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

A Strategy for Improved Phone-Level Lyrics-to-Audio Alignment for Speech-to-Singing Synthesis

David Ayllón, Fernando Villavicencio, Pierre Lanchantin

ObEN Inc., CA, USA
david@oben.com

Abstract

Speech-to-Singing refers to techniques that transform speech to a singing voice. A major performance factor of this process relies on the precision to align the phonetic sequence of the input speech to the timing of the target singing. Unfortunately, the precision of existing techniques for phone-level lyrics-to-audio alignment has been found insufficient for this task. We propose a complete pipeline for automatic phone-level lyrics-to-audio alignment based on an HMM-based forced-aligner and singing acoustics normalization. The system obtains phone-level precision in the range of a few tens of milliseconds as we report in the objective evaluation. The subjective evaluation reveals that the smoothness of the singing voice generated with the proposed methodology was found close to the one obtained using manual alignments.

Index Terms: lyrics alignment, singing synthesis, text-to-speech, automatic speech recognition

1. Introduction

Current state-of-the-art of Text-to-Speech technology (TTS) producing highly natural speech brings the old challenge of singing voice generation with renewed enthusiasm. Although there exists valuable work on singing synthesis [1–6], including successful commercial products such as [7], the automatic generation of natural-like singing voice is still an open topic under study. An alternative for singing synthesis can be seen in Speech-To-Singing (STS) approaches in which a spoken version of the lyrics of a singing sequence is transformed into a sung version [8,9]. A transformation of this nature relies greatly in properly aligning the input speech to match the timing (e.g. duration) and pitch of the target singing. Lyrics-to-audio alignment deals with the automatic synchronization between music and lyrics with several consumer-oriented applications such as lyrics display functions in karaokes [10], content-based music information retrieval [11] and the generation of audio thumbnails [12]. In these applications, time synchronization at a lyrics line or phrase level is sufficient and automatic methods may provide a convenient solution. However, for STS purposes, precise phone-level alignment generally requires human verification [13]. It has been observed that timing errors in the order of a few tens of milliseconds may produce perceived artifacts in the generated singing voice.

A common approach for lyrics-to-audio alignment is the use of a Forced-Aligner (FA) based on Automatic Speech Recognition (ASR) techniques [14–16]. The Acoustic Model (AM) is typically trained on large speech datasets. However, ASR resources are not optimized for singing voices, which have higher variation of phoneme duration and contains in general larger vocal phenomena than speech (in terms of pitch range, pronunciation, voice quality, vibrato, etc.). The availability of large a capella singing corpora for AM training is limited, and the amount of available data is only suitable for AM adaptation [17]. The work in [18] proposes to transform speech training data into 'song-like' using time stretching and pitch shifting. Recently in [19] the authors proposed a novel method to obtain segments of singing voice aligned to their corresponding lyrics, using Google speech-to-text API and a lyrics matching algorithm. They used this data to adapt the AM for automatic transcription of sung lyrics. Many works on lyrics-to-audio alignment deals with polyphonic audios where the singing voice is mixed with musical instruments. In this case audios are preprocessed by a singing isolation algorithm prior to FA [20, 21]. Unfortunately, voice isolation algorithms may not remove all the instrumental content and usually modify the original voice spectrum, which reduce the performance of the ASR-based FA.

Alternative approaches rely in the use of musical structure information such as chords and chorus [22–24]. For instance, [12] proposes the integration of a chord-to-audio alignment method with an ASR-based system. However, such methods are more suitable for singing mixed with instrumental accompaniment. Others focus on the alignments of vowels. For instance, the work in [25] combines non-negative matrix factorization and canonical time warping to discover repetitive acoustic patterns of vowels. In [26] the authors proposed to train vowel likelihood models to identify vowels and then to align syllables. This approach can be useful for applications where word or line level alignment is sufficient but not for STS where accurate timing of both vowels and consonants is required.

