合成單元與問題集之定義於隱藏式馬可夫模型中文歌聲合成系統之建立
Synthesis Unit and Question Set Definition for Mandarin HMM-based Singing Voice Synthesis

Ju-Yun Cheng, Yi-Chin Huang, and Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
E-mail: carrie771221@gmail.com ychin.huang@gmail.com chunghsienwu@gmail.com

Long Abstract

The fluency and continuity properties are very important in singing voice synthesis. In order to synthesize smooth and continuous singing voice, the Hidden Markov Model (HMM)-based synthesis approach is employed to build our Mandarin singing voice synthesis system. The system is designed to generate Mandarin songs with arbitrary lyrics and melodies in a certain pitch range. We also build a singing voice database for system training and synthesis, which is designed based on the phonetic converge of Mandarin speech. In addition, the acoustic feature extraction using STRAIGHT algorithm is employed to generate satisfactory vocoded singing voices.

The purpose of this paper is to elaborate the construction of Mandarin singing voice synthesis system by defining the synthesis model and question set for HMM-based singing voice synthesis. In addition, we implemented two techniques, including pitch-shift pseudo data extension and vibrato post-processing, to make synthesized singing voice more natural.

The proposed system framework consists of two main phases, the training phase and the synthesis phase. In the training phase, excitation, spectral and aperiodic factors are extracted from a singing voice database. The lyrics and notes of songs in the singing voice corpus are considered as contextual information for generating context-dependent label sequences. Then, the sequences are clustered with context-dependent question set and then the context-dependent HMMs are trained based on the clustered phone segments. In the synthesis phase, the input musical score and the lyric are converted into a context-dependent label sequence. The label sequence, consisting of excitation, spectrum and aperiodic factors, for the given song is constructed by concatenating the parameters generated from the context-dependent HMMs. Finally, the generated parameter sequences are synthesized using Mel Log Spectrum Approximation (MLSA) filter to generate the singing voice.

The approaches used in this study are to improve the model accuracy by defining the question set, extending the singing voice database through generating pitch-shift pseudo data, and adding the vibrato singing skill using signal post-processing. The selection of question set is crucial to generate proper synthesis models. In the baseline system, the most frequently used questions of F0 and mel-cepstral clustering trees are sub-syllables types, position of note and phrase level. Since the recorded singing database is not large enough to contain each combination of contextual factors. Thus, only essential and suitable questions are defined compared to the traditional method. Besides, the extended pitch-shift pseudo data are helpful to cover the missing pitch information of sub-syllables and increase the size of the training data. Based on the analysis results of the defined pitch range (C4~B4) of the recorded singing corpus, shifting the frequency of a note too much would change the timbre. Thus, the missing pitch information of sub-syllables of the recorded corpus is compensated using the nearby notes from other songs, and shifting the frequency of signal to the corresponding Hertz by a pitch-to-frequency mapping table. The vocal vibrato is a natural oscillation of musical pitch and the singers generally employ vibrato as an expressive and musically useful aspect of the performance. So adding vibrato can make synthesized singing voice more natural and expressive. The frequency and the amplitude can be considered since the two fundamental parameters affect the singing voice with vibrato effect. The method to create vibrato is to vary the time delay periodically and use the principle of Doppler Effect. Our system implemented this phenomenon by a delay line and a low frequency oscillator (LFO) to vary the delay.

For evaluation, the singing voice signals were sampled at a rate of 48 kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained from STRAIGHT-extracted spectra. The feature vectors consist of spectrum, excitation and aperiodic factor. The spectrum parameter vectors consist of 49th-order STRAIGHT mel-cepstral coefficients including the zero-th coefficient, their delta, and delta-delta coefficients. The excitation parameter vectors consist of log F0, its

delta, and delta-delta. A five-state, left-to-right Hidden Semi-Markov Models (HSMM) was employed in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution. The excitation stream was modeled with multi-space probability distributions HSMM (MSD-HSMM), each of which consists of a Gaussian distribution for "voiced" frames and a discrete distribution for "unvoiced" frames. We evaluate the nature of synthesized singing voice with the long duration model, and the result show that the system with long duration model obtained 62% preference higher than 38% for the system without long duration model. It shows that long duration model can actually improve the nature of phones with longer duration. Besides, the experimental results show that suitable question set definition can improve the quality and intelligibility of synthesized singing voice, and pitch-shift pseudo data and vibrato post-processing can successfully improve the quality and naturalness of the synthesized singing voice.

In conclusion, a corpus-based Mandarin singing voice synthesis system based on HMM framework was implemented in this paper. We defined the Mandarin synthesis models and the question set for model clustering and construction. In the context-dependent HMM, linguistic information and musical information are considered. Music information such as pitch, duration, is included to model the singing characteristics. Furthermore, we used three methods to refine our system, i.e. question set definition, pitch-shift pseudo data extension and vibrato post-processing. Experimental results show that our system can synthesize singing voice successfully and the refinements can actually improve the fluency and continuity of the proposed Mandarin singing voice synthesis system.

References

[1] H. Kenmochi, H. Ohshita, "VOCALOID-Commercial singing synthesizer based on sample concatenation", in *INTERSPEECH*, pp.4009-4010, 2007.

[2] S.-S. Zhou, Q.-C. Chen, D.-D. Wang, X.-H. Yang, "A Corpus-Based Concatenative Mandarin Singing voice Synthesis System", in *Machine Learning and Cybernetics, 2008 International Conference on*, vol.5, no., pp.2695-2699, 2008.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis", in *EUROSPEECH*, vol.5, pp.2347-2350, 1999.

[4] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent Development of the HMM-based Singing Voice Synthesis System-Sinsy", in *7th ISCA Speech Synthesis Workshop*, pp.211-216, 2010.

[5] H.-Y. Gu, H.-L. Liau, "Mandarin Singing Voice Synthesis Using an HNM Based Scheme," in *International Congress on Image and Signal Processing (CISP)*, vol.5, no., pp.347-351, 2008.

[6] T. Saitou, M. Goto, M. Unoki, M. Akagi, "Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices", in *Applications of Signal Processing to Audio and Acoustics Workshop on*, vol., no., pp.215,218, 2007

[7] J. Li, H. Yang, W. Zhang, L. Cai, "A Lyrics to Singing Voice Synthesis System with Variable Timbre", in *Applied Informatics and Communication Communications in Computer and Information Science*, vol.225, pp.186-193, 2011.

[8] Y. E. Kim, "Singing Voice Analysis/Synthesis", *Massachusetts Institute of Technology*, 2003.

[9] C.-C. Hsia, C.-H. Wu and J.-Y. 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, November 2010, pp. 1994~2003.

[10] Y.-C. Huang, C.-H. Wu, S.-T. Weng, "Hierarchical prosodic pattern selection based on Fujisaki model for natural mandarin speech synthesis", in *Chinese Spoken Language Processing (ICASSP), 2012 8th International Symposium on* , vol., no., pp.79-83, 2012

[11] Y.-C. Huang, C.-H. Wu, and Y.-T. Chao, "Personalized Spectral and Prosody Conversion using Frame-Based Codeword Distribution and Adaptive CRF," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 21, No. 1, January 2013, pp. 51~62.

[12] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda "An HMM-based Singing Voice Synthesis System", in *International Conference on Spoken Language Processing (ICSLP)*, pp. 1141-1144, 2006.