AUTOMATIC LYRICS TRANSCRIPTION IN POLYPHONIC MUSIC: DOES BACKGROUND MUSIC HELP?

Chitralekha Gupta, Emre Yılmaz, Haizhou Li

Department of Electrical and Computer Engineering, National University of Singapore

ABSTRACT

Background music affects lyrics intelligibility of singing vocals in a music piece. Automatic lyrics transcription in polyphonic music is a challenging task because the singing vocals are corrupted by the background music. In this work, we propose to learn music genre-specific characteristics to train polyphonic acoustic models. For this purpose, we firstly study and compare several automatic speech recognition pipelines for the application of lyrics transcription. Later, we present the lyrics transcription performance of these music-informed acoustic models for the best-performing pipeline, and systematically study the impact of music genre and language model on the performance. With this genre-based approach, we explicitly model the characteristics of music, instead of trying to remove the background music as noise. The proposed approach achieves a significant improvement in performance in the lyrics transcription and alignment tasks on several well-known polyphonic test datasets, outperforming all comparable existing systems.

Index Terms— Lyrics transcription, lyrics alignment, acoustic modeling, music genre, automatic speech recognition

1. INTRODUCTION

Lyrics is an important component of music, and people often recognize a song by its lyrics. Lyrics contribute to the mood of the song [1,2], affect the opinion of a listener about the song [3], and even help in foreign language learning [4,5]. Automatic lyrics transcription is the task of recognizing the sung lyrics from polyphonic audio. It is the enabling technology for various music information retrieval applications such as generating karaoke scrolling lyrics, music video subtitling, query-by-singing [6,7], keyword spotting, and automatic indexing of music according to transcribed keywords [8,9].

Automatic lyrics transcription of singing vocals in the presence of background music is a challenging task, and an unsolved problem. One of the earliest studies [11] conducted frame-wise phoneme classification in polyphonic music where the authors attempted to recognize three broad-classes of phonemes in 37 popular songs using acoustic features such as MFCCs, and PLP. Mesaros et al. [12,13] adopted an automatic speech recognition (ASR) based approach for phoneme and word recognition of singing vocals in monophonic and polyphonic music. A vocal separation algorithm is applied to separate the singing vocals from polyphonic music. They achieved the best word error rate (WER) of 87% on singing vocals in polyphonic music. This high error rate is mainly due to a lack of available lyrics annotated music data for acoustic modeling.

Many studies have focused on the simpler task of lyrics-to-audio alignment, where the singing vocal is temporally synchronized with the textual lyrics. Kruspe [14] and Dzhambazov [15] presented systems for the lyrics alignment challenge in MIREX 2017, where ASR systems trained on a large publicly available solo-singing dataset called DAMP [16] were used to forced-align lyrics to singing vocals. Gupta et al. [17] adapted DNN-HMM speech acoustic models to singing voice with this data, that showed 36.32% word error rate (WER) in a free-decoding experiment on short solo-singing test phrases from the same solo-singing dataset. In [18], these singing-adapted speech models were further enhanced to capture long duration vowels with a duration-based lexicon modification, that reduced the WER to 29.65%. However, acoustic models trained on solo-singing data result in a significant drop in performance when applied to singing vocals in the presence of background music.

Singing vocals are often highly correlated with the corresponding background music, resulting in overlapping frequency components [19]. To suppress the background accompaniment, some approaches have incorporated singing voice separation techniques as a pre-processing step [12,15,20,21]. However, this step makes the system dependent on the performance of the singing voice separation algorithm, as the separation artifacts may make the words unrecognizable. Moreover, this requires a separate training setup for the singing voice separation system. Recently, Gupta et al. [22] trained acoustic models on a large amount of solo singing vocals and adapted them towards polyphonic music using a small amount of in-domain data – extracted singing vocals, and polyphonic audio. They found that domain adaptation with polyphonic data outperforms that with extracted singing vocals. This suggests that adaptation of acoustic model with polyphonic data helps in capturing the spectro-temporal variations of vocals+background music, better than adaptation with extracted singing vocals that introduces distortions.

One can imagine that the background music might enhance or degrade lyrics intelligibility for humans. Schulz and Huron [23] found that genre-specific musical attributes such as instrumental accompaniment, singing vocal loudness, syllable rate, reverberation, and singing style influence human intelligibility of lyrics.

