INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

A Combination of Model-based and Feature-based Strategy for Speech-to-Singing Alignment

Bidisha Sharma, Haizhou Li

Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
{s.bidisha, haizhou.li}@nus.edu.sg

Abstract

Speech and singing are different in many ways. In this work, we propose a novel method to align phonetically identical spoken lyric with a singing vocal in a speech-singing parallel corpus, that is needed in speech-to-singing conversion. We attempt to align speech to singing vocal using a combination of model-based forced alignment and feature-based dynamic time warping (DTW). We first obtain the word boundaries of speech and singing vocals with forced alignment using speech and singing adapted acoustic models, respectively. We consider that speech acoustic models are more accurate than singing acoustic models, therefore, boundaries of spoken words are more accurate than sung words. By searching in the neighborhood of the sung word boundaries in the singing vocal, we hope to improve the alignment between spoken words and sung words. Considering the word boundaries as landmark, we perform speech-to-singing alignment at frame-level using DTW. The proposed method is able to achieve a 47.5% reduction in terms of word boundary error over the baseline, and subsequent improvement of singing quality in a speech-to-singing conversion system.

Index Terms: Speech-to-singing alignment, speech-to-singing convesrion, singing adapted ASR, dynamic time warping.

1. Introduction

One of the important tasks in speech-to-singing conversion is to align the spoken words in the speech to the sung words in the singing vocal. Speech-to-singing is the process of converting spoken lyrics to singing, that adapts the particular singing prosody, the linguistic content and the speakers voice. In doing so, we need to obtain the alignment to effect the spectral and prosodic conversion [1].

Speech and singing differ from each other in several aspects. Although we use the same underlying physiological mechanism for production of both speech and singing, we manipulate the sounds and prosody in different ways. This leads to significantly different speaking rate, pitch range, loudness, and voice quality between speech and singing, which makes it challenging to temporally align the two signals.

Due to different laryngeal height in professional singer's singing style, there is addition of singer's formant and extra formant in the singing voice spectrum. The authors of [2] explains that the lowering the larynx causes the lengthening of the vocal-tract which in turn causes most formants of singing generally to shift downward from the corresponding formants of the speaking voice. According to the musical score, the singers manipulate duration of each phoneme and add pitch fluctuations overshoot, vibrato, preparation and fine fluctuation, unlike speaking [3, 4]. The fine variation of power in singing also makes it different from speech. The ratio of vowel to consonant duration is higher in singing than in speech [5], which results in a non-linear temporal relation between the two. Due to these differences, applying signal processing methods to align speech and singing may not be always successful and reliable, specially when songs are sung with high pitch and much lowered larynx position.

We note that text-to-speech alignment has been well studied [6]. The counterpart in singing, which is lyrics-to-singing alignment has also been widely studied [7–11], which aims to automatically align the phonemes/words in the lyrics to the singing vocal. However, the problem of speech-to-singing alignment is entirely different in the sense that we have to align two signals and not signal to text. It also significantly differs from aligning singing voice with MIDI melody [12, 13] and singing-to-singing alignment [14–16], which were attempted using phoneme and musical context recognizers and dynamic time warping (DTW), respectively. The initial work in speech-to-singing conversion [3] used duration information from the input MIDI file and, [4] used manual alignment in phoneme level between speech and singing.

As a solution to this problem, the authors of [17] proposed an alignment method, where, initially speech and singing signals are divided into several segments manually, and further fine alignment is performed using DTW between the Mel frequency cepstral coefficients (MFCCs). To overcome the errors in DTW based alignment technique and avoid manual processing, Karthika et al [18] proposed a dual alignment scheme. Considering the availability of parallel speech corresponding each singing utterance, they obtained alignment between source speech and singer's speech in the first pass. In the second pass, singer's speech is aligned to original singing template, which in turn provides the temporal alignment between source speech and original singing template. Yet, they assumed that manual word boundary alignment is available for the original singing template. Another fully automatic improved alignment method using knowledge of commonalities between speech and singing with DTW is proposed in [19]. In this work, we leverage the idea of forced alignment with speech and singing adapted acoustic models. We use Montreal forced aligner (MFA) [6, 20] to align the speech and lyrics, and obtain word boundaries. In [21], lyrics-to-audio alignment is performed using singing adapted acoustic models, which has a reliable performance on solo-singing vocals. In this case, the difference between speech and signing is incorporated during the adaptation of speech models to singing, using singing data. We hypothesize that instead of directly using DTW over two very long segments, we can first obtain the word boundaries for source speech and target singing, using forced alignment with speech and singing adapted acoustic models, respectively. Further, refining the singing word boundaries by taking cues from speech words would certainly help for later stage. To obtain fine level align-

