Speaker-independent Spectral Mapping for Speech-to-Singing Conversion

Gao Xiaoxue
National University of Singapore
3 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Xiaohai Tian
National University of Singapore
29 PUBLICATIONS **182** CITATIONS
SEE PROFILE

Rohan Kumar Das
National University of Singapore
59 PUBLICATIONS **273** CITATIONS
SEE PROFILE

Yi Zhou
National University of Singapore
5 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Speaker-independent Spectral Mapping for Speech-to-Singing Conversion

Xiaoxue Gao, Xiaohai Tian, Rohan Kumar Das, Yi Zhou and Haizhou Li
Department of Electrical and Computer Engineering, National University of Singapore, Singapore
{xiaoxue.gao, yi.zhou}@u.nus.edu, {eletia, rohankd, haizhou.li}@nus.edu.sg

***Abstract*—Speech-to-Singing (STS) conversion aims at converting one's reading speech into his/her singing vocal. The prior work was mainly focused on transforming the prosody of speech to singing, however, there exist prominent differences between the spectra of speech and singing, which need to be transformed as well. In this paper, we propose to make use of parallel multi-speaker speak-sing data to develop a speaker-independent spectral mapping model, which is conditioned on i-vector to generate target speaker/singer identity. The model is therefore called speaker conditioned spectral mapping model. The converted singing spectra are then used together with prosody features to synthesize the target singing. We investigate the effectiveness of i-vector based average model adaptation to model the differences between speech and singing spectra for a specific speaker. The proposed model does not require parallel speak-sing data from target speakers during training. The experimental results conducted on NUS-48E and NUS-HLT-SLS database indicate that the proposed approach significantly outperforms the baselines in terms of quality and similarity.**

I. Introduction

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, STS conversion is not trivial [2], as it requires careful manipulation of prosody and proper mapping of acoustic characteristics from speech to singing signals [3].

The STS conversion aims to convert the reading speech into singing according to the reference prosody while preserving the speaker identity. The basic idea of STS conversion is to find a mapping function to transform the prosody and spectral features from reading speech to those of reference singing.

Two major methods have been studied for prosody transformation, namely, template-based speech-to-singing conversion (TSTS) [1], [4]–[6], and musical score based speech-to-singing conversion (MSTS) [3], [7]. In TSTS, the reference prosody is obtained from a singing template by a trained singer. The lyrics of a song read by a user are first aligned with singing template by different alignment techniques [1], [4], [5], [8]. The aligned spectral features from speech and fundamental frequency (F0) contour extracted from the singing template are then used to generate converted singing. Alternatively, the reference prosody in MSTS is obtained from synthetic musical scores, such as MIDI files [3], [7]. The prosody transformation is conducted by building singing F0 contour from musical scores in the way of controlling F0 fluctuations [3], [7] or performing vibrato modeling by single Gaussian Mixture Model (GMM) [9]. Spectral parameters of reading speech are then aligned by duration control models [3], [7], [10]. Singing output is synthesized by modifying aligned speech spectrum [7] together with transformed prosody.

Spectral transformation is as important as prosody transformation in STS conversion, since significant spectral differences exist between one's speech and singing including the singing formant [11]–[13] and the resonance tuning by singing F0 [7], [14]. In zero effort approaches, the prosody of the synthesized singing follows the lyrical alignment of the template, while the speech spectrum is directly used for synthesizing singing [4], [5] in TSTS. Some spectral conversion techniques were studied, for example, to adjust the speech spectrum according to the vibrato information of template singing F0 [9], or to make use of the weighted linear and shifting functions [7], [10] to convert the spectra of vowels from speech to singing. However, the spectral control model in [7] 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 [7].

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 a speaker-conditioned spectral mapping model for STS conversion. 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 [10], [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 main contributions of this paper include (a) we propose a data-driven approach to learning a speaker-conditioned spectral mapping function that is a departure from the hand-crafting, frequency warping or speaker-dependent spectral mapping; (b) the proposed approach does not need parallel speak-sing data from target speakers during training, which is more practical; (c) the proposed model better retains target speaker's identity by augmenting the network input with i-vectors [28].