Our primary goal is STS processing, and more in particular, the template-based schema presented in [27] in which an a capella performance of the target singing, denoted as "template", is used to guide the transformation [28]. Since we have access to clean a capella recordings as templates, the voice separation step is not required. However, phone-level timing precision is required to properly align the acoustic information between speech and singing sequences since misalignments larger than a few tens of milliseconds may be audible or result in artifacts at the singing sequence generated as output. Accordingly, we present in this paper a novel method to increase the precision of ASR-based lyrics alignment based on: a) bringing the main acoustic conditions of singing (e.g. pitch, vocal-tract) closer to those of speech, b) adjusting the duration model of a speech-based forced-aligner, c) training an AM with two silence models, and d) applying silence correction. As a result our alignment system outperforms scores of a baseline ASR system given the results of a study conducted on Chinese songs.

2. Proposed system

Figure 1 shows an overview of the proposed system for lyrics-to-audio alignment. It includes a pre-processing step in which the audio and the lyrics are prepared for alignment. The main processing block can be seen in the forced-aligner which is

http://dx.doi.org/10.21437/Interspeech.2019-3049

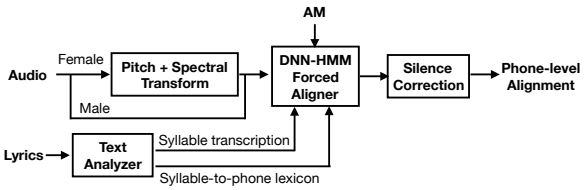

Figure 1: *System overview.*

based on an Deep Neural Network Hidden Markov Model (DNN-HMM) speech recognizer. Finally, a post-processing step is performed in which some phone boundaries are adjusted.

2.1. Pre-processing

2.1.1. Audio

One of the main aspects limiting the performance of acoustic models trained on speech for the lyrics-to-audio alignment can be seen in a significant variation of the spectral information given by the change of the vocal-tract and excitation conditions of singing voices. Studies as [29–31] denote that phenomena such as the singer formant and a progression of both source and filter conditions of the vocal system across singing pitch ranges have an impact in the featured spectra. Note that whereas average pitch values of male and female speech can be typically found within a range near 110 Hz and 210 Hz respectively, the corresponding values may jump to 250 Hz and 350 Hz and beyond given our informal observations on pop-music singing. We also observed that higher pitch singing was in general worse aligned resulting in notably lower performance scores on female singing than on male ones. We therefore apply a pre-processing step to perform a global normalization of source and filter conditions on female singing voice to bring both pitch and main spectral features closer to speech ranges. Pre-processing consists of decreasing the pitch one octave (i.e. one half of its value) and applying a fixed global frequency warping transformation to also adjust the perceived timbre to a male voice range [32]. Note that the frequency warping transformation was fixed for all female voices since a precise adjustment was not found to significantly impact the overall performance. Pitch estimations are performed using the Spectral Amplitude Autocorrelation (SAC) algorithm described in [33].

Finally, given the pitch range differences between speech and singing we also investigated the effect of the analysis window used for FA. Since three to four periods length is typically accepted as a sufficient criterion for spectral envelope estimation we adjusted the window size in the proposed system from 25 to 10 ms after having observed higher scores in our experiments for both male and female voices.

Figure 2 shows an example of the alignment of a female singing voice with the system proposed in this paper. The first phone sequence (REF) shows the ground truth boundaries (manually annotated), the second one (ORI) shows the boundaries obtained from the original audio, and the third one (PRE) shows the boundaries obtained from the transformed audio. Pitch estimation is presented on the lower panel. Whereas the alignment fails with the original audio, precision notably increases on the transformed audio.

2.1.2. Lyrics

Alignment is performed at the phone-level with optional silence insertions between syllables, which was found experimentally

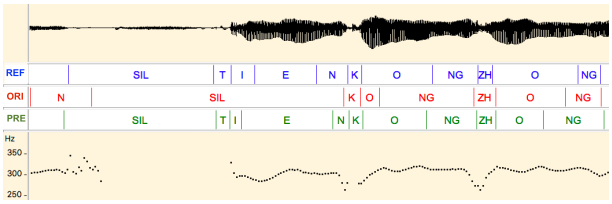

Figure 2: *Alignment of a female voice before (ORI) and after pre-processing (PRE) compared to the ground truth (REF).*

to be the best suited for singing voices. Thus, a syllable transcription of the input word sequence is obtained using a text analyzer. In addition, a syllable-to-phone lexicon is generated with the syllables contained in each particular song. The lexicon is used by the FA to map syllables to phones.

