LYRICS-TO-AUDIO ALIGNMENT WITH MUSIC-AWARE ACOUSTIC MODELS

Chitralekha Gupta **Emre Yılmaz** **Haizhou Li**

Electrical and Computer Engineering Dept., National University of Singapore, Singapore
{chitralekha, emre, haizhou.li}@nus.edu.sg

ABSTRACT

We describe the algorithm that we have submitted for the MIREX 2019 task of Automatic Lyrics-to-Audio Alignment. The goal is to automatically detect word boundaries in English pop music using an automatic speech recognition system, given the mixed singing audio (singing voice + musical accompaniment) and lyrics as inputs. The key component of the submission is the music-aware acoustic models that learn music genre-specific characteristics to train polyphonic acoustic models. With this genre-based approach, we explicitly model the characteristics of music by using genre-specific acoustic models, instead of trying to suppress the background music. Moreover, to account for the long duration vowels in singing, we have modified the lexicon with longer duration vowel pronunciation variants. We use the final ASR to forced-align lyrics-to-audio and obtain word boundaries. Experimental results have shown that the proposed modifications provided considerable improvements in the alignment quality.

1. OVERVIEW AND MOTIVATION

In automatic speech recognition (ASR) tasks, word or phone-level segmentation is obtained by forced-aligning the transcription to the speech using acoustic models trained with speech data. In this MIREX task, we apply the same concept to align lyrics to music audio. However, since singing vocals in the presence of background music is different from speech, we introduce music related information to train acoustic models for this task.

Singing vocals are often highly correlated with the corresponding background music, resulting in overlapping frequency components [17]. To suppress the background accompaniment, some approaches have incorporated singing voice separation techniques as a pre-processing step [4,5,7,12]. 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. [8] 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.

Stoller et al. [18] 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 [18] that uses more than 44,000 songs with line-level lyrics annotations from Spotify's proprietary music library. Unfortunately, publicly available resources for polyphonic music are limited.

Table 1. Training dataset description.

Name	Content	Lyrics Ground-Truth	Genre distribution
DALI [13]	3,913 songs	line-level boundaries, 180,034 lines	hiphop:119, metal:1,576, pop:2,218

Recently, a multimodal DALI dataset [13] was introduced, that provides open access to 3,913 English polyphonic songs with note annotations and weak word-level, line-level, and paragraph-level lyrics annotations. In our recent work [9], we train genre-informed polyphonic acoustic models for automatic lyrics transcription and alignment using this openly available polyphonic audio resource. In this work, instead of treating the background music as noise that corrupts the singing vocals, we trained acoustic models induced with genre information that captured the acoustic variability across different genre, thus showing improvement in both the tasks of lyrics transcription and alignment in polyphonic music. In this MIREX task, we submit the best performing system for the task of lyrics alignment from our work [9], that we will briefly discuss in the next sections.

2. SYSTEM DESCRIPTION

2.1 Datasets

As shown in Table 1, the training data for acoustic modeling contains 3,913 audio tracks.[1] English polyphonic songs from the DALI dataset [13], consisting of 180,034 lyrics-transcribed lines with a total duration of 134.5 hours.

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 [1] which has 80% classification accuracy, to get their genre tags. We applied the genre groupings from Table 2 to assign a genre broadclass to every song. The distribution of the number of songs across genres in the training data is skewed towards *pop*, while *hiphop* 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.

2.2 ASR Framework

The ASR system used in these experiments is trained using the Kaldi ASR toolkit [14]. 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 [15] 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 [10]. This augmented training data is used for training the neural network-based acoustic model.

[1] There are a total of 5,358 audio tracks in DALI, out of which only 3,913 were English language and audio links were accessible from Singapore.

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 [16], 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 [6].

2.3 Genre-informed acoustic modeling

Genre of a music piece is characterized by background instrumentation, rhythmic structure, and harmonic content of the music [19]. 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 [2].

