Analysis of Speech and Singing Signals for Temporal Alignment

Karthika Vijayan, Xiaoxue Gao and Haizhou Li
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
E-mails: vijayan.karthika@nus.edu.sg, e0204976@u.nus.edu, haizhou.li@nus.edu.sg

***Abstract*—Accurate alignment between singing signal and its spoken lyrics at frame-level is imperative to several applications in singing signal processing. As the acoustic characteristics of speech and singing signals differ significantly, finding the temporal alignment between them is not easy. In this paper, we study the characteristics of speech and singing signals to identify their common properties to facilitate temporal alignment. We observe that: (i) the characteristics of excitation source in human voice production mechanism largely vary with speaking and singing and, (ii) for the same linguistic content, speaking and singing signals present very different formant patterns. Based on these observations, we formulate a set of tandem features that represent only those characteristics consistent between speech and singing signals. Such tandem features are used in dynamic time warping for temporal alignment, and in a speech-to-singing conversion experiment. In both objective and subjective evaluations, we show that the proposed tandem features are significantly superior to the baseline features in temporal alignment.**

Index Terms: Singing voice, Singing formant, Low-time cepstrum, Tandem features, Temporal alignment.

I. Introduction

Temporal alignment between speech and singing signals of the same linguistic content is extremely crucial in various applications including speech-to-singing (STS) voice conversion, audio search of songs, content retrieval from songs databases and singing signal processing of amateur singers. Particularly in STS, the temporal alignment between source speech and target singing template/musical score significantly affects the processes of prosody and spectral transformations from speech to singing signals. As STS is becoming a key enabling technology for many innovative applications such as personalized singing, speech therapy and training & assessment of singing, there is a strong call for such accurate temporal alignment.

However, temporally aligning speech and singing signals is not a trivial problem. Though the human voice production mechanism generating speech and singing signals is the same, its characteristics vary vividly with voice styles. The particularly high placement of larynx forming singing formant [1], [2], the abrupt and consistent changes in subglottal pressure [3], carefully obtained pitch contractions [4], etc. are prominent features of voice production system used by a trained singer to produce different musical notes [5], [6]. In addition, fine variations in pitch and intensity like, overshoot, preparation, vibrato, tremolo, etc. are also used by singers to improve the naturalness of singing [4], [5]. These features play a substantial role in establishing the unique characteristics of singing, which differentiate singing from speech spectra, and make speech to singing alignment difficult.

Lyrics-singing signals alignment has been extensively studied in music information retrieval with different techniques such as adapted hidden Markov model (HMM), or aligning vowel onsets to lyrical notes [7]–[12]. The strategy of detecting and matching vowel onset points was used for aligning singing signals to musical notes [13], [14]. Singing vocals-musical notes alignment was also performed using dynamic time warping (DTW) using spectral features [15] and by matching carefully designed performance trajectories [16]. The alignment of two singing signals was attempted for voice conversion using DTW and phoneme & musical context recognizers [10], [17], [18]. All these methods deal with aligning singing audio with lyrics, musical notes or other singing signals. However, for STS conversion, singing signals have to be aligned with speech signals of the same linguistic content.

One way of time-synchronizing speech to singing signals is by manual segmentation and labeling [19]–[21]. A dual pass alignment between speech and singing, exploiting speaker similarities, was also proposed for increased accuracy [22], [23]. In this method, an arbitrary segment of speech uttered by a speaker is aligned with speech signal produced by a singer (first pass). Later, the singer's speech is temporally aligned with the singing signal, by levying on speaker similarity (second pass). Yet, the synchronization information in the second pass of the dual alignment scheme was assumed to be available by manual intervention [24]. Such schemes are exhaustive and expensive for temporal alignment. There has been an attempt to automate the temporal alignment by using singing signal adapted HMMs [4], [25]. Unfortunately, this is possible only when HMM acoustic models are available.

In this paper, we present an automatic temporal alignment strategy between speech and singing signals, which neither requires any manual intervention nor relies on any speaker similarity characteristics. We attempt to solve the alignment problem using signal processing techniques, without using any statistical acoustic modeling tools. Here, we study the common and different characteristics between speech and singing signals. We then formulate the tandem features of signals that capture only the consistent characteristics thereby, effectively nullifying the effects of inconsistent characteristics between speech and singing signals. Temporal alignment with DTW using the tandem features is performed for STS conversion.