2.2. DNN-HMM Forced-aligner

The alignment is based on a DNN-HMM speech recognizer and is performed on the audio of the whole song. The Viterbi algorithm is used to align the acoustic features to the text. The inputs are the pre-processed audio, the pre-processed lyrics, and a pre-trained AM. The proposed FA includes AM speaker adaptation. In addition, the standard HMM topology is modified to improve vowel alignment.

2.2.1. Acoustic model

A tri-phone DNN-HMM AM is trained on top of a GMM-HMM system using Mel-Frequency Cepstral Coefficients (MFCCs) and delta features. The dimension of the feature vector is reduced to 40 using Linear Discriminant Analysis (LDA), and then Maximum Likelihood Linear Transform (MLLT) is applied for each speaker. Speaker Adaptive Training (SAT) [34] is performed. The DNN model is trained on top of the SAT model with the same set of training data. The DNN configuration and training algorithm are described in [35]. To train the AM we used two types of speech corpora, an ASR corpus and a TTS corpus. The main difference is that ASR corpora are designed to train robust ASR systems and contain noisy speech data, while TTS corpora are designed to train TTS systems and hence contain clean speech.

Alignment is performed on the whole song and no Voice Activity Detection (VAD) is used to detect speech segments. Thus, songs may contain long silences, for instance, at the beginning of a song, or between a chorus and the verse, which are longer that optional silences found between speech units. To improve the alignment of those long silences, we train a different silence model considering silences longer than 0.5 s. Long silences are explicitly annotated in the training corpus using a VAD to detect long silences at the beginning and at the end of each training utterance. During lyrics alignment, the position of long silences are manually annotated into the lyrics, at the beginning, end, and between choruses.

2.2.2. Modified HMM transition probabilities to increment vowel duration

The left-right HMM topology for the phones contains three emitting states. The transition probabilities control the way the hidden state at time t is chosen given the hidden state at time $t-1$. Each state has two types of transition probabilities: the *self-loop* transition probability associated with self-state transition (P_{SL}) and the *left-to-right* transition probabilities associ-

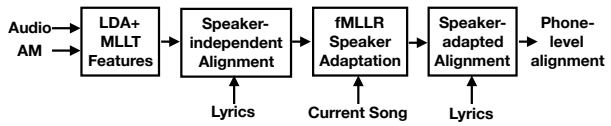

Figure 3: *DNN-HMM forced aligner.*

ated with transitions to the next states (P_{LR}). We can apply different scales to the self-loop and left-to-right probabilities. Let us define p and q as the initial self-loop and left-to-right transitions probabilities, and a and b their respective scales. Since our HMM topology has a single output transition, $p + q = 1$ and $b = 1$. According to this, the transition log-probabilities can be expressed as:

$$P_{SL} = a * \log(p)$$
$$P_{LR} = a * \log(1-p) + b * \log(q) = (a+1) * \log(1-p)$$

Considering that self-loop probabilities are higher than left-to-right probabilities, the increment of the self-loop scale a increments the probability associated with a transition to the same state, which forces the phones to be longer. In our system we propose to apply a different self-loop scale to vowels and consonants to allow the FA to produce longer vowels than the ones learnt during training from speech data, thus improving the alignment of singing vowels.

2.2.3. Alignment steps

During alignment, a 2-pass decoding strategy is used, involving Feature-space Maximum Likelihood Linear Regression (fMLLR) speaker adaptation [36]. Figure 3 shows the alignment procedure, which is performed in four steps. First, the LDA+MLLT acoustic features are extracted from input audio. Second, speaker-independent forced-alignment is performed. Then the AM is adapted to the current speaker using the obtained phonetic alignment, and finally speaker-adapted forced alignment is performed. Forced-alignments are based on the Viterbi algorithm.