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 [2], 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 [2], 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 acoustic models for singing voice in polyphonic audio and 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 silence acoustic model is separately modeled for better alignment and recognition. Non-vocal segments or musical interludes are also frequently occurring in songs, especially between verses. However, in polyphonic songs, these non-vocal segments consist of different kinds of musical accompaniments that differ across genres. For example, a metal song typically consists of a mix of highly amplified distortion guitar, and emphatic percussive instruments, a typical jazz song consists of saxophone and piano, and a pop song consists of guitar and drums. The spectro-temporal characteristics of the combination of instruments vary across genres, but are somewhat similar within a genre. Thus, we propose to train genre-specific non-vocal or "silence" models to characterize this variability of instrumentation across genres.

2.3.3 Acoustic Models

We train 3 different types of acoustic models corresponding to the three genre broadclasses (Table 2), 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 of the corresponding genre songs. 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.

3. RESULTS

Lyrics alignment shows an improvement in performance with genre-informed silence + phone models over those with no genre info and genre-informed silence models [9]. The mean absolute word alignment error is less than 220 ms across three test datasets – Hansen [2], Mauch, and Jamendo. This indicates that the genre-informed phone models trained on limited data are able to capture the transition between phones well. We compare our best results with the most recent prior work (Table 3), and find that our strategy provides the best results for lyrics alignment on the three test datasets. Our proposed strategies show a way to induce music knowledge in ASR to address the problem of lyrics alignment in polyphonic audio.

Table 2. 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 3. Comparison of lyrics alignment (mean absolute word alignment error (seconds)) performance with existing literature.

Test	MIREX 2017		MIREX 2018	ICASSP 2019		Interspeech2019	
Datasets	AK [11]	GD [3, 4]	CW [20]	DS [18]	CG [7]	CG [8]	Ours [9]
Mauch	9.03	11.64	4.13	0.35	6.34	1.93	0.21
Hansen	7.34	10.57	2.07	-	1.39	0.93	0.22
Jamendo	-	-	-	0.82	-	-	0.22

4. REFERENCES

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

[2] N. Condit-Schultz and D. Huron. Catching the lyrics: intelligibility in twelve song genres. *Music Perception: An Interdisciplinary Journal*, 32(5):470–483, 2015.

[3] G. Dzhambazov. *Knowledge-based probabilistic modeling for tracking lyrics in music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2017.

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

[5] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

[6] C. Gupta, H. Li, and Y. Wang. Automatic pronunciation evaluation of singing. *Proc. INTERSPEECH*, pages 1507–1511, 2018.

[7] 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*, pages 396–400. IEEE, 2019.

[8] C. Gupta, E. Yılmaz, and H. Li. Acoustic modeling for automatic lyrics-to-audio alignment. In *Proc. INTERSPEECH*, Sept. 2019.

[9] C. Gupta, E. Yılmaz, and H. Li. Automatic lyrics transcription in polyphonic music: Does background music help? *arXiv preprint arXiv:1909.10200, eess.AS*, 2019.

[10] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur. Audio augmentation for speech recognition. In *Proc. INTERSPEECH*, pages 3586–3589, 2015.

[11] A. M. Kruspe. Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *ISMIR*, pages 358–364, 2016.

[2] excluding the song "clock" due to errors in the ground-truth alignment.

[12] A. Mesaros and T. Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):546047, 2010.

[13] 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.

[14] A. Povey, D.and 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 *in Proc. ASRU*, 2011.

[15] 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*, pages 3743–3747, 2018.

[16] 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*, pages 2751–2755, 2016.

[17] M. Ramona, G. Richard, and B. David. Vocal detection in music with support vector machines. In *2008 Proc. ICASSP*, pages 1885–1888. IEEE, 2008.

[18] D. Stoller, S. Durand, and S. Ewert. End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *Proc. ICASSP*, pages 181–185. IEEE, 2019.

[19] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

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