The objective and subjective experiments reported in this paper exhibit the effectiveness of the proposed tandem features over the baseline features.

The rest of the paper is organized as follows: In Section II, we explain a detailed study of various characteristics of speech and singing signals. In Section III, we elaborate the extraction of tandem features from the consistent set of characteristics between speech and singing signals for temporal alignment. The Section IV reports the experiments on temporal alignment and its' application in STS conversion. In Section V, we conclude the study and summarize its' contributions to STS conversion.

II. Analysis of Speech and Singing Signals

The human voice production system, for both speech and singing signals, is widely described by the source-filter model [26]. The vocal tract system takes the excitation from lungs as input and, generates speech or singing as outputs. Here, we use the source-filter model to compare and contrast characteristics of speech and singing signals.

A. Source characteristics

The excitation source in human voice production is constituted by the air flow from lungs, modulated at the glottis [26]. In singing signal production, the air consumption is much higher than that in speech production, resulting in larger lung volumes and hence longer breath cycles [5]. This enables a trained singer to vary subglottal pressure rapidly and accurately to produce target musical notes of intended loudness. The subglottal pressure plays a significant role in controlling the fundamental frequency (F0) and intensity in singing, whereas, it does not have any considerable role in deciding the F0 in speaking [5].

The source characteristics largely vary across singing and speaking voice styles. Short segments of speech and singing signals are shown in Figure 1(a) and (b), respectively, and the corresponding F0 contours are shown in Figure 1(e) and (f), respectively. This figure reveals the large margin of variation among F0 contours of speech and singing, as F0 in singing signal is regulated by target melody, while F0 in speech is not. Also, the F0 contours in singing are affected by vibrato, overshoot and preparation, as can be seen from Figure 1(f) around 0.5 s (vibrato), 2.25 and 2.5 s (overshoot) and 2 and 3 s (preparation). In addition, the loudness of speech and singing signals differ from each other, owing to the wide range of intensity in singing. As we are in search of the signal properties consistently appearing across singing and speaking signals, we consider that the source characteristics (F0 contours) and short-time energy, cannot be included as features for temporal alignment.

B. System characteristics

The vocal tract system in human voice production is constituted by the glottis, oral and nasal cavities [26]. Particular positions of larynx, acting as a resonator by itself and contributing to the formant structure, are manifested as the most

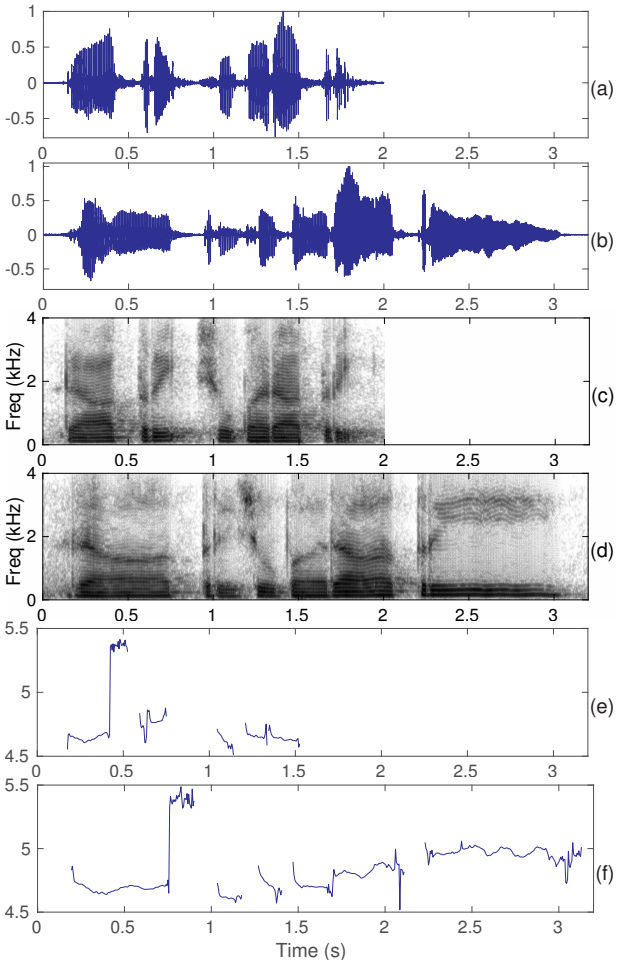

Fig. 1. Illustration of differences in characteristics between speech and singing signals: (a) Speech signal, (b) Singing signal, (c) Speech spectrogram, (d) Singing spectrogram, (e) log(F0) contour - speech and (f) log(F0) contour - singing.

