A Dual Alignment Scheme for Improved Speech-to-Singing Voice Conversion

Karthika Vijayan*, Minghui Dong† and Haizhou Li*†
* Department of Electrical and Computer Engineering, National University of Singapore, Singapore
E-mails: vijayan.karthika@nus.edu.sg, haizhou.li@nus.edu.sg
† Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore
E-mails: mhdong@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—In speech-to-singing (STS) voice conversion, the source speech signals from a speaker are used to generate his/her singing voice. Such a process requires accurate detection of boundaries between phonemes and words in the speech signal. The computation and modification of analysis parameters of speech signals with respect to the target musical scores or singing templates, largely depend upon estimation of phoneme durations. In this paper, an improved dual alignment scheme for speech and singing voices in template-based STS (TSTS) systems is proposed. The subsequence dynamic time warping (subDTW) is employed to match source speech to singer's speech in the first pass of dual alignment. We assume that an accurate correspondence between singer's speech and target singing vocals has been established as part of the singing template development. Therefore, once the source speech is aligned with the singer's speech, it is automatically aligned with singing template, that we call the second pass of dual alignment. The proposed scheme delivers a relative reduction of 95.8% in word alignment error, over the baseline dynamic time warping (DTW) approach. Also, it provides a relative improvement of 38.7% in mean opinion scores of synthesized singing voices in subjective studies, over the same baseline. We demonstrate that the proposed dual alignment with the subDTW is effective in STS conversion applications.**

I. Introduction

Speech-to-singing (STS) voice conversion systems find extensive applications in the entertainment industry. The karaoke systems can employ STS mechanism to perfect the singing of individuals with limited singing abilities [1]. The STS systems enable training and evaluation of singing skills of vocal prodigy [2]. Apart from entertainment industry, the STS system is also useful in medical applications. It can assist the evaluation of verbal communication capability of persons with stammering problem, or other ailments like autism. The characteristics of human voice production are different while speaking and singing. And, the STS system renders an efficient pathway for analyzing the relationship between speaking and singing voices by elucidating the varying characteristics of voice production mechanism [3]–[5]. Hence, devising an adept STS system is significant for music information processing.

Singing voices are characterized by distinctive properties like the peculiar variations in fundamental frequency (F0) of glottal vibrations, the singing formant, frequency modulated source-spectral interactions, pitch dependence of the timbre, etc. [4]–[10]. Effective modeling and synthesis of singing voices should attempt to capture these salient features. The model-based singing synthesis captures spectral information in singing voices with the aid of control information derived from excitation characteristics [11]–[18]. Concatenative singing synthesis employs sampling method for selection and boundary smoothing of phonetic units [19]–[21]. The model-based approaches often suffer from degraded naturalness of synthesized signals, whereas the concatenative approaches compromise on flexibility and expressivity of control parameters. A comprehensive comparison of model-based and concatenative approaches for singing voice synthesis was presented in [22].

The STS systems convert the speech signals from a user (source speech) to corresponding singing voices. The major approaches for STS conversion can be broadly classified as score-based STS systems and template-based STS systems. These systems use the target musical scores or singing templates from professional singers to modify the prosody, or equivalently the excitation characteristics, of speech signals. The timbre, or spectral properties, of the source speech are preserved and the output singing voices are synthesized. The basic methodology for STS conversion is illustrated in Fig. 1.

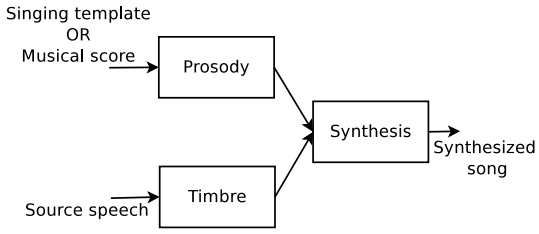

Fig. 1. Basic methodology for STS conversion.

In score-based STS systems, the target musical scores (eg: MIDI files) were utilized to modify the excitation characteristics of speech signals [23]–[25]. The spectral and excitation parameters of source speech were extracted, followed by modifications of phoneme durations, spectral and F0 parameters to match with those of target musical scores. For nullifying the degradations caused by a vocoder in singing synthesis, the transformation of voices was attempted in the time domain using a spectral differential control and F0 control on the singing waveform itself [26]–[28]. As the parameters of singing in score-based systems were derived synthetically from speech

parameters, the resulting singing voices are less natural [29].

The template-based STS (TSTS) system employs singing templates recorded from professional singers. The parameters of spoken vowels were varied with reference to the corresponding parameters obtained from human singing templates [30]. The control models for F0, note durations and spectral properties were learned from a database of natural singing voices. Similarly the excitation characteristics, or equivalently the prosody, of singers were directly used for synthesizing singing voices while keeping the timbre of source speech intact [31]. As the excitation parameters are extracted from human singing, the synthesized singing voices from TSTS systems are more natural [29].

The task of estimating phone boundaries in speech signals is extremely important for both score-based and template-based STS systems. The accurate estimation of phoneme durations serves as a prerequisite for control models to effectively compute and modify the spectral, F0 and durations of each phoneme in speech signals with respect to the target musical score or singing template. The source speech has to be aligned with target musical scores in score-based systems, whereas the source speech signals are aligned with singing templates in template-based systems. Thus an alignment technique delivering phoneme matching between speech signals and target score/template is crucial to STS conversion. The alignment of lyrics with vocals was previously attempted using Viterbi alignment or likelihood-based scoring with phonetic models [11], [32], [33]. These methods works well only on vocals isolated from accompanying polyphonic music and particularly on vowel sounds.