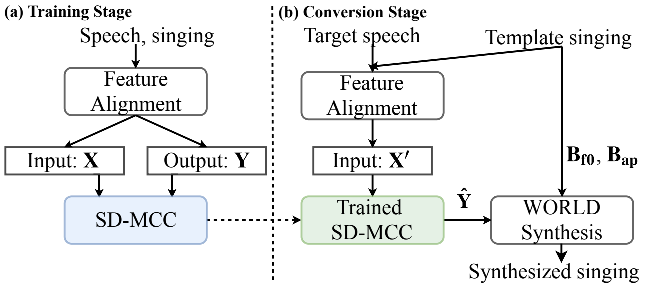

Fig. 1. Block diagram of the training and conversion stages for Speaker Dependent spectral mapping model (SD-MCC).

The rest of this paper is organized as follows. Section II demonstrates two common techniques for spectral conversion. In Section III, we motivate the idea of speaker-conditioned spectral mapping between speech and singing, and devise the model by specific deep learning architecture. Evaluations of both objective and subjective results will be presented in Section IV. In Section VI, we summarize the contributions of this paper.

II. Spectral Mapping from Speech to Singing

We first study two simple techniques projecting the spectral features from speech to singing.

A. Zero-effort transfer

We review the TSTS without spectral mapping as zero-effort baseline. Given parallel speech and template singing utterances, we first extract Mel Cepstral Coeficients (MCCs), fundamental frequency (F0), aperiodic component (AP) and low-time cepstra (LTC). We then apply dynamic time warping (DTW) [29] to the LTC features [8] to obtain the alignment between speech and singing. With the frame-wise alignment, we can copy over the MCCs from speech to singing, and use the F0 and AP from the singing template to synthesize singing output.

B. Speaker-dependent Spectral Mapping (SD-MCC)

If the speak-sing parallel data are available for a user (target speaker), we can build a speaker-dependent spectral mapping, that we refer to as SD-MCC in Fig. 1. The spectral mapping can be implemented with deep learning techniques.

1) Training Stage: During training in Fig. 1 (a), given a collection of parallel speak-sing utterances from the target, the singing MCCs ($\mathbf{B}_m \in R^{D_m \times N}$) and aligned speech MCCs ($\mathbf{A}_m \in R^{D_m \times N}$) are first obtained by feature alignment, as described in Section II-A. D_m denotes the dimensions of MCC features, and N denotes the frame number. Then, the aligned speech MCCs $\mathbf{A}_m$ and singing MCCs $\mathbf{B}_m$ serve as the model input $\mathbf{X} = \mathbf{A}_m$ and output $\mathbf{Y} = \mathbf{B}_m$ respectively. We can train a model by applying a supervised training to predict $\mathbf{Y}$.

$$\mathbf{Y} = F(\mathbf{X}) + \mathbf{e}, \tag{1}$$

where $\mathbf{e}$ is the prediction error.

2) Run-time Conversion Stage: At run-time in Fig. 1 (b), given parallel data form target's speech and template singing, we first obtain AP ($\mathbf{B}'_{ap} \in R^{D_{ap} \times N}$) and F0 ($\mathbf{B}'_{f0} \in R^{D_{f0} \times N}$) features from template singing and aligned MCCs (denoted as $\mathbf{A}'_m$) from speech. D_{f0} and D_{ap} denote the dimensions of F0 and AP features, respectively. Then, the constructed input $\mathbf{X}' = \mathbf{A}'_m$ are taken by the trained mapping model to predict the converted MCCs $\hat{\mathbf{Y}}$ (where $\hat{\mathbf{Y}} = \hat{\mathbf{B}}_\mathbf{m} \in R^{D_m \times N}$) as:

$$\hat{\mathbf{Y}} = F(\mathbf{X}'). \tag{2}$$

The converted MCCs $\hat{\mathbf{Y}}$ are then used together with F0 ($\mathbf{B}'_{f0}$) and AP ($\mathbf{B}'_{ap}$) parameters of singing template to synthesize singing output.

3) Limitation: Although SD-MCC model learns a speaker dependent spectral mapping from speech to singing successfully, it generally requires a number of speak-sing parallel data from a target, which is not always available in real-life applications. For each target, its SD-MCC model need to be trained, which is computationally expensive and inconvenient.

