AUTOMATIC LYRICS-TO-AUDIO ALIGNMENT ON POLYPHONIC MUSIC USING SINGING-ADAPTED ACOUSTIC MODELS

Bidisha Sharma[*1] *Chitralekha Gupta*[*1,2] *Haizhou Li*[3] *Ye Wang*[1,2]

[1] School of Computing, [2] NUS Graduate School for Integrative Sciences and Engineering,
[3] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
s.bidisha@nus.edu.sg, chitralekha@u.nus.edu,{haizhou.li, dcswangy}@nus.edu.sg

ABSTRACT

Lyrics-to-audio alignment is to automatically align the lyrical words with the mixed singing audio (singing voice+musical accompaniment). Such alignment can be achieved with an automatic speech recognition (ASR) system. We propose to adapt the acoustic model of a speech recognizer towards solo singing voice. This avoids the hurdles of annotating a large polyphonic music training dataset. Moreover, a lexicon-modification based duration modelling has been incorporated to account for the long duration vowels in singing. As practical application demand the alignment on polyphonic music, we study the effect of different singing vocal separation methods in the task of lyrics-to-audio alignment in polyphonic music. The extracted vocals are forced-aligned with the singing-adapted models. We demonstrate that the use of audio source separation method and effective end-pointing of the songs has a high impact on the alignment performance through the experiments. We report a mean average absolute error of 3.87 seconds, which is comparable with the state-of-the-art lyrics-to-audio alignment system that is trained on a large polyphonic music database.

Index Terms— Lyrics-to-audio alignment, polyphonic music, ASR, audio source separation

1. INTRODUCTION

Automatic lyrics-to-audio alignment has various applications such as the automatic generation of karaoke scores, song-browsing by lyrics, and the generation of audio thumbnails. Given audio signal of singing voice and corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as a problem of estimating the temporal relationship between them.

Many studies on lyrics-to-audio alignment exploit the knowledge of the musical structure of the song to align the lyrics [1, 2]. One of the earliest studies [2] (LyricAlly) uses the structural information of popular songs to align the chorus and the verse sections of lyrics to the music audio. Mauch et al. [3] incorporated time-aligned chord information along with the lyrics to improve the alignment. The limitation of these methods is that the music structure may vary with genre, and they need manually transcribed chord labels with reliable temporal information corresponding to the lyrics.

In ASR, word or phone level segmentation is obtained by forced-aligning the transcription to the speech using acoustic models trained with speech data. The same idea has been applied to align lyrics to music audio [3–7]. In [4], singing vocal is separated from polyphonic music, and maximum likelihood linear regression (MLLR) is used for adapting the speech phone models to the singing vocal. These adapted phone models achieved a low word alignment accuracy of 46.4%. Mesaros et al. [7] used 49 fragments of songs, 20-30 seconds long, along with their manually acquired transcriptions to adapt Gaussian mixture model (GMM)-hidden Markov model (HMM) speech models for singing. Using these singing-adapted speech models, they reported a phoneme error rate of 80%. These works provide a direction for solving the problem of lyrics alignment in music, but they suffer from manual post-processing and the models are based on a small number of annotated singing samples.

A major problem in building a lyrics alignment system is the lack of availability of lyrics-annotated dataset. In [8], Gupta et al. designed an algorithm to automatically obtain lyrics annotations for solo-singing data by leveraging on speech models to force-align $\sim$50 hours of solo-singing audio from a karaoke singing dataset [9] with the lyrics. They iteratively adapted the speech models to singing voice, while automatically refining the training data by removing the bad quality audio and lyrics, and improving the lyrics annotations for the songs. These models showed 36% word error rate (WER) in a free-decoding experiment on solo-singing [8]. Kruspe [10] and Dzhambazov [11] also attempted the alignment task in MIREX 2017, but did not account for the bad audio recordings and refinement of the training data. However, all these models are not expected to perform well in polyphonic audio.

Recently, in MIREX 2018 [12], the systems submitted by Wang [13] achieved a mean average absolute error (ASE) of 2.7 seconds for Hansen's polyphonic music dataset [14] and 4.12 seconds for Mauch's polyphonic dataset [3]. They used 7,300 annotated English songs (more than 300 hours) from KKBOX Inc.'s music library to train HMM based models. In pre-processing, they segmented the audio files according to the position of blank lines in lyrics and performed vocal detection. Despite of the good performance, they used a large amount of annotated polyphonic data to train the models. Such copyrighted large polyphonic music audio dataset is not available to the research community. Moreover, obtaining human annotations of polyphonic music is a tedious task and extending such a system to an under-resource language will be challenging.

