TEMPLATE-BASED PERSONALIZED SINGING VOICE SYNTHESIS

Ling Cen, Minghui Dong, Paul Chan

Institute for Infocomm Research (I^2R), A*STAR,
1 Fusionopolis Way, Singapore 138632

ABSTRACT

In this paper, a template-based personalized singing voice synthesis method is proposed. It generates singing voices by means of conversion from the narrated lyrics of a song with the use of template recordings. The template voices are parallel speaking and singing voices recorded from professional singers, which are used to derive the transformation models for acoustic feature conversion. When converting a new instance of speech, its acoustic features are modified to approximate those of the actual singing voice based on the transformation models. Since the pitch contour of the synthesized singing is derived from an actual singing voice, it is more natural than modifying a step contour to implement pitch fluctuations such as overshoot and vibrato. It has been shown from the subjective tests that nearly natural singing quality with the preservation of the timbre can be achieved with the help of our method.

Index Terms— Singing voice synthesis, music score, pitch, spectrum, alignment

1. INTRODUCTION

There has been a constant increase in the direct impact of computer-based music technology towards the entertainment industry. This has attracted a lot of research interest in the field of singing voice synthesis over the recent years [1–9]. Apart from the direct development of practical music applications, this also contributes towards understanding the mechanism underlying the perception and production of human singing voices [6]. Singing voice synthesis can be classified into two categories based on the source used in the generation of singing.

Singing voices in the first category are synthesized from the lyrics of a song, which is called Lyrics-to-Singing synthesis (LTS) [1, 2, 5]. In this category, corpus-based methods, such as wave concatenation synthesis and Hidden Markov Model (HMM) synthesis, are mostly used. This is more practical than traditional systems using methods such as vocal tract physical modeling and formant-based synthesis.

Singing voices in the second category are generated from spoken utterances of the lyrics of the song. This is called Speech-to-Singing (STS) synthesis [3, 6]. One example of the STS applications is to modify the singing from an unprofessional singer by correcting the imperfect parts to improve the quality of his/her voice. This makes it possible to create professional-quality singing voice for poor singers. Since the synthesized singing voices are converted directly from the speech, it preserves the timbre of the speaker and it can, thus, sound like it is being sung by the speaker. In this way, we can synthesize personalized singing voices.

To convert speech into singing, it is necessary to understand the differences between the singing and speaking voices. In previous studies, the perceptual effects of the pitch contour [4, 7] and spectrum [3] have been investigated. In [3], the singing voice is generated by manually modifying the F0 contour, phoneme duration, and spectrum of a speaking voice. An automatic STS system is proposed in [6], where the F0, phoneme duration, and spectrum are controlled and modified based on the information from the music score of the song and its tempo.

The STS synthesis methods proposed in [3, 6] control the conversion from speech to singing with the guidance of the music score of a song. In this paper, a template-based STS method is proposed. This derives the transformation models for the conversion of acoustic features by analyzing a pair of template speaking and singing voices that are recorded from a professional singer. The features of the new speaking voice reading the lyrics of the same song as the template voices will be modified to approximate those of the singing voice based on the transformation models. Compared to score-based STS, this derives two main advantages. Firstly, it omits the need of the music score, simplifying the operation of the system. Secondly, the pitch contour is derived from an actual singing voice, which is more natural than modifying a step contour to account for pitch fluctuations such as overshoot and vibrato. This sets out to improve the naturalness and quality of the synthesized singing.

The remaining part of this paper is organized as follows. The proposed template-based STS system is presented in Section 2, where the overall system, alignment between speech and singing and the transformation models of acoustic features are presented. The experimental results are presented in Section 3, where the subjective tests on naturalness and similarity were conducted. The concluding remarks and future work are given in Section 4.

2. TEMPLATE-BASED SPEECH-TO-SINGING (STS) SYSTEM

2.1. Overall System

The template-based STS system converts speaking voices into singing voices by automatically modifying the acoustic features of the speech with the help of pre-recorded template voices. The entire system can be broken down into three stages, namely, the learning, transformation and synthesis stages.