III. Speaker-Conditioned Spectral Mapping

In this section, we present the proposed speaker-conditioned spectral mapping with i-vectors. Fig. 2 shows the training and run-time conversion stages of speaker-conditioned spectral mapping model (SC).

A. i-vector Feature Extraction

We propose to condition a speaker independent model on a speaker i-vector to maintain the speaker identity between speaking and singing. We note that at run-time, we only have spoken lyrics, therefore, the speaker i-vector is only extracted from spoken lyrics both during training and at run-time. An i-vector is a compact representation of a speaker model that possesses the dominant speaker characteristics [28], [30]. It is derived by a factor analysis approach from a total variability space that is trained on a large amount of background data.

In this study, we extract an i-vector for each speaker to represent the speaker attribute for speech-to-singing conversion. In this way, the network learns to distinguish the speakers during training. At run-time conversion, we extract an i-vector from the user (target), on which the conversion network is conditioned to generate the singing voice with the target voiceprint.

B. Methodology

1) Training Stage: During training in Fig. 2 (a), given parallel speak-sing utterances from multiple speakers $\{c_1,...,c_j\}$, we first extract i-vectors features $\{\mathbf{I}^{c_1},...,\mathbf{I}^{c_j}\}$ where $\mathbf{I}^{c_j} \in R^{D_I \times N}$ from multi-speaker speech. D_I denotes the dimension of i-vector and N demotes the number of frames. Then we obtain singing F0 $\{\mathbf{B}^{c_1}_{f0},...,\mathbf{B}^{c_j}_{f0}\}$, singing AP $\{\mathbf{B}^{c_1}_{ap},...,\mathbf{B}^{c_j}_{ap}\}$ and singing MCCs $\{\mathbf{B}^{c_1}_{m},...,\mathbf{B}^{c_j}_{m}\}$. Aligned speech MCCs $\{\mathbf{A}^{c_1}_{m},...,\mathbf{A}^{c_j}_{m}\}$ are also obtained by feature alignment.

Then, singing F0 (${\mathbf{B}^{c_j}_{f0}}^T$) and AP (${\mathbf{B}^{c_j}_{ap}}^T$) are augmented to the aligned speech MCCs (${\mathbf{A}^{c_j}_{m}}^T$) as the acoustic features

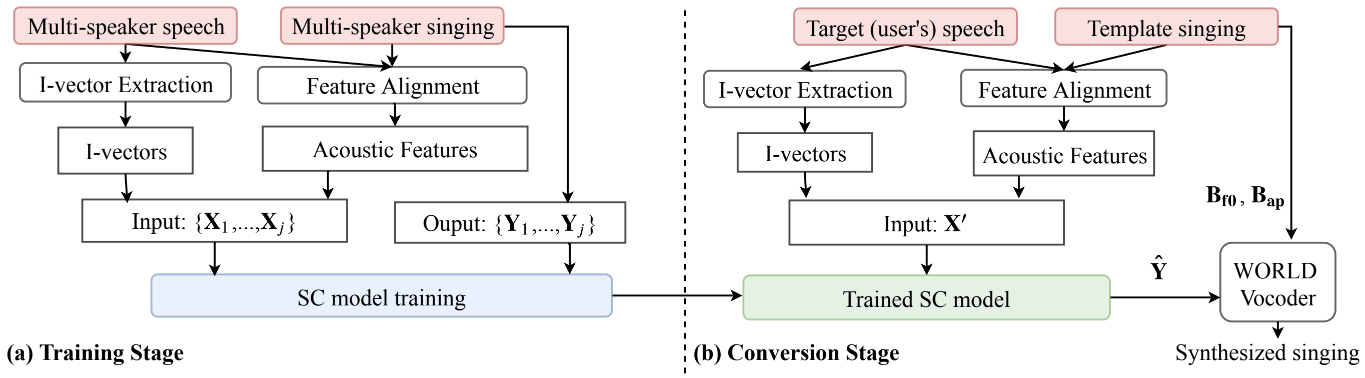

Fig. 2. Block diagram of the training and conversion stages for the proposed Speaker-Conditioned spectral mapping model (SC).