2.3. Post-processing

After analyzing the output of the FA, we found some systematic errors that can be corrected in a post-processing step. We focus on errors that affect the quality of the generated singing voice the most. One of the most problematic error pattern is when optional silences span over voiced phones, bringing a perceived degradation at the beginning and ending parts. Voiced segments are characterized by having perceptible pitch values, whereas silence segments should have pitch values close to zero. We propose to use a pitch estimator to correct the boundaries of optional silences. When pitched frames are found inside a silence and next to any of its boundaries, they are appended to the adjacent phone in case that it is a voiced phone. This adjustment does not introduce important improvements in alignment metrics but it does improve the quality of the generated singing voice, according to informal listening tests that we conducted.

3. Experimental work

To our knowledge, only few studies of lyrics-to-audio alignment have been conducted on Chinese songs [37, 38]. Our evaluation singing voice corpus is composed of 16 mandarin pop songs, 9 male voices and 7 female voices, with a total of 67 minutes of audio. The mean pitch is 260 Hz for male voices and 330 Hz for female voices. The average phone duration is 0.269 s for vowels, 0.117 s for consonants, and 0.29 s for silences. The singing voices and instrumental accompaniments are in separated audio tracks. The original lyrics are in Chinese characters. During the pre-processing step, the lyrics are converted into pinyin, assuming that pinyin symbols are equivalent to syllables for optional silence insertion. Audio files are downsampled to 16 kHz for the experiments. We use the Kaldi toolkit [39] to build our ASR system. Two different AMs are trained. The first one is trained with the GALE corpus [40] which is an ASR corpus comprised of approximately 134 hours of Mandarin Chinese broadcast news speech. The second AM is trained on the RASC863 corpus [41], which is a multi-speaker TTS corpus composed of 25 hours of spontaneous and read speech. The silence probability is set to 0.2, the self-loop scale for consonants is 0.1 and for vowels 0.5. These optimal values were found experimentally.

Table 1: *Forced-Alignment F-score*

	Male			**Female**			**M+F**
Model	All	Vow	Cons	All	Vow	Cons	All
Baseline	0.56	0.54	0.58	0.53	0.52	0.53	0.54
+RASC	0.82	0.80	0.84	0.74	0.73	0.75	0.78
+FPre	0.82	0.80	0.84	0.79	0.77	0.81	0.81
+AnWin	0.85	0.83	0.87	0.83	0.82	0.84	0.84
+DM	**0.87**	0.85	0.88	**0.84**	0.83	0.85	**0.86**

3.1. Objective evaluation

In order to evaluate the quality of the automatic alignments, the 16 songs in the corpus have been manually annotated by professional native linguists. The metric used to evaluate the quality of the alignments is the F-score, which reflects the number of phones matching inside a match window of a determined length. With a default match window of 100 ms, a phone matches when the addition of the onset and offset mismatch with respect to the manual annotations is lower than 100 ms. In addition, we include the metrics used in the Automatic Lyrics-to-Audio Alignment task of MIREX 2018 challenge (Mandarin pop dataset) [42]: Average Absolute Error (ASE), Percentage of Correct Segments (PCS), and Percentage of Correct Estimates according to a Tolerance Window of 0.3 s (PCETW).

Table 1 shows the mean F-score calculated with a 100 ms match window obtained with different versions of the FA system. Baseline is a system which consists of an AM trained with a single silence model on GALE (ASR) corpus, no preprocessing of female audios, 25 ms analysis window, and similar self-loop scale applied to all phones. +RASC is the previous system with an AM trained with 2 silence models on RASC (TTS) corpus. +FPre incorporates female voice pre-processing to the previous system. +AnWin further reduces the analysis window to 10 ms. Finally, +DM includes a different self-loop scale for vowels and consonants to modify the duration model. F-scores are shown for all phones (All), vowels (Vow) and consonants (Cons), and are further grouped into male and female songs. Using the RASC corpus instead of GALE for AM train-

Table 2: *Phone and pinyin level metrics*

metric	phone-level	pinyin-level
Mean Fscore (100 ms)	0.86	0.83
Mean ASE	0.19	0.14
Mean PCS	0.77	0.84
Mean PCETW (300 ms)	0.94	0.96