In the learning stage, the template singing and speaking voices are analyzed to extract the Mel-Frequency Cepstral Coefficients (MFCC), short-time energy, voice and unvoiced (VUV) information, F0 contour, and spectrum. MFCC, energy and VUV are used as acoustic features in the alignment of the singing and speech in order to accommodate to their differences in timing and achieve optimal mapping between them. The transformation models for the F0 contour and phoneme duration are then derived based on the synchronization information obtained. In the transformation stage, features are extracted for the new speaking voice which is usually uttered by a different person from the template speaker. These are modified to approximate those of the singing voice based on the transformation models. After these features have been modified, the singing voice is synthesized in the last stage. For enhancing the musical effect, the backing track and reverberation effect may be added to the synthesized singing. In our implementation, the analysis of speech and singing voices as well as singing voice synthesis are carried out using STRAIGHT [10].

2.2. Speech-Singing Alignment

It is certain for the point of entry and duration of each phoneme in a singing voice to be different from those in a speaking voice. The two voices should be aligned before deriving the transformation models and carrying out acoustic feature conversion. The quality of the synthesized singing voice is largely dependent on the accuracy of the alignment results. A 2-step DTW-based alignment method using multiple acoustic features is employed in our system, which is elaborated below.

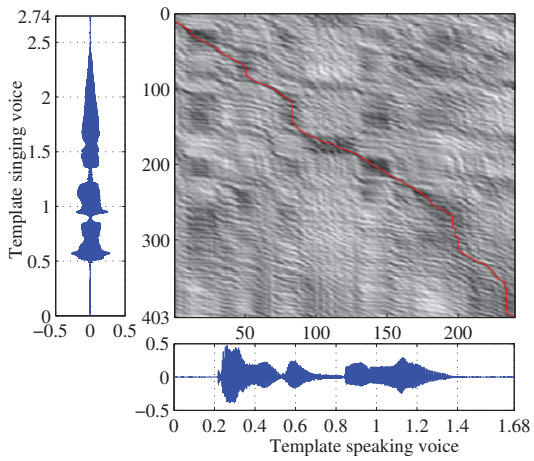
2.2.1. Acoustic features

Prior to alignment, the silence is removed from the signals to be aligned, which is detected based on energy and spectral centroid. This aims to improve the accuracy of alignment.

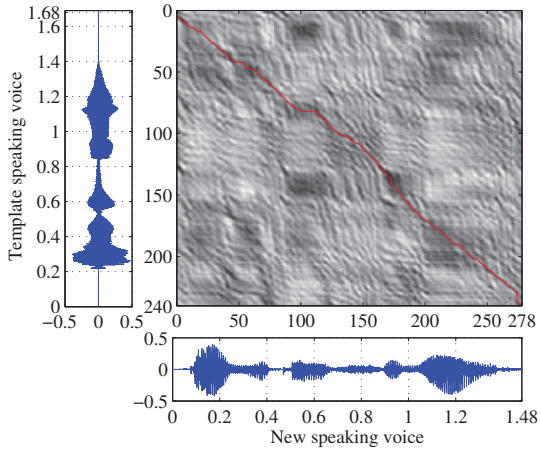
MFCC, short-time energy and voiced/unvoiced regions are then extracted as acoustic features for deriving aligned data. MFCC computes the cosine transform of the real logarithm of the short-time power spectrum on a Mel-warped frequency scale, which has become the most popular features used in Automatic Speech Recognition (ASR). Since the same lyrics having equal syllables are uttered in both singing and speech, the voiced and unvoiced regions can provide useful information for alignment. VUV is, hence, extracted as a feature in our implementation too.

Besides the raw features, the Delta and Acceleration (Delta-Delta) of features are calculated too. Frame- and parameter-level normalization are carried out on the features to reduce the acoustic variation across different frames and different parameters. Normalization is performed by subtracting the mean and dividing by the standard deviation of the features.

2.2.2. Two-Step DTW-based alignment



(a) Aligning the template singing and speaking voices.



(b) Aligning the template speech with the new speaking voice.

Fig. 1. Alignment of singing and speaking voices reading the same lyrics.