$\tilde{\mathbf{X}}^{c_j} = [\mathbf{A}_m^{c_j T}, \mathbf{B}_{f0}^{c_j T}, \mathbf{B}_{ap}^{c_j T}]^T$ for a part of the model training input. The i-vectors ($\mathbf{I}^{c_j T}$) are also augmented to the obtained acoustic features, and final training input features are represented as $\{\mathbf{X}_1,...,\mathbf{X}_j\}$, where $\mathbf{X}_j = [\tilde{\mathbf{X}}^{c_j T}, \mathbf{I}^{c_j T}]^T$. $\mathbf{X}_j$ refers to the features of j-th speaker. The paired input features and singing MCCs features $\mathbf{Y}_j$= $\mathbf{B}_m^{c_j}$ from all speakers, $\{\{\mathbf{X}_1, \mathbf{Y}_1\},...,\{\mathbf{X}_j, \mathbf{Y}_j\}\}$, are utilized to train the SC model.

2) Run-time Conversion Stage: At run-time as shown in Fig. 2 (b), target i-vector ($\mathbf{I}'$) is first extracted from target speech. Given target (user's) speech and template singing, acoustic features are also constructed by concatenating template singing F0 ($\mathbf{B}'_{f0}$), AP ($\mathbf{B}'_{ap}$) and the aligned speech MCCs ($\mathbf{A}'_m$) following the same process as in Section II-B2. Then the trained SC model is used to convert $\mathbf{X}' = [\mathbf{A}'^T_m, \mathbf{B}'^T_{f0}, \mathbf{B}'^T_{ap}, \mathbf{I}'^T]^T$ to $\hat{\mathbf{Y}}$ by Eq. 2, as shown in Fig. 2 (b).

We note that SD-MCC and SC models are applicable to transform spectral features in both template-based speech-to-singing (TSTS) and musical score based speech-to-singing (MSTS). In TSTS, we obtain the F0 and AP parameters from singing template, while in MSTS, we obtain the F0 and AP parameters from control models [3].

IV. Experiment

We conducted several experiments to validate the proposed speaker-conditioned spectral mapping model.

A. Database

Two databases, NUS-48E corpus [31] and NUS-HLT SLS corpus [32], with parallel speak-sing data were used for system implementations. The NUS-48E consists of 48 English songs from 12 speakers/singers, including 6 males and 6 females, where each speaker has 4 songs. The NUS-HLT SLS contains 100 English songs from 10 speakers/singers, including 5 females and 5 males, where each speaker has 10 songs. All speech and singing data were resampled at 16kHz.

All above data, including 148 songs from 22 speakers (11 male and 11 female), were used in our experiments. For multi-speaker training, 18 speakers were used, including all 12 speakers from NUS-48E and 6 singers (3 male and 3 female) from NUS-HLT SLS. For multi-speaker testing and speaker-dependent model training, 4 non-overlap target speakers were selected from NUS-HLT SLS, including 2 female speakers (Jesica and Nichole) and 2 male speakers (Understand and Kenza).

B. Experimental Setup

We now compare several different spectral mapping models that we propose and their variants in the experiments.

- **Zero-effort**: the template-based speech to singing model without spectral mapping described in Section II-A.
- **SD-MCC**: the speaker-dependent spectral mapping model, as described in Section II-B where the input only consisted of MCC features (120-dim) formed by aligned speech MCCs (40-dim) and their delta and delta-delta coefficients.
- **SD**: a variant of the speaker-dependent spectral mapping model. The dimension of model input was 127, including aligned speech MCCs (40-dim), singing *log*(F0) (1-dim), singing AP (1-dim) with their delta and delta-delta coefcients and the voiced/unvoiced flag (1-dim).
- **SI**: a variant of the proposed speaker-conditioned spectral mapping model. The model input (127-dim) was same as SD. We consider this multi-speaker training model as speaker-independent model (SI).
- **SC**: the proposed speaker-conditioned spectral mapping model with i-vectors. The dimension of model input was 187, including aligned speech MCCs (40-dim), singing *log*(F0) (1-dim), singing AP (1-dim) with their delta and delta-delta coefficients, the voiced/unvoiced flag (1-dim) and i-vectors (60-dim).

A summary of the framework description with their training sets can be found in Table I. All the objective and subjective results were reported based on the 8 songs (174 utterances) from 4 target speakers, where each target speaker provided 2 songs.