Instead of treating the background music as noise that corrupts the singing vocals, we hypothesize that acoustic models induced with music knowledge will help in lyrics transcription and alignment in polyphonic music. In this study, we propose to build genre-specific polyphonic acoustic models, that captures the acoustic variability across different genres. However, a limitation to this approach is the lack of lyrics annotated polyphonic data.

Stoller et al. [24] presented a data intensive approach to lyrics transcription and alignment. They proposed an end-to-end system based on the Wave-U-Net architecture that predicts character probabilities directly from raw audio. However, end-to-end systems require a large amount of annotated training polyphonic music data to perform well, as seen in [24] that uses more than 44,000 songs with line-level lyrics annotations from Spotify’s proprietary music library, while publicly available resources for polyphonic music are
2. LYRICS TRANSCRIPTION

Our goal is to build a designated ASR framework for automatic lyrics transcription and alignment. We explore and compare various approaches to understand the impact of background music, and the genre of the music on acoustic modeling. We detail the design procedure followed to gauge the impact of different factors in the following subsections.

2.1. Singing vocal extraction vs. polyphonic audio

Earlier approaches to lyrics transcription have used acoustic models that were trained on solo-singing audio. Singing vocal extraction was then applied on the test data [12,20,26,27]. Such acoustic models can be adapted to a small set of extracted vocals to reduce the mismatch of acoustic models between training and testing [22]. Now that we have available a relatively large polyphonic lyrics annotated dataset (DALI), we explore two approaches for acoustic modeling for the task of lyrics transcription and alignment: (1) to apply singing vocal extraction from the polyphonic audio as a pre-processing step, and train acoustic models with the extracted singing vocals, and (2) to train acoustic models using the lyrics annotated polyphonic dataset directly. Approach (1) treats the background music as the background noise and suppresses it. On the other hand, approach (2) observes the combined effect of vocals and music on acoustic modeling. With these two approaches, we would like to answer the question whether background music helps in acoustic modeling for lyrics transcription and alignment.

2.2. Standard ASR vs. end-to-end ASR

We further compare the performance of different ASR architectures on the lyrics transcription task in the presence of polyphonic music. For this purpose, we first train a standard ASR system with a separate acoustic and language modeling (LM) and pronunciation lexicon. This ASR system is compared with an end-to-end ASR approach which learns how to map spectral speech features to characters without explicitly using a language modeling and pronunciation lexicon [25]. Initial ASR results provided by an end-to-end ASR system trained on a much larger amount of polyphonic audio is presented in [24].

For the standard ASR architecture, a time-delay neural network with additional convolutional layers with senones as targets has been trained for acoustic modeling. Both polyphonic audio data and extracted vocals have been used for training acoustic models to investigate the impact of front-end music separation on the ASR performance in terms of lyrics transcription of unseen songs.

To exploit the linguistic constraints in music lyrics such as rhyming patterns, connecting words, and grammar [31]. The lyrics recognition grammar using a finite state automaton (FSA) built from the lyrics in the queried database. In [8], the LM and dictionary was constructed separately for each test song. However, these methods have been tested only on small solo-singing datasets, and their scalability to a larger vocabulary recognition of polyphonic songs needs to be tested. In this work, we investigate the performance of standard N-gram techniques, also used for large vocabulary ASR, for lyrics transcription in polyphonic songs. We train two N-gram models: (1) an in-domain LM (henceforth referred to as the lyrics LM) trained only on the lyrics from the training music data and (2) a general LM trained on a large publicly available text corpus from extracted from different resources.

The lyrics transcription quality of the standard ASR architecture is compared with an end-to-end system that is trained on the same polyphonic training data. The end-to-end ASR system is trained using a multiobjective learning framework with a connectionist temporal classification (CTC) objective function and an attention decoder appended to a shared encoder [30]. A joint decoding scheme has been used to combine the information provided by the hybrid model consisting of CTC and attention decoder components and hypothesize the most likely recognition output.

2.3. Genre-informed acoustic modeling

Genre of a music piece is characterized by background instrumenta- tion, rhythm structure, and harmonic content of the music [32]. Factors such as instrumental accompaniment, vocal harmonization, and reverberation are expected to interfere with lyric intelligibility, while predictable rhyme schemes and semantic context might improve intelligibility [23].