important characteristics of singing signal spectrum. When a trained singer delivers loud singing, he/she does not rely on the rise in subglottal pressure alone, as it is impossible for any human being to produce extremely high pressure beneath vocal folds. Instead, the singer raises the larynx making the high frequency formants (F3,F4 and F5) to form a tight cluster, delivering a noticeable peak around 3 kHz, termed as the 'singing formant' [1], [5]. The presence of singing formant forces the decay of spectral slope of singing signals to be much slower than that in speech signals. This enables the singer to produce loud singing, enough to be heard in between the musical accompaniment [5].

The singing formant and associated slower spectral decay can be observed from the spectrogram of singing signal shown in Figure 1(d) (between 2.25 and 3 s), which are absent in the corresponding speech spectrum given in Figure 1(c). In addition to the singing formant, the spectrum of singing signals

are characterized by tighter coupling between pitch harmonics and formant tracks, as compared to speech signals.

Yet, the spectra of speech and singing signals are in common in many ways, which can be utilized for temporal alignment. As the configurations of articulators do not change significantly upon production of same phonemes in different voice styles, the low frequency formants (F1 and F2) closely match with each other in speech and singing spectra. From Figure 1(c) and (d), it can be observed that the formant tracks of F1 and F2 in speech and singing spectrograms closely follow each other, despite the elongation of duration of these tracks in singing signal. Thus, we propose to use the low frequency region in speech and singing spectra, which are consistent and not affected by the singing formant. Based on the findings from analysis of source and system characteristics between speech and singing signals, we propose

1) normalizing the short-time energy across segments of speech and singing to nullify intensity variations
2) performing source-filter decomposition to obtain smoothed spectral envelope, thus removing the source characteristics
3) restricting the smoothed spectrum to low frequency region to avoid the singing formant

Feature extraction will be performed from the low frequency constrained smoothed spectrum of speech and singing signals.

III. Features for Temporal Alignment

We attempt feature extraction from speech and singing signals, that describe only the consistent set of characteristics between them. A sequence of feature extraction steps is followed to realize the consistency of characteristics between the signals at-hand, as per the study reported in Section II. We concatenate different features from these steps to form a unique feature vector, that we call the 'tandem feature'.

A. VAD and energy normalization

As the linguistic content in singing signals and their spoken lyrics to be temporally aligned are the same, the number of voiced and unvoiced segments are also the same. Hence we use voiced activity decisions (VAD) as part of the proposed tandem features. We perform zero-frequency filtering of speech/singing signals generating zero-frequency signals, whose positive zero crossings indicate epochs in voice production. The periodicity and energy of epochs are used to deduce robust VAD information [27]. Also, the short-time energy of segments of speech and singing signals are normalized to unity in order to avoid mismatches due to intensity variations.

B. Low-time cepstrum

The most commonly used features in speech signal processing are mel frequency cepstral coefficients (MFCC). Typically the MFCC features consist of cepstral, delta and acceleration coefficients and, were used for temporal alignment between speech and singing [22], [24]. The delta and acceleration coefficients represent the dynamic characteristics of audio signals, and are very different between speech and singing. Hence we do not use them here. The lower order cepstral coefficients represent the smoothed spectral envelope of vocal audio. Thus, we perform a 24th order cepstral analysis of speech and singing signals and choose the 12 lower order coefficients representing vocal tract characteristics, that we call the low-time cepstrum (LTC), to be part of the tandem features. We note that the high-time cepstrum reflects the source characteristics, which are therefore isolated out to avoid inconsistency between speech and singing.

To reduce the spectral mismatch between speech and singing signals due to the singing formant, as discussed in Section II-B, we propose to restrict the spectral range to low frequency region, thereby excluding the singing formant.

C. LP and STRAIGHT analyses

The source-filter decomposition from speech and singing signals is performed using (i) linear prediction (LP) analysis and (ii) STRAIGHT analysis. The LP analysis models vocal tract system as an all-pole model and captures the source characteristics in the LP residual [28]. The STRAIGHT analysis performs pitch adaptive windowing and smoothing of speech/singing signals to obtain smoothed spectral envelope and source characteristics [29].

