INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

NUS Speak-to-Sing: A Web Platform for Personalized Speech-to-Singing Conversion

Chitralekha Gupta, Karthika Vijayan, Bidisha Sharma, Xiaoxue Gao, Haizhou Li

Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{chitralekha, vijayan.karthika, s.bidisha}@nus.edu.sg, xiaoxue.gao@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

Singing like a professional singer is extremely appealing to the general public. However, many individuals are not able to sing like a singer who has received formal training over several years. We develop a web platform, where users can perform personalized singing synthesis. A user has to read and record the lyrics of a song in our web platform, and enjoy good quality singing vocals synthesized in his/her own voice. We perform a template-based speech-to-singing voice conversion at the backend of the web interface, that uses the prosody characteristics of the song derived from good quality singing by a trained singer and retains the speaker characteristics from the respective user. We utilize an improved temporal alignment scheme between speech and singing signals using tandem features, and employ a deep-spectral map to incorporate singing spectral characteristics into user's voice. The singing vocals are later synthesized by a vocoder. Using this web platform, we advocate that 'everyone can sing as they desire'.

1. Introduction

Personalized singing synthesis is an attractive application, catering to a wide range of population. Many of us desire to sing our favorite songs, yet unable to do that due to limited training, inadequate singing capability and various other reasons. It is extremely appealing to have a singing synthesis system, which corrects the faults in our singing and generates good quality singing vocals in our own voice. This desire is the motivation behind the development of *'NUS Speak-to-Sing'*.

The history of singing synthesis is very rich, ranging from IBM's first computer to sing the rhyme 'daisy bell' to applications like Vocaloid and Realivox [1–3]. The techniques employed for singing synthesis span from concatenative synthesis and statistical modeling to WaveNet-based deep learning architecture. Comprehensive studies of techniques utilized in singing voice processing can be found in [4–6]. However, the majority of existing singing synthesizers are limited to generation of singing vocals in a set of predetermined voices.

The demand for personalized singing synthesizers stem from one's desire to sing. This desire of singing and exhibiting singing talent can be observed from the increasing popularity of applications, like Smule [7]. Towards personalizing singing synthesis addressing this desire of users, speech-to-singing voice conversion was coined [8, 9]. Speech-to-singing (STS) voice conversion is the task of converting the lyrics of a song read by a user to good quality singing vocals in the user's own voice. The main requirements of STS conversion are (i) transformation of prosody characteristics of spoken lyrics to resemble those of good quality singing, and (ii) retention of spectral characteristics of spoken lyrics to preserve speaker identity of the user [10].

The methodologies proposed for STS conversion mainly fall into two categories, namely, model-based and template-based STS schemes. They differ in the way in which reference prosody is generated for synthesized singing. The model-based scheme rely on synthetic musical scores (eg: MIDI files) of songs to model realistic pitch contours as reference prosody [8]. The template-based scheme extracts pitch contours from good quality singing templates and utilizes these as reference prosody [9]. In either case, the spectral characteristics are retained from spoken lyrics of the song by users.

We employ the template-based scheme for STS conversion in our web platform, *NUS Speak-to-Sing*. As natural singing templates are used to extract reference prosody, gross and fine characteristics of prosody including pitch, vibrato, overshoot, preparation and fine variations are well preserved in the synthesized singing. Also, we utilized an improved temporal alignment scheme between speech and singing signals, as well as, a deep-spectral map to transform the spectra of speech to that of singing, without distorting the speaker identity. Thus we are able to generate enhanced quality singing vocals from read lyrics of songs.

2. *NUS Speak-to-Sing*

The *NUS Speak-to-Sing* uses template-based scheme for STS conversion. The block diagram of the template-based STS conversion in *NUS Speak-to-Sing* is shown in Figure. 1. The lyrics of a song read by a user (user speech) is converted to singing vocals (synthesized singing) by employing prosody characteristics obtained from good quality singing template. The templates database is carefully prepared by including singing templates corresponding to a set of predetermined songs. These templates are sung by trained singers in a professional studio environment, and are manually labeled for their sentence boundaries. The pitch (fundamental frequency (F0) contours) are computed from the singing templates, which form the reference prosody for synthesized singing.

The user speech can be recorded in any real-world setting. The short-time spectra are computed from the user speech and are used for generating the synthesized singing. However, the spectra of user speech are not used as they are, owing to the fact that there exist multiple differences between spoken and sung spectra of signals of the same linguistic content. The singing formant, pitch harmonics, modulation of formant frequencies by pitch harmonics, etc. are some of these spectral differences between spoken and sung signals. To incorporate the characteristics of singing into spectrum of user speech, we utilize