2.3.1. Genre-informed phone models

One main difference between genres that affects lyric intelligibility is the relative volume of the singing vocals compared to the background accompaniment. For example, as observed in [23], in metal songs, the accompaniment is loud and interferes with the vocals, while is relatively softer in jazz, country, and pop songs. Another difference is the syllable rate between genres. In [23], it was observed that rap songs, that have a higher syllable rate, show lower lyric intelligibility than other genres. We expect that these factors are important for building an automatic lyrics transcription algorithm.

Figure 1 shows the spectrogram of a song segment from the genres (a) pop, (b) metal, and (c) hip hop, containing singing vocals in the presence of background music. The pop song contains acoustic guitar and drums with loud singing vocals with visible singing voice harmonics in the spectrogram. The metal song, on the other hand, has a highly amplified distortion on electric guitar, and loud beats, with relatively soft singing vocals, as seen in the dense spectrogram. The hip hop song has clear and rapid vocalization corresponding to a rhythmic speech in presence of beats. We hypothesize that genre-specific acoustic modelling of phones would capture the combined effect of background music and singing vocals, depending on the genre.

2.3.2. Genre-informed “silence” models

In speech, there are long-duration non-vocal segments that include silence, background noise, and breathing. In an ASR system, a si-
2.3.3. Genre broadclasses

Music has been divided into different genres in many different and overlapping ways, based on a shared set of characteristics. Tzanetakis [32] provided one of the pioneering works in music classification into 10 genres - classical, country, disco, hip hop, jazz, rock, blues, reggae, pop, and metal. Although they are different in some ways, many songs are tagged with more than one genres because of shared characteristics. Schultz and Huron [23] found that across 12 different genres, the overall lyrics intelligibility for humans is 71.7% (i.e. the percentage of correctly identified words from a total of 25,408 words), where genre-specific musical attributes such as instrumental accompaniment, singing vocal loudness, and syllable rate influence human intelligibility of lyrics. They also found that “Death Metal” excerpts received intelligibility scores of zero, while “Pop” excerpts achieved scores close to 100%.

To build genre-informed acoustic models, we consider the shared characteristics between genres that affect lyrics intelligibility, such as type of background accompaniments, and loudness of vocals, and group all genres to three broad genre classes: pop, hiphop, and metal. Table 1 summarizes our genre broadclasses. We categorize songs containing some rap along with electronic music under hiphop broadclass, which includes genres such as Rap, Hip Hop, and Rhythms & Blues. Songs with loud and dense background music are categorized as metal, that includes genres such as Metal and Hard Rock. Songs with clear and louder vocals under genres Pop, Country, Jazz, Reggae etc. are categorized as pop broadclass.

3. EXPERIMENTAL SETUP

3.1. Datasets

All datasets used in the experiments are summarized in Table 2. The training data for acoustic modeling contains 3,913 audio tracks of English polyphonic songs from the DALI dataset [25], consisting of 180,034 lyrics-transcribed lines with a total duration of 134.5 hours.

We evaluated the performance of lyrics transcription and alignment on three test datasets - Hansen’s polyphonic songs dataset (9 songs) [33], Mauch’s polyphonic songs dataset (20 songs) [34], and Jamendo dataset (20 songs) [35]. Hansen’s and Mauch’s datasets were used in the MIREX lyrics alignment challenges of 2017 and 2018. These datasets consist mainly of Western pop songs with manually annotated word-level transcription and boundaries. The Jamendo dataset consists of English songs from diverse genres, along with their lyrics transcription and manual word boundary annotations.

Genre tags for most of the songs in the training dataset (DALI) is provided in their metadata, except for 840 songs. For these songs, we applied a state-of-the-art automatic genre recognition implementation [35] which has 80% classification accuracy, to get their genre tags. We applied the genre groupings from Table 1 to assign a genre broadclass to every song. For the songs in the test datasets, we scanned the web to find their genre tags, as well as manually listened

Table 1: Genre broadclasses grouping

| Genre Broadclasses | Characteristics | Genres |
|--------------------|-----------------|--------|
| hiphop             | rap, electronic music | Rap, Hip Hop, R&B |
| metal              | loud and many background accompaniments, a mix of percussive instruments, amplified distortion, vocals not very loud, rock, psychedelic | Metal, Hard Rock, Electro, Alternative, Dance, Disco, Rock, Indie |
| pop                | vocals louder than the background accompaniments, guitar, piano, saxophone, percussive instruments | Country, Pop, Jazz, Soul, Reggae, Blues, Classical |