Thus in our study, the smoothed spectral envelope of speech/singing signals are obtained by a parametric modeling and a nonparametric modeling approaches, namely LP and STRAIGHT analyses. The spectrograms obtained from smoothed short-time spectra of singing signal, using STRAIGHT and LP analyses, are shown in Figure 2(b) and (c), respectively. It can be seen that the first two formant tracks in these spectrograms are almost identical to the original tracks in singing spectrogram, given in Figure 2(a). Similar figure can be observed in case of speech signals also. The STRAIGHT spectrogram and LP spectrogram thus obtained are restricted between 0 and 3 kHz, in order to avoid the effects of singing formant. A 16th order cepstral analysis is carried out to produce STRAIGHT and LP cepstral coefficients (STR and LPC) from the respective constrained spectra.

We formulate the set of tandem features constituting of sequential feature extraction schemes including VAD, LTC and STR & LPC steps, which represent the consistent characteristics across speech/singing signals while neglecting the inconsistent characteristics. These tandem features are utilized for temporal alignment between speech and singing using DTW algorithm.

An example of the temporal alignment thus obtained between segments of speech and singing signals is shown in Figure 3. The vertical red lines in Figure 3(a) and (b) represent ground truth for word boundaries, obtained by manual labeling. The vertical black lines in Figure 3(a) represent the word boundaries in singing signal which are aligned with those in speech signal, estimated by DTW algorithm using the tandem features. The timing difference between the red and black vertical lines in Figure 3(a) represent the word-boundary alignment error produced by the proposed features. From visual inspection of Figure 3, it can be observed that the

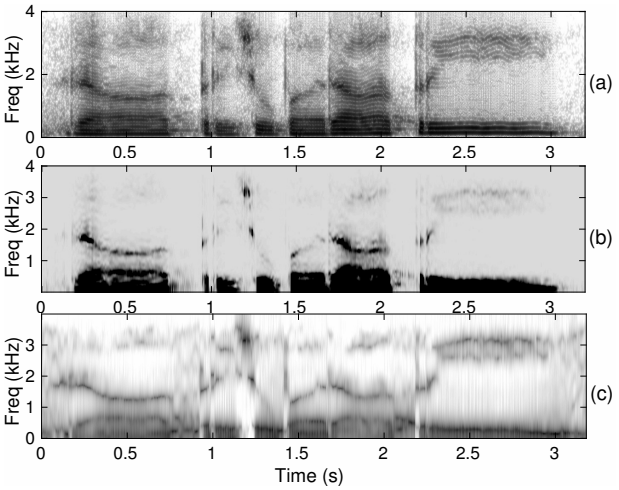

Fig. 2. Smoothed spectrograms obtained from STRAIGHT and LP analyses (a) Singing spectrogram, (b) STRAIGHT spectrogram and (c) LP spectrogram.

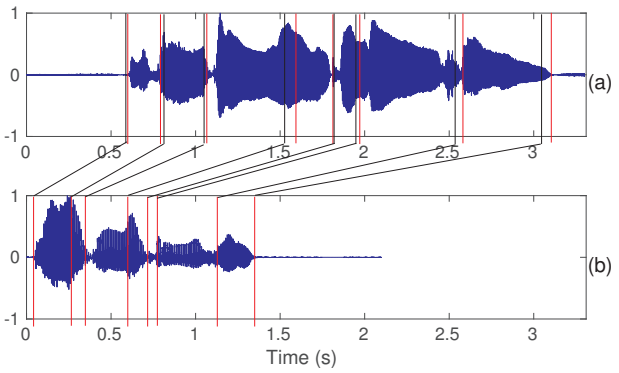

Fig. 3. Word boundary alignment obtained using the proposed tandem features. (a) Singing signal and (b) speech signal for the sentence: 'Now I don't want to lose you'. The vertical red lines and vertical black lines show the original and estimated word boundaries, respectively. Diagonally oriented lines show the boundary alignments between speech and singing.

error committed by the proposed features in aligning speech and singing signals is minimal.

