Speaker-independent Speech-to-Singing Conversion and Singing Synthesis

Xiaoxue Gao

Department of Electrical and Computer Engineering, National University of Singapore, Singapore
xiaoxue.gao@u.nus.edu

1. Motivation

Speech-to-Singing (STS) conversion potentially enables various innovative applications in music production and entertainment. Synthesizing personalized singing just by reading lyrics of a song is appealing to users, especially to those who are not talented singers [1]. However, speech-to-singing conversion is not trivial [2], as it requires careful manipulation of prosody and proper mapping of acoustic characteristics from speech to singing signals [3].

Speech-to-singing (STS) conversion aims at converting one's reading speech into his/her singing vocal, in which the reading speech converted into singing according to the reference prosody while preserving the speaker identity. The basic idea of Speech-to-Singing conversion is to find a mapping function to transform the prosody and spectral features from reading speech to those of reference singing. Many previous studies focus on transforming the prosody of speech to singing [1, 4–6], however, there exist prominent differences between the spectra of speech and singing, which need to be transformed.

Prior studies have shown that significant spectral differences exist between one's speech and singing, such as the singing formant [7–9] and the resonance tuning by singing F0 [10, 11], which can be characterized by a speaker-specific mapping function. In particular, singing formant is a peak around 3kHz in the singing spectrum, formed by clustering of the third, fourth and possibly fifth formants that represents the energy concentration [7–9, 12]. Singing formant can always be found in singing spectrum of trained singers [2, 6, 8], while it is usually absent in speech spectrum. Therefore, spectral mapping from speech to singing has become imperative to enhance the performance of STS conversion.

In the literature of spectral mapping in STS conversion, the prosody of the singing follows the lyrical alignment of the template, while the speech spectrum is directly used for singing [4, 5] in TSTS. Some spectral conversion techniques were also studied, for example, to adjust the speech spectrum according to the vibrato information of template singing F0 [13], or to make use of the weighted linear and shifting functions [10, 14] to convert the spectra of vowels from speech to singing. However, the spectral control model in [10] requires empirical and hand-crafted settings of parameter values, which is not suitable for large scale deployment. In addition to these mathematical adjustments of speech spectrum, GMM and weighted frequency warping [2] voice conversion methods have also been adopted for STS spectral mapping. We note that the results reported in [2] show that they do not outperform the spectral control model [10].

Inspired by the success of average modeling approach to voice conversion [15–21], text-to-speech [22–24] and speaker adaptation [25] technique, we propose to learn a speaker-independent spectral mapping (SI-ivector) between speech and singing spectra from multiple speakers using i-vectors for speaker identity representation. To preserve speaker identity during the conversion, we augment one's speech spectra with her/his speaker identity features (i-vectors) in the network input. According to the studies in [11, 14, 26, 27], the amplitude of formants in singing voices is modulated in synchronization with the vibratos in singing F0 contours. Hence, we introduce the singing F0 and AP as joint features to train the spectral mapping model. The converted singing spectra are then used together with prosody features to synthesize the target singing.

Despite speech-to-singing (STS) conversion has been widely studied, a large database for this task has not been constructed yet. We present a Spoken Lyrics and Singing (SLS) corpus developed at NUS-HLT that can be useful for STS conversion. This database contains 3,058 utterances of 90 English songs from 10 professional singers collected in a recording studio environment. The spoken lyrics corresponding to the songs are also recorded from the singers to create the database, which we refer to as NUS-HLT SLS corpus. We highlight a few potential applications where this corpus can be used for future studies in our paper [28].

2. Contributions

The main contributions of my research include a) we propose a data-driven approach to learn a speaker-independent spectral mapping function that is a departure from the hand-crafting, simple functional warping or speaker-dependent spectral mapping; b) the proposed spectral mapping approach does not need the speech and singing data from a target speaker during training, which is more practical; c) the proposed spectral mapping better retains target speakers' identity by augmenting i-vectors as network input; d) the proposed model significantly improves the naturalness and quality of synthesized singing in comparison with baseline approaches in both subjective and objective evaluations; e) the proposed NUS-HLT SLS corpus has the high-quality recording of parallel speech and singing with sizable number of songs that is suitable for many application.

3. Methodology

We propose to condition a speaker independent model on a speaker i-vector [29, 30] to maintain the speaker identity between speaking and singing. During training, given parallel speak-sing utterances from multiple speakers, we first extract i-vectors features from multi-speaker speech. Then we obtain singing F0, singing AP and singing MCCs. Aligned speech MCCs are also obtained by feature alignment between one's speak-sing LTC features [6] using dynamic time warping (DTW) [31]. Then, singing F0, AP and i-vectors are augmented to the aligned speech MCCs as final training input features. The paired input features and singing MCCs features from all speakers are utilized to train SI-ivector model, which consisted of two DBLSTM layers with 512 hidden units in each layer.