The TSTS system synthesizes singing voices by using spectral vectors from short-time frames of source speech and prosody from corresponding frames of singing templates. The prosody characteristics from singing templates deliver the rhythm, tempo, etc. to the synthesized singing from artistic expertise of a professional singer. Any misalignment between the frames of speech and singing can cause annoying distortions in the synthesized voice. As the nature of phonemes changes with the mode of speaking (speaking or singing), direct alignment between speech and singing signals is not feasible. Notice that, TSTS systems presented in [31], [34] use the dynamic time warping (DTW) algorithm for temporal alignment [35].

The conventional DTW fails to identify accurate correspondence between short-time frames of speech and singing signals, as they exhibit significantly distinct characteristics. The large difference in durations between speech and singing, which is largely elongated for its soothing effects, also contributes to the failure of DTW. And, the DTW is not generally effective in aligning two continuous signals compared to aligning isolated words. Another challenge with employing DTW for alignment in the TSTS system arises from the fact that the source speech may be contaminated with noise. Though the singing templates are recorded in a quiet studio environment, the source speech may be recorded anywhere from a normal room to a car stopped in heavy traffic. There could be multiple noises and crosstalk present in the recording environment of source speech. Thus an additional word, unintentionally recorded into the source speech due to crosstalk, can readily force the DTW algorithm to fail in alignment of speech and singing template. Hence, an efficient temporal alignment technique is essential for the effective functioning of the TSTS system.

In this paper, we propose a highly competent temporal alignment methodology for TSTS systems. We present the dual alignment scheme to match the source speech to the singer's speech and then the singer's speech to original singing template, in two passes. As the speech signals demonstrate similar properties among themselves, dissimilar to those of singing voices, it is advantageous to match source speech with singer's speech. The key contribution of this paper is the use of subsequence DTW (subDTW) to match segments in source speech to singing template in two passes [36]. The subDTW will not only render better alignment paths than the conventional DTW, but also reduce the risk of misalignment by the presence of crosstalk in source speech. Thus the use of dual alignment with subDTW renders the TSTS system to produce singing voices with improved naturalness and lessened perceivable distortions as demonstrated in the objective and subjective studies reported in this work.

The rest of the paper is organized as follows: In Section II, we review the baseline TSTS system. Here, we will also highlight the challenges needed to be addressed for improving the baseline system. The proposed dual alignment scheme with subDTW is presented in Section III, illustrating its efficacy in addressing the shortcomings discussed in Section II. In Section IV, we explain the subjective and objective experiments conducted to validate the role of the proposed alignment scheme in improving the singing voice synthesis using TSTS system. In Section V, we summarize the contributions of this paper towards the synthesis of singing voices of high perceptual quality.

II. Template-based speech-to-singing (TSTS) system

The TSTS system makes use of singing templates from professional singers for STS conversion process. The singing voice output is synthesized by retaining the prosody of singing templates and replacing the spectral characteristics with those of the vocally untrained user. As excitation characteristics of the synthesized singing voice in TSTS system are derived from those of professional singer's data, naturalness of the synthesized singing is preserved to a great extend [29], [31]. In this section, we review the functioning of the baseline TSTS system and highlight the possibilities of its failure in synthesizing good quality singing.

A. The singing templates database

To construct singing templates, multiple songs sung by professional singers were recorded. The choice of songs were made by the singers themselves, according to their individual singing skills [25]. The songs were recorded in a noise-proof studio environment and were segmented into lyrical sentences.

Together with the singing data, the speech data obtained by the singers reading out the lyrics of the songs were recorded. Thus the templates database consists of singing data (singing template), singer's speech and corresponding sentence-level transcriptions of multiple songs.

B. The STS conversion

STS conversion in the TSTS system is expressed as a three-stage process, namely, learning, transformation and synthesis [31]. In the learning stage, mel frequency cepstral coefficients (MFCC) and voiced-unvoiced (VUV) decisions were extracted. The alignment between source speech and singing templates was performed in two steps, where the MFCC features from short-time frames of singer's speech and singing templates were aligned followed by the short-time frames alignment between source speech and singer's speech. Thus the source speech was aligned with singing template, via singer's speech. The alignment was carried out using DTW algorithm operating upon cosine distance metric between MFCC vectors [35]. The synchronization information (sync info) between frames of signals were saved for further processing. The alignment process proposed in [31], [34] is illustrated as a block diagram in Fig. 2.

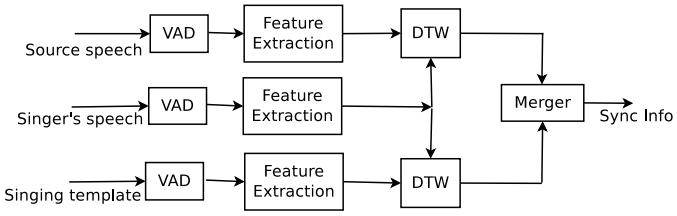

Fig. 2. Alignment scheme proposed in [31] for TSTS systems.