In this work, we propose to use singing-adapted speech acoustic models trained on a relatively small solo-singing dataset in conjunction with audio source separation to obtain word-level lyrics alignment boundaries for polyphonic audio. The acoustic models are trained to handle the differences between speech and singing, such as long duration of vowels, and the pitch dynamics. Furthermore, we incorporate audio source separation as a pre-processing step to extract the singing vocals, and conduct a comparative study of the effect of different audio source separation methods on the performance of our lyrics-to-audio alignment system. We also study the impact of reliable vocal detection on lyrics alignment.

[*]The first two authors contributed equally

The rest of the paper is organized as follows: we describe the framework of the proposed system in Section 2. The experiments are presented in detail in Section 3. We summarize our results in Section 4.

2. FRAMEWORK FOR LYRICS-TO-AUDIO ALIGNMENT

In this work, we would like to build a framework to automatically align lyrics to the polyphonic music audio, as shown in Figure 1. The idea is to use trained solo-singing-adapted speech acoustic models to force-align lyrics with the corresponding music audio. But there is a training and test data mismatch. We can bridge this gap in two ways: (a) by improving the acoustic models further by training on polyphonic data, and (b) by making the test data closer to the trained solo-singing acoustic models. As a large polyphonic music dataset with aligned lyrics is not publicly available for training, in this work we consider the latter. Our framework consists of three main components: singing vocal separation, vocal begin- and end-point detection, and singing-adapted speech acoustic models.

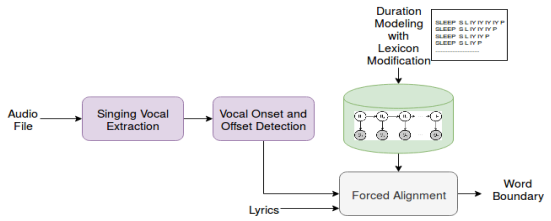

Fig. 1: Framework for automatic lyrics-to-audio alignment.

2.1. Singing-adapted acoustic models

To reduce the mismatch between singing and speech, speech acoustic models are adapted to singing voice using speaker adaptation methods [7, 8]. In [8], the authors applied a semi-supervised speaker adaptive training (SAT) method, with lyrics-aligned solo-singing dataset to adapt speech models to singing voice. They use feature-space maximum likelihood linear regression (fMLLR) for the adaptation of the speech models. One major difference between speech and singing voice is the duration of vowels. The vowels in singing could be longer in duration than spoken vowels. Therefore they introduced pronunciation variants in the lexicon to model the longer duration of vowels in singing. This modification reduced the WER from 36% to 29.65% in solo-singing data [15]. However, these singing-adapted models were not evaluated for lyrics-to-audio alignment, introduction of these modified models will contribute to achieve good performance. However, these models may not be well-suited for polyphonic music because the presence of background music introduces noisy components that results in mismatch between these solo-singing trained models and the test data.

2.2. Singing vocal separation

To overcome the differences between the trained models and the test data, we incorporate a source separation module to extract the singing vocals from polyphonic songs. We study the effect of three different audio source separation methods on our lyrics alignment algorithm: harmonic/percussive, convolutional neural network (CNN) based, and U-Net based. Percussion component in the background accompaniment introduces vertical lines in the spectrogram, which makes it noisy. Therefore, we first attempted to remove these using the traditional harmonic/percussive method [16], which is reported to be simple and effective. This method uses median filters individually in the horizontal and vertical directions to separate the harmonic

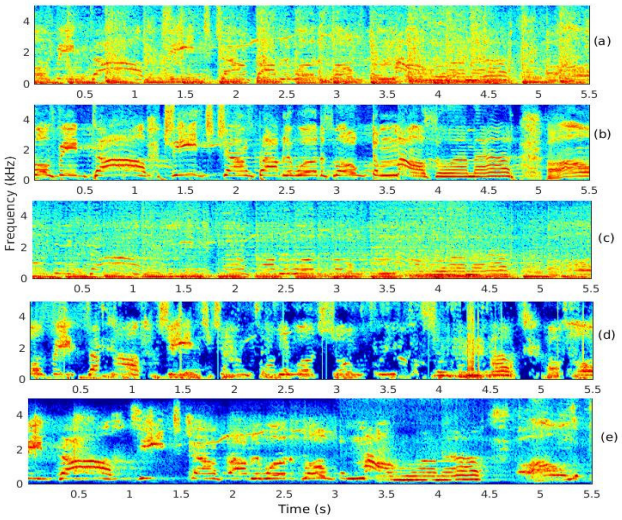

Fig. 2: Comparison of spectrograms for different audio source separation methods for "*this afternoon*" song from Hansen's dataset, (a) original mixed audio, (b) original clean audio, extracted vocal using (c) harmonic/percussive, (d) CNN based, (e) U-Net based audio source separation method.

