ENHANCED VIRTUAL SINGERS GENERATION BY INCORPORATING SINGING DYNAMICS TO PERSONALIZED TEXT-TO-SPEECH-TO-SINGING

Kantapon Kaewtip, Fernando Villavicencio, Fang-Yu Kuo, Mark Harvilla, Iris Ouyang, Pierre Lanchantin

ObEN Inc., Pasadena, California, U.S.A.

ABSTRACT

We present in this work a strategy to enhance the quality of Text-to-Speech (TTS) based Singing Voice generation. Speech-to-singing refers to techniques transforming a spoken voice into singing, mainly by manipulating the duration and pitch of a spoken version of a song's lyrics. While this strategy efficiently preserves the speaker identity, the generated singing is not always perceived fully natural since the vocal conditions generally change between spoken and singing voice. By incorporating speaker-independent natural singing information to TTS-based Speech-to-Singing (STS) we positively impact the sound quality (e.g. reducing hoarseness), as it is shown in the subjective evaluation reported at the end of this paper.

Index Terms— Singing Synthesis, Speech-to-Singing, Text-to-Singing, TTS

1. INTRODUCTION

Current commercial TTS systems are able to generate high-quality speech. These systems are generally limited to the generation of spoken content of a single voice; however, an interest is emerging on techniques for identity transformation such as Voice Conversion and Speaker Adaptation.

Although there have not been many attempts to extend TTS capacity to singing voice generation, there has been some work done in what has been referred to as a Speech-to-Singing transformation (STS). In a pioneering work [1], psycho-acoustical aspects referred to as *vibration* and *ringing-ness* were found to significantly affect the *singing-ness* of the voice and a STS schema was proposed using a music score and F0, spectral, and duration controls based on STRAIGHT analysis and synthesis [2]. In [3], [4], and [5] there are extensive studies of this work addressing additional aspects such as F0 and duration modeling or even on the inverse task (singing to speech transformation) [6].

We can also find in [7] a method based on training F0 and spectral envelope transformations from singing data using a single vowel. Later, pairing spoken and sung sequences with Dynamic Time Warping (DTW) was proposed in [8], followed by a 2-step procedure in [9], in which the music information is extracted from the audio of actual singing performances (called templates) instead of using a music score. This strategy called Template-based STS (TSTS) was found to positively impact the perceived naturalness. A mobile application using a phonetic-based alignment version of this approach was also presented in [10]. The speech and singing alignment problem is addressed in [11], proposing a semi-automatic procedure consisting of a manual correction of a Forced Alignment (FA) between spoken and sung versions of singing lyrics to improve the alignment with a new spoken sequence. This strategy was reported to effectively impact the quality of the generated singing sequence and an extensive study on a frame-level DTW-based alignment was also presented in [12]. It is worth mentioning other prominent contributions to singing synthesis such as those found in [13–20], among others.

In this work we use TTS as the input speech on a TSTS-like schema to build what we denote for simplicity as Template-based Text-to-Singing (TTSing). Moreover, we propose: 1) enhanced singing generation by integrating singer-independent features from natural singing to a baseline TTSing engine, and 2) to use a personalized TTS system (i.e. a target speaker identity is applied) as input speech so that new *Virtual Singers* can be easily generated from small adaptation data. This paper is structured as follows: Section 2 describes our baseline TTSing system based in TSTS, and Section 3 describes the singing enhancement method. Section 4 presents a subjective evaluation assessing the perceptual benefits of the proposed enhancement and 5 concludes this paper.

2. TTSING: PERSONALIZED TEMPLATE-BASED TEXT-TO-SINGING

Our baseline system comprises the entire framework in Fig. 1 but without the shaded region. The framework follows the main TSTS configuration proposed in [10] and [11], except that the input speech is generated by a TTS system based on [21]. In our personalized TTS, the voice model is pre-trained from a large speaker corpus and then adapted to a new voice by parameter adaptation using one hour of speech. In our work, we employ the WORLD vocoder for the TTS system and waveform generation [22]. The acoustic features in WORLD include Mel-generalized cepstrum (MGC), band aperiodicity (BAP), and fundamental frequency (F0).