The learning stage is followed by the transformation stage, in which the speech and singing signals were analyzed using a vocoder (eg: STRAIGHT) to extract short-time spectral, F0 and aperiodicity parameters [37], [38]. The F0 transformation model directly modified the F0 contour of source speech by replacing it with the F0 contour of singing template. Phoneme durations in source speech were estimated by DTW, and later compressed/elongated to match with singing template durations [34]. The matching of phoneme durations was done based on the synchronization information acquired in the learning stage. The spectral parameters of the source speech were preserved intact, and spectral vectors were replicated or deleted according to the modified phoneme durations. In the synthesis stage, singing voices corresponding to the source speaker were synthesized by a vocoder with the modified parameters [38]. The functioning of the TSTS system is illustrated in Fig. 3, where the temporal alignment module is constituted by the alignment scheme illustrated in Fig. 2.

C. Challenges in speech to singing voice alignment

The human voice production mechanism behind singing voices has certain distinctive characteristics, which are not present in the speech production system. Depending upon

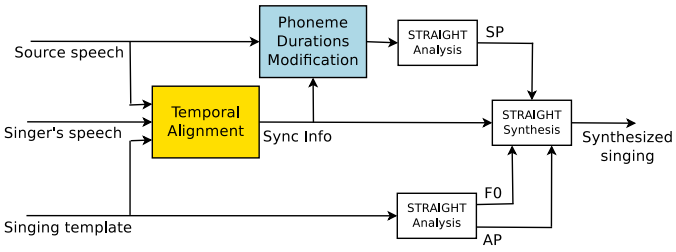

Fig. 3. The Template-based STS (TSTS) system.

the style of singing, mode of phonation producing different singing expressions and varying loudness, the subglottal pressure changes considerably [5]. The changes in subglottal pressure forces F0 to change to a much larger extend in comparison with speech sounds. Together with variations in F0, the epochal strengths of glottal flow, peak glottal flow derivative and the overall sound pressure level also change [3], [5]. To beautify the singing, singers often introduce quasi periodic pitch changes known as vibrato. Also with natural singing, the F0 contour exhibit pitch changes between musical notes above a certain threshold (overshoot), pitch changes in a direction opposite to the musical note changes (preparation) and other fine fluctuations [23]. As human singing is a natural process, the F0 contours obtained from singers will always be different from ideal synthetic pitch tracks corresponding to musical notes [39]. All these unique characteristics of excitation signal in voice production mechanism make singing to exhibit distinctive differences from the corresponding spoken sounds.

Together with the excitation characteristics, the singing voices exhibit peculiar spectral properties as well. The singing formant is the most prominent property, which is an emphasized peak in spectral envelope around 3 kHz [4]. Also it is observed that the formants in singing spectra undergo certain frequency modulations depending on variations in F0 contour [8]. Due to these characteristics, the singing voices can be very different from spoken voices even if the underlying linguistic information are the same. Hence, the conventional DTW may fail in aligning speaking and singing voices. Particularly, if there exist a difference in linguistic content due to the presence of crosstalk, the DTW alignment shows high mismatch. The alignment procedure shown in Fig. 2 uses DTW to match speech-speech and speech-singing signals. This will result in accumulation of DTW errors at the 'merger' shown in Fig. 2. Thus the alignment scheme presented in [31], [34] will force large errors, causing voice quality degradations in synthesized sounds. Hence, there exist a persistent requirement for an improved alignment algorithm, which can work even in adverse situations.

III. Dual alignment scheme with subDTW

In this section, we elaborate the proposed dual alignment scheme for aligning speech and singing voices. The problem of frame-by-frame alignment of speech to singing signals is not trivial due to the reasons explained in Section II-C. Instead,

aligning short-time frames of two speech signals is a more straightforward problem as they share similar characteristics.

Fig. 4 shows the alignment between singing and speech signals, as well as the alignment between two speech signals using the conventional DTW technique. The DTW algorithms discussed in this work use cosine distance metric between MFCC vectors from audio signals. The ground truths regarding word boundaries in the audio signals are manually marked for reference. It can be observed from Fig. 4 that the alignment between two speech signals are more reliable than the same between speech and singing signals. The speech-singing signal alignment using DTW forces errors as large as about 1 second, in matching word boundaries. But the timing errors due to mismatch of word boundaries made by the DTW algorithm for two speech signals are considerably less than the same with speech-singing alignment, as can be seen from Fig. 4(b). Hence we choose to use a dual alignment scheme for TSTS conversion systems, different from the one proposed in [31], [34] as detailed below.