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1942

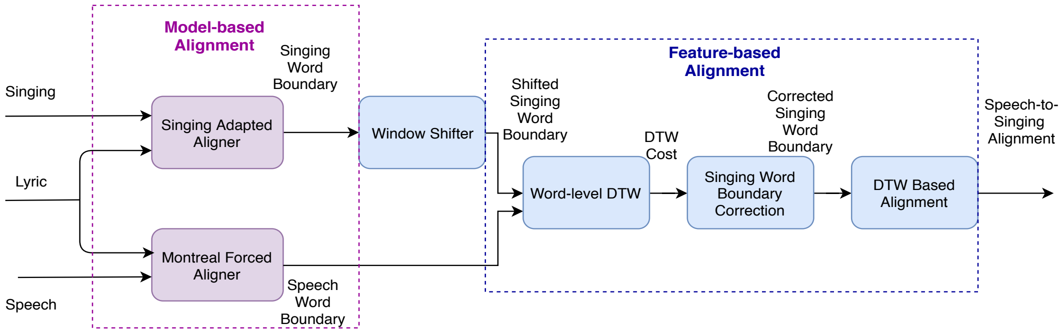

Figure 1: *Block diagram of the proposed combined model-based and feature-based speech-to-singing alignment strategy.*

ment, we can apply DTW between word boundaries of speech and singing. This method can take advantage from both acoustic models and DTW, which may reduce the propagated error compared to the traditional DTW-based alignment performed between two utterances. We note that forced alignment is a model-based technique, while DTW is a feature based technique. We propose a framework to leverage the use of two different techniques for the best performance.

The rest of the paper is organized as follows, Section 2 describes the lyric-to-audio alignment method. In Section 3, we describe the proposed speech-to-singing alignment. The database and experimental set up is explained in Section 4. We draw conclusion of this study in Section 5.

2. Lyrics-to-audio Alignment Using Acoustic Models

We suppose that we have a pair of speech-singing parallel utterances, and the corresponding text. We use MFA to align the spoken words in the speech to the text in word level, that we call text-to-speech alignment.

MFA uses a standard Gaussian mixture model (GMM)/hidden Markov model (HMM) architecture, that is built with Kaldi recipes [22, 23]. As MFA relies on an acoustic model, we call it a model-based approach. To build such a model, monophone GMMs are trained in an iterative manner and used to generate a basic alignment. Then, the triphone GMMs are trained to take surrounding phonetic context into account, along with clustering of triphones to combat sparsity, which are used to generate the alignments. The triphone models are then used for learning acoustic feature transforms on a per-speaker basis, in order to make the models more applicable to speakers in other datasets. We have used their pre-trained English acoustic model trained on the Librispeech corpus [23]. The reported mean word boundary error of MFA is 24 msec, which is comparable to inter-transcriber reliability.

In view of the differences between speech and singing, it is not appropriate to use the same speech models to align singing vocal to text. As a solution to this, the speech acoustic models can be adapted to singing voice using speaker adaptation methods [24, 25]. Studies [25, 26] have suggested ways to adapt speech acoustic models towards singing acoustic models. These singing adapted models are applied to lyrics-to-audio alignment in [21], which achieves a 400 ms mean word boundary error on solo-singing vocals. However, the performance of forced alignment using the singing adapted acoustic model remains to be improved for effective speech-to-singing conversion. In this work, we propose to use the forced alignment boundaries at the word level, and use dynamic time warping to refine the word boundaries at the frame level.