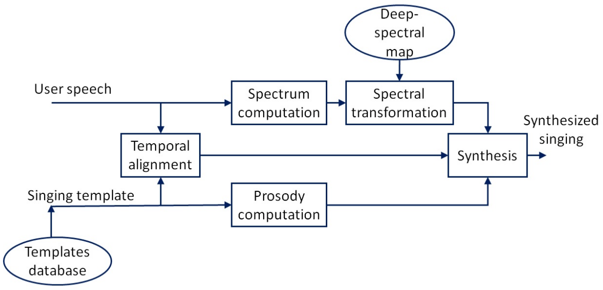

Figure 1: *Template-based scheme for STS conversion in NUS Speak-to-Sing.*

a deep-spectral map. This map is trained as a deep learning model using parallel spoken and sung utterances. The spectra of user speech are transformed using this map to resemble singing spectra by the same user. The transformed spectral vectors are then fed into a vocoder, together with reference prosody from singing template, to generate synthesized singing. The temporal alignment information between speech and singing is obtained using dynamic time warping (DTW) executed over a set of tandem features, and is used for combining spectral vectors with corresponding prosody parameters for singing synthesis [11].

We have currently implemented the template-based STS conversion as an interactive web platform. A user can log in, choose a song from our templates database, indicate their gender, and read the lyrics of the song. Our algorithm, hosted in the server, processes the input speech audio based on the corresponding gender-specific singing template, and generates the synthesized singing voice as the output. This output singing is played back, along with musical accompaniment.

2.1. Comparison with existing applications

The *NUS Speak-to-Sing* differs from existing singing synthesis/conversion applications in a multitude of ways. The VOCALOID performs singing synthesis from lyrics in a set of predetermined voices [2]. *NUS Speak-to-Sing* is able to synthesize singing vocals in any user's voice. The Auto-Tune corrects errors in pitch contour of an amateur singing to match with the quality of trained singing [12]. However, when the pitch differences between the two vocals are large (like speech and singing), Auto-Tune suffers from severe quality degradation. The karaoke apps like Smule, StarMaker, etc. [7, 13] provide facility for a user to sing a song in synchronization with the corresponding musical accompaniment. These apps do not attempt conversion of user speech to good quality singing, as the *NUS Speak-to-Sing* does.

The I^2R Speech2Singing performs STS conversion and is able to produce good quality singing vocals when the temporal alignment between user speech and singing template is accurate [14]. As the temporal alignment is obtained by DTW executed over spectral features, and there exist multiple differences between spectral features of speech and singing signals with the same lyrics, the temporal alignment can often go wrong. In *NUS Speak-to-Sing*, we employ DTW with tandem features for accurate temporal alignment between user speech and singing template. Also, we employ a deep-spectral map in *NUS Speak-to-Sing*, which captures the fine characteristics of singing spectra for singing synthesis to improve naturalness of the output and preserve user's voice identity, as opposed to the I^2R Speech2Singing.

3. Conclusions

We present a novel and innovative web platform for STS conversion, namely, the *NUS Speak-to-Sing*. It is a simple and user-friendly web platform, where a user can read and record the lyrics of a song. The web platform synthesizes good quality singing vocals in the user's own voice and plays it back with musical accompaniment. This web platform caters to one's desire to sing like a trained singer, by taking only the read lyrics of the song as input. By this web platform, we advocate that 'everyone can sing' their favorite songs as they desire.

4. Acknowledgements

This research is supported by Ministry of Education, Singapore AcRF Tier 1 Grant FY2016 for the project 'Non-parametric approach to voice morphing'.

5. References

[1] J. Kelly and C. Lochbaum, "Speech synthesis paper G42," in *4th. Intl. Congr. of Acoustics*, 1962, pp. 1–4.

[2] H. Kenmochi and H. Ohshita, "VOCALOID - commercial singing synthesizer based on sample concatenation," in *INTERSPEECH*, Antwerp, Belgium, Aug 2007, pp. 4009–4010.

[3] "Realivox," https://en.wikipedia.org/wiki/Realivox.

[4] 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.

[5] M. Umbert, J. Bonada, M. Goto, T. Nakano, and J. Sundberg, "Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges," *IEEE Sig. Proc. Mag.*, vol. 32, no. 6, pp. 55–73, Nov 2015.

[6] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer," in *INTERSPEECH*, 2017, pp. 4001–4005.

[7] "Smule-connecting the world through music." [Online]. Available: https://www.smule.com/

[8] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using STRAIGHT," in *INTERSPEECH*, Antwerp, Belgium, Aug 2007, pp. 4005–4006.

[9] 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)*, Kyoto, Japan, Mar 2012, pp. 4509–4512.

[10] K. Vijayan, H. Li, and T. Toda, "Speech-to-singing voice conversion: The challenges and strategies for improving vocal conversion processes," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 95–102, Jan 2019.

[11] K. Vijayan, X. Gao, and H. Li, "Analysis of speech and singing signals for temporal alignment," in *APSIPA Annual Summit and Conference*, Hawaii, USA., Nov 2018.

[12] "Auto-tune & vocal processing tools by antares audio technologies," www.antarestech.com.

[13] "Starmaker: Connect the world through music !" [Online]. Available: https://www.starmakerstudios.com/#/

[14] 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*, Singapore, Sep 2014, pp. 2148–2149.