Parallel Speak-Sing Corpus of English and Chinese Songs for Speech-to-Singing Voice Conversion

Karthika Vijayan, Haizhou Li
Department of Electrical and Computer Engineering
National University of Singapore, Singapore
{vijayan.karthika, haizhou.li}@nus.edu.sg

Abstract

We present a continuing data collection effort towards building a rich database of English and Chinese pop songs for efficient speech-to-singing (STS) voice conversion. Parallel recordings of lyrics of songs, sung and read by professional singers, are recorded in a professional studio environment using high quality recording equipments under the supervision of a trained sound engineer. Sentence-level and word-level labeling of the read and sung audio files are performed manually. Then temporal alignment between frames of words in the read lyrics and singing is performed automatically using dynamic time warping (DTW) with carefully crafted features. The accuracy of temporal alignment of frames of speech and singing voices is crucial for STS conversion, as it decides the effectiveness of mapping of parameters from speech signals to those of singing. The temporally aligned frames of speech and singing voices are used to map characteristics for STS conversion. The presented database of parallel recordings of speaking and singing voices of same linguistic content assist in facilitating efficient STS conversion, in addition to providing valuable resorts to singing voice analysis, understanding differences in production-perception of speech and singing voices and, evaluation of singing quality.

Keywords: Singing voice analysis, Speech-to-singing, Parallel speak-sing corpus

1. Introduction

The speech-to-singing (STS) voice conversion is a relatively recent application, gaining momentum now-a-days due to the extensive interest from the entertainment industry. In STS conversion, the read lyrics of a song is converted to perfect singing, while retaining the speaker identity of the person reading the lyrics. This task involves mapping the prosody of read lyrics to that of singing, preserving the timbre of the read speech (Saitou et al., 2007a),(New et al., 2010). The STS conversion finds numerous applications related to training and evaluation of singing skills of music students or amateur singers, beautifying singing in karaoke systems, music compositions, singing voice analysis and modeling, better understanding of the relationship between speaking and singing voice styles, etc (Vijayan et al., 2017). Hence building resources to devise efficient schemes for STS conversion will benefit various aspects of singing voice processing.

The basic technique behind STS conversion is demonstrated in the Figure 1. The perfect singing is synthesized by combining the melody of singing extracted from the prosody of a reference singing template/musical score and, speaker characteristics of the person reading the lyrics extracted from the timbre of read speech (Saitou et al., 2007b), (New et al., 2010). The reference singing template consists of a professional singer rendering high quality singing vocals and, the reference musical score consists of the target melody of singing denoted in standard MIDI format. Either reference singing template or target musical score is required for deriving the melody of singing for STS conversion. In this paper, we consider that the reference singing template from a professional singer is available for template-based STS conversion. The linguistic content in the read lyrics and reference singing template are temporally aligned to ensure that correct timbre of a frame of read lyrics is being combined with the prosody from corresponding frames of reference singing template. Later the prosody and timbre from aligned frames of reference singing template and read lyrics, respectively, are combined together to produce synthesized singing in an STS conversion system.

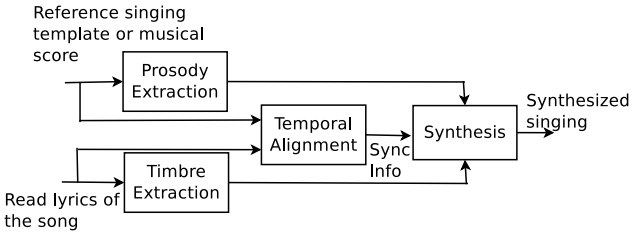


Figure 1: Basic technique for STS conversion.