3. Speech-to-Singing Alignment using Acoustic Models

As shown in Figure 1, we pass the spoken lyrics and corresponding text through MFA, to obtain the speech-to-text alignment. Similarly, we apply singing vocal and corresponding linguistic information to the singing adapted lyrics-to-audio alignment module, to obtain word boundaries in singing vocals. We believe that the spoken word boundaries are more accurate, whereas the sung word boundaries obtained from singing adapted models are less precise [21]. Nevertheless, it is always useful for us to get some intuition regarding word boundaries in singing vocals without any manual intervention. To refine these word boundaries, we propose to use the more accurate spoken word boundaries as the landmark points.

3.1. Singing word boundary correction

We consider that there are N number of words in each of spoken and singing utterance. The beginning and ending word boundaries for speech can be represented as $(T^b_{\text{sp},1}, T^e_{\text{sp},1}), (T^b_{\text{sp},2}, T^e_{\text{sp},2})...(T^b_{\text{sp},N}, T^e_{\text{sp},N})$. The same word boundaries for singing can be represented as $(T^b_{\text{sn},1}, T^e_{\text{sn},1}), (T^b_{\text{sn},2}, T^e_{\text{sn},2})...(T^b_{\text{sn},N}, T^e_{\text{sn},N})$. We consider speech word boundaries as constant and we vary the singing word boundaries by a factor h. We use different values of $h(k)$ as shown in Equation 1 and determine the h_{optimal} based on minimum DTW cost (D) between speech ($\boldsymbol{W}_{\text{sp}}$) and shifted singing ($\boldsymbol{W}_{\text{sn},\boldsymbol{h}(k)}$) words. Based on h_{optimal}, we find the shifted and modified singing word boundary ($T^b_{\text{sn},i,\text{optimal}}$, $T^e_{\text{sn},i,\text{optimal}}$).

$$\boldsymbol{h}(k) = k \times 20 \text{ ms}; \;\; \boldsymbol{k} = -5, -4, ..., 4, 5, \tag{1}$$

$$\boldsymbol{D}(k) = DTW(\boldsymbol{W}_{\text{sp}}, \boldsymbol{W}_{\text{sn},\boldsymbol{h}(k)}), \tag{2}$$

$$h_{\text{optimal}} = argmin(\boldsymbol{D}(k)), \tag{3}$$

$$T^b_{\text{sn},i,\text{optimal}} = T^b_{\text{sn},i} + h_{\text{optimal}}, \tag{4}$$

$$T^e_{\text{sn},i,\text{optimal}} = T^e_{\text{sn},i} + h_{\text{optimal}}, \tag{5}$$

where, $T^b_{\text{sn},i,\text{optimal}}$ and $T^e_{\text{sn},i,\text{optimal}}$ and represent the optimal beginning and end singing word boundaries for i^{th} word.

After refinement of the singing word boundaries, we performed DTW between corrected singing words and original speech words to obtain fine level (frame level) alignment between the two. The alignment procedure is illustrated in Figure 2. In this manner, for the entire utterance we obtain frame level alignment for all word pairs. In between two words, if there a silence segment in both singing and speech, we derive the corresponding alignment in the identical way.

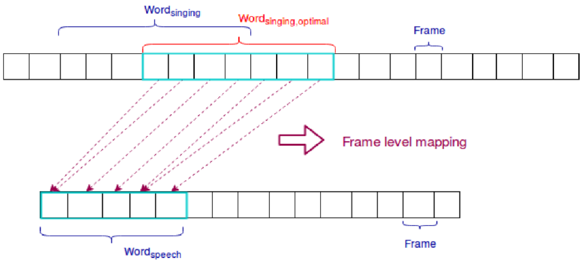

Figure 2: *Illustration of singing word boundary correction.*

Table 1: *Details of speech singing parallel database.*

Category	Spoken lyrics	Sung lyrics
Number of songs	80	80
Number of singers	8	8
Number of utterances	2,314	2,314
Number of words	17,017	17,017
Total duration (mins)	116	237
Sampling rate (kHz)	48	48
Bits/sample	32	32

4. Experiments

In this section, we would like show the impact of the proposed alignment method in a speech-to-singing conversion system.