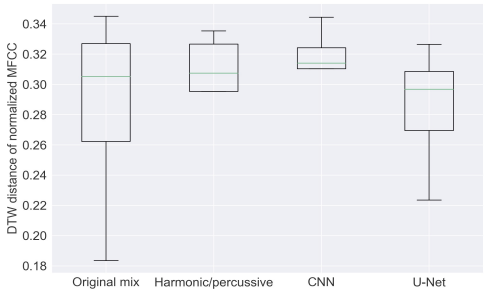

Fig. 3: Boxplot of the distribution of DTW distances between normalized MFCCs extracted from solo singing audio and corresponding original mixed audio, extracted vocals using various audio source separation.

and the percussive events. This separation method is integrated in the widely used audio and music analysis library Librosa [17].

As an alternative to the traditional methods and owing to the advantages and success of CNN, we are also interested to apply the audio source separation method proposed in [18], which achieves the same performance as that of multilayer perceptron based audio source separation with less time complexity and compact representation. In this case, we use the model trained on iKala dataset, for voice, bass, and drums separation. Although CNN based audio source separation is undoubtedly successful, it would be interesting for us to investigate another method recently proposed by Jansson et al. [19], that uses U-Net architecture (initially developed for medical imaging). The architecture builds upon the fully convolutional network [20] with symmetric down-sampling and up-sampling path. It has the capacity for recreating the fine, low-level detail required for high-quality audio reproduction. We have used the pre-trained models corresponding to iKala dataset and the implementation available in [21], which uses Chainer framework.

In Figure 2, we show a comparison of the vocals obtained from the three audio source separation methods. Figure 2(a) and (b) show the spectrograms (with 20 ms frame-size, 10 ms frame-shift, sampling rate 10 kHz) corresponding to original mixed audio and

solo-singing audio for a 5.5s segment of the song "*this afternoon*" from Hansen's dataset [14]. The extracted vocals for the same audio segment using harmonic/percussive, CNN and U-Net based audio source separation are shown in Figure 2(c),(d),(e) respectively. If we compare each of these with Figure 2(b) we can observe that, using harmonic/percussive method, the percussive component is removed, however the other components are preserved in the spectrogram, similar to the original mixed audio shown in Figure 2(a). After applying CNN based source separation, although the vocal specific characteristics are preserved in the spectrogram, as shown in Figure 2(c), there are some glitches present in the boundaries of the phonemes. This distortion is also evident for the songs with high intensity background accompaniment during informal listening. Figure 2(e) shows that the extracted vocal from U-Net based source separation has highest similarity with that of clean speech and is least distorted.

To further analyze the deviation of the extracted vocals from clean solo-singing audio, we perform DTW between the normalized Mel-frequency cepstral coefficients (MFFCs) (13-dimensional) of the solo-singing, and the extracted vocals from the three source separation methods shown in Figure 3. It can be observed that the mean distance with solo-singing audio is significantly smaller in case of U-Net based source separation method, with a reasonable standard deviation. This gives us the intuition that the U-Net method will reduce the train and test data deviation the most.

Table 1: Mean average absolute error/deviation (ASE) and percentage of correct segments (PCS) for lyrics-to-audio alignment systems using GMM-HMM (SAT) model after applying different audio source separation methods.