The accuracy of the synchronization information (Sync Info) from temporal alignment plays crucial role in deciding the accuracy of synthesis. The problem of temporal alignment of linguistic content of read lyrics and reference singing template is not a straight-forward problem. Even though the linguistic content in the read lyrics and reference singing template is the same, the signals corresponding to them vastly differ from each other. The spoken and sung vocals produced by human voice production mechanism exhibit many similar characteristics due to the similarity in the vocal tract system producing them. But there exist some unique properties of the vocal tract system configuration and excitation source, that distinguish between the speaking and singing voice styles (Lindblom and Sundberg, 2007). Due to these differences between speaking and singing, the problem of temporal alignment between read lyrics and reference singing template becomes very challenging.

In this paper, we present a parallel speak-sing database, which can provide rich resource to learn the prosody char-

acteristics from reference singing templates and timbre characteristics from read lyrics for efficient STS conversion. We explain an ongoing data collection process in which professional singers are hired to sing good quality song vocals and also read the lyrics of the songs in natural manner. We then proceed to explain the temporal alignment of frames of read speech to the corresponding frames in reference singing template. The resultant database of English songs will either be released in public domain or be shared upon request, tentatively by the end of March, 2018.

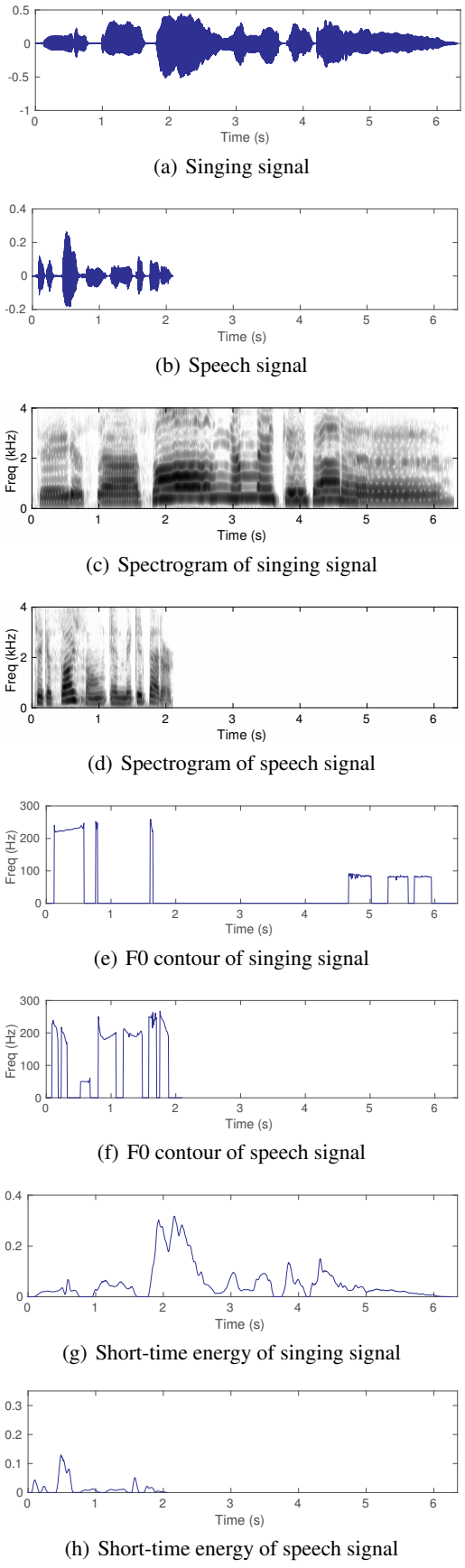
The rest of the paper is organized as follows: In Section 2., we explain the differences between speech and singing voice styles that make the process of temporal alignment between them a difficult task. We also discuss the significance of the parallel speak-sing database in STS conversion and details the requirement for accurate temporal alignment between read lyrics and reference singing template. Section 3. explains the parallel speak-sing database under preparation. In Section 4., we elaborate the method devised for effective temporal alignment of read lyrics and reference singing template. In Section 5., we summarize the contributions of this paper towards STS conversion and indicate other significant applications of the presented database, in addition to STS conversion.