4.1. Database

For the analysis and experiments, we use speech and singing parallel corpus, which is an ongoing effort at Human Language Technology (HLT) Laboratory, National University of Singapore (NUS), towards development of a publicly available database for different research problems, related to speech and singing. Both speech and singing audio are recorded in a professional studio environment with high quality recording equipment, under the supervision of a trained and experienced sound engineer. Currently, we have 80 (12 unique) English songs, that comes from 8 singers (4 male and 4 female). The singers are either professionally trained or have at least three years of public singing experience. For this available data, we have manually marked utterance and word boundaries. The details of the database is mentioned in Table 1.

4.2. Experimental Setup

In this work, our aim is to obtain alignment between source speech and target singing utterances, with same linguistic context. To derive text-to-speech alignment we use MFA, with pre-trained model for English that has been trained on the LibriSpeech corpus using MFCC features [23].

As discussed in Section 2, we use singing adapted speech acoustic models, trained on solo-singing dataset [27] to force-align lyrics with singing vocals. The baseline speech acoustic model is a tri-phone GMM-HMM trained on Librispeech corpus [23] using MFCC features on Kaldi toolkit [22]. To make the Viterbi alignment algorithm operate over the long duration, we set the alignment retry-beamwidth to a high value of 4000 and allow for optional silence between words. We obtain word boundaries of the singing utterance with forced alignment using these singing adapted models.

To compensate the inaccuracy of the singing adapted aligner in estimation of the word boundaries, we refine them using cues from segmented speech words as described in Section 3. To obtain the proposed alignment between speech and singing signals, we use these word boundaries as landmark points and perform DTW based alignment between speech and singing words. To perform DTW, we use 13-dimensional MFCCs and their delta & delta delta features (39-dimensional) derived from both speech and singing signals. To handle the silences, we have also mapped the silence segments in between two words, if the silence is present in both speech and singing. To justify the word boundary correction strategy, we also obtain the same frame-level alignment between the speech and singing words without singing word boundary correction.

We compare our proposed alignment strategy with the work done in [19], where the same speech-to-singing alignment is performed in utterance level using DTW. Based on the commonalities between speech and singing, they used features that include one-dimensional voice activity detection (VAD), 12-dimensional low-time cepstrum (LTC), 16th order cepstrum obtained from STRAIGHT and 16 linear prediction (LP) cepstral coefficients. In total, they used 45-dimensional tandem feature set obtained from speech and singing to perform DTW alignment between two utterances. Unlike our approach, they didnot use any acoustic model for this baseline alignment.

In DTW, we find the lowest-cost path between the two signals. We added the DTW costs over all the words corresponding to an utterance and normalized by total number of frames. In a similar manner, we obtain DTW cost of alignment for the baseline method [19], which performs alignment in the utterance level. As depicted in Table 2, for the baseline method (*Baseline*) the DTW cost is 0.69, which is 0.55 for the proposed method (*Forced alignment+Word boundary correction+DTW*). The same cost for word-level alignment without singing word boundary correction is 0.60, which referred as *Forced alignment+DTW* in Table 2. This shows that instead of performing alignment between two utterances, we can perform the alignment based on the landmark points. In this case, our landmarks are the corrected word boundaries obtained from acoustic models. The DTW cost along with mean of word boundary errors (WBE) are depicted in Table 2. The WBE is computed by comparing the word boundaries obtained from proposed and baseline methods with manual word boundaries. We can observe that employing forced alignment using acoustic models to obtain word boundaries gives us wide improvement over baseline method. The proposed combined model-based and feature-based strategy further reduces both the evaluation parameters.

4.3. Speech-to-singing conversion

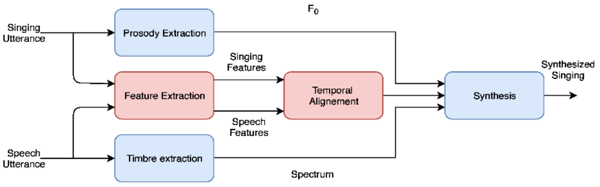

Figure 3: *Speech-to-singing conversion framework.*

As mentioned earlier, the alignment between source speech and target singing has crucial importance on a speech-to-singing conversion system. The quality of synthesized singing voice severely degrades if the alignment between source speech and target singing is not correct.