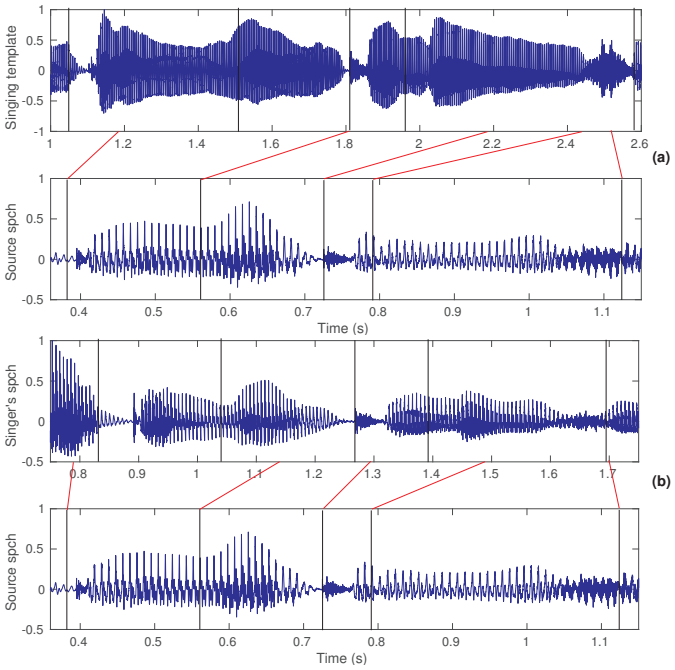

Fig. 4. Alignment using conventional DTW technique: (a) Speech-singing alignment and (b) Speech-speech alignment. The vertical lines on the waveforms represent word boundaries and the red lines show the alignment between word boundaries in signals.

While building the singing templates database, the singer's speech and singing templates from professional singers are recorded. The signals are segmented into lyrical sentences and the corresponding sentence-level transcriptions are made available. For the STS conversion through STRAIGHT analysis-modification-synthesis [38], the estimation of phoneme boundaries of speech and singing signals has to be done accurately. The TSTS system in [31] blindly uses the DTW technique for aligning the phoneme boundaries, which is not a reliable procedure. We employ a semi-automated process for this task. The initial phoneme boundaries are marked using forced alignment by an automatic speech recognizer (ASR). In this work, we have used the Montreal forced aligner based on pretrained ASR models [40]. Later the phoneme boundaries delivered by ASR forced aligner are manually verified and corrected for any errors. Even though this task demands exhaustive manual effort, it is a part of our database preparation and has to be done only once for numerous repeated usages in future. Since the accuracy of marking phoneme boundaries plays a very crucial role in the performance of TSTS systems, we believe that this process is advantageous. Thus the phoneme boundaries for singer's speech and singing templates are already placed, as accurately as it can be, and the subsequent alignments are prepared.

We propose to do automatic alignment of source speech to singer's speech as this has to be done multiple times in run-time, unlike the database preparation. The automatic alignment forms the first pass in the dual alignment scheme. We use the manually corrected alignment already available between singer's speech and singing template from the database as the second pass to complete the dual alignment. But we refrain from using the conventional DTW technique for aligning speech signals. Even though the error produced by the DTW technique was lesser in speech-speech alignment than in speech-singing alignment, as illustrated in Fig. 4, it is far from an optimum solution for matching two continuous speech signals. Also, the crosstalk attacks in real world scenario can cause the DTW algorithm to fail miserably in aligning speech signals, which can result in poor performance of the TSTS system. Hence we choose to employ a variant of DTW, termed as subsequence DTW (subDTW) in this work [36].

The subDTW was previously used for the task of query word detection, in which an isolated audio query word will be searched in a continuous speech database [41]. It inherently assumes that the query word is considerably shorter in duration than the search database. In TSTS system, the source speech can be segmented into isolated words to form queries and the singer's speech can act as the search sentence. The segmented words from source speech are subsequences of the singer's speech sentences, and can be searched for using subDTW. We have used the Montreal forced aligner to mark word boundaries of source speech and then segment the continuous speech into word units. While segmentation, we allow ten additional frames on either ends of words to accommodate for any alignment error. There is no manual intervention for error correction at this stage of the process. The subDTW compensates for errors in word boundaries by giving nearly optimal alignment paths. Any residual error will be nullified by the concatenation of alignment paths from subDTW of subsequent words, by allowing to choose the nearest neighboring frame in singer's speech at the overlapped word boundaries. Thus subDTW and the associated post processing of alignment paths successfully outperform the conventional DTW in aligning continuous speech signals.

Also, the subDTW addresses the problem of unintentionally recorded additional words in source speech. If the query

word is present in the search database, then the subDTW will give a 'hit' and the corresponding alignment path with least cost. If the query word is absent in the search database, then the subDTW will give a 'miss'. This specific characteristic of subDTW will help in directly nullifying the crosstalk/babble attacks in the source speech. We assume that the crosstalk/babble will not contain the poetic lyrics of songs and they are not overlapping with the source speech. Even if the crosstalk contains similar words to those in lyrical sentences, it is observed that the subDTW remains robust to a considerable extend as it always deliver 'hits' of query words with alignment paths of least cost.

Fig. 5 illustrates the comparison between alignments of word boundaries between segments of speech signals, delivered by DTW and subDTW algorithms. It is clearly demonstrated that the word boundary errors produced by the subDTW are lesser than those with DTW. The frame-by-frame alignment obtained between source speech and singer's speech using subDTW technique is utilized to pick the corresponding manually corrected alignment between singer's speech and singing templates from the database. Thus the dual alignment between source speech and singing template is realized in two passes as (i) automated subDTW-based alignment between source speech and singer's speech and (ii) picking corresponding alignment between singer's speech and singing template from the true alignment available in the database.

The overall alignment between source speech and singing template in two passes is illustrated in Fig. 6. The time differences between adjacent black solid lines and black dotted lines in the 'singer's spch' in Fig. 6 represent alignment errors produced by subDTW in the first pass. As nearly accurate timing information is used for alignment in the second pass, no additional error is generated. Thus the overall alignment error in the TSTS system is incurred only by the subDTW in the first pass.