The acoustic features of different signals are aligned with each other using the DTW algorithm [11]. It measures the similarity between 2 sequences which vary in time or speed,

aiming to find an optimal match between them. The similarity between the acoustic features of two signals is measured using cosine distance as follows:

$$s = (x_i \cdot y_j) \,/\, \|x_i\| \, \|y_j\| \,, \tag{1}$$

where s is the similarity matrix, and x_i and y_j are the feature vectors of i-th and j-th frames in two signals, respectively.

To improve the accuracy in aligning the new speaking utterance to be converted with the template singing voice that is sung by a different speaker, a two-step alignment is implemented.

Step 1. We align the template singing and speaking voices that are from the same speaker. The alignment data are used to derive the mapping models of the acoustic features between singing and speech.

Step 2. The template speech and the new speaking voice are then aligned. The synchronization information derived from the alignment data together with that acquired from aligning the template voices is used to find the optimal mapping between the template singing and the new speech.

An example of the results in the two-step alignment is shown in Fig. 1(a) and 1(a). In each figure, the waveforms on the left and bottom represent the two voices to be aligned. The middle plot (red line) indicates the optimal warping path in the time warping matrix.

2.3. Transformation Models of Acoustic Features

After the mapping between singing and speaking voices is achieved via alignment, the transformation models are derived based on the template voices. The acoustic features of the new speech are then modified to obtain the features for the synthesized singing.

Prior to transformation, interpolation and smoothing are carried out on the acoustic features to be converted if their lengths are different from those of the short-time features used in alignment. In view of accuracy and computational load, template voices are divided into several segments. The transformation models are trained separately for each segment. When a new instance of speech is converted into singing using the trained transformation models, it need to be segmented similarly to the template speech [12].

In the proposed system, the F0 contour of the speaking voice is modified by acquiring a natural F0 contour from the template singing voice. In doing so, we do not need to modify a step contour to account for F0 fluctuations such as overshoot and vibrato. The synthesized singing voice could be more natural with the F0 contour of the actual singing.

The phoneme durations of the speaking voice are different from that in the singing voice. It should be lengthened or shortened during the transformation according to the latter. Unlike the STS system in [6], the musical score is not required as an input to derive the duration of each phoneme in singing, and we also do not need to carry manual segmentation for each phoneme of the speaking voice before conversion. Instead, the synchronization information from aligning the template voices and the converted speech is used to determine the transformation for phoneme duration. The duration of each phoneme in the speech is modified to be equal to that in the template singing. To implement this, the VUV, spectral envelope and aperiodicity (AP) index estimated using STRAIGHT [10] are compressed or elongated according to the transformation model of phoneme duration.

Let us use an example to illustrate the modified duration of the phonemes. The first figure in Fig. 2 shows the spectrogram of the template singing voice, the second one is the spectrogram of the converted speech, and the third one is the spectrogram with modified duration of phonemes. Via modification, the phoneme durations in the synthesized singing can be approximated to those in the template singing.

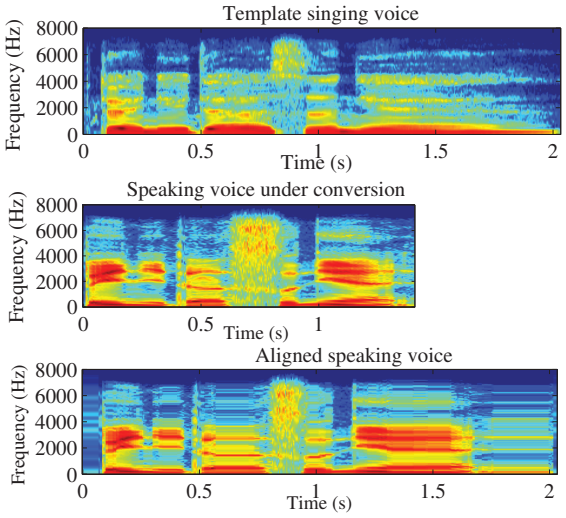


Fig. 2. Spectrogram of template singing voice, converted speaking voice, and aligned voice with modified duration of phoneme.

3. EVALUATION

To evaluate the performance of the proposed system, a popular Chinese song titled "why do you bear to hurt me so" and an English song titled "Fly me to the moon" have been synthesized. The first 6 stanzas in both songs were sung and the lyrics were read by a male and a female singer, which were used as template singing and speaking voices.