Figure 3 shows a conventional template-based speech-to-singing conversion system [1], where we extract the prosody (F_0) information from the target singing and timber (spectral) information from the source speech. The singing voice output is synthesized by retaining the prosody of singing template, re-

Table 2: *Performance of the proposed alignment method in terms of mean word boundary error (WBE) and DTW cost.*

Method	WBE (ms)	DTW cost
Baseline	315.50	0.69
Forced alignment+DTW	193.60	0.60
Forced alignment+Word boundary correction+DTW	165.35	0.55

placing the spectral characteristics from speech template, according to the alignment information. In this work, we use WORLD analysis/synthesis framework [28] to extract F_0, spectrum and synthesize singing voice. The implementation provided in [29] is followed for this task.

We perform subjective listening test to evaluate the performance of the proposed alignment on speech-to-singing conversion. We consider 11 subjects and provide them 3 sets, each with 10 samples of synthesized singing. The 3 sets correspond to speech-to-singing conversion using baseline alignment, proposed alignment (*Forced alignment+Word boundary correction+DTW*) and vocoded original singing. The 10 samples in each set consist of 5 examples of female singing obtained from female speech, and 5 examples of male singing obtained from male speech. The listeners are asked to give their opinion scores on a scale of 1 to 5, where 1 represents unacceptable, 2-poor, 3-fair, 4-good and 5 denotes excellent, based on the quality of synthesized singing. As shown in Table 3, the mean opinion score (MOS) corresponding to the baseline alignment is 2.89, while for the proposed (*Forced alignment+Word boundary correction+DTW*) alignment the MOS is 4.33. The distributions of opinion scores provided by all the subjects corresponding to each method are shown in Figure 4, with the median values indicated by red lines. Corresponding mean values are noted in Table 3. For the same set of examples, we also note the log spectral distance (LSD) between original and converted singing as shown in Table 3. Besides, we perform a preference test, where we provide 20 pairs of synthesized singing, each pair consists of examples from proposed and baseline alignment methods. The same 11 subjects are asked to listen and provide their preference among the two or mark as no preference. The percentage of preference for baseline (7.17%) and proposed (87.08%) methods are shown in Table 3. This shows the efficacy of the proposed combined model-based and feature-based alignment, which is reflected on the performance of an speech-to-singing conversion system. The examples of speech-to-singing converted samples for proposed and baseline alignment are provided in this link [1].

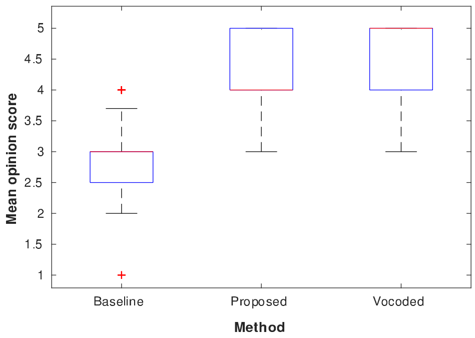

Figure 4: *Boxplot of mean opinion scores for synthesized singing obtained using baseline & proposed (Forced alignment+Word boundary correction+DTW) alignment methods, and vocoded singing.*

Table 3: *Subjective evaluation results in terms of mean opinion score (MOS), log spectral distance (LSD) and preference %.*

Method	MOS	LSD	Preference (%)
Baseline	2.89	1.63	7.17
Forced alignment+Word boundary correction+DTW	4.33	1.42	87.08
Vocoded	4.52	–	–
No preference	–	–	5.74

5. Discussion and Conclusion

In this work, we present a fully automatic method to align phonetically identical spoken lyric with a singing vocal in a speech-singing parallel corpus. We propose to combine both model-based and feature-based approaches to bring out a better alignment strategy. The novel idea of employing word boundaries obtained from forced alignment as landmark points using speech and singing adapted acoustic models, certainly helps us to overcome the fine level alignment error to a large extent. This module would be definitely helpful in avoiding the manual task of word segmentation and the errors in frame-level alignment. The landmark based frame alignment using DTW, between short segments (words) of speech and singing further helps to reduce the accumulated error compared to alignment between longer segments (utterances) of speech and singing.