The proposed scheme of dual alignment is shown in the Fig. 7. For implementation of the TSTS system, the alignment module (shown as yellow box) in Fig. 3 will be constituted by the proposed scheme shown in Fig. 7. Also, we used the synchronous overlap-add fixed synthesis (SOLAFS) algorithm [42] to modify phone durations in the TSTS system (shown as blue box in Fig. 3), as opposed to direct DTW alignment and spectral replications/deletions used in [31]. This choice of phoneme durations modification, together with the proposed dual alignment scheme, have helped in improving the quality of synthesized singing.

IV. Experimental Evaluation

To validate the efficacy of the proposed dual alignment scheme with subDTW, we conduct objective and subjective evaluations. We chose a database of 3 English songs, namely, 'I dont want to lose you', 'Stars shining bright above you' and 'Fly me to the moon', sung by a male and a female singer. The read speech data of each singer reading out the lyrics of the three songs are also recorded. The three songs collectively contributed 80 lyrical sentences, totaling to 604 words, for

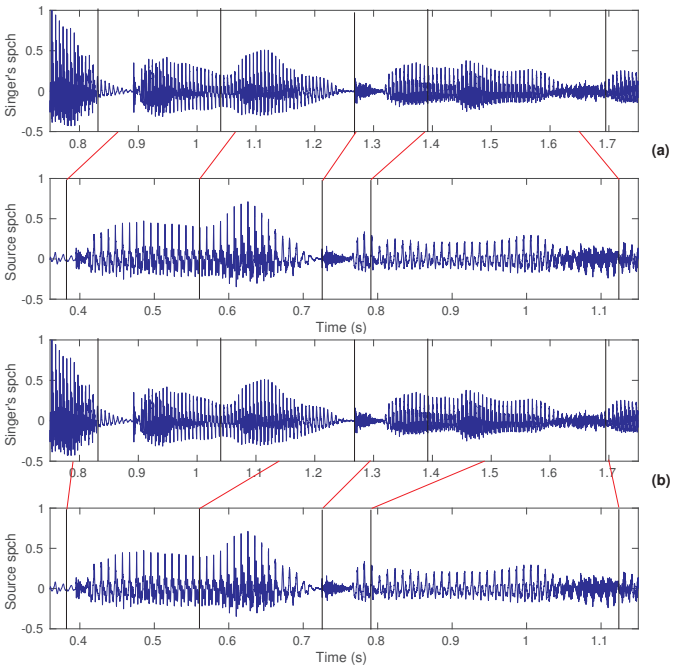

Fig. 5. Alignment between source speech and singer's speech (a) using subDTW algorithm and (b) using conventional DTW algorithm. The vertical lines on the waveforms represent word boundaries and the red lines show the alignment between word boundaries in signals.

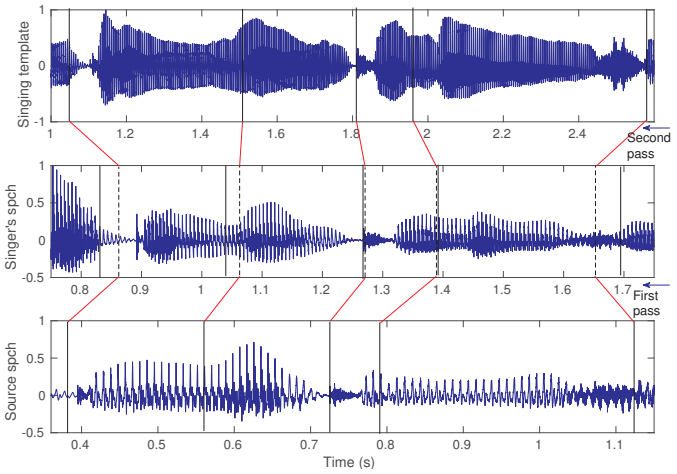

Fig. 6. Dual alignment between source speech and singing template. The vertical lines on the waveforms represent word boundaries and the red lines show the alignment between word boundaries in signals. The vertical dotted lines on 'Singer's spch' represent the estimated boundaries by subDTW alignment.

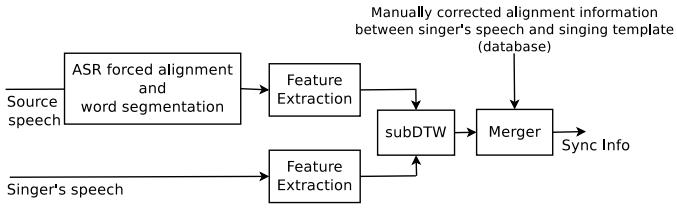

Fig. 7. The proposed dual alignment scheme with subDTW for TSTS system.

each singer. The average spoken duration of the sentences is about 3 seconds. The phoneme-level transcriptions of singing templates and singer's speech data are prepared as described in Section III. Later, the source speech corresponding to the 80 sentences are recorded by a male and a female user. The word-level transcriptions of the source speech sentences are also prepared. These transcriptions are used only as references for experimental evaluations, not for synthesizing voices from the TSTS system. Speech signals are converted to singing voices using the TSTS system shown in Fig. 3, where the alignment module is constituted by the dual alignment scheme illustrated in Fig. 7.

