LYRICS TRANSCRIPTION AND LYRICS-TO-AUDIO ALIGNMENT WITH MUSIC-INFORMED ACOUSTIC MODELS

Xiaoxue Gao **Chitralekha Gupta** **Haizhou Li**

Department of Electrical and Computer Engineering, National University of Singapore, Singapore

xiaoxue.gao@u.nus.edu, {chitralekha, haizhou.li}@nus.edu.sg

ABSTRACT

We present the systems that we submitted for MIREX 2020 of Lyrics Transcription and Lyrics-to-Audio Alignment tasks. Instead of separating the singing vocals from mixed audios (singing voice + musical accompaniment), we jointly train and optimize the acoustic models directly on mixed audios using music-informed acoustic models. The music-aware acoustic models are able to better capture music genre-specific characteristics during the polyphonic acoustic model training. The trained acoustic model is used to forced-align lyrics to audio for the lyrics alignment task. To take advantage of the available lyrics textual resources, we interpolate a general purpose language model with an in-domain language model to improve lyrics transcription. Experimental results have shown that our methods achieve substantial improvements over the prior work in both lyrics alignment and recognition.

1. INTRODUCTION

In recent years, there has been an increasing interest in lyrics-to-audio alignment as well as lyrics transcription. These tasks have a great potential in applications such as the automatic generation of karaoke lyrical content, music video subtitling and query-by-singing. In MIREX, this year, there are two simultaneous tasks - lyrics-to-audio alignment and lyrics transcription. Lyrics-to-audio alignment aims at automatically detecting the word boundaries in polyphonic music audios given the corresponding lyrics, while the goal of lyrics transcription is to recognize the sung lyrics from mixed music and singing vocal audios.

Knowing that background music in polyphonic audio may interfere with the lyrics intelligibility, singing voice separation techniques were utilized as a pre-processing strategy with lyrics transcription techniques [2, 4, 6, 13] to remove background accompaniment. However, these approaches make the performance of lyrics recognition highly dependent on the accuracy of singing voice separation algorithms, and require additional training procedures of separation methods.

Without the usage of singing voice separation techniques, an end-to-end system was also proposed [19] for lyrics transcription and alignment. The system based on the Wave-U-Net architecture was able to predict character probabilities directly from raw audio, but it required a large amount of annotated training polyphonic music data (more than 44,000 songs along with line-level lyrics annotations) and they are not publicly available.

In this work, we apply the standard automatic speech recognition (ASR) pipeline, consisting of acoustic model, language model, and pronunciation model for the tasks of lyrics-to-audio alignment and lyrics transcription. Instead of suppressing the background music, we incorporate music genre-specific information from polyphonic audios to train acoustic models [9] using a multimodal DALI dataset [14]. Additionally, we incorporate duration-based lexicon modification to accommodate the presence of the long duration vowel in singing [5]. In this MIREX submission, we further investigate the interpolation of language model between the in-domain lyrics and a high resource speech corpus, which yields better performing system for the task of lyrics transcription.

For Mirex 2020, we submit two systems - GGL1 and GGL2. Both these systems should be used for the lyrics-to-audio alignment and lyrics transcription tasks. We describe their specifics in the following sections.

Table 1. Dataset description.

Name	Content	Lyrics Ground-Truth	Total Duration
DALI [14]	3,913 songs	line-level boundaries, 180,034 lines	208.6 hours
Proprietary	517 songs	line-level boundaries, 26,462 lines	27.0 hours
DALI-dev [8, 14]	100 songs	line-level boundaries, 5,356 lines	3.9 hours

2. SYSTEM DESCRIPTION

2.1 Training Dataset

As shown in Table 1, the training data for acoustic modeling consist of DALI [14], that has 3,913 English polyphonic audio tracks [1]. It comprises of 180,034 line-level audio and lyrics transcription with a total duration of 208.6 hours. System GGL2 uses only DALI dataset for acoustic model training.

For System GGL1, along with the DALI dataset, we also use a small proprietary dataset consisting of 517 popular English songs, for acoustic model training. Line-level lyrics boundaries of this dataset was obtained automatically with the help of our previous system [9]. This dataset consists of 26,462 line-level audio and lyrics transcription with a total duration of 27.0 hours.

We also used 100 songs from DALI dataset [8] which is not present in its training dataset, as a development set. We fine-tune our language model on this dev set, as discussed in Section 2.3.

2.2 Acoustic Model

The ASR system used in these experiments is trained using the Kaldi ASR toolkit [15]. The two submitted systems differ in their acoustic model architecture.

- **GGL1**: has a factorized time-delay neural network (TDNN-F) architecture [16], and
- **GGL2**: has a factorized time-delay neural network (TDNN-F) model with additional convolutional layers (2 convolutional, 10 time-delay layers followed by a rank reduction layer) [9].

Both of these acoustic models were trained according to the standard Kaldi recipe (version 5.4), where the default setting of hyperparameters provided in the standard recipe was used and no hyperparameter tuning was conducted during the acoustic model training. An augmented version of the polyphonic training data

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

Table 2. Comparison of lyrics alignment (mean absolute word alignment error (seconds)) and lyrics transcription (WER%) performance with existing literature.