To generate a singing voice of a particular song, the baseline system requires an acappella version of the song, its corresponding instrumental content, and its lyrics. The acappella version is phonetically labeled in order to obtain template timing. The template pitch contour is extracted from the acappella with a robust estimator for singing voice, SAC [23]. The TTS system takes the lyrics as input, and produces their corresponding phonetic timing information and acoustic features. The TTS-based features are aligned in duration per phoneme to match the template timing in a linear time warping fashion (block A). The singing is generated using the WORLD vocoder with time-aligned features (MGC and BAP) and the template pitch contour (shown in block E). The short-term energy contour of the synthesized singing is scaled to match that of the template (block F). Finally, the resulting singing voice is mixed with the corresponding instrumental content. Note that building blocks in the shaded region will be explained later in Section 3.

The TTSing system just described above has enabled us to generate singing voices with a personalized identity (using a small amount of training data) as long as the lyrics and timing information are available. However, some overall hoarseness and a lack of energy at vowels preceding pauses or silences can be perceived in the

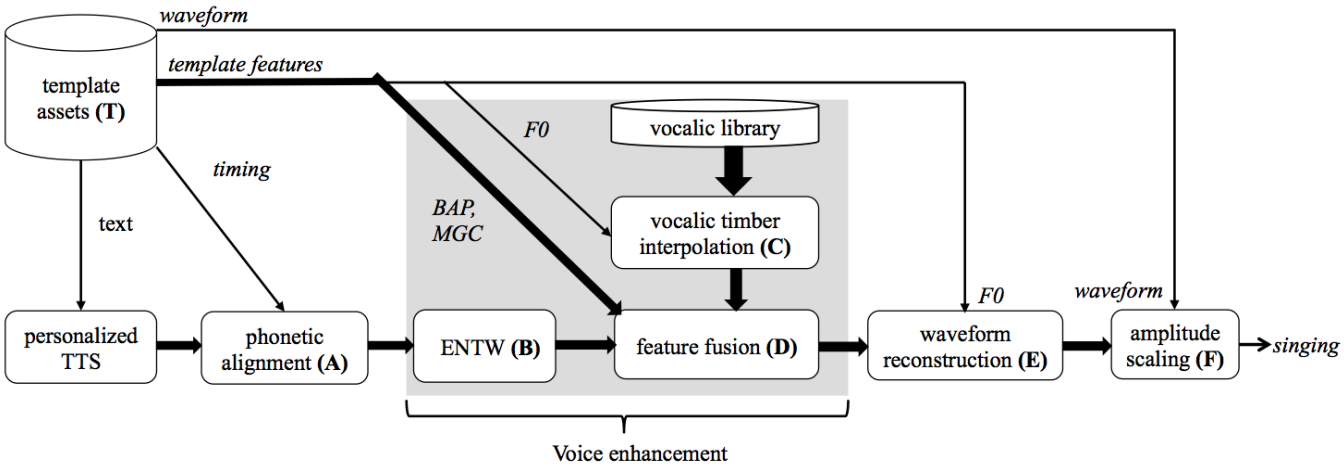

Fig. 1. Proposed enhanced TTSing system. The components shown in the shaded region are the voice enhancement. Our baseline system comprises the entire framework but without the shaded region (i.e, from block A directly to block E).

generated singing.

The acoustic features of the baseline are based on the signals generate from TTS, which model spoken signals. The energy and voicing of a natural spoken voice tends to fade toward the end of a phrase or a sentence. While these characteristics are natural for spoken voices, they do not generally match the flow observed on singing voices. When people sing (especially professional singers), wide openings of the voice source and vocal track make the voice sound richer or fuller than a typical spoken voice. In TTSing, a vowel segment is stretched uniformly to match the vowel duration of the singing template. This may elongate low-energy frames found at the beginning and at the end, which have been perceived as unnatural and shortened vowels on the generated singing.