At run-time, target i-vector is first extracted from target speech. Given target (user's) speech and template singing, input features are also constructed by concatenating template singing F0, AP, aligned speech MCCs and target i-vectors. Then the trained SI-ivector model is used to predict the converted MCCs, which are then used together with F0 and AP parameters of singing template to synthesize singing output.

4. Results

We conducted several experiments to validate the performance of the proposed spectral mapping approach (SI-ivector), as shown in Fig. 1 and Fig. 2. Zero-effort and SD-MCC denote the template-based STS without spectral mapping and the speaker-dependent spectral mapping baselines. SD (trained with singing F0 and AP) and SI (trained without i-vectors) denote the variants of SD-MCC baseline and the proposed SI-ivector.

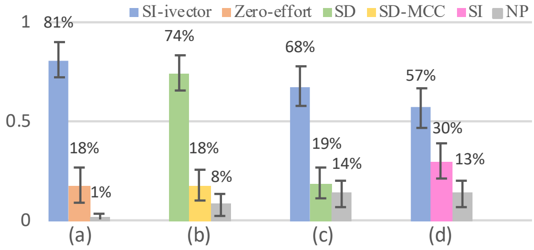

Figure 1: *AB preference results with 95% confidence intervals for singing quality and naturalness of zero-effort, SD-MCC, SD, SI and SI-ivector models; NP stands for no preference. (a) SI-ivector vs. Zero-effort; (b) SD vs. SD-MCC; (c) SI-ivector vs. SD; (d) SI-ivector vs. SI.*

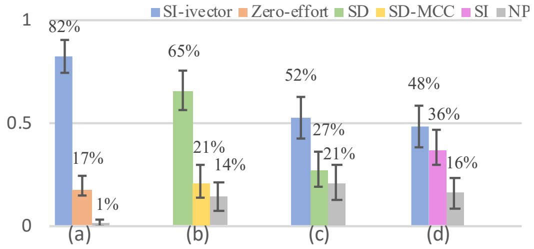

Figure 2: *XAB preference results with 95% confidence intervals for synthesized singing similarity of zero-effort, SD-MCC, SD, SI and SI-ivector models; NP stands for no preference. (a) SI-ivector vs. Zero-effort; (b) SD vs. SD-MCC; (c) SI-ivector vs. SD; (d) SI-ivector vs. SI.*

Both AB and XAB preference tests indicate that the proposed SI-ivector outperforms zero-effort, SD-MCC baselines, SD and SI. This confirms the effectiveness of the proposed model in terms of improving both singing quality and speaker similarity. This also suggests the proposed model with i-vectors can be beneficial to the preservation of target speakers' identities. Additionally, the superiority of SD over SD-MCC indicates the effectiveness of incorporating singing F0 and AP for spectral mapping. The synthesized singing samples for different models can be found in the website [1].

5. Future Directions

We have successfully developed spectral mapping technique that addresses the issues in Speech-to-Singing conversion. However, the approach was developed with some specific constrains, such as the need of parallel speak-singing data in spectral mapping, and the certain distortion in the process of synthesizing singing by WORLD vocoder. To generalize the proposed method and to enhance the quality of synthesized singing, some issues need to be addressed in the future work.

5.1. Parallel-data Free STS Conversion by Singing PPGs

It is difficult to collect parallel speak-sing database in real world applications for the proposed SI-ivector model. Thus, we propose a parallel-data free STS conversion by making use of singing Phonetic PosteriorGrams(PPGs) [32], which hasn't been explored in STS conversion. We extract singing MCCs and corresponding singing PPGs to train a conversion model by DBLSTM. The singing PPGs can be obtained by singing ASR system or applying speech ASR directly. At run-time, the user's speech PPGs, that is extracted from the speech ASR and then aligned with singing template, will be fed into the trained conversion model to obtain the converted MCCs. As the model is trained by singing data, we expect it can capture singing correlations between PPGs and MCCs in the conversion process. There is no parallel speech and singing data required for training, and only user's speech and singing template are parallel for test.

5.2. Singing Synthesis by WaveNet Vocoder in STS conversion

The WORLD vocoder is utilized to synthesize singing in the proposed spectral mapping for STS conversion, but the WORLD vocoder suffers from the problems that it loses phase information and temporal structure of the synthesized singing, and has many prior assumptions. WaveNet [33] was proposed to generate speech audios directly without any assumptions. WaveNet recovers the lost phase and temporal information of speech voices, thus generating higher quality speech output [34, 35] in voice conversion. Therefore, we apply WaveNet vocoder to replace WORLD vocoder in STS conversion, aiming to obtain more natural sounding singing output. The usage of WaveNet for singing synthesis has never been investigated, and the incorporation of WaveNet vocoder with the proposed spectral mapping model improves the quality of synthesized singing.