Furthermore, to compensate the error in singing word boundary prediction, we attempt to shift them by taking cues from the speech word boundaries, which is explained in Section 3. This refinement of singing word boundaries reduces the DTW cost and in turn the alignment error. The histogram of duration of the singing word boundary shift for one singer is shown in Figure 5(a). We can observe that most of the boundaries are shifted by 100 ms towards right, compared to the boundaries obtained from forced alignment. With this shift, the average DTW cost comes down from 0.69 (baseline) to 0.54 (proposed), which is shown in Figure 5(b).

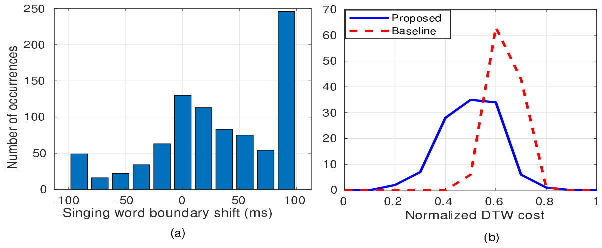

Figure 5: *(a) Histogram showing word boundary shift for singing words, (b) continuous histogram distribution of normalized DTW cost corresponding to baseline and proposed method, for one singer.*

Although we achieve certain reduction in the DTW cost which represents alignment error for the entire utterance, we note that this optimization of boundaries may not be favourable for obtaining the global least cost path for an utterance. In our future work, we would like to improve this optimization by considering global strategy for singing word boundary correction, instead of the method used.

6. Acknowledgements

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

[1] https://bidishasharma.github.io/alignment/

7. References

[1] 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, 2019.

[2] 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.

[3] 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*, 2007.

[4] S. Aso, T. Saitou, M. Goto, K. Itoyama, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "Speakbysinging: Converting singing voices to speaking voices while retaining voice timbre," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.

[5] Y. E. Kim, "Singing voice analysis, synthesis, and modeling," in *Handbook of Signal Processing in Acoustics*. Springer, 2008, pp. 359–374.

[6] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[7] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 2146–2149.

[8] R. Gong, P. Cuvillier, N. Obin, and A. Cont, "Real-time audio-to-score alignment of singing voice based on melody and lyric information," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[10] Y.-R. Chien, H.-M. Wang, S.-K. Jeng, Y.-R. Chien, H.-M. Wang, and S.-K. Jeng, "Alignment of lyrics with accompanied singing audio based on acoustic-phonetic vowel likelihood modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 11, pp. 1998–2008, 2016.

[11] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Systems*, vol. 12, no. 4-5, pp. 307–323, 2007.

[12] M. Dong, P. Chan, L. Cen, and H. Li, "Aligning singing voice with midi melody using synthesized audio signal," in *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 2010, pp. 95–98.

[13] 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, 2007.

[14] P. Cano, A. Loscos, J. Bonada, M. De Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications." in *ICMC*, 2000.

[15] K. Kobayashi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[16] K. Kobayashi, T. Toda, and S. Nakamura, "Implementation of f0 transformation for statistical singing voice conversion based on direct waveform modification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5670–5674.

[17] L. Cen, M. Dong, and P. Chan, "Segmentation of speech signals in template-based speech to singing conversion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2011.

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

[19] K. Vijayan, X. Gao, and H. Li, "Analysis of speech and singing signals for temporal alignment," in *Proc. APSIPA Annu. Summit and Conf.*, 2018.

[20] "Montreal-forced-aligner," https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner, [Online; accessed 26-Nobvember-2018].

[21] B. Sharma, C. Gupta, H. Li, and Y. Wang, "Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models," in *Accepted in ICASSP, 2019*, 2019.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, p. 4, 2010.

[25] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *International Society for Music Information Retrieval (ISMIR)*, 2018.

[26] C. Gupta, H. Li, and Y. Wang, "Automatic pronunciation evaluation of singing," *Proc. Interspeech 2018*, pp. 1507–1511, 2018.

[27] S. Sing!, "Smule.digital archive mobile performances(damp)," https://ccrma.stanford.edu/damp/, 2010 (accessed March 15, 2018).

[28] 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.

[29] "World vocoder," https://github.com/mmorise/World, [Online; accessed 26-Nobvember-2018].