Table 2: Dataset description.

| Name          | Content | Lyrics Ground-Truth | Genre distribution |
|---------------|---------|---------------------|--------------------|
| Training data |         |                     |                    |
| DALI          | 3,913 songs | line-level boundaries, 180,034 lines | hiphop:1,919, metal:1,576, pop:2,218 |
| Hansen        | 9 songs | word-level boundaries, 2,212 words | hiphop:1, metal:3, pop:5 |
| Mauch         | 20 songs | word-level boundaries, 5,052 words | hiphop:0, metal:8, pop:12 |
| Jamendo       | 20 songs | word-level boundaries, 5,677 words | hiphop:4, metal:7, pop:9 |
| Test data     |         |                     |                    |
| Hansen        | 9 songs | word-level boundaries, 2,212 words | hiphop:1, metal:3, pop:5 |
| Mauch         | 20 songs | word-level boundaries, 5,052 words | hiphop:0, metal:8, pop:12 |
| Jamendo       | 20 songs | word-level boundaries, 5,677 words | hiphop:4, metal:7, pop:9 |
to the songs to categorize them into the three genre broadclasses according to their lyric intelligibility characteristics. The distribution of the number of songs across the three genre broadclasses for all the datasets is shown in Table 2. The distribution of the number of songs across genres in the training data is skewed towards pop, while hip hop is the most under-represented. However, we are limited by the amount of available data available for training, with DALI being the only resource. Therefore, we assume this to be the natural occurring distribution of songs across genres.

3.2. Vocal separated data vs. polyphonic data
As discussed in Section 2.1, we compare the strategies of vocal extracted data vs. polyphonic data to train the acoustic models, as a way to find out if the presence of background music helps. We use the reported best performing models M4, from the state-of-the-art Wave-U-Net based audio source separation algorithm [36][37] for separating vocals from the polyphonic audio.

3.3. ASR framework: standard ASR vs. end-to-end ASR
The ASR system used in these experiments is trained using the Kaldi ASR toolkit [38]. A context dependent GMM-HMM system is trained with 40k Gaussians using 39 dimensional MFCC features including the deltas and delta-deltas to obtain the alignments for neural network training. The frame rate and length are 10 and 25 ms, respectively. A factorized time-delay neural network (TDNN-F) model [39] with additional convolutional layers (2 convolutional, 10 time-delay layers followed by a rank reduction layer) was trained according to the standard Kaldi recipe (version 5.4). An augmented version of the polyphonic training data (Section 2.1) is created by reducing (x0.9) and increasing (x1.1) the speed of each utterance [40]. This augmented training data is used for training the neural network-based acoustic model. The number of context-dependent phone states varies between 5616, 5824, and 5448 for the ASR system with no genre information, genre-specific silence modeling and genre-specific phone modeling, respectively.

The default hyperparameters provided in the standard recipe were used and no hyperparameter tuning was performed during the acoustic model training. The baseline acoustic model is trained using 40-dimensional MFCCs as acoustic features. During the training of the neural network [41], the frame subsampling rate is set to 3 providing an effective frame shift of 30 ms. A duration-based modified pronunciation lexicon is employed which is detailed in [42]. Two language models are trained using the transcriptions of the DALI dataset and the open source text corpus [43] released as a part of the Librispeech corpus [42]. The perplexities provided by the in-domain (Lyrics LM) and general models (general LM) on different test sets (Hansen, Mauch, and Jamendo) are given in Table 3.

The end-to-end system is trained using the ESPNet toolkit [43]. The shared encoder is a combination of two VGG [44] layers followed by a BLSTM with subsampling [29] with 5 layers and 1024 units. The attention-based decoder is a 2-layer decoder with 1024 units with coverage attention [45]. The batchsize is set to 20 to avoid GPU memory overflow. The rest of the hyperparameters are consistent with the standard Librispeech recipe available in the toolkit (version 0.3.1). In pilot experiments, using a language model during the decoding with the default language model weight provided worse results than decoding without a language model. Therefore, no LM is used during the decoding step to avoid parameter tuning on the test data.