In this paper, we present a technique to stretch a vowel segment in such a way that is suitable for singing. In order to make the voice sound richer, we employ additional recordings of vowels enunciated at several pitch levels and incorporate their acoustic information to enhance the timbre of the singing voice. In addition, we utilize acoustic information from the template to further balance the voicing features and energy contours to reduce hoarseness and energy fading. The implementation details are described in the next section.

3. ENHANCING STS TRANSFORMATION BY INCORPORATING SINGING DYNAMICS

In Fig. 1 the enhancement system is depicted in the shaded region. The system takes the acoustic features after the phonetic alignment in the baseline framework (block A) and modifies the features so that they observe acoustic conditions more suitable for a singing voice. Throughout this paper, for any building block, say block H, notations X_H and Y_H refer to MGC and BAP as outputs from Block H, respectively. The first index refers to frame number and the second index refers to the MGC order for X_H and BAP order for Y_H. Subscript T denotes the features of the template song.

3.1. Energy-based Nonlinear Time Warping (ENTW)

Component B applies a non-linear time warping function $d(n)$ to MGC and BAP in such a way that vowel elongations is concentrated near the middle of a spoken vowel (assumed to be more acoustically consistent) to avoid over-lengthening the bordering ones (generally exhibiting lower energy and/or weaker spectral features). Utilizing the relationship of the first coefficient and the summation of the logarithmic of the filter bank energy, we approximate the relative energy contour using $C0$.

For a given vowel segment, let N_1 be the first frame and N_2 be the last frame of the segment, our warping function is defined as

$$d(n) = N_1 + \frac{\sum_{m=N_1+1}^{n} e^{X_0(m,0)}}{\sum_{m=N_1+1}^{N_2} e^{X_0(m,0)}}(N_2 - N_1). \quad (1)$$

If $d(n)$ is not an integer, the value of $X_B(d(n), k)$ is approximated using linear interpolation. BAP and F0 are warped in the same fashion as MGC using the same function $d(n)$. Intuitively, the high energy frames are stretched while the lower energy frames are compressed in such a way that the segment length remains the same, as $d(N_1) = N_1$ and $d(N_2) = N_2$. In other words, the warping affects only vowel segments.

The effect of ENTW is illustrated in Fig. 2 and in Fig. 3. Fig. 2 shows waveform outputs and Fig. 3 shows the corresponding spectrograms. Plots a, b, and c are the signal outputs of blocks A, B, and D, respectively. We also include the template waveform as a reference (Fig. d). The timing information indicates that the signal has three vowel segments (I, II, and III), whose boundaries are indicated in both figures. In Fig. 2, we can see that the amplitude of both vowels I and III fades toward the end of the segments, which is a characteristic of poor singing voice quality compared to the template song shown in Fig. 2d.

After ENTW, the high-energy interval elongates (Fig. 2b) and this is also shown in the frequency domain (Fig. 3b). In Fig. 3a, the high frequency content of vowel III fades after 5s but appears fuller in 3b. In other words, the high-energy part is stretched and the low-energy is compressed in such a way that the length of both vowels remain the same. It can be seen that the spectral content of the last parts of both vowels are fuller than the signal without ENTW.

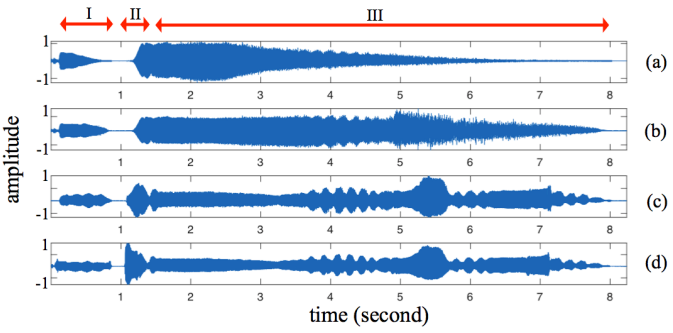

Fig. 2. Intermediate outputs in time domain: (a) - (c) are intermediate outputs and (d) is the template reference.