2. Speaking and Singing Voice Styles

Vocal sounds are produced as the response of a time-varying vocal tract system to a time-varying glottal excitation signal (Rabiner and Schafer, 1978). Spoken and sung vocals are produced by the same voice production mechanism. Hence they exhibit several similar characteristics like, the consistency in lower order formants. However, the voice production characteristics vary considerably across speaking and singing. A prominent difference between speaking and singing is in the duration of phones. It can be observed that the voiced sounds are largely elongated or compressed depending upon the melody of singing, whereas, the unvoiced sound durations are relatively retained in singing with respect to those in speaking. Also, the dynamic range of singing amplitude is much larger than the same in speaking, resulting in the vast difference in energy of singing voice with respect to spoken sounds (Titze and Sundberg, 1992).

The spectral characteristics of singing vocals differ from those in speaking. Particular positioning of larynx while producing loud singing results in clustering of higher order formants to form the 'singing formant', which serves as the major resource for production of loud and high pitch singing (as in Opera) without the need for increasing the subglottal pressure beyond the capacity of a human singer. The characteristics of excitation source also change vastly in singing. The fundamental frequency of glottal vibrations (F0), equivalently the pitch, is aided by the target melody of singing. But, the pitch in speaking stays relatively flat. As the pitch and/or loudness of singing increases, the subglottal pressure increases as well (Sundberg et al., 1993), (Sundberg et al., 2005), (Sundberg, 2009). To summarize, the major differences between speaking and singing voice styles are,

- Duration
- Energy



(a) Singing signal

(b) Speech signal

(c) Spectrogram of singing signal

(d) Spectrogram of speech signal

(e) F0 contour of singing signal

(f) F0 contour of speech signal

(g) Short-time energy of singing signal

(h) Short-time energy of speech signal

Figure 2: Difference between speaking and singing voices.

- Pitch
- High frequency spectrum (singing formant)

The differences between singing and speaking in terms of these factors can be observed in Figure 2.

2.1. Significance of a parallel speak-sing database

The differences between speech and singing signals as discussed above, present several difficulties in STS conversion such as mapping of dissimilar characteristics, automatic temporal alignment, etc. A parallel speak-sing database of same linguistic content from the same singer can be proven beneficial in such scenario. As the same professional singer sings and reads the lyrics of a song, the speaker identity is preserved in the parallel recordings. This can provide additional advantages, alongside to the same linguistic content, for accurate automatic temporal alignment. Once the read lyrics and reference singing template from the professional singer are aligned, the temporal alignment of read lyrics by an unknown user to reference singing template will be reduced to a speech-speech alignment problem (Vijayan et al., 2017). This problem can be dealt with effectively using dynamic time warping (DTW). Also, the parallel recordings of read lyrics and reference singing template will render an easy resource for learning the speaker-dependent mapping of characteristics from speech signals to singing signals. Thus the parallel speak-sing database construction effort is valuable for designing an efficient STS conversion system.

3. Parallel Speak-Sing Database

In this section, we detail an ongoing data collection effort for designing an efficient STS conversion system. We record singing and speech audio signals in a professional studio environment, employing high quality recording devices under the supervision of a trained and experienced sound engineer. We hired professional singers who either have a diploma in vocal training or have an experience of more than three years in public singing. We provide them with a list of English pop songs and Chinese pop songs, from which they choose 10 songs according to their singing capabilities and vocal range. Special attention was taken to choose singers who can speak and sing in English language without prominent mother tongue influence. The list of English songs provided to the singers is given in Table 1.

The recording of read lyrics is performed by instructing the singers to read the lyrics of songs in their natural speaking manner, without taking long pauses in between. The recording of singing vocals is performed in synchronization with the background music of the corresponding song. The background music is played via headphones to the singer and, each singer is instructed to sing vocals with respect to the background score. Multiple takes are recorded whenever necessary to ensure the pronunciation is correct, the vocals are in-tune with the background music, etc. Currently, two male and two female singers have completed their recording of read lyrics and singing vocals corresponding to three English songs, namely, 'I dont want to lose you', 'Stars shining bright above you', and 'Fly me to the moon'.