3.4. Genre-informed acoustic modeling
We train 3 different types of acoustic models corresponding to the three genre broadclasses, for (a) genre-informed "silence" or non-vocal models and (b) genre-informed phone models. We extract the non-vocal segments at the start and the end of each line in the training data to increase the amount of frames for learning the silence models. The symbols representing different genre-specific silence models are added to the ground truth training transcriptions so that they are explicitly learned during the training phase. For the genre-informed phone models, we append the genre tag to each word in the training transcriptions to be able to make genre-specific ASR performance analysis. These genre-specific words are mapped to genre-specific phonetic transcriptions in the pronunciation lexicon which enables learning separate phone models for each genre. For the alignment task, we use the same genre-informed phone models that are mapped to the words without genre tags, i.e. the alignment system chooses the best fitting phone models among all genres during the forced alignment, to prevent the additional requirement of genre information for songs in the test sets.

4. RESULTS AND DISCUSSION
We conduct four sets of experiments to demonstrate and compare different strategies for lyrics transcription and alignment: (1) train acoustic models using extracted vocal and polyphonic audio and compare the ASR performance, (2) compare different ASR architectures, namely a standard ASR system and an end-to-end ASR, (3) compare the performance of genre-informed silence and phone models compared to the baseline genre-agnostic models, and finally (4) explore the impact of different language models.

4.1. Singing vocal extraction vs. polyphonic audio
We compare the lyrics transcription performance of standard ASR trained on extracted singing vocals and polyphonic audio in Table 4. We see an improvement in performance with polyphonic data on all the test datasets. This indicates that there is value in modeling the combination of vocals and music, instead of considering the background music as noise and suppressing it. Although we have used the state-of-the-art vocal extraction algorithm, these techniques are still not perfect, and introduce artifacts and distortions in the extracted vocals, which is the reason for poor performance of the models trained with extracted vocals.

When we compare the lyrics alignment performance with acoustic models trained with extracted singing vocals and polyphonic audio, we see a trend similar to the transcription performance, as shown in Table 5. The mean absolute word alignment error has reduced to less than 350 ms in all the test datasets with the polyphonic models given in the third column. We observe a large improvement in the alignment accuracy on the Mauch’s dataset, that consists of many songs with long musical interludes, where the extracted vocals models fail to align the lyrics around the long non-vocal sections because of erratic music suppression. Polyphonic models, on the other
Table 4: Lyrics transcription WER (%) comparison of vocal extracted vs. polyphonic data trained acoustic models

| Test Datasets | Vocal extracted | Polyphonic  |
|---------------|----------------|------------|
| Mauch         | 76.31          | 54.08      |
| Hansen        | 78.85          | 60.77      |
| Jamendo       | 71.83          | 66.58      |

Table 5: Mean absolute word alignment error (seconds). The acoustic models are trained on extracted vocals, polyphonic data with no genre information, genre-informed silence, and silence+phone models.

| Test Datasets | Vocal Extracted | Polyphonic: No Genre Info | Polyphonic: Genre Silence | Polyphonic: Genre Silence+Phone |
|---------------|----------------|---------------------------|---------------------------|--------------------------------|
| Mauch         | 3.62           | 0.25                      | 0.28                      | 0.21                           |
| Hansen        | 0.67           | 0.16                      | 0.25                      | 0.18                           |
| Jamendo       | 0.39           | 0.34                      | 0.42                      | 0.22                           |

Table 7: Comparison of lyrics transcription WER (%) with acoustic models trained on polyphonic data with no genre information, genre-informed silence, and silence+phone models, with general and lyrics LMs.

| Test Datasets | No Genre Info | Genre Silence | Genre Silence+Phone |
|---------------|---------------|---------------|---------------------|
| Mauch         | 54.08         | 52.45         | 53.74               |
| Hansen        | 66.77         | 59.10         | 62.71               |
| Jamendo       | 66.58         | 64.42         | 67.91               |

Fig. 2: Comparison of (a) lyrics transcription WER (%), and (b) lyrics alignment mean absolute word alignment error (seconds)—for different genre broadclasses across all the test datasets. The acoustic models are trained on polyphonic data with no genre information, genre-informed silence, and silence+phone models, with general LM.