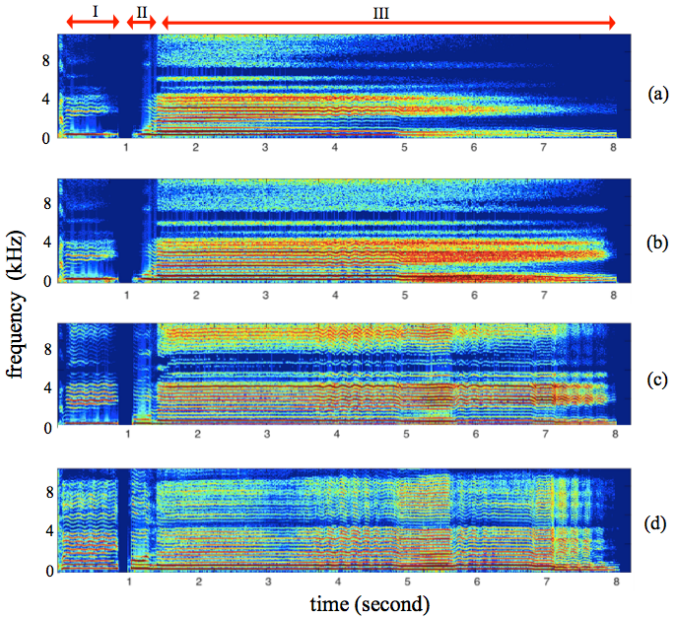

Fig. 3. Intermediate outputs in frequency domain: (a) - (c) are intermediate outputs and (d) is the template reference.

3.2. F0-driven vocalic timbre interpolation

Supplementary recordings are obtained from another speaker to add some fullness to vowel segments. We refer to the speaker's voice as the *supplementary voice* and refer to each vowel recording or sample as a *vowel exemplar* or simply *exemplar*. Our *vocalic library* refers to a collection of vowel exemplars where the supplementary speaker enunciates each vowel at different pitch levels (e.g, low, mid, and high). We found that recordings at different pitch levels have different spectral envelopes, so exemplars at several pitches are needed for accuracy. The recording process is done offline once and the vocalic library can be used with any singing voice.

For each vowel segment, the phonetic label (extracted from the template timing) is used to query which vowel exemplars to use from the vocalic library. In Block C, the pitch sequences from the song template ($F0_T$) determines which pitch level(s) of that vowel is used to construct the exemplar features that match the pitch of each frame. Since, the limited number of pitch levels cannot cover all pitch values, we estimate the MGC features $X_C(n,k)$ at a certain pitch by linear interpolation from the exemplars whose $F0$ averages closest to the $F0$ of that particular frame. It is possible that the vocoder may detect a voiced frame as unvoiced, so we select the minimum BAP value of the exemplars (higher voicing degree) , i.e., $Y_C(n,k) = \min\{Y_1(n,k), Y_2(n,k)\}$ for each frequency bin k and frame n.

3.3. Acoustic feature fusion

This component (block D) combines features from three sources of information: TTS-based features, vowel exemplars, and the template. We keep the first K coefficients of $X_B(n,k)$. The rest of the coefficients are replaced with $X_C(n,k)$. From our inspection, we found that K = 30 is an appropriate order that adds some spectral content (from the exemplar voice) to high frequencies while still maintaining the identity of the virtual singer. Note that this procedure is only executed in vowel frames.

