–288.

[2] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, “Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, no. 4, pp. 1109–1116, Jul. 2006.

[3] C.-H.Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, “Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 18, no. 6, pp.1394–1405, Aug. 2010.

[4] D. Erro, A. Moreno, and A. Bonafonte, “INCA algorithm for training voice conversion systems from nonparallel corpora,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 18, no. 4, pp. 944–953, Jul. 2010.

[5] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, “Nonparallel training for voice conversion based on a parameter adaptation approach,” *IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, no. 3, pp. 952–963, May 2006.