	MIREX 2017		MIREX 2018	ICASSP 2019		Interspeech 2019	MIREX 2019	Ours	
	AK [12]	GD [1, 2]	CW [20]	DS [19]	CG [6]	CG [7]	CG [10]	GGL1	GGL2 [9]
Lyrics Alignment									
Mauch	9.03	11.64	4.13	0.35	6.34	1.93	0.21	0.24	**0.20**
Hansen	7.34	10.57	2.07	-	1.39	0.93	0.22	0.22	**0.22**
Jamendo	-	-	-	0.82	-	-	0.22	0.30	**0.20**
Lyrics Transcription									
Mauch	-	-	-	70.9	-	-	-	**43.7**	45.6
Hansen	-	-	-	-	-	-	-	**50.5**	51.1
Jamendo	-	-	-	77.8	-	-	-	**56.7**	61.2

is created by reducing (x0.9) and increasing (x1.1) the speed of each utterance [11], which is used for the training of the acoustic model. The acoustic model is trained using 40-dimensional MFCCs as acoustic features. During the training of the neural network [17], the frame subsampling rate is set to 3 providing an effective frame shift of 30 ms. A duration-based modified pronunciation lexicon is employed to achieve a longer duration vowel [5].

Genre-information was provided in the lexicon at the time of acoustic model training for system GGL2 to capture the genre-specific behaviours in polyphonic audios. The details are explained in the paper [9].

2.3 Language Model

For the task of lyrics transcription, we explore the impact of different language models. In order to better capture the linguistic characteristics of lyrics of songs such as connecting words and rhythmic patterns [3], we propose to use an interpolated language model (LM) that bridges between a small in-domain sung lyrics corpus and a large vocabulary speech corpus text for lyrics transcription.

The in-domain lyrics LM is built using the lyrics corpus of the songs in training datasets 1. The general LM is 3-gram ARPA LM, pruned with threshold 3e-7 obtained from the open source of LibriSpeech language models [2]. The interpolated LM is an interpolation between lyrics LM and general LM, where the interpolation weight yields the lowest perplexity on the DALI development set (Table 1). The interpolated LM used is standard 3-grams with interpolated Kneser-Ney smoothing using SRILM toolkit [18]. In our preliminary investigation, we found that the interpolated LM outperforms the lyrics LM and general LM in terms of word error rate (WER %) for lyrics transcription task. Both the submitted systems GGL1 and GGL2 use the interpolated LM for the task of lyrics transcription.

2.4 System Description

2.4.1 Task 1: Lyrics-to-Audio Alignment

Given the polyphonic song audio and the lyrics as inputs, the system detects the word-level onset and offset boundaries. Both our submitted systems, GGL1 and GGL2, use Viterbi forced-alignment to align the lyrics to the audio. The two systems differ in their training data and acoustic model architecture, as discussed in the previous sections.

2.4.2 Task 2: Lyrics Transcription

Given the polyphonic song as the input, the system transcribes the sung lyrics of the song. In free-decoding mode, both our submitted systems GGL1 and GGL2 in combination with the interpolated language model, generate a string of words for any given input polyphonic song.

3. RESULTS

To assess the quality of lyrics-to-audio alignment, we calculate the mean absolute word boundary error averaged over all songs using the Mirex evaluation toolkit [1]. To assess the quality of lyrics transcription, we compute word error rate (WER), a standard metric of evaluation of ASR, which is the percentage of the total number of insertions, substitutions, and deletions with respect to the total number of words.

We compare the performance of our submitted systems GGL1 and GGL2 with the recent prior work (in Table 2) on three test datasets – Hansen [3], Mauch, and Jamendo. The test datasets were obtained from the respective authors for our research.

System GGL2 shows alignment error of less than or equal to 220 ms across all the three test datasets, and outperforms the previous systems for lyrics alignment task. Both our systems with the interpolated LM show considerable improvements in lyrics transcription performance compared to previous work [19]. This indicates that the integrated LM used in genre-informed acoustic model is effective in improving the accuracy of lyrics transcription in polyphonic audio. System GGL2 outperforms GGL1 by 1-4% showing that the model architecture along with increased amount of training data can improve the transcription performance.

4. REFERENCES

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

[2] 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.

[3] J. Fang, D. Grunberg, D. T. Litman, and Y. Wang. Discourse analysis of lyric and lyric-based classification of music. In *ISMIR*, pages 464–471, 2017.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] C. Gupta, E. Yılmaz, and H. Li. Automatic lyrics alignment and transcription in polyphonic music: Does background music help? In *ICASSP*, pages 496–500, 2020.

[2] http://www.openslr.org/11/

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

[10] Chitralekha Gupta, Emre Yılmaz, and Haizhou Li. Lyrics-to-audio alignment with music-aware acoustic models.

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

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

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, 2002.

[19] 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.

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