To reduce an abrupt change of the MGC values at the vowel segment boundaries, we gradually increase the effect of the exemplar coefficient values when transitioning from non-vowel frames to vowel frames. We achieve this by using a ramp function with 4 defining points as (M_1,M_2,M_3,M_4) in order (shown in Fig. 4). To transition between a vowel and a non-vowel segment, we use a ramp function $r_V(n)$ defined by (N_1, $N_1 + L$, $N_2 - L$, N_2) with the ramp length L. In other words, the MGC at this stage is defined as $X_V(n,k) = r_V(n)X_C(n,k) + (1 - r_V(n))X_B(n,k)$ for $k \geq K$ and $X_V(n,k) = X_B(n,k)$ for $k < K$.

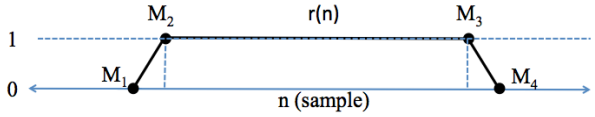

Fig. 4. Ramp function with defining points (M_1,M_2,M_3,M_4)

In addition, we utilize the energy contour and spectral tilt of the template to further enhance the features. To do so, we take the average of X_V and X_T instead of only X_T in order to avoid amplitude instability in the reconstructed waveform, which can occur when the modified C0 contour significantly differs from the original values. We found that applying the same process for the second coefficient $C1$ also makes the output have more singing characteristics.

We found that the above process works well for sonorant phonemes (e.g, vowels, semivowels, or nasals). However, obstruent phonemes (plosives, fricatives, and affricates) are short and turbulent, making the process unreliable. For this reason, we keep the intervals near the boundaries close to the output of the baseline with a margin ramp of length M as a leeway when applying a ramp function to transitions between obstruents and sonorants. The ramp function $r_D(n)$ is defined by $(N_1 - L - M, N_1 - M, N_2 + M, N_2 + M + L)$ where L, N_1 and N_2 are ramp length, the first and last sample of the obstruent segments, respectively. Or mathematically, $X_D(n,k) = r_D(n)X_B(n,k) + (1 - r_D(n))(\frac{X_V(n,k)+X_T(n,k)}{2})$ for k = 0 and 1.

Fig. 3c shows that the spectral and voice characteristics improve after feature fusion. By comparing 3b and 3c, the high frequency content (>4 kHz) in all vowel segments is more visible indicating that the high-frequency energy and formants are enhanced, resulting in a richer or fuller voice. We can notice a difference in segment II where the spectral content is barely visible in the baseline but relatively full in the enhanced output. In segment III, the spectral content above 6 kHz has higher energy, clearer formants, and higher voicing.

	Method effect		Gender effect	
	t-value	p-value	t-value	p-value
Base vs. Enh-SD	2.079	0.0387 *	0.180	0.8571
Base vs. Enh-SI	3.035	0.00266 **	-0.267	0.78949
Enh-SD vs. Enh-SI	0.773	0.4404	-0.303	0.7619

Table 1. Significance tests for the subjective evaluation. The '*' symbol indicates significant results.

In 5s - 8s, the harmonic structure in Fig. 3c is much clearer than the baseline, suggesting voicing, which is expected in a vowel segment. The overall spectral characteristics are also similar to that of the template (Fig. 3d). The improvement is also evident in the time domain (Fig. 2). The overall amplitude envelope of the enhanced signal (Fig. 2c) is similar to that of the template waveform (Fig. 2d) with modulation details that make the singing voice more pleasant. The amplitude of vowel segment II is also dramatically lifted. Note that the output shown in Fig. 2c is without amplitude scaling in the time domain. The similarity between the amplitude contour of the enhanced output and that of the template song therefore suggests the effectiveness of the feature fusion technique.