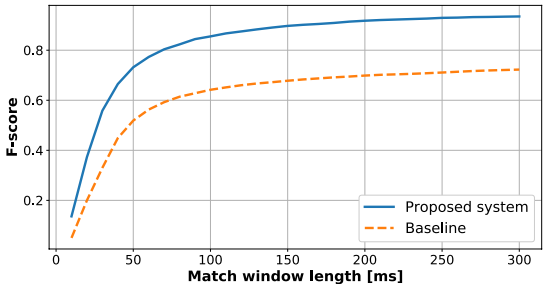

Figure 4: *Fscore vs match bound.*

ing improves drastically the performance, with a 44% of relative improvement. Regarding female voices, the F-score obtained by the proposed transformation together with the use of a smaller analysis window is notably higher than the one obtained with the original audios: the F-score of female songs is increased from 0.74 to 0.83 (12% of relative improvement). The variation of the duration model (self-loop scale) introduces little improvement in the mean F-score, but it improves the detection of important phones transitions.

In addition, Table 2 includes the best F-score and the three MIREX metrics for comparison. Results are calculated for the alignment of both phone and pinyin boundaries (the latter proposed in MIREX). We want to highlight that we did not have access to the MIREX 2018 evaluation dataset in order to perform a direct comparison to our method. Nevertheless, metrics obtained with our corpus for pinyin-level are slightly better than the ones reported by the CW1 submission of MIREX 2018 (mean PCS = 0.82, mean PCETW = 0.94). Authors are aware that this is not a fair comparison but it is the best option to place the system within the state-of-the-art.

Finally, Figure 4 shows the average F-score as a function of the match window length (from 10 to 300 ms), for the full proposed system (+DM) and the +RASC system. For 100 ms, the average F-score obtained by the proposed system is 0.86, as previously shown in Table 1. For larger windows, the F-score seems to reach a maximum value of 0.95. For 50 ms window length, the F-score is 0.74, which is still a good value for such a small tolerance. For values lower than 50 ms, the F-score drops notably. In any case, the values obtained by the proposed system are higher than the ones obtained by the baseline system.

3.2. Subjective evaluation

The different alignments of a capella recordings were used to generate singing voice following the baseline STS schema of [27]. Briefly, given the lyrics and pitch contour information, the acoustic features generated by a Chinese TTS were time-aligned to replicate singing with a different voice. The singing streams generated using the manual alignment followed closely the resynthesized version of the a capella recordings whereas the majority of perceived and observed issues (e.g. artifacts, missing content) were found when the baseline FA was used. An example is shown in Figure 5 where the baseline FA based output (third row) shows significant losses in comparison to the proposed FA system (last row). Note that it was observed that in the few cases were the proposed FA performed worse it generally resulted in smaller artifacts.

To assert the perceptual effect of issues as the one previously described and shown in Figure 5, we carried out a simple listening test on extracts showing visible differences between the baseline and proposed FA. The three STS outputs

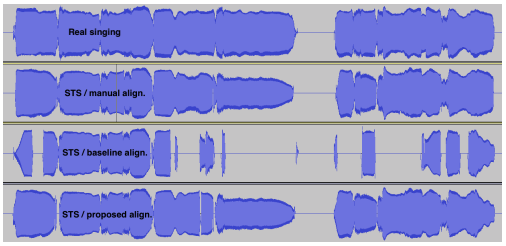

Figure 5: *Real singing resynthesis and STS generated waveforms (shown in dB) using manual and both baseline and proposed FA based alignment (from top to bottom).*

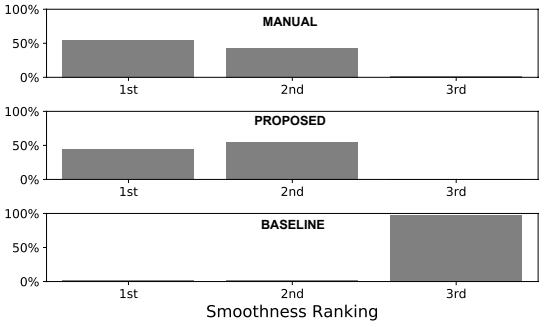

Figure 6: *Ranking of the perceived smoothness of the STS generated waveforms for the different alignments.*

generated from the different alignments (manual, baseline, proposed) from sixteen extracts of a female voice song and a male voice song (eight per each) were presented to 16 native listeners who ranked them according to the perceived smoothness of the singing stream (256 evaluation points in total per alignment case). Results are shown in Figure 6 in terms of the percentage of times each system was ranked at each position. Clearly, the baseline FA was found systematically to be the worst, whereas the resulting confusion in the preference between the manual and proposed FA suggests no significant difference between both alignments. Despite a few cases in which the proposed system performed the worst, in an informal observation among different singing content these trends did not change overall.