Database	Source separation method							
	No source separation		Harmonic/percussive		CNN		U-Net	
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's [14]	33.81	0.12	24.56	0.08	10.99	0.23	**9.48**	0.36
Mauch's [3]	26.94	0.13	24.83	0.04	12.28	0.13	**6.43**	0.25

2.3. Intro and outro non-vocal suppression

We would like to avoid misalignments due to the presence of long musical segments in the intro and the outro of a song. In an ideal case, the non-vocal segments in these sections should appear as silence in the extracted sung vocals. However, due to error in singing vocal separation method, the instrumental accompaniments are suppressed only to some extent. Therefore, to detect these suppressed non-vocal segments, we divided the spectrum of each frame (framesize 25 ms, frameshift 5 ms, sampling frequency 44.1 kHz) into four equal sub-bands. The energy corresponding to the 2^{nd} sub-band shows a prominent difference between the segments with vocals and without vocals. A threshold based on the average 2^{nd} sub-band energy is set to classify the frames into vocal and non-vocal. The non-vocal segments with very long duration corresponding to intro and outro are removed from the audio as a pre-processing step.Based on the above discussion, we propose the framework for lyrics-to-audio alignment, as shown in Figure 1.

3. EXPERIMENTAL EVALUATION

We develop the lyrics-to-audio alignment framework at different stages to observe the efficiency and significance of each component. We use Hansen's [14] and Mauch's datasets [3] for evaluation of the alignment systems. Hansen's dataset contains 9 pop music songs in English with annotations of both begin- and end-timestamps of each word [14]. The audio has two versions: the original with instrumental accompaniment and a *capella* singing voice only. Mauch's dataset contains 20 pop music songs in English with annotations of begin-timestamps of each word [3]. We use two different metrics for the evaluation, which are ASE and percentage of correct segments from total audio duration (PCS) [22].

As discussed in Section 2.1, we use singing-adapted speech acoustic models (SAT) trained on solo-singing dataset [9] to force-align lyrics with the audio. The baseline speech acoustic model is a tri-phone GMM-HMM trained on Librispeech corpus [23] using MFCC features on Kaldi toolkit [24]. To make the Viterbi alignment algorithm operate over the long duration of songs (4-5 minutes), we set the alignment retry-beamwidth to a high value of 4000. Also the flag for optional silence was on to handle the possibility of pauses. To avoid misalignment due to the presence of long duration silence, we apply an energy-based algorithm for intro and outro non-vocal suppression as mentioned in Section 2.3.

3.1. Effect of vocal separation and non-vocal suppression

We first tested the system with solo-singing versions of the songs from Hansen's data, which gives average ASE of 0.4 second. Using the polyphonic version of the same dataset, we obtain an average ASE of 33.81 seconds as shown in Table 1. This shows that our singing-adapted acoustic models are well-suited for solo-singing data. However, as expected, they do not perform as well on polyphonic music. Therefore, we use different source separation methods to extract the singing vocal for which the performance of the system is depicted in Table 1. We observe that the average ASE values are best for the system with U-Net, which is 9.48 seconds for Hansen's data and 6.43 seconds for Mauch's data. The harmonic/percussive method gives a relatively poor performance, which implies that the presence of other non-percussive instruments has an impact on the performance of singing voice models.

In order to reduce the alignment error further, we performed intro and outro non-vocal suppression as described in Section 2.3. The average ASE and PCS values for the systems with different audio source separation methods and after applying silence removal are depicted in Table 2. The U-Net source separation performs the best. We achieve ASE of 1.39 and 6.34 secs for Hansen's and Mauch's data respectively. This is a significant improvement over MIREX 2017 [25] best system. We also note that our system performance is comparable to MIREX 2018 [12] best system which are 2.07 and 4.13 secs respectively for the two datasets. Moreover, the performance of all the systems have significantly improved after removal of beginning and ending silences. To summarize, the average ASE value obtained from our best system is 3.87 secs.

It can be also noted the PCS values shown in Table 2 are not consistent with ASE values. The PCS measure [22], which captures the percentage overlap of the aligned segments with ground-truth, is lower for U-Net compared to CNN-based source separation method. Our further investigation shows that in case of CNN, the incorrect boundaries deviate from the ground-truth by a large amount, which results in higher ASE values. However, the correct boundaries are aligned accurately, resulting in high value of PCS. On the other hand, in case of U-Net, only a few instances of the resultant boundaries show a large deviation, but there is a small alignment error in most of the boundaries. This leads to lower PCS values for U-Net. A demo of the presented results is given in `https://www.comp.nus.edu.sg/~chitrale/LyricsAlignmentDemo.html`.

3.2. Result Analysis

To show the results obtained for all the songs, we plotted distributions in terms of boxplot shown in Figure 4 corresponding to all the songs for each system. We observe that the performance of

Table 2: Mean average absolute error/deviation (ASE) and percentage of correct segments (PCS) for lyrics-to-audio alignment systems using GMM-HMM (SAT) model after applying different audio source separation methods and removal of beginning and ending silences.