3.1. Naturalness Test

In the naturalness test, the lyrics of both songs were read by one male and one female speaker. The recorded utterances

were converted into singing voices based on the learning models from template voices using the proposed method. Ten subjects were tasked to rank the naturalness of the synthesized singing voices on a scale of 1 to 5, i.e. 1-highly unnatural, 2-unnatural, 3-neutral, 4-natural, 5-highly natural. The average scores of synthesized singing voices were tabulated as shown in Table 1. In this table, a larger value means a better score. It can be seen from Table 1 that the average score of the 2 speakers for both songs is almost 4 representing natural rank.

Table 1. Scores of Ranking for Synthesized Singing Voices

Gender	Male	Female
Naturalness (Chinese song)	3.6	3.8
Naturalness (English song)	3.9	3.8

3.2. Similarity Test

Similarity test is carried out to verify the similarity of the timbre of the synthesized singing voices to the original singers' speaking voices. The similarity aspect describes the amount of resemblance that the synthesized singing bears to the original speaking voice. In this test, 10 subjects were tasked to listen to 20 synthesized singing voices (10 for each song) and identify the original speech from a list of 20 spoken utterances uttered by 20 speakers . There are only 6 mistakes among 200 identifications. This indicates that our method successfully preserved the timbre of the synthesized singing to be close to that of the original speaking voice.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a template-based speech-to-singing synthesis system that is able to generate a singing voice from the uttered lyrics of a song based on the transformation models derived from the template voices. First, the template singing and speaking voices of the same person are analyzed to model the transformation of the acoustic features of speech to singing. The F0 contour, spectral envelope and aperiodicity index (AP) as well as voiced and unvoiced regions (VUV) of the new speaking voice uttered by a different person are then modified to approximate those of the template singing according to the transformation models. The results of the listening tests where the singing of a popular Chinese and an English song were synthesized show that the singing voices synthesized by the proposed system is perceived to be almost natural. At the same time, the timbre of the synthesized singing is preserved to be close to that of the original speaking voice.

The greatest challenge is the spectrum transformation from speech to singing voices. In order to improve the naturalness of the singing, the spectrum of the speaking voice should be modified to be similar to that of the singing voice. However, this may give rise to two problems. Firstly, it is usually inevitable that at least some cacophony is introduced to the synthesized singing when the spectrum is modified. Secondly, the similarity aspect may be reduced if the spectrum is modified according to that of the singing voice. The timbre of the generated singing is closer to that of the template singing voice. More work will be carried out to develop an effective way for spectral transformation in the future.

5. REFERENCES

[1] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, March 2007.

[2] YAMAHA Corporation, "Vocaloid: New singing synthesis technology," http://www.vocaloid.com/en/index.html.

[3] T. Saitou, N. Tsuji, M. Unoki, and M. Akagi, "Analysis of acoustic features affecting 'singing-ness' and its application to singing-voice synthesis from speaking-voice," in *Proc. Interspeech*, 2004, vol. 3, pp. 1929–1932.

[4] T. Saitou, M. Unoki, and M. Akagi, "Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis," *Speech Commun.*, vol. 46, pp. 405–417, 2005.

[5] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "Hmm-based singing voice synthesis system," in *Proc. Interspeech*, Sept. 2006, pp. 1141–1144.

[6] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Speech-to-sing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2007, pp. 215–218.

[7] H. B. Rothman and A.A. Arroyo, "Acoustic variability in vibrato and its perceptual significance," *J. Voice*, vol. 1, no. 2, pp. 123–141, 1987.

[8] N. Fonseca, A. Ferreira, and A. P. Rocha, "Concatenative singing voice resynthesis," in *International Conference on Digital Signal Processing*, July 2011, pp. 1–4.

[9] X. Rodet, "Synthesis and processing of the singing voice," in *Proc. the First IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, Nov. 2002.

[10] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based f0 extraction," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[11] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[12] L. Cen, M. H. Dong, and P. Chan, "Segmentation of speech signals in template-based speech to singing conversion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Xian, China, Oct. 2011.