4. Conclusions

In this paper we proposed a methodology based on pre/post-processing a customized DNN-HMM forced-aligner to automatically align lyrics and singing for STS purposes. Unlike other lyrics-to-audio alignment applications STS requires accurate phone-level precision. From the experiments carried out the following conclusions can be drawn. First, an acoustic adjustment (pitch and spectral features) towards the speech range and the adaptation of the analysis window introduced important improvements. Second, the consideration and modeling of different types of silence led to better detection of their boundaries. Third, the use of a TTS corpus for AM training drastically increased the alignment performance. Fourth, a modification of the HMM transition probabilities alleviated the lower performance observed in the alignment of vowels due to duration differences between speech and singing. Finally, although the MIREX dataset could not be considered in our evaluation, our performance was found to be comparable to state-of-the-art for word-level lyrics-to-audio alignment.

As future work the authors are considering the generalization of the results on English, Korean, and Japanese as well as the use of actual singing voices for AM training.

5. References

[1] P. Depalle, G. Garcia, and X. Rodet, "A Virtual Castrato," in *Proc. ICMC*, 1994, pp. 357–360.

[2] P. R. Cook, "Singing Voice Synthesis: History, Current Work, and Future Directions," *Computer Music Journal*, vol. 20, no. 3, pp. 38–46, 1996.

[3] J. Bonada and X. Serra, "Synthesis of the Singing Voice by Performance Sampling and Spectral Models," *IEEE Signal Process. Magazine*, vol. 24, no. 2, pp. 67–79, 2007.

[4] K. Saino, H. Zen, Y. Nankaku, A. Lee, and K. Tokuda, "An HMM-based Singing Voice Synthesis System," in *Proc. ICSLP*, 2006.

[5] T. Nose, M. Kanemoto, T. Koriyama, and T. Kobayashi, "HMM-based Expressive Singing Voice Synthesis with Singing Style Control and Robust Pitch Modeling," *Computer Speech & Language*, vol. 34, no. 1, pp. 308–322, 2015.

[6] M. Blaauw and J. Bonada, "A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs," *Applied Sciences*, vol. 7, no. 12, p. 1313, 2017.

[7] H. Kenmochi and H. Ohshita, "Vocaloid-Commercial Singing Synthesizer Based on Sample Concatenation," in *Proc. ISCA*, 2007.

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

[9] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," in *Proc. ICME*, 2010, pp. 1421–1426.

[10] S. W. Won and J. Scott, "Word Level Lyrics-Audio Synchronization using Separated Vocals," in *Proc. ICASSP*, 2017, pp. 646–650.

[11] A. M. Kruspe, "Retrieval of Textual Song Lyrics from Sung Inputs," in *Proc. Interspeech*, 2016, pp. 2140–2144.

[12] M. Mauch, H. Fujihara, and M. Goto, "Integrating Additional Chord Information into HMM-based Lyrics-to-Audio Alignment," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 20, no. 1, pp. 200–210, 2012.

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

[14] A. Mesaros, "Singing voice recognition for music information retrieval," *Tampereen teknillinen yliopisto. Julkaisu-Tampere University of Technology. Publication; 1064*, 2012.

[15] 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 Process.*, vol. 5, no. 6, pp. 1252–1261, 2011.

[16] A. M. Kruspe and M. Goto, "Retrieval of Song Lyrics from Sung Queries," in *Proc. ICASSP*, 2018.

[17] A. Mesaros and T. Virtanen, "Adaptation of a Speech Recognizer for Singing Voice," in *Proc. 17th EUSIPCO*, 2009.