Database	Source separation method								MIREX 2017 best system		KKBOX System (MIREX 2018 best system)	
	No source separation		Harmonic/percussive		CNN		U-Net					
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's	18.78	0.19	20.69	0.09	3.14	0.29	**1.39**	0.13	**7.34**	0.25	**2.07**	0.45
Mauch's	25.42	0.08	25.83	0.03	17.74	0.06	**6.34**	0.07	**9.03**	0.15	**4.13**	0.35

the system using U-Net source separation has better performance compared to best system of MIREX 2017 and comparable performance with MIREX 2018 best system. In this case, the system with CNN based source separation method also achieves a comparable result. For Mauch's dataset similar distributions are shown in Figure 4 where the system with CNN based source separation performs poorer than U-Net source separation. In this case also it is evident that the proposed lyrics-to-audio alignment system has comparable performance with the best system of MIREX 2018.

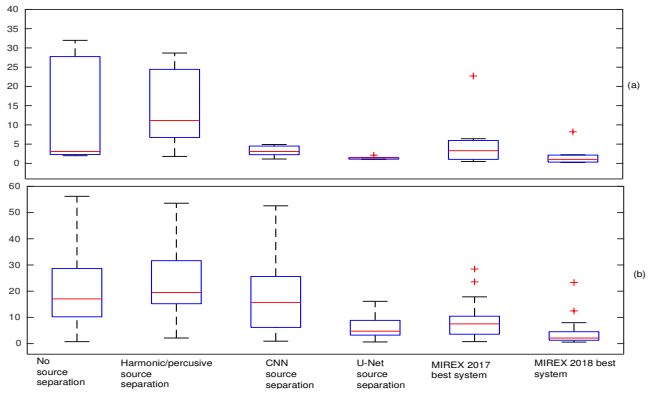

Fig. 4: Boxplot showing the distribution of ASE values for all the songs from (a) Hansen's data (b) Mauch's data, using different systems shown in Table 2.

To analyze the above mentioned results further we show the histogram distributions of absolute word onset deviation between the boundaries obtained from our algorithm and ground truth in Figure 5 for both the datasets. The histograms in Figure 5 show that CNN source separation method leads to a larger spread of onset deviation errors than U-Net. We can observe that in lower bins the number of instances are prominently high in case of Figure 5(a). However, more number of instances show large onset deviation in CNN-based (Figure 5(a)) than in U-Net-based (Figure 5(b).

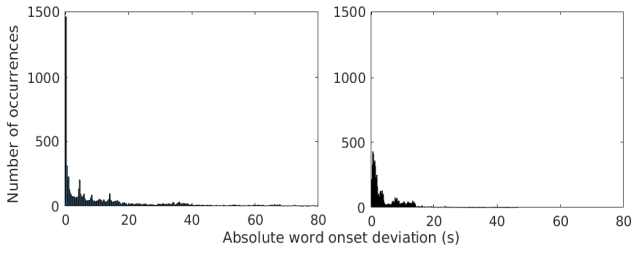

Fig. 5: Histogram showing absolute word onset deviation for the alignment obtained using (a) CNN, (b) U-Net based vocal extraction.

3.3. Effect of DNN-SAT singing-adapted models

In [8], it is reported that the use of deep neural network (DNN) model led to a significant reduction in the WER. A DNN model [26] is trained on top of the SAT model with the same set of training data. During DNN training, temporal splicing is applied on each frame with left and right context window of 4. The SAT+DNN model has 3 hidden layers and 2,976 output targets. In this work, apart from the GMM-HMM (SAT) models, we also applied the SAT+DNN model for lyrics-to-audio alignment for which the ASE and PCS values are shown in Table 3. Although the SAT+DNN model is reported to improve the lyrics recognition performance [8], it does not show an improvement for the alignment task. In the typical acoustic modeling for speech recognition, a baseline GMM-HMM system is first trained and applied to produce an initial word alignment, with which a DNN-HMM system is further trained for recognition. Some researchers have made similar observations with ours, and have conjectured that the objective function for training the DNN models does not force them to produce good alignments, as they are only optimized for good sequence of phonemes [27–29]. For forced-alignment, a GMM-based model is generally recommended as it is more effective and efficient [27].

Table 3: ASE and PCS for lyrics-to-audio alignment systems using SAT+DNN model after applying different audio source separation methods and removal of beginning and ending silences.