The continuing effort of data collection aims at recording the read lyrics and singing vocals by five male and five female singers, each recording 10 songs. Thus we aim to build a rich database of atleast 100 songs each, from English and Chinese pop genre. This database will aid our ongoing research of implementing an efficient STS conversion system. A similar database of parallel recordings of spoken lyrics and singing vocals can be found here (Duan et al., 2013).

4. Temporal Alignment

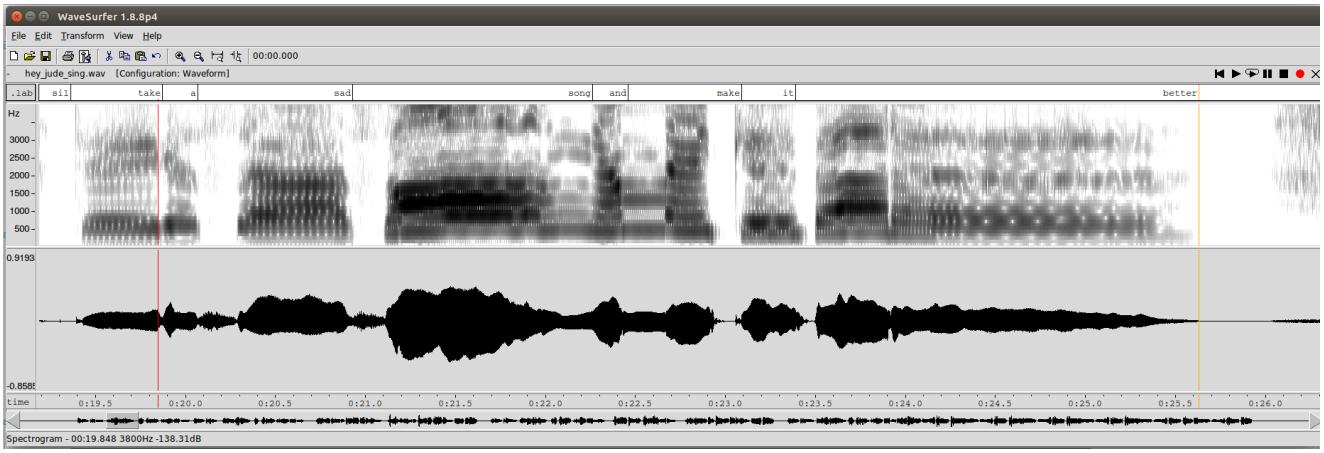
As a part of our database development, we attempt to provide the temporal alignment between read lyrics and singing vocals. The recorded singing vocals will become the reference singing templates for STS conversion. The temporal alignment between read lyrics from an unknown speaker/user of the STS conversion system (user speech) and the reference singing template is extremely crucial for the effectiveness of the STS conversion. The accuracy of temporal alignment will have decisive role in the mapping of characteristics from speech signals to singing. Any error in temporal alignment of linguistic content between speech and singing will result in mismatched combination of timbre from user speech and prosody from reference singing template, consequently producing annoying distortions in the synthesized singing. In our data collection effort, we attempt to temporally align the ready lyrics by the professional singer to the reference singing template. The read lyrics by the singer can act as a bridge between the user speech and the reference singing template in run-time STS conversion (Vijayan et al., 2017).

We perform manual labeling of the read lyrics and singing vocals, at sentence-level and word-level. Researchers and students from our lab, who have working knowledge of speech processing, inspected the waveforms and spectrograms of audio recordings using the tool wavesurfer (Sjolander and Beskow, 2000). They manually label the sentence- and word-level boundaries, creating the required label files. The screenshots of word boundary labeling using wavesurfer for read lyrics and singing vocal are shown in Figure 3.