5.3. CycleGAN-based alignment-free speaker-independent STS conversion

To directly utilize speech and singing data for STS conversion, we propose to use CycleGAN to find optimal pseudo pair from unpaired speech and singing data. The CycleGAN-based STS conversion will not require any alignment technique, which is still a difficult task in STS conversion. We aims to make use of multi-speaker speak-sing data for CycleGAN training. However, if several speakers data is used, the speaker identity is difficult to keep. Hence, we propose to design a trainable speaker identity network inside CycleGAN and further device the loss function accordingly.

6. Acknowledgements:

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing. The study is supervised by Professor Haizhou Li, National University of Singapore.

[1] http://xiaoxue1117.github.io/sample

7. 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 *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 2148–2149.

[2] S. W. Lee, Z. Wu, M. Dong, X. Tian, and H. Li, "A comparative study of spectral transformation techniques for singing voice synthesis," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 2499–2503.

[3] 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 *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007, pp. 215–218.

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

[5] K. Vijayan, M. Dong, and H. Li, "A dual alignment scheme for improved speech-to-singing voice conversion," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1547–1555.

[6] K. Vijayan, X. Gao, and H. Li, "Analysis of speech and singing signals for temporal alignment," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1893–1898.

[7] E. Joliveau, J. Smith, and J. Wolfe, "Acoustics: tuning of vocal tract resonance by sopranos," *Nature*, vol. 427, no. 6970, p. 116, 2004.

[8] J. Sundberg, "The level of the singing formant and the source spectra of professional bass singers," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 4, pp. 21–39, 1970.

[9] B. Lindblom and J. Sundberg, "The human voice in speech and singing," *Springer handbook of acoustics*, pp. 669–712, 2007.

[10] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using straight," in *Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2007, pp. 4005–4006.

[11] N. Henrich, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones," *The Journal of the Acoustical Society of America*, vol. 129, no. 2, pp. 1024–1035, 2011.

[12] 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.

[13] S. W. Lee and M. Dong, "Singing voice synthesis: Singer-dependent vibrato modeling and coherent processing of spectral envelope," in *Twelfth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 2001–2004.

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

[15] J. Wu, Z. Wu, and L. Xie, "On the use of i-vectors and average voice model for voice conversion without parallel data," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2016, pp. 1–6.

[16] Y. Zhou, X. Tian, H. Xu, R. K. Das, and H. Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[17] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average modeling approach to voice conversion with non-parallel data," in *Proceedings of Odyssey The Speaker and Language Recognition Workshop*, 2018, pp. 227–232.

[18] M. Zhang, B. Sisman, S. S. Rallabandi, H. Li, and L. Zhao, "Error reduction network for dblstm-based voice conversion," in *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 823–828.

[19] T. Hashimoto, D. Saito, and N. Minematsu, "Many-to-many and completely parallel-data-free voice conversion based on eigenspace dnn," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 332–341, 2018.

[20] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit, "Cross-language voice conversion based on eigenvoices," in *IEEE Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009, pp. 1635–1638.

[21] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[22] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for hmm-based speech synthesis," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 86, no. 8, pp. 1956–1963, 2003.

[23] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, and S. Renals, "Robust speaker-adaptive hmm-based text-to-speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1208–1230, 2009.

[24] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4475–4479.

[25] S. H. K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom, and S. Garimella, "fmllr based feature-space speaker adaptation of dnn acoustic models," in *Sixteenth annual conference of the international speech communication association (INTERSPEECH)*, 2015, pp. 3630–3634.

[26] P. Oncley, "Frequency, amplitude, and waveform modulation in the vocal vibrato," *The Journal of the Acoustical Society of America*, vol. 49, no. 1A, pp. 136–136, 1971.

[27] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The production of speech*. Springer, 1983, pp. 39–55.

[28] X. Gao, B. Sisman, R. K. Das, and K. Vijayan, "Nus-hlt spoken lyrics and singing (sls) corpus," in *Proc. Int. Conf. Orange Technologies (ICOT)*, 2018.

[29] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[30] R. K. Das, S. Abhiram, S. M. Prasanna, and A. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 1836–1840.

[31] H. Sakoe, S. Chiba, A. Waibel, and K. Lee, "Dynamic programming algorithm optimization for spoken word recognition," *Readings in speech recognition*, vol. 159, p. 224, 1990.

[32] X. Chen, W. Chu, J. Guo, and N. Xu, "Singing voice conversion with non-parallel data," *arXiv preprint arXiv:1903.04124*, 2019.

[33] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." *SSW*, vol. 125, 2016.

[34] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent wavenet vocoder." in *INTERSPEECH*, 2017, pp. 1118–1122.

[35] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.