4. EXPERIMENTAL EVALUATION

4.1. Evaluation framework

We observed that the proposed enhancement techniques effectively reduced hoarseness and appropriately elongated the vowel segments. To validate these observations, we carried out a listening test using the Comparsion Mean Opinion Score (CMOS) approach to evaluate three different methods against one another: Base (baseline TTSing), Enh-SD (proposed enhancement where the supplementary voice is the virtual singer), and Enh-SI (proposed enhancement where the supplementary voice is the opposite gender of the virtual singer). The purpose of our subjective evaluation is to determine: 1) how cleaner and *healthier* (without noticeable hoarseness or lack of energy) both singing enhancement approaches are perceived compared to the baseline, 2) if there is a significant difference in the same perceptual aspects between the speaker-dependent and speaker-independent approaches.

Twelve short singing excerpts (6 per gender, selected from four Chinese songs) were used to generate samples with the three methods, resulting in a total of 36 unique audio clips. Twenty-two native speakers of Chinese were presented with pairs of audio clips and asked to compare how clean or healthy the two clips are relative to each other on a 7-point scale: -3 (the one on the left is much better), -2 (the one on the left is better), -1 (the one on the left is slightly better), 0 (the two audios sound the same, or their differences are not pertinent to the question), 1 (the one on the right is slightly better), 2 (the one on the right is better), and 3 (the one on the right is much better). The locations of the audios in each pair were randomly swapped to avoid the effect of presentation order.

The instrumental accompanying music at a low volume was included since it was found helpful for the identification of the singing flow in short excerpts. The listeners were asked to perform the test using good-quality headphones in a quiet room. One participant was excluded from the statistical analysis because she or he did not complete the test. Finally, we analyzed the participants' responses for each comparison (Base vs. Enh-SD; Base vs. Enh-SI; Enh-SD vs. Enh-SI) using linear regression.

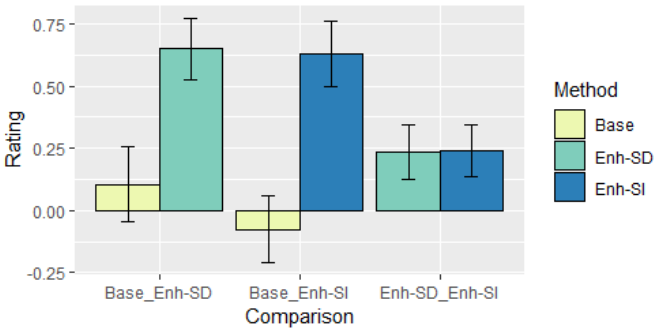

Fig. 5. Results of the subjective evaluation. Error bars show standard errors of the mean.

4.2. Results

The results of the listening test are shown in Fig. 5 and Table. 1. As expected, the enhanced singing is significantly better than the baseline (p-values <0.05). This pattern holds for both speaker-dependent and speaker-independent methods regardless of gender. This confirms our claim about the perceptual benefits of the proposed enhancement strategy. Furthermore, the speaker-dependent method (Enh-SD) is not significantly different from the speaker-independent method (Eng-SI) (p-value >0.05). This indicates that the cepstrum-based timbre interpolation between TTS features and natural singing actually shows speaker independent conditions since it leads to similar benefits despite the use of acoustic features from different voices.

5. CONCLUSIONS

We have presented a TTS-based singing framework as well as techniques to enhance the singing voice output. The energy-based nonlinear time warping (ENTW) algorithm appropriately stretches and compresses different portions in each vowel to reduce low-energy intervals. The timbre of the signals are enhanced by supplementary vowel recordings from our vocalic library. The feature fusion algorithm combines the information from the enhanced timbre, the ENTW output, and the reference template to improve the contours of energy and aperiodicity of the singing voice. The listening test validates that the enhanced singing was perceived with higher quality than the baseline framework without the enhancement techniques. Additionally, the enhancement techniques are flexible to use with different voices. Future work will include validating the system with more languages. In addition, we plan to further investigate the different characteristics between speech and singing such as the dynamics of formant frequencies, aperiodicity, and consonants. We also plan to develop other enhancement techniques and utilize other useful information from the template reference to further improve the quality of the singing voices.