A. Objective studies

We utilized DTW to align 160 source speech sentences with 160 singing sentences in the database (spch-sing), belonging to both the male and the female singers. Also, the alignment of source speech sentences with corresponding singer's speech sentences (spch-spch) was attempted. We did not perform any cross gender trials. The mean of absolute values of alignment errors in marking word boundaries was computed with respect to the reference transcriptions. The mean error in seconds for both spch-sing and spch-spch alignments computed over 160 lyrical sentences (1208 words), are reported in Table I. The DTW-based word alignment error in spch-sing alignment is almost twice the error from spch-spch alignment. Hence, we proposed a dual alignment scheme for TSTS systems, in which the source speech will be aligned with singer's speech and the manually corrected transcriptions in the database will be used to align singer's speech with singing templates. Thus the overall alignment error of the system will simply be the error occurring in the first pass of dual alignment, as the second pass utilizes nearly accurate annotations of singer's speech and singing templates.

TABLE I
MEAN WORD-BOUNDARY ALIGNMENT ERROR USING CONVENTIONAL DTW ALGORITHM.

Alignment	Mean error (s)
spch-sing	1.5747
spch-spch	0.7867

Instead of using conventional DTW technique to align source speech to singer's speech, we propose to use the subDTW algorithm in the dual alignment. As illustrated in Fig. 5, the subDTW outperforms DTW in aligning continuous speech signals. To quantitatively validate this observation, we report the mean word alignment error produced by DTW and subDTW in aligning 160 source speech sentences with singer's speech sentences in Table II. The subDTW (subDTW (spch-spch)) indeed outperforms conventional DTW (DTW (spch-spch)). In fact the alignment error produced by subDTW is merely about 8% of the error produced by the DTW algorithm, demonstrating its superiority. We also analyzed the capability of subDTW in directly aligning source speech to singing template (subDTW (spch-sing)) and, the subDTW had unambiguously delivered better performance than the DTW algorithm in this task also. The mean word-boundary alignment error produced by the proposed dual alignment scheme is 0.07 seconds, as opposed to the error of 1.57 seconds produced by the baseline DTW approach.

TABLE II
MEAN WORD-BOUNDARY ALIGNMENT ERROR USING DIFFERENT VARIANTS OF DTW ALGORITHM.

Algorithm	Mean error (s)
Alignment of source speech to singer's speech	
DTW (spch-spch)	0.7867
subDTW (spch-spch)	**0.0662**
Alignment of source speech to singing template	
DTW (spch-sing)	1.5747
subDTW (spch-sing)	0.3850

B. Subjective studies

We conducted subjective experiments using singing voices synthesized by the TSTS system from source speech, employing different alignment techniques as (i) subDTW for direct alignment of source speech to singing template (subDTW (spch-sing)), (ii) the proposed dual alignment scheme (subDTW (spch-spch)), and (iii) conventional DTW for direct alignment of source speech with singing template (DTW (spch-sing)). We have also included the baseline TSTS system using DTW for aligning both passes in the dual alignment [31]. Notice that, the perceptual quality of synthesized singing depends on several factors including the accuracy of temporal alignment, the proper conversion of spectral characteristics, pitch mapping incorporating excitation features like vibrato, overshoot, etc. In this work, we only study the contribution of alignment to perceptual quality of synthesized singing through subjective studies.

Fifteen neutral listeners, aged 15 to 35, with normal hearing ears had volunteered for the subjective study. Each volunteer had listened to three sets of audio files, each containing four singing voices corresponding to the three alignment techniques mentioned above, together with the baseline. The audio files were played to the listeners monoaurally through headphones in a normal room environment. They rated the singing voices generated from TSTS systems based on naturalness, distortions and overall voice quality. The ratings are given on a scale of 1 to 5, where 1 denotes unacceptable, 2-poor, 3-fair, 4-good and 5 denotes excellent. The opinion scores provided by the listeners were averaged over all trials including male and female singing. The mean opinion scores (MOS) are reported in Table III. The proposed dual scheme with subDTW had outperformed all the other alignment techniques in TSTS systems, providing a relative improvement of 38.7% in MOS scores over the baseline system.

The MOS reported in Table III shows only the average value of opinion scores awarded by the listeners, which is a very limited representation. To illustrate the efficiency of the proposed alignment method based on the entire set of opinion scores, the boxplot is shown in Fig. 8. The boxplot takes into account of individual values of opinion scores and employs

a notched box to enclose all values between the 25^{th} and 75^{th} percentiles of opinion scores. The notch of the boxes (shown as red line in Fig. 8) represents the median value of opinion scores, and the ends of whiskers over each box represent the extreme values. It can be observed from Fig. 8 that the set of individual opinions scores for the proposed alignment algorithm (subDTW (spch-spch)) is distinguishably larger than those corresponding to the other techniques. Thus its effectiveness is unequivocally demonstrated, not just in terms of mean or median values of opinion scores.

In MOS scoring, the listeners were asked to listen to the audio samples and rate them based on their individual perceptual quality. The listeners did not compare the samples or discriminate them from each other. While interpreting MOS scores, we manually compared the individual scores given by the listeners using average values and boxplots. In order to evaluate the comparative perceptual quality of audio samples, all of the 15 listeners were asked to choose the best and worst sounding singing from each of the three sets of 4 singing voices. Each listener had listened to the individual singing voices within a set, which were played multiple times in different orders. And, the best and worst samples from each set were chosen by the listeners.