IV. Experimental Evaluation

To demonstrate the superiority of tandem features, we perform quantitative evaluation of temporal alignment between speech and singing signals from two different individuals (user and singer) using DTW algorithm. A database of parallel recordings of source speech and target singing, consisting of 24 popular English songs sung by 3 male and 3 female singers (6 singers $\times$ 4 songs) and read by 3 male and 3 female users, is chosen for the evaluation [30]. All the signals are recorded at a sampling rate of 16 kHz. This database includes 728 lyrical sentences having average spoken duration of 4 s and, has a total of 4,236 words in both speech and singing. Manual labeling of word boundaries of all recordings is performed to build the ground truth for evaluation. The mean value of timing error between the ground truth and estimated word-boundaries using DTW, evaluated over the entire database, is computed as the performance measure for temporal alignment. This measure is termed as the average word-boundary error (AWBE).

The AWBE in DTW-based temporal alignment using different features are reported in Table I. The MFCC features are widely used in speech signal processing and, were utilized for temporal alignment in STS conversion [22]–[24]. Apart from manual labeling, we identify two baseline temporal alignment techniques used in STS systems based on MFCC features as (i) Direct alignment (MFCC+Δ + $\Delta\Delta$_1) [23] and (ii) Dual alignment- source speech aligned with singer's speech, which is later aligned with target singing (MFCC+Δ + $\Delta\Delta$_2) [22]. As discussed earlier, the MFCC features include delta and acceleration coefficients that represent the dynamic characteristics of audio signals. As speech and singing are very different in terms of dynamic characteristics, inclusion of delta and acceleration coefficients in features adversely affects the DTW alignment. Upon removal of these coefficients from MFCC features, the AWBE is reduced to an extent, as shown in Table I.

TABLE I
Objective and subjective evaluation of temporal alignment using different features.

Features	AWBE (s)	MOS	BWS
MFCC+Δ + $\Delta\Delta$_1 [23]	0.268	1.64	-0.21
MFCC+Δ + $\Delta\Delta$_2 [22]	0.636	1.22	-0.75
MFCC	0.224	2.12	-0.18
LTC	0.148	3.16	0.08
Tandem features	**0.122**	**3.72**	**0.88**

We investigated the efficiency of LTC over MFCC in temporal alignment. It was observed that the LTC delivered a relative improvement of 33.93% in AWBE over MFCC, that confirms its efficacy in representing vocal tract characteristics and nullifying source characteristics of signals. The cumulative improvement in temporal alignment by the incorporation of consistent characteristics among singing and speaking styles (by sequential addition of features upon LTC to form the 45-dimensional tandem features (1VAD+12LTC+16STR+16LPC)) can be observed from Table I. The relative improvement in AWBE over 4,236 words in the database, provided by the tandem features over MFCC is a noticeable 45.54%. Also, the proposed tandem features outperform both baseline techniques, by delivering relative improvements of 54.48% and 80.82% over MFCC+Δ+$\Delta\Delta$_1 and MFCC+Δ + $\Delta\Delta$_2, respectively. This clearly demonstrates the superiority of the tandem features in overcoming the differences between speech and singing signals, thereby providing accurate temporal alignment.

In order to study the effects of accuracy of temporal alignment upon perception of synthesized singing from STS conversion, we implement a baseline STS system and conduct subjective tests. The voice conversion module in STS systems formed by STRAIGHT analysis-modification-synthesis [22],

[24] is fixed for different alignment modules, and hence, the difference in subjective scores can be attributed to the alignment module alone. 18 neutral listeners have participated in the subjective test, who listened to 5 sets of synthesized singing from STS conversion systems. Each set consists of 5 singing samples obtained using 5 temporal alignment modules with different features. The listeners were 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. The mean opinion scores (MOS) across all listeners over all sets of samples are reported in Table I. The proposed tandem features had produced perceptually superior synthesized singing to the baseline systems using MFCC features, as indicated in Table I.

In order to remove ambiguities in scoring of perceptually similar samples, the listeners were asked to choose the best and worst sounding singing samples from each set, after listening to them multiple times in different orders. We perform best-worst scoring (BWS) of samples as [31], [32]: $\text{BWS}_i = \frac{(B_i - W_i)}{N_i}$ where, B_i and W_i denote the number of times the item i is chosen as 'best' and 'worst', respectively by listeners. N_i represents the total number of appearances of item i in the set of trials [31]. A more positive BWS score points to a more perceptually appealing singing sample, and vice versa. And, the proposed tandem features had delivered the most pleasing singing samples to the listeners participated in this study, as is evident from Table I.