6. ACKNOWLEDGMENTS

We would like to thank our team colleagues Sandesh Aryal, Sam Kang, Gilles Degottex, and Carol Figueroa for their active participation in the development of our singing synthesis technology research project.

7. REFERENCES

[1] T. Saitou, N. Tsuji, Unoki M., and M. Akagi, "Analysis of acoustic features affecting "singing-ness" and its application to singing-voice synthesis from speaking voice," in *Proc. of INTERSPEECH*, 2004.

[2] H. Kawahara, M. Katsuse, and A. Cheveigne, "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, pp. 187–207, 1999.

[3] T. Saitou, M. Unoki, and M. Akagi, "Development of an f0 control model based on f0 dynamic characteristics for singing-voice synthesis," *Speech Communication*, vol. 46, 2005.

[4] Saitou. T., M. Goto, M. Unoki, and M. Akagi, "Vocal conversion from speaking voice to singing voice using straight," in *Proc. of INTERSPEECH*, 2007.

[5] 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 *Proc. of WASPAA*, 2007.

[6] S. Aso, M. Saitou, T. Goto, K. Itoyama, T. Takahashi, K. Komatani, T. Ogata, and H. Okuno, "Speakbysinging:converting singing voices to speaking voices while retaining voice timbre," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx)*, 2010.

[7] T.L. Nwe, M. Dong, P. Chan, Xi. Wang, B. Ma, and H. Li, "Voice conversion: from spoken vowels to singing vowels," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2010.

[8] L. Cen, M. Dong, and P. Chan, "Segmentation of speech signals in template-based speech to singing conversion," in *Proc. of APSIPA Annual Summit and Conference (APSIPA-ASC)*, 2011.

[9] L. Cen, M. Dong, and P. Chan, "Template-based personalized singing voice synthesis," in *Proc. of ICASSP*, 2012.

[10] M. Dong, S.W. Lee, H. Li, P. Chan, J.W. Peng, X. Ehnes, and D. Huang, "Iir speech2singing perfects everyone's singing," in *Proc. of INTERSPEECH (show & tell session)*, 2014.

[11] K. Vijayan, M. Dong, and H. Li, "A dual alignment scheme for improved speech-to-singing voice conversion," in *Proc. of APSIPA Annual Summit and Conference (APSIPA-ASC)*, 2017.

[12] Vijayan. K. and H. Li, "Parallel speak-sing corpus of english and chinese songs for speech-to-singing voice conversion," in *Proc. of the Language Resources Evaluation Conference (LREC)*, 2018.

[13] P. Depalle, G. Garcia, and X. Rodet, "A virtual castrato(!?)," in *Proc. of the International Computer Music Conference (ICMC)*, 1994.

[14] P.R. Cook, "Singing voice synthesis: history, current work, and future directions," *Computer Music Journal*, vol. 20, no. 3, 1996.

[15] Bonada J. and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, 2007.

[16] H. Kenmochi and H. Oshita, "Vocaloid commercial singing synthesizer based on sample concatenation," in *Proc. of INTERSPEECH'07*, Antwerp, Belgium, 2007.

[17] K. Saino, H. Zen, Y. Nankaku, Lee A., and K. Tokuda, "An hmm-based singing voice synthesis system," in *Proc. of INTERSPEECH*, 2006.

[18] T. Nose, T. Kanemoto, M. Koriyama, and T. Kobayashi, "Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling," vol. 34, no. 1, 2015.

[19] L. Feugere, C. D'Alessandro, B. Doval, and O. Perrotin, "Cantor digitalis: chironomic parametric synthesis of singing," 2017.

[20] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences (Sound and Music Computing)*, vol. 7, no. 12, 2017.

[21] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[22] Masanori Morise, Fumiya Yokomori, and Kenji 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.

[23] Emilia Gómez and Jordi Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.