We perform a best-worst scoring (BWS) to quantify the inter-relationships in perceptual quality of samples and to nullify any confusions by listeners in evaluating perceptually similar segments [43]. We selected the BWS score on the aggregate level, that is with respect to the entire set of trials. The necessary experimental conditions for BWS scoring are (i) each trial should contain the same number of items to be studied and (ii) each item should have equal number of samples across the entire set of trials. Also, samples within a set should be played in all possible permutations to listeners before they choose the best and worst sounding voices [43]. These conditions are satisfied in our subjective study and the resultant BWS score for each item i, is computed as:

$$(\text{BWS})_i = \frac{B_i - W_i}{N_i} \tag{1}$$

where B_i and W_i denote the number of times the item i is chosen as 'best' and 'worst', respectively by listeners. N_i denotes the number of times the item i is appearing in the entire set of trials and, $N_i = N, \forall i$ based on the necessary conditions for BWS scoring on aggregate level [43]. The most positive BWS score indicate that the item is most appealing to the listeners and vice versa. The BWS scores computed for different alignment techniques in the TSTS systems are reported in Table. III, from which it can be observed that the BWS score for the proposed alignment technique is the most positive and consequently the most appealing to the listeners. Both the MOS and BWS scores for the proposed alignment technique are significantly better than the other techniques under consideration. Also, the proposed dual alignment scheme with subDTW clearly outperforms the baseline system.

TABLE III
SUBJECTIVE EVALUATION OF DIFFERENT ALIGNMENT TECHNIQUES USED IN TSTS SYSTEMS.

Algorithm	MOS	BWS
subDTW (spch-sing)	2.2111	-0.2667
subDTW (spch-spch)	**3.7311**	**0.7111**
Baseline system	2.5333	-0.1333
DTW (spch-sing)	1.9667	-0.3111

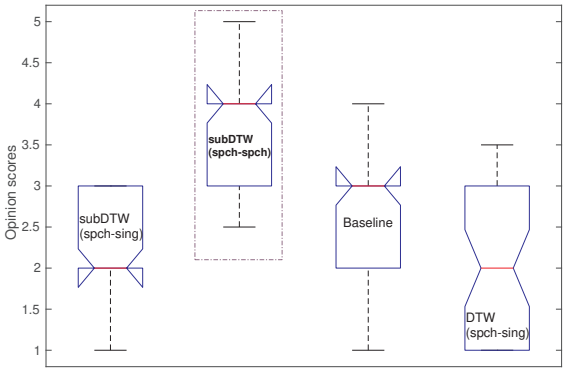

Fig. 8. The boxplot of opinion scores for synthesized singing obtained using different alignment techniques. The boxplot within the dotted rectangle denotes the proposed dual alignment.

C. Comparison with existing techniques

The baseline TSTS system uses silence removal and voiced-unvoiced decisions to preprocess signals and uses conventional DTW algorithm to align source speech to singing template [31]. The idea of dual alignment is intuitively used in the baseline system, but it was implemented using DTW in both passes resulting in cumulating the errors in DTW along each pass. There was no usage of manually corrected phoneme transcriptions in the baseline TSTS system. We use subDTW to match source speech with singer's speech in the first pass, as it is a more efficient tool than DTW for aligning spoken sounds. Motivated from the fact that DTW, or even subDTW, are not good enough to directly align speech to singing signals, we made use of correct transcriptions to align singer's speech to singing templates in the second pass. Thus we proposed a more efficient strategy in comparison with the baseline, as illustrated in the subjective and objective studies. In addition, we do not use silence removal or voicing decisions for preprocessing as subDTW delivers alignment paths which actually matches to segments in singer's speech.

Compared to score-based systems, we use human singing templates to modify parameters of speech and hence synthesized singing will be more natural. Also, the score-based systems need to force align the speech signals every time it is being used. Manual corrections may be repeatedly required to ensure the accuracy of the durations and boundaries of phonemes to avoid erroneous synthesis. In this work, we perform accurate alignment of templates in the database only once during the database preparation, with no manual intervention in any other stage of processing.

V. Conclusions

In this paper, we proposed a dual alignment scheme for speech-singing alignment in TSTS systems. We employed a preferable variant of DTW, namely the subDTW algorithm, to align source speech with singer's speech in the first pass of dual alignment. And we used the manually verified alignment information computed during database preparation for TSTS systems, to align singer's speech with singing templates in the second pass of dual alignment. The proposed alignment scheme had consistently outperformed the temporal alignment used in the baseline TSTS systems in both objective and subjective evaluations. The utilization of improved alignment techniques is bound to improve the perceptual quality and naturalness of synthesized singing.

Acknowledgement

The authors would like to acknowledge the NUS Start-up grant with WPS number: R-263-000-C35-133/731 for supporting this work.

References

[1] M. Dong, S. W. Lee, H. Li, P. Chan, X. Peng, J. W. Ehnes, and D. Huang, "I2R speech2singing perfects everyone's singing," in *Interspeech*, 2014, pp. 2148–2149.

[2] M. Goto, T. Saitou, T. Nakano, and H. Fujihara, "Singing information processing based on singing voice modeling," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 5506–5509.

[3] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[4] J. Sundberg, "The level of the 'singing formant' and the source spectra of professional bass singers," *Quarterly Progress and Status Report: STL-QPSR*, vol. 11, no. 4, pp. 21–39, 1970.

[5] J. Sundberg, I. R. Titze, and R. Scherer, "Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source," *Journal of Voice*, vol. 7, no. 1, pp. 15 – 29, 1993.