It is understood that temporal alignment has a direct impact on the quality of STS outputs. The proposed tandem features extracted from the consistent set of characteristics across speech and singing signals have proven to be beneficial in temporal alignment. Notice that these tandem features are formulated based on the observations from analysis on speech and singing signals and, they are not a simple concatenation of randomly chosen features. Also, use of these features had rendered accurate temporal alignment between two very different signals by means of signal analysis and processing steps, rather than complex statistical modeling tools.

V. Conclusions

In this paper, we presented a detailed study of source and spectral characteristics of speech and singing signals to identify a set of features, which are consistent across speech-singing voice styles to perform temporal alignment between them. We have identified the differences between speech and singing signals, that include variations in energy and F0 contour regulated by the target melody in singing. Also, the existence of singing formant in the high frequency singing spectra presents a major distinction between speech and singing. Motivated by the findings, we propose tandem features that describe the common traits between speech and singing signals. We have shown that the proposed features are effective in neglecting the vivid differences between speech and singing voice styles and hence, provide accurate temporal alignment between them as compared to other traditional features. The objective and subjective experiments reported in this paper reveal the superiority of the proposed tandem features over the MFCC features in aiding the quality of temporal alignment. As the accuracy of alignment is crucial for STS synthesis, the proposed features abet in generating good quality synthesized singing from STS.

VI. Acknowledgements

The authors would like to acknowledge the AcRF Tier 1 NUS Startup grant FY2016 for the project- Non-parametric approaches to voice morphing, supported by Ministry of Education, Singapore.

References

[1] J. Sundberg, "The level of the 'singing formant' and the source spectra of professional bass singers," *Quarterly Progress and Status Report: STL-QPSR*, vol. 11, no. 4, pp. 21–39, 1970.

[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] J. Sundberg, I. R. Titze, and R. Scherer, "Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source," *Journal of Voice*, vol. 7, no. 1, pp. 15 – 29, 1993.

[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 *International Conference on Digital Audio Effects*, Sept 2010, pp. 1–8.

[5] B. Lindblom and J. Sundberg, "The human voice in speech and singing," in *Springer Handbook of Acoustics*, Jan 2007, pp. 703–746.

[6] Y. E. Kim, *Singing Voice Analysis, Synthesis, and Modeling*. New York, NY: Springer New York, 2008, pp. 359–374.

[7] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *Proc. ICASSP*, March 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 *Interspeech*, Dresden, Germany, Sep 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, Oct 2011.

[10] P. Cano, A. Loscos, J. Bonada, M. D. Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications," in *In Proceedings of the ICMC*, 2000.

[11] 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*, vol. 24, no. 11, pp. 1998–2008, Nov 2016.

[12] C. H. Wong, W. M. Szeto, and K. H. Wong, "Automatic lyrics alignment for cantonese popular music," *Multimedia Syst.*, vol. 12, no. 4-5, pp. 307–323, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1007/s00530-006-0055-8

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

[14] D. Iskandar, Y. Wang, M.-Y. Kan, and H. Li, "Syllabic level automatic synchronization of music signals and text lyrics," in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM '06. ACM, 2006, pp. 659–662.

[15] 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*, Nov 2010, pp. 95–98.

[16] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, March 2007.

[17] K. Kobayashi and T. Toda, "Statistical singing voice conversion with direct waveform modification based on the spectrum differential," in *Interspeech*, 2014, pp. 2514–2518.

[18] 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)*, March 2016, pp. 5670–5674.

[19] 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 *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 215–218.

[20] ——, "Vocal conversion from speaking voice to singing voice using STRAIGHT," in *Interspeech*, 2007, pp. 4005–4006.

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

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

[23] ——, "Segmentation of speech signals in template-based speech to singing conversion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2011.

[24] 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)*, Dec 2017, pp. 1547–1555.

[25] S. W. Lee, S. T. Ang, M. Dong, and H. Li, "Generalized F0 modelling with absolute and relative pitch features for singing voice synthesis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 429–432.

[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[27] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 8, November 2008, pp. 1602–1613.

[28] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[29] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modication and synthesis framework," *Sadhana*, vol. 36, no. 5, pp. 713–727, Oct. 2011.

[30] 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 *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct 2013, pp. 1–9.

[31] T. Flynn and A. Marley, *Best-worst scaling: theory and methods*. Cheltenham, UK: Edward Elgar Publishing, Inc., 2014. [Online]. Available: //www.elgaronline.com/9781781003145.00014.xml

[32] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2017, pp. 677–684.