Study On End-to-End Lyrics Alignment Training

Seri Lee
Seoul National University
sally20921@snu.ac.kr

***Abstract*—Time-aligned lyrics can enrich the music listening experience by enabling text-based song retrieval and other applications. Compared to text-to-speech alignment, lyrics alignment remains highly challenging. In this study, a system is presented based on the recent studies on lyrics alignment that predicts character probabilities from raw audio.**

***Index Terms*—Lyric alignment, end-to-end neural networks, CTC training**

I. Introduction

In the last decade, there has been considerable interest in digital music services that display the lyrics of songs that are synchronized with their audio. An automatic lyrics-to-audio alignment system could reduce the huge amount of time and labor required to manually construct such time stamps. The given lyrics-text needs to be temporally aligned to a corresponding song. However, state of the art systems still remains to have low alignment accuracy. In paper [1], the authors present a method employing a multi-scale neural network that predicts character probabilities end-to-end directly from raw audio. In paper [2], the authors discover repetitive acoustic patterns of vowels in the target audio by referencing vowel patterns appearing in lyrics. We will be exploring both techniques and try to come up with the best way to create an end-to-end lyrics alignment network.

II. Preliminaries

A. Voice Separation

A low-rank source separation method based on **robust principal component analysis** is used, using a spectrogram as an input. RPCA-based method generates a binary time-frequency mask for separation of the singing voice. Then the singing voice signal from the processed spectrogram is reconstructed.

B. Feature Extraction

mel-frequency cepstral coefficients (MFCCs) are extracted that can compactly represent the spectral envelopes of vowel features.

C. Text Processing

The goal of the text processing is to generate a vowel sequence matrix from the textual lyrics of each target song. Assume a sequence allowing repetition and that we must pick one vowel each time. The sequence P with time index m, where $m = \{1, 2, \cdots, M\}$ can be denoted as

$$P_m = \{p_1, p_2, \cdots, p_M\} \tag{1}$$

III. Studied Model

A. Acoustic Model

For the acoustic model f, we adapt the architecture of paper [1]. Paper [1] adapts the model of the Wave-U-Net (variant M4). The model of originally proposed for singing voice separation and was designed to model highly non-stationary vocals. Similar to paper [1], our work employs a series of blocks combining a 1D-convolution and a downsampling layer, whose receptive field grows exponentially with the number of layers and thus enables efficient translation of low-level into higher level features.

B. Training

To enable the use of weakly aligned lyrics data, intermediate models are employed to force-align the lyrics to corresponding audio recordings. Frame-level annotations are then generated from the aligned lyrics. The **connectionist temporal classification** (CTC) loss offers an alternative to Viterbi training and is used today in many state-of-the-art speech recognition systems in some form. The CTC loss is essentially a simplified version of the forward-backward procedure used to calculate the posterior marginals of the states in a hidden Markov model. The CTC loss takes the character probability distributions P generated by the acoustic model f to calculate a "soft alignment" between the time slices in P and each character in a given target sequence y, which can be represented as a probability distribution over all possible alignments. To use this loss directly, however, the model would need to make predictions for the entire song x. This creates memory issues, we thus calculate character probabilities for a chunk of audio with a fixed size, and apply the CTC loss with individual lyrical lines as target sequence, using only slices of P corresponding to a time position between the start and end times of the lyrical line.

C. Alignment Procedure

Given a music recording x and corresponding lyrics y, we employ the trained acoustic model in a procedure resembling BViterbi forced-alignment. Assuming x contains T time slices, our goal is to find an aligned sequence with maximum probability under the acoustic model predictions P.

IV. Experiments

A. Dataset

For training, we use an internal dataset comprised of 44,232 songs with English lyrics and varied Western genres such as Pop, Rock and Hip-Hop. The songs are annotated with start and end times of each lyrical line. We use 39232 songs for training and 5000 songs for validation.

B. Hyper-parameters

We optimize the CTC loss using the ADAM Optimizer and a batch size of 32.

C. Alignment Results

As we directly output character probabilities and not phonemes as commonly done, we do not have to convert the lyrics into phonemes using a pronunciation dictionary. Since these are usually built for speech and assume that every word has exactly one pronunciation, they are less suitable for lyrics due to t he way pronunciation is often extensively varied in singing voice and since they do not contain rules for vocalizations such as "aah" and "ooh".

V. Conclusion

We proposed a modified architecture that employs learnt multi-scale representations to predict character probabilities directly from the waveform of music. In contrast to most existing systems, the system can be trained end-to-end to avoid complex non-optimized component inter-dependencies and requires only weak.

References

[1] Stoller, Daniel, Simon Durand, and Sebastian Ewert. "End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[2] Chang, Sungkyun, and Kyogu Lee. "Lyrics-to-audio alignment by unsupervised discovery of repetitive patterns in vowel acoustics." IEEE Access 5 (2017): 16635-16648.