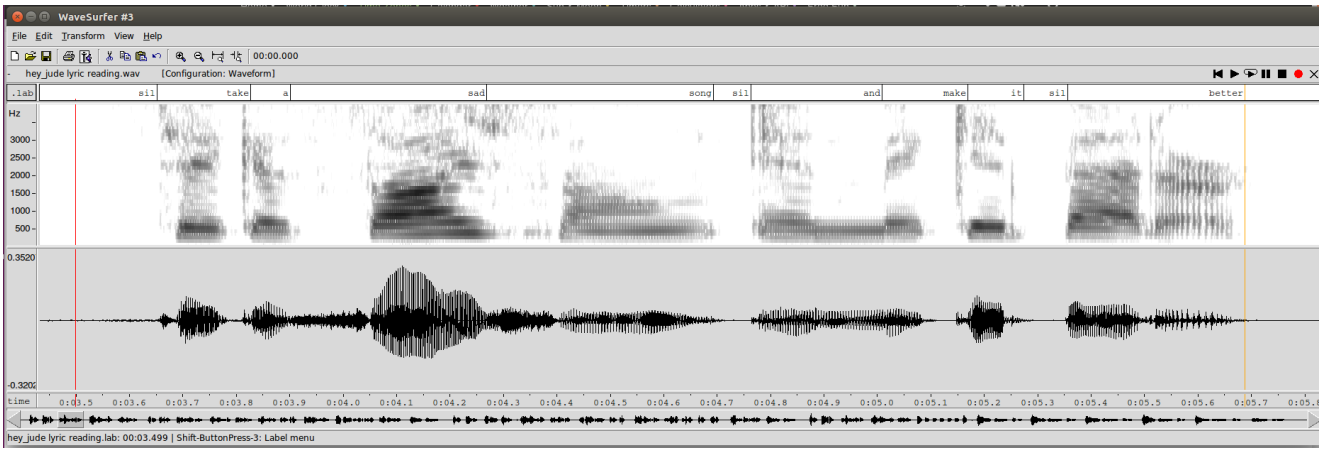
Once the word-level boundaries are accurately marked, we attempt the frame-level alignment of words from read lyrics to singing vocals. The DTW algorithm was employed for temporal alignment between the speech and singing signals (Sakoe and Chiba, 1978). Generally mel frequency cepstral coefficients (MFCC), together with the delta and acceleration values are used as features of speech signals for DTW alignment. We note that the delta and acceleration values represent the dynamic characteristics of speech signals and, these values are varying vastly across speaking and singing voice styles. Hence we choose to neglect the delta and acceleration values from our feature set, as they will adversely affect the accuracy of temporal alignment. Instead of using MFCC features, we perform a 24th order cepstral analysis from 1024-point Fast Fourier Transform (FFT) corresponding to the read lyrics and singing vocals. Then we choose to retain the first 12 coefficients, termed as low-time cepstral

Table 1: The list of English pop songs included in the database.

S. No:	Song	Artist	Year
1	I Will Always Love You	Dolly Porton	1973
2	My Heart Will Go On	Celine Dion	1997
3	Linger	The Cranberries	1993
4	How Do I Live	LeAnn Rimes	1997
5	Foolish Games	Jewel	1995
6	Billie Jean	Michael Jackson	1983
7	Total Eclipse of the Heart	Bonnie Tyler	1983
8	Take My Breath Away	Berlin	1986
9	Poker Face	Lady Gaga	2008
10	Let it be	The Beatles	1970
11	I don't want to lose you	Tina Turner	1989
12	Staying Alive	Bee Gees	1977
13	Dancing Queen	ABBA	1976
14	How Deep is Your Love	Bee Gees	1978
15	You Light Up My Life	Debby Boone	1977
16	Hey Jude	The Beatles	1968
17	Stars shining bright above you	Ozzie Nelson	1931
18	Fly me to the moon	Kaye Ballard	1954
19	Yesterday	the Beatles	1965
20	Stand by Me	Ernest Tubb	1965



