LYRICS-TO-AUDIO ALIGNMENT FOR DYNAMIC LYRIC GENERATION

Bin Zhang **WuCheng Wang** **Ethan Zhao** **Simon Lui**
Tencent Music Entertainment Group (TME)
{robingzhang, wuchengwang, ethanzhao, nomislui}@tencent.com

ABSTRACT

This short paper introduces our algorithm submitted to the MIREX 2020 task of Automatic Lyrics-to-Audio Alignment in 2020:Lyrics Transcription. Lyric-to-audio alignment can be applied to dynamic lyric generation, which locates the timestamp of each word. The goal is aimed to align the lyrical content at word-level with the corresponding solo-singing vocal in polyphonic music automatically. We present an approach that uses the singer adapted of trip-phone GMM-HMM and chain-TDNN acoustic models trained by the kaldi-toolkit. Firstly, we use an audio source separation tool to extract the solo-singing vocal for acoustic model training. Secondly, we build a customized lexicon with huge amount of annotated lyrics which cover the missing words in free pronouncing dictionary such as CMUdict. Thirdly, we add <NOISE> tag between sentences to absorb silent non-vocal segments and act as delimiter to generate LRC or QRC file.

1. INTRODUCTION

Automatic lyrics alignment in polyphonic music is a challenging task because the singing vocal is well mixed with the background music. Some previous works have incorporated singing voice separation techniques as a preprocessing step in order to suppress the background accompaniment [1]. We use a state-of-the-art source separation tool [2]to extract the solo-singing vocal [3] from the polyphonic music. The extracted solo-singing vocal can be used to train the model and hence to improve the model intelligibility. Lyrics-to-Audio Alignment automatically generates LRC or QRC file which saves a lot of manpower. It is very useful in many applications such as karaoke scrolling lyrics and dynamic lyrics generation.

2. LYRICS-TO-AUDIO ALIGNMENT

2.1 Data Collection

Tencent Music Entertainment Group(TME) is the leading online music entertainment platform in China. QQ Music is a music streaming service owned by the TME and we serve 800M+ users globally. We have a huge music library with an extensive collection of albums. In this work, we used 7864 sentence-annotated polyphonic songs as training data for acoustic modeling, which consists of 568721 lyrics-transcribed sentence with a total duration of 524 hours. We used a source separation tool [2] to extract the solo-singing vocal tracks.

2.2 Method

An overview of the complete framework is depicted in Figure 1. Our system is inspired by a baseline ASR speech alignment system: a tri-phone GMM-HMM acoustic model trained with the Librispeech corpus using MFCC extracted with the Kaldi toolkit. We trained our model with solo-singing vocal extracted by spleeter [2]. We used 39-MFCC including the deltas and delta-deltas to obtain the alignments for training. The frame rate and length were 10 and 25 ms respectively. We used feature-space maximum likelihood linear regression to compute transformations of the singing feature vectors. These transformations were applied at the time of training for the adaptation of the acoustic models to singing voice using solo-singing data, called SAT. Then we used SAT alignment results to train the Chain model of TDNN using high-resolution MFCC features called CHAIN. The pronunciation lexicon used in the training is based on our own music lyrics library with around 200,000 song. It maps words to phonemes by the Grapheme-to-Phoneme (G2P) [4] model, for example, the word “hello” refers to the phonemes “[HH EH L OW]”. Some music have a very long music intro or outro without singing vocals [5], which can have serious negative effect on lyrics alignment. A common solution is to apply a voice activity detection (VAD) algorithm over the extracted vocals to detect the vocal and non-vocal segments. We solve this problem by adding <NOISE> tags between sentences in the prediction stage, which help to locate the vocal part precisely.

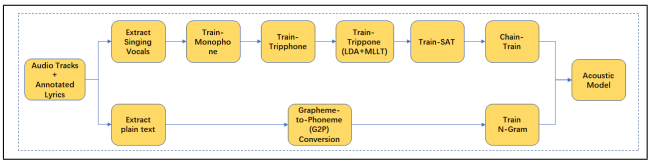

Figure 1. Framework for automatic lyrics-to-audio alignment at training stage.

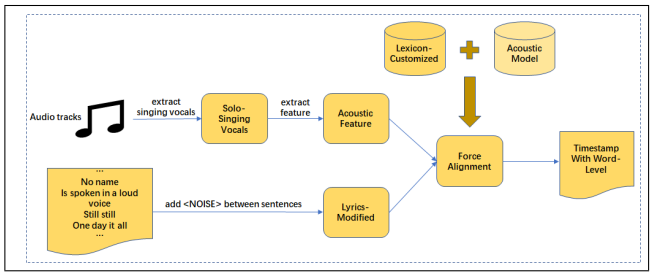

Figure 2. Diagram of automatic lyrics-to-audio alignment at predicting stage.

3. EXPERIMENT

We used both word-level and sentence-level annotated music as training data from the TME music library. We evaluated the performance of the lyrics alignment system with the DALI [6] dataset which is a relatively large polyphonic lyrics annotated dataset from the MIREX website. The DALI dataet contains 4,018 English and 1340 other language audio tracks. We adapt the lyrics-to-audio alignment evaluation metrics according to the MIREX Lyrics-to-audio challenge [7].

	Mono	Trip	SAT	CHAIN
Mean AAE	1.093	0.996	0.677	**0.442**
Median AAE	0.239	0.297	**0.125**	0.129
PCS	0.486	0.523	**0.559**	0.557
PCETW(0.3s)	0.722	0.785	0.842	**0.853**
PCETW(0.5s)	0.786	0.829	0.878	**0.896**
PCETW(1.0s)	6.61	0.879	0.913	**0.933**

Table 1. Mean/Median average absolute error (AAE) , percentage of correct segments (PCS) and Percentage of correct estimates according to a tolerance window (PCETW) for lyrics-to-audio alignment systems using mono-phone, trip-phone, singer adapted of trip-phone GMM-HMM (SAT, ZWZL1) and basic Chain acoustic(CHAIN, ZWZL3) model after applying spleeter as audio source separation algorithm.

After we compared the performance of our 4 models as illustrated in Table 1, we choose the SAT and CHAIN models as our submissions to the Automatic Lyrics-to-Audio Alignment in MIREX 2020 (subtask 2 of 2020:Lyrics Transcription) , named ZWZL1 , ZWZL2 and ZWZL3 respectively. ZWZL2 and ZWZL3 are actually the similar CHAIN models except for internal network while the former has a more complex network structure which adds six CNN layers before TDNN.

4. CONCLUSIONS

We developed a system to align the lyrical content at word-level with the corresponding solo-singing vocal in polyphonic music automatically. We exploited this system to generate LRC and QRC lyrics timestamp file which helped to to build an AI-empowered music library. We also compared the performance of different acoustic models trained by Kaldi. Finally, we demonstrate that the SAT and CHAIN acoustic models perform better than other models compared in this paper.

5. REFERENCES

[1] C. Gupta, E. Yılmaz and H. Li. "Automatic Lyrics Alignment and Transcription in Polyphonic Music: Does Background Music Help?," *CASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 496-500, 2000.

[2] https://github.com/deezer/spleeter.

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

[4] https://github.com/cmusphinx/g2p-seq2seq

[5] B. Sharma, C. Gupta, H. Li and Y. Wang. "Automatic Lyrics-to-audio Alignment on Polyphonic Music Using Singing-adapted Acoustic Models," *CASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pp. 396-400, 2019.

[6] 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," *Proc. ISMIR*, 2018.

[7] https://github.com/georgid/AlignmentEvaluation