Table 6: Comparison of lyrics transcription WER (%) of Standard ASR vs. End-to-end ASR

| Test Datasets | Standard | End-to-end |
|---------------|----------|------------|
| Mauch         | 54.08    | 75.2       |
| Hansen        | 60.77    | 80.1       |
| Jamendo       | 66.58    | 87.9       |

5. CONCLUSIONS

In this work, we introduce a music-informed strategy to train polyphonic acoustic models for the tasks of lyrics alignment and transcription in polyphonic music. We model the genre-specific characteristics of music and vocals, and study their performance with different ASR frameworks, and language models. We find that this
music-informed strategy learns the background music characteristics that affect lyrics intelligibility, and shows improvement in lyrics alignment and transcription performance over others with music suppression. We also show that with limited available data, our strategy of genre-informed acoustic modeling as well as lyrics constrained language modeling in a standard ASR pipeline is able to outperform all existing systems for both lyrics transcription and alignment task.

### 6. REFERENCES

[1] S. O. Ali and Z. F. Peynircioğlu, “Songs and emotions: are lyrics and melodies equal partners?,” *Psychology of Music*, vol. 34, no. 4, pp. 511–534, 2006.

[2] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. Nieminen, and M. Tervaniemi, “A functional mri study of happy and sad emotions in music with and without lyrics,” *Frontiers in psychology*, vol. 2, 2011.

[3] B. Anderson, D. G. Berger, R. S. Denisoff, K. P. Etzkorn, and P. Hesbacher, “Love negative lyrics: Some shifts in stature and alterations in song,” *Communications*, vol. 7, no. 1, pp. 3–20, 1981.

[4] H. Nakata and L. Shockey, “The effect of singing on improving syllabic pronunciation–vowel epenthesis in japanese,” in *International Conference of Phonetic Sciences*, 2011.

[5] A. J. Good, F. A. Russo, and J. Sullivan, “The efficacy of singing in foreign-language learning,” *Psychology of Music*, vol. 43, no. 5, pp. 627–640, 2015.

[6] C.-C. Wang, J.-S. R. Jang, and W. Wang, “An improved query by singing/humming system using melody and lyrics information,” in *ISMIR*, 2010, pp. 45–50.

[7] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and J. H. Witten, “Lyrics recognition from a singing voice based on finite state automaton for music information retrieval,” in *ISMIR*, 2005, pp. 532–535.

[8] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, “An autoregressive, non-stationary excitation signal parameter estimation method and an evaluation of a singing-voice recognition,” in *Proc. ICASSP*. IEEE, 2005, vol. 1, pp. 1–237.

[9] H. Fujihara, M. Goto, and J. Ogata, “Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics,” in *ISMIR*, 2008, pp. 281–286.

[10] M. Müller, F. Kurth, D. Damm, C. Fremeney, and M. Clausen, “Lyrics-based audio retrieval and multimodal navigation in

### Table 8: Comparison of lyrics alignment (mean absolute word alignment error (seconds)) and transcription (WER%) performance with existing literature.

|                      | MIREX 2017 | MIREX 2018 | ICASSP 2019 | INTERSPEECH2019 |
|----------------------|------------|------------|-------------|-----------------|
|                      | AK [14] | GD [15,46] | CW [7] | JS [4] | CG [20] | CG [22] | Ours |
| **Lyrics Alignment** | Mauch Hansen Jamendo | 9.03 | 7.34 | 10.57 | 4.13 | 7.09 | 6.34 | 7.09 | 7.08 | 1.93 | 0.82 | 0.82 | 0.93 | 0.82 | 0.18 |
|                      | Mauch Hansen Jamendo | - | - | - | - | - | - | 7.09 | 4.10 | 4.10 | 59.6 |
| **Lyrics Transcription** | Mauch Hansen Jamendo | - | - | - | - | - | - | 7.09 | 4.10 | 4.10 | 59.6 |

[8] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, “An automatic recognition of lyrics in singing,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 546047, 2010.

[9] H. Fujihara, M. Goto, and J. Ogata, “Automatic alignment of music audio and lyrics,” in *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

[10] A. M. Kruspe, “Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing,” in *ISMIR*, 2016, pp. 358–364.

[11] M. Gruhne, C. Dittmar, and K. Schmidt, “Phoneme recognition in popular music.” in *ISMIR*, 2007, pp. 369–370.