Database	Source separation method							
	No source separation		Harmonic/percussive		CNN		U-Net	
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's	23.51	0.06	27.49	0.01	9.26	0.14	2.71	0.12
Mauch's	30.27	0.01	26.79	0.01	23.64	0.03	6.99	0.04

4. SUMMARY

In this work, we present lyrics-to-audio alignment systems using singing adapted GMM-HMM acoustic models. The GMM-HMM speech models are adapted to singing voice using a relatively small set of solo-singing data. We use different audio source separation methods to extract the singing voice and obtain alignment for polyphonic songs. Three different audio source separation methods are compared and observed that the efficacy of singing vocal extraction has a high impact on alignment accuracy. The U-Net based audio source separation performs best for our system. After removal of the begin and end non-vocal sections, the system performance improves further. Our best system has lower ASE value compared to MIREX 2017 best system and comparable to that of MIREX 2018 best system, which is trained on large polyphonic database. This study demonstrates that by using a relatively small solo-singing database for adaptation of speech models to singing voice, along with a reliable audio-source separation, we can develop a high performing lyrics-to-audio alignment system.

5. REFERENCES

[1] Kyogu Lee and Markus Cremer, "Segmentation-based lyrics-audio alignment using dynamic programming," in *International Society for Music Information Retrieval (ISMIR)*, 2008, pp. 395–400.

[2] Ye Wang, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin, "Lyrically: automatic synchronization of acoustic musical signals and textual lyrics," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 212–219.

[3] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto, "Integrating additional chord information into hmm-based lyrics-to-audio alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 200–210, 2012.

[4] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," in *Eighth IEEE International Symposium on Multimedia*, 2006, pp. 257–264.

[5] Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi 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.

[6] Matt McVicar, Daniel PW Ellis, and Masataka Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3117–3121.

[7] Annamaria Mesaros and Tuomas Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 4, 2010.

[8] Chitralekha Gupta, Rong Tong, Haizhou Li, and Ye Wang, "Semi-supervised lyrics and solo-singing alignment," in *International Society for Music Information Retrieval (ISMIR)*, 2018.

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

[10] Anna M Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *International Society for Music Information Retrieval (ISMIR)*, 2016, pp. 358–364.

[11] Georgi Bogomilov Dzhambazov and Xavier Serra, "Modeling of phoneme durations for alignment between polyphonic audio and lyrics," in *12th Sound and Music Computing Conference*, 2015, pp. 281–286.

[12] "Mirex 2018," `https://www.music-ir.org/mirex/wiki/2018:Automatic_Lyrics-to-Audio_Alignment_Results`, [Online; accessed 28-October-2018].

[13] Chung-Che Wang, "Mirex2018: Lyrics-to-audio alignment for instrument accompanied singings," in *MIREX 2018*, 2018.

[14] Jens Kofod Hansen, "Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients," in *9th Sound and Music Computing Conference (SMC)*, 2012, pp. 494–499.

[15] Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic pronunciation evaluation of singing," *Proc. Interspeech 2018*, pp. 1507–1511, 2018.

[16] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," in *3th International Conference on Digital Audio Effects (DAFX10), Graz, Austria*, 2010.

[17] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015, pp. 18–25.

[18] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez, "Monoaural audio source separation using deep convolutional neural networks," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 258–266.

[19] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep U-Net convolutional networks," in *International Society for Music Information Retrieval (ISMIR)*, 2017.

[20] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[21] "A chainer implementation of u-net singing voice separation model," `https://github.com/Xiao-Ming/UNet-VocalSeparation-Chainer`, [Online; accessed 28-October-2018].

[22] Georgi Dzhambazov, *Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals*, Ph.D. thesis, Universitat Pompeu Fabra, 2017.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[25] "Mirex 2017," `https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results`, [Online; accessed 28-October-2018].

[26] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[27] "nnet3 alignment issues," `https://groups.google.com/d/msg/kaldi-help/cSAm5iXGhZo/ZUEqzVZqCgAJ`, [Online; accessed 12-February-2019].

[28] "Good ASR result, bad alignment result (lstm and dnn)," `https://groups.google.com/d/msg/kaldi-help/BkEub9VTUmk/ZJH8wz7pCAAJ`, [Online; accessed 12-February-2019].

[29] "Montreal forced aligner," `https://montreal-forced-aligner.readthedocs.io/en/latest/alignment_techniques.html#deep-neural-networks-dnns`, [Online; accessed 12-February-2019].