C. Model Training and Conversion

All spectral mapping models shared the same neural network architecture, which consisted of two DBLSTM layers with 512 hidden units in each layer. The network output dimension was 120, including singing MCCs (40-dim) with their delta and delta-delta coefficients. Adam optimizer was

TABLE I

THE SUMMARY OF EXPERIMENT SETUPS OF VARIOUS SPECTRAL MAPPING MODELS, AND THE OBJECTIVE EVALUATION RESULTS.

Models	**Training set**	**# Training utterances**	**Spectral mapping model**	**Speaker independent**	**I-vectors included**	**Feature dimension for model training**	**MCD (dB)**
Zero-effort	N/A	N/A	No	No	N/A	N/A	8.900
SD-MCC	32 songs from 4 target speakers (each 8 songs)	977	Yes	No	No	120 dims	7.124
SD	32 songs from 4 target speakers (each 8 songs)	977	Yes	No	No	127 dims	6.041
SI	108 songs from 18 speakers	2917	Yes	Yes	No	127 dims	6.265
SC	108 songs from 18 speakers	2917	Yes	Yes	Yes	187 dims	6.212

used for model training with the learning rate and momentum of 0.002 and 0.9, respectively. The minibatch size was set to 10.

During the conversion phase, the Maximum Likelihood Parameter Generation (MLPG) algorithm was employed to refine the spectral parameter trajectory [33], followed by a spectral enhancement post-filtering in the cepstral domain. The APs and F0s features from singing template were used to generate the singing signal. Additionally, Merlin toolkit [34] was used for model training.

D. Feature Extraction

The WORLD vocoder [35] was used to extract the spectrum (513-dim), AP (1-dim) and F0 (1-dim) for both speech and singing utterances with 5 ms frame step. 40-dimensional MCCs were computed from the spectrum using Speech Signal Processing Toolkit (SPTK) [1]. The frame alignment between speech and singing was obtained by performing dynamic time warping (DTW) [29] on LTC features [8] with word-level annotations.

V. Evaluations

Both objective and subjective evaluations were conducted to compare the proposed approaches with baselines.

A. Objective Evaluation

Mel-cepstral distortion (MCD) [36] was employed as the objective measure, which was defined as:

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{D}(c_d - c_d^{converted})^2} \qquad (3)$$

where c_d and $c_d^{converted}$ are d-th dimension of singing and the converted MCCs respectively. D denotes the dimension of MCCs. A lower MCD value indicates a smaller distortion.

We summarize five spectral mapping models and their average MCD results in Table I. We first examine the effect of spectral mapping for STS conversion. It shows that all approaches using spectral mapping give lower MCD values than zero-effort (8.900 dB) approach. This suggests that the spectral mapping is essential in STS conversion. Secondly, we evaluate the effectiveness of F0 and AP for model training. It is showed that MCD drops from 7.124 dB (SD-MCC model) to 6.041 dB (SD model), by incorporating F0 and AP into the input feature. The results suggest that the spectral mapping benefits from singing F0 and AP parameters. Hence, SD model will be used for comparisons in the rest of the experiments

We further compare the speaker-independent mapping with the speaker-dependent mapping. It is observed that SI model obtains a comparable result to SD model, with MCDs of 6.265 dB and 6.041 dB respectively. We also observe that SC model (6.212) slightly outperforms SI model (6.265) in terms of MCD, which indicates that i-vector further enhances the performance of synthesized singing.

B. Subjective Evaluation

The Mean Opinion Score (MOS), AB preference tests were conducted for subjective evaluations [37]–[40] on the quality and naturalness of synthesized singing outputs. The XAB preference tests was employed to assess the similarity between synthesized singing and target singing. 12 listeners whose ages range from 20 to 35 with normal hearing abilities participated in all tests. For each test, 10 samples were randomly selected from each system. In the MOS tests, listeners were asked to rate the quality and naturalness ranging from 1 (bad) to 5 (excellent) for all systems. In AB preference tests, one singing sample was picked from system A and another was from system B. A listener was asked to select the better sample in terms of the quality of synthesized singing. In XAB preference tests, given reference target singing X, listeners were asked to choose which one was more similar to X from samples of system A and B.