(a) Singing signal



(b) Speech signal

Figure 3: Word-boundary labeling using wavesurfer.

coefficients (LTCC), as the features for DTW alignment. Notice that the low-time cepstrum denotes the vocal tract system properties and high-time cepstrum denote the excitation source characteristics of the voice production system. As the source characteristics are expected to vary vividly across speech and singing voice styles, we choose to isolate out the high-time cepstrum, thereby retaining only the contribution of vocal tract system in the feature set. This strategy helps to preserve the consistent properties of vocal tract system across speech and singing, and rule out the inconsistent properties of excitation source.

The resultant temporal alignment of words between read lyrics and singing vocals, using DTW algorithm with LTCC features, was evaluated on the database presented in (Duan et al., 2013). The word-boundary alignment errors of the automatic alignment were computed against the manually marked transcriptions available with this database. It was observed that the proposed temporal alignment scheme was providing near-accurate synchronization information required for efficient STS conversion.

5. Conclusions

In this paper, we presented an ongoing data collection effort to record parallel speak-sing corpus for English and Chinese pop songs. This database is expected to aid the development of an efficient STS conversion system. The audio recordings of read lyrics and singing vocals are performed in a professional studio environment and trained professional singers are hired to sing the songs and read the lyrics in natural manner. The sentence-level and word-level transcriptions of the audio recordings are labeled manually. Later, automatic temporal alignment between frames within words of read lyrics and singing vocals is performed using DTW algorithm with low-time cepstral features. As temporal alignment is a crucial requirement for a successful STS conversion system, the accuracy of frame-level alignment of words in read lyrics to singing vocals is enhanced using DTW with specially designed features.

Apart from STS conversion, the parallel speak-sing corpus can be proven as a valuable resource for singing voice analysis and understanding. The parallel recordings of same linguistic content by the same speaker in reading and singing voice styles can provide rich resorts to understand the differences in production and perception of speech and singing signals. Thus, the presented database will be advantageous in development of new modeling strategies for singing voices. Also, it can assist in singing quality evaluation, vocal training of music students, etc.

6. Bibliographical References

Duan, Z., Fang, H., Li, B., Sim, K. C., and Wang, Y. (2013). The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9, Oct.

Lindblom, B. and Sundberg, J. (2007). The human voice in speech and singing. In *Springer Handbook of Acoustics*, pages 703–746. Jan.

New, T. L., Dong, M., Chan, P., Wang, X., Ma, B., and Li, H. (2010). Voice conversion: From spoken vowels to singing vowels. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1421–1426, July.

Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, USA.

Saitou, T., Goto, M., Unoki, M., and Akagi, M. (2007a). 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*, pages 215–218, Oct.

Saitou, T., Goto, M., Unoki, M., and Akagi, M. (2007b). Vocal conversion from speaking voice to singing voice using straight. In *INTERSPEECH*, pages 4005–4006.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, Feb.

Sjolander, K. and Beskow, J. (2000). Wavesurfer - an open source speech tool.

Sundberg, J., Titze, I. R., and Scherer, R. (1993). 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*, 7(1):15 – 29.

Sundberg, J., Fahlstedt, E., and Morell, A. (2005). Effects on the glottal voice source of vocal loudness variation in untrained female and male voices. *The Journal of the Acoustical Society of America*, 117(2):879–885.

Sundberg, J. (2009). Voice source studies of register differences in untrained female singing. *Logopedics Phoniatrics Vocology*, 24:76–83, July.

Titze, I. R. and Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, 91(5):2936–2946.

Vijayan, K., Dong, M., and Li, H. (2017). A dual alignment scheme for improved speech-to-singing voice conversion. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC)*, December.