[12] G. B. Dzhambazov and X. Serra, “Modeling of phoneme durations for alignment between polyphonic audio and lyrics,” in *12th Sound and Music Computing Conference*, 2015, pp. 281–286.

[13] G. B. Dzhambazov and X. Serra, “Modeling of phoneme durations for alignment between polyphonic audio and lyrics,” in *12th Sound and Music Computing Conference*, 2015, pp. 281–286.

[14] Smule, “Digital Archive Mobile Performances (DAMP),” [https://ccrma.stanford.edu/damp/](https://ccrma.stanford.edu/damp/)

[15] Smule, “Digital Archive Mobile Performances (DAMP),” [https://ccrma.stanford.edu/damp/](https://ccrma.stanford.edu/damp/)

[16] Smule, “Digital Archive Mobile Performances (DAMP),” [https://ccrma.stanford.edu/damp/](https://ccrma.stanford.edu/damp/)

[17] C. Gupta, R. Tong, H. Li, and Y. Wang, “Semi-supervised lyrics and solo-singing alignment,” in *Proc. ISMIR*, 2018.

[18] C. Gupta, H. Li, and Y. Wang, “Automatic pronunciation evaluation of singing,” in *Proc. INTERSPEECH*, pp. 1507–1511, 2018.

[19] M. Ramona, G. Richard, and B. David, “Vocal detection in music with support vector machines,” in *2008 Proc. ICASSP*. IEEE, 2008, pp. 1885–1888.

[20] C. Gupta, B. Sharma, H. Li, and Y. Wang, “Automatic lyrics-to-audio alignment on polyphonic music using singing-adapted acoustic models,” in *Proc. ICASSP*. IEEE, 2019, pp. 396–400.

[21] M. Fujihara, H. and Goto, J. Ogata, and H. G. Okuno, “Lyric-synchronizer: 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.

[22] C. Gupta, E. Yilmaz, and H. Li, “Acoustic modeling for automatic lyrics-to-audio alignment,” in *Proc. INTERSPEECH*, Sept. 2019, pp. 2040–2044.

[23] N. Condit-Schultz and D. Huron, “Catching the lyrics: intelligibility in twelve song genres.” *Music Perception: An Interdisciplinary Journal*, vol. 32, no. 5, pp. 470–483, 2015.

[24] D. Stoller, S. Durand, and S. Evert, “End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model,” in *Proc. ICASSP*. IEEE, 2019, pp. 181–185.

[25] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, “Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm,” in *Proc. ISMIR*, 2018.

[26] X. Hu, J. H. Lee, D. Bainbridge, K. Choi, P. Organisciak, and Organisciak, “Semi-supervised lyrics and transcription,” in *Proc. ISMIR*, vol. 68, no. 1, pp. 97–112, Jan. 2017.

[27] G. Dzhambazov, Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals, Ph.D. thesis, Universitat Pompeu Fabra, 2017.
[28] A. Graves and N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. 2014, ICML’14, pp. II–1764–II–1772, JMLR.org.

[29] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in Proc. ICASSP, March 2016, pp. 4960–4964.

[30] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/Attention architecture for end-to-end speech recognition,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240–1253, Dec 2017.

[31] J. Fang, D. Grunberg, D. T. Litman, and Y. Wang, “Discourse analysis of lyric and lyric-based classification of music,” in ISMIR, 2017, pp. 464–471.

[32] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293–302, 2002.

[33] J. K. 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.

[34] M. Mauch, H. Fujihara, and M. 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.

[35] “Musical genre recognition using a cnn,” https://github.com/thomas-bouvier/music-genre-recognition.git [Online; accessed 5-July-2019].

[36] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” in Proc. ISMIR, 2018.

[37] “Implementation of the wave-u-net for audio source separation,” https://github.com/f90/Wave-U-Net.git [Online; accessed 5-July-2019].

[38] A. Povey, D. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, and K. Vesely, “The Kaldi speech recognition toolkit,” in in Proc. ASRU, 2011.

[39] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, “Semi-orthogonal low-rank matrix factorization for deep neural networks,” in Proc. INTERSPEECH, 2018, pp. 3743–3747.

[40] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio augmentation for speech recognition,” in Proc. INTERSPEECH, 2015, pp. 3586–3589.

[41] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” in Proc. INTERSPEECH, 2016, pp. 2751–2755.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in Proc. ICASSP, April 2015, pp. 5206–5210.