[18] A. M. Kruspe and I. Fraunhofer, "Training phoneme models for singing with" songified" speech data." in *Proc. ISMIR*, 2015, pp. 336–342.

[19] C. Gupta, R. Tong, H. Li, and Y. Wang, "Semi-supervised lyrics and solo-singing alignment," in *Proc. ISMIR*, 2018, pp. 600–607.

[20] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *Proc. ISMIR*, 2017.

[21] K. W. E. Lin, B. Balamurali, E. Koh, S. Lui, and D. Herremans, "Singing voice separation using a deep convolutional neural network trained by ideal binary mask and cross entropy," *Neural Computing and Applications*, pp. 1–14, 2018.

[22] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "Lyrically: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 16, no. 2, pp. 338–349, 2008.

[23] M. Mauch, H. Fujihara, and M. Goto, "Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations," in *Proc. SMC*, vol. 27, 2010, pp. 28–38.

[24] K. Lee and M. Cremer, "Segmentation-based lyrics-audio alignment using dynamic programming." in *Proc. ISMIR*, 2008, pp. 395–400.

[25] S. Chang and K. Lee, "Lyrics-to-Audio Alignment by Unsupervised Discovery of Repetitive Patterns in Vowel Acoustics," *IEEE Access*, vol. 5, pp. 16 635–16 648, 2017.

[26] 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 Trans. on Audio, Speech, and Language Process.*, vol. 24, no. 11, pp. 1998–2008, 2016.

[27] K. Kaewtip, F. Villavicencio, F. Kuo, M. Harvilla, and I. Ouyang, "Enhanced virtual singers generation by incorporating singing dynamics to personalized text-to-speech-to-singing," in *Proc. ICASSP*, 2019.

[28] L. Cen, M. Dong, and P. Chan, "Template-based Personalized Singing Voice Synthesis," in *Proc. ICASSP*, 2012, pp. 4509–4512.

[29] J. Sundberg and T. D. Rossing, "The Science of Singing Voice," *The Journal of the Acoustical Society of America*, vol. 87, pp. 462–463, 1998.

[30] I. R. Titze and B. H. Story, "Acoustic Interactions of the Voice Source with the Lower Vocal Tract," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2234–2243, 1997.

[31] J. Wolfe, M. Garnier, and J. Smith, "Vocal Tract Resonances in Speech, Singing, and Playing Musical Instruments," *HFSP journal*, vol. 3, no. 1, pp. 6–23, 2009.

[32] F. Villavicencio, J. Yamagishi, J. Bonada, and F. Espic, "Applying spectral normalisation and efficient envelope estimation and statistical transformation for the voice conversion challenge 2016." in *Proc. Interspeech*, 2016, pp. 1657–1661.

[33] F. Villavicencio, J. Bonada, J. Yamagish, and M. Pucher, "Efficient pitch estimation on natural opera-singing by a spectral correlation based strategy," IPSJ Technical Report, Tech. Rep., 2015.

[34] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker Adaptive Training," in *Proc. ICSLP*, 1996.

[35] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. ICASSP*. IEEE, 2014, pp. 215–219.

[36] M. J. F. Gales and P. C. Woodland, "Mean and Variance Adaptation Within the MLLR Framework," *Computer Speech and Language*, vol. 10, 1996.

[37] C. G. 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.

[38] G. Dzhambazov, Y. Yang, R. Repetto, and X. Serra, "Automatic Alignment of Long Syllables in a Cappella Beijing Opera," in *Proc. FMA*, 2016.

[39] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.

[40] C. Plahl, B. Hoffmeister, M.-Y. Hwang, D. Lu, G. Heigold, J. Loof, R. Schlüter, and H. Ney, "Recent Improvements of the RWTH GALE Mandarin LVCSR System," in *Proc. Interspeech*, 2008.

[41] A.-J. Li and Z.-G. Yin, "Standardization of speech corpus," *Data Science Journal*, no. 6, pp. S806–S812, 2007.

[42] D. G. Bogomilov, "Knowledge-based probabilistic modeling for tracking lyrics in music audio signals," *Universitat Pompeu Fabra*, 2017.