MIREX2018: Lyrics-to-Audio Alignment for Instrument-Accompanied Singings

Chung-Che Wang
KKBOX Inc.
chungchewang@kkbox.com

ABSTRACT

This extended abstract presents our submission to the MIREX 2018 Automatic Lyrics-to-Audio Alignment subtask 2, which uses instrument-accompanied singing to train models for this subtask.

1. INTRODUCTION

Automatic lyrics-to-audio alignment is useful in many applications, such as scrolling lyrics and karaoke lyrics display. Currently most of the methods use clean speech or a cappella singing to train the model, and vocal separation is usually needed for such methods [1]. In this work, we try to use instrument-accompanied singing for training the model, and perform the alignment without using source separation.

2. LYRICS-TO-AUDIO ALIGNMENT

2.1 Data Collection

KKBOX is the leading music streaming service in Asia, holding over 45 million songs with varieties of languages and genres. There are also varieties of user-generated content in KKBOX. For instance, a user can submit the singing lyrics with time alignment information (i.e. scrolling lyrics), which will be only briefly reviewed for trivial errors like malicious content before being available for all users.

In this work, 7,300 English songs which having scrolling lyrics were randomly selected from KKBOX's music library for training acoustic models. We did check the selected songs using the announced track information of Mauch's dataset used in MIREX 2017, and nothing with same artist name and song name were found.

2.2 Method

The first step is to segment the audio files according to the position of blank lines in lyrics. This step was found being very useful in our preliminary experiments. Second, since pre-trained acoustic models are not available, we transfer the words of lyrics to phonemes directly using a dictionary instead of using pre-trained models to perform alignment. Third, a vocal detection module, which was also trained using scrolling lyrics, is added to our alignment system. This module is used to cut the non-vocal part out. Finally, the acoustic models were trained by using HTK [2]. Two different kinds of models are used in our submissions for instrument-accompanied singings: mono-phone and right-context-dependent (RCD) bi-phone.

2.3 Experiments

We tested our model using 13 songs from KKBOX's music library, which have same artist name and song name as those in Mauch's dataset. The results are shown in Table 1 together with the training parameter (number of Gaussian mixtures), where "absolute error" is the absolute error of sentences' beginning, and "ratio of overlap" is at sentence level (due to the fact that word level annotation is not available, we can only perform sentence level evaluation). From these results, we choose mono-phone and bi-phone acoustic models using 24 Gaussian mixtures as our submissions for subtask 2 of Automatic Lyrics-to-Audio Alignment in MIREX 2018, named CW2 and CW3, respectively.

Model settings	Absolute error (sec)	Ratio of overlap (%)
Mono-phone 16 Mixtures	9.65	47.80
Mono-phone 24 Mixtures	8.60	51.43
RCD bi-phone 16 Mixtures	12.51	45.37
RCD bi-phone 24 Mixtures	8.95	48.75

Table 1. Table captions should be placed below the table.

3. REFERENCES

[1] 2017:Automatic Lyrics-to-Audio Alignment Results. Available: http://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results.

[2] HTK Web-Site. Available: http://htk.eng.cam.ac.uk/