Three sets of listening tests were conducted. 1) Quality tests: MOS test of Zero-effort, SD-MCC, SD, SI and SC; 2) Quality tests: AB preference tests of SC vs. Zero-effort; SD vs. SD-MCC; SC vs. SD and SC vs. SI; 3) Similarity tests: XAB preference tests of SC vs. Zero-effort; SD vs. SD-MCC; SC vs. SD and SC vs. SI.

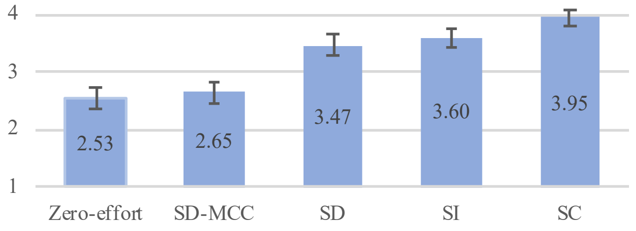

Fig. 3. MOS results with 95% confidence intervals for synthesized singing quality of Zero-effort, SD-MCC, SD, SI and SC.

1) MOS tests: Fig. 3 reports the results of MOS tests. It shows that zero-effort gives the lowest MOS value with 2.53

[1] https://sourceforge.net/projects/sp-tk/

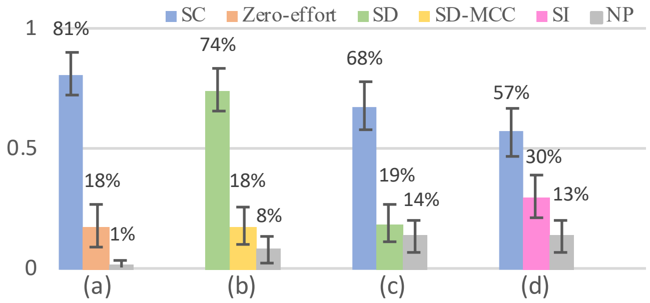

Fig. 4. AB preference results with 95% confidence intervals for singing quality and naturalness of zero-effort, SD-MCC, SD, SI and SC models; NP stands for no preference. (a) SC vs. Zero-effort; (b) SD vs. SD-MCC; (c) SC vs. SD; (d) SC vs. SI.

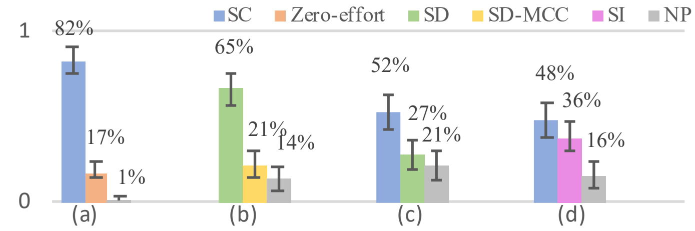

Fig. 5. XAB preference results with 95% confidence intervals for synthesized singing similarity of zero-effort, SD-MCC, SD, SI and SC models; NP stands for no preference. (a) SC vs. Zero-effort; (b) SD vs. SD-MCC; (c) SC vs. SD; (d) SC vs. SI.

compared with other models using spectral mapping. It is also showed that the SI model (3.60) outperforms SD model (3.47), which suggests that SI model benefits from multi-speaker training data. We also observe that SC model (3.95) achieves the best performance in terms of quality.

2) AB preference tests: Fig. 4 presents the results of AB preference tests with 95% confidence intervals. We first examine the effect of spectral mapping. It is observed in Fig. 4 (a) that the proposed SC model (81%) significantly outperforms zero-effort baseline, which demonstrates the effectiveness of spectral mapping in terms of improving singing quality and naturalness. Then we examine the effectiveness of including singing F0 and AP features for training, as shown in Fig. 4 (b). The SD model (68%) trained with singing F0 and AP outperforms SD-MCC baseline (18%) significantly, hence SD model will be used for later comparison.

We further compare speaker-conditioned model SC with speaker-dependent model SD from Fig. 4 (c). It is observed that SC model is superior to SD model for singing quality significantly. This suggests that the SC model benefits from the training on multi-speaker database that is larger than the speaker dependent database.