[6] S. Wang, "Singer's high formant associated with different larynx position in styles of singing," *Journal of the Acoustical Society of Japan (E)*, vol. 7, no. 6, pp. 303–314, 1986.

[7] M. L. Erickson, S. Perry, and S. Handel, "Discrimination functions: Can they be used to classify singing voices?" *Journal of Voice*, vol. 15, no. 4, pp. 492 – 502, 2001.

[8] I. Arroabarren and A. Carlosena, "Vibrato in singing voice: The link between source-filter and sinusoidal models," *EURASIP Journal on Advances in Signal Processing*, no. 2004:720342, pp. 1007–1020, June 2004.

[9] M. Sakaguchi, M. Kobayashi, R. Nisimura, T. Irino, and H. Kawahara, "Sectrally estimated vocal tract lengths of singing voices and their contributing factor," in *Proc. MAVEBA*, 2013, pp. 121–124.

[10] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *Eurospeech*, 2005, pp. 1141–1144.

[11] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno, "Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals," in *Eighth IEEE International Symposium on Multimedia (ISM'06)*, Dec 2006, pp. 257–264.

[12] J. S. Sung, D. H. Hong, H. W. Koo, and N. S. Kim, "Factored maximum likelihood kernelized regression for HMM-based singing voice synthesis," in *Interspeech*, 2013, pp. 359–363.

[13] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis and its application to Japanese and English," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 265–269.

[14] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, "HMM-based expressive singing voice synthesis with singing style control and robust pitch modeling," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 308–322, Nov. 2015.

[15] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638–648, March 2010.

[16] Y. Ohishi, H. Kameoka, D. Mochihashi, H. Nagano, and K. Kashino, "Statistical modeling of f0 dynamics in singing voices based on gaussian processes with multiple oscillation bases," in *Interspeech*, 2010, pp. 2598–2601.

[17] W. H. Lai and S. F. Liang, "An f0 control model for singing synthesis based on proportional-integral-derivative controller," in *Tenth IEEE International Symposium on Signal Processing and Information Technology*, Dec 2010, pp. 182–185.

[18] Y. R. Chien, H. M. Wang, and S. K. Jeng, "An acoustic-phonetic model of f0 likelihood for vocal melody extraction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1457–1468, Sept 2015.

[19] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr 1997, pp. 435–438.

[20] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," in *Interspeech*, 2007, pp. 4009–4010.

[21] F. Villavicencio and J. Bonada, "Applying voice conversion to concatenative singing-voice synthesis," in *Interspeech*, 2010, pp. 2162–2165.

[22] 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.

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

[24] ——, "Vocal conversion from speaking voice to singing voice using straight," in *Interspeech*, 2007, pp. 4005–4006.

[25] S. W. Lee and M. Dong, "Singing voice synthesis: Singer-dependent vibrato modeling and coherent processing of spectral envelope," in *Interspeech*, 2011, pp. 2001–2004.

[26] K. Kobayashi and T. Toda, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Interspeech*, 2014, pp. 2514–2518.

[27] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion based on direct waveform modification with global variance," in *Interspeech*, 2015, pp. 2754–2758.

[28] K. Kobayashi, T. Toda, and S. Nakamura, "F0 transformation techniques for statistical voice conversion with direct waveform modification with spectral differential," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec 2016, pp. 693–700.

[29] M. Dong, N. Chen, and H. Li, "Speech synthesis perfects everyone's singing," *IEEE Signal Processing Society Newsletter*, 2014.

[30] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *2010 IEEE International Conference on Multimedia and Expo*, July 2010, pp. 1421–1426.

[31] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4509–4512.

[32] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06. ACM, 2006, pp. 659–662.

[33] Y. R. Chien, H. M. Wang, and S. K. Jeng, "Alignment of lyrics with accompanied singing audio based on acoustic-phonetic vowel likelihood modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1998–2008, Nov 2016.

[34] L. Cen, M. 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*, 2011.

[35] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb 1978.
[36] M. Müller, *Information Retrieval for Music and Motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.
[37] P. Zolfaghari, Y. Atake, K. Shikano, and H. Kawahara, "Investigation of analysis and synthesis parameters of straight by subjective evaluation," in *ICSLP*, 2000.
[38] 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: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187 – 207, 1999.
[39] D. Gerhard, "Pitch track target deviation in natural singing." in *ISMIR*, 2005, pp. 514–519.
[40] M. Michael, M. Socolof, S. Mihuc, M. Wagner, , and M. Sonderegger, "Montreal forced aligner: an accurate and trainable aligner using Kaldi," in *91st Annual Meeting of the Linguistic Society of America*, Austin, TX, 2017.
[41] K. Rout, P. R. Reddy, and K. S. R. Murty, "Experimental studies on effect of speaking mode on spoken term detection," in *2015 Twenty First National Conference on Communications (NCC)*, Feb 2015, pp. 1–6.
[42] D. Hejna and B. R. Musicus, "The SOLAFS time-scale modification algorithm," BBN, Tech. Rep., Jul. 1991.
[43] T. Flynn and A. Marley, *Best-worst scaling: theory and methods*. Cheltenham, UK: Edward Elgar Publishing, Inc., 2014. [Online]. Available: //www.elgaronline.com/9781781003145.00014.xml