Last, we evaluate the effect of augmenting i-vectors for spectral mapping training, as shown in Fig. 4 (d). We observe that SC model outperforms SI model in terms of singing quality (57% vs. 30%). This indicates the SC model with i-vectors achieves higher singing quality than the speaker-independent model (SI model) without i-vectors.

3) XAB preference tests: Fig. 5 presents the results of XAB preference tests with 95% confidence intervals. The results of XAB preference tests are consistent with that of AB preference tests, where SC model is superior over zero-effort baseline and SD model outperforms SD-MCC model. In Fig. 5 (c), we observe that speaker-conditioned model SC (52%) achieves better performance than speaker-dependent model SD (27%) in terms of preserving user's speaker identity. Last, we evaluate the effect of introducing i-vectors in Fig. 5 (d), where SC model outperforms SI model with preference scores of 48% vs. 36%. This suggests the SC model with i-vectors can be beneficial to the preservation of target speakers' identities.

C. Summary of Evaluation Results

Both objective and subjective evaluations indicate that the proposed SC model outperforms zero-effort, SD-MCC baselines and SI model. This confirms the effectiveness of the proposed model in terms of both singing quality and speaker similarity.

Moreover, the proposed SC model achieves better performance than speaker-dependent model (SD model) in subjective evaluations, and obtains comparable results with SD model in objective evaluations. Although the subjective results of SD model vs . SC model are not consistent with objective evaluations, it is understandable since human perception may not be well consistent with objective scores. Such inconsistent results were also reported in prior works [15], [17], [41], [42]. Additionally, as the target speaker's data is not accessible in many applications, the proposed SC model is more practical than the target speaker dependent approaches (SD and SD-MCC models). The synthesized singing samples for different models can be found in the website [2].

VI. Conclusions

This paper presented a speaker-conditioned spectral mapping model using i-vectors for speaker-independent speech-to-singing conversion. The proposed approach benefited from multi-speaker data with their i-vectors to model speaker-specific differences between speech and singing spectra. Both objective and subjective results on NUS-48E corpus and NUS-HLT SLS corpus databases showed that the proposed speaker-conditioned spectral mapping approach outperforms the zero-effort and speaker-dependent spectral mapping baselines in terms of the naturalness, quality and speaker similarity. Experimental results also confirmed the effectiveness of involving singing F0 and AP, and augmenting i-vectors to adapt target speaker identities for spectral mapping.

Acknowledgements: This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing.

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

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] S. W. Lee, Z. Wu, M. Dong, X. Tian, and H. Li, "A comparative study of spectral transformation techniques for singing voice synthesis," in *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 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 APSIPA ASC*, 2017, pp. 1547–1555.

[6] L. Cen, M. Dong, and P. Chan, "Segmentation of speech signals in template-based speech to singing conversion," in *IEEE APSIPA ASC*, 2011, pp. 1–4.

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

[8] K. Vijayan, X. Gao, and H. Li, "Analysis of speech and singing signals for temporal alignment," in *IEEE APSIPA ASC*, 2018, pp. 1893–1898.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[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 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 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 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 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 *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] 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.

[29] 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.

[30] R. K. Das, Abhiram B, S. R. M. Prasanna, and A. G. Ramakrishnan, "Combining source and system information for limited data speaker verification," in *INTERSPEECH*, 2014, pp. 1836–1840.

[31] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *IEEE APSIPA ASC*, 2013, pp. 1–9.

[32] 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.

[33] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *IEEE ICASSP (Cat. No. 00CH37100)*, vol. 3, 2000, pp. 1315–1318.

[34] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *The 9th ISCA Speech Synthesis Workshop (SSW)*, 2016, pp. 202–207.

[35] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[36] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128.

[37] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder." in *Interspeech*, 2018, pp. 1978–1982.

[38] B. Çişman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *IEEE ASRU*, 2017, pp. 677–684.

[39] B. Şişman, H. Li, and K. C. Tan, "Transformation of prosody in voice conversion," in *IEEE APSIPA ASC*, 2017, pp. 1537–1546.

[40] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 282–289.

[41] D. Erro, A. Moreno, and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2009.

[42] X. Tian, Z. Wu, S. W. Lee, N. Q. Hy, E. S. Chng, and M. Dong, "Sparse representation for frequency warping based voice conversion," in *IEEE ICASSP*, 2015, pp. 4235–4239.