MIREX 2020: LYRICS-TO-AUDIO ALIGNMENT

Chun-Hao Chiu
chunhao.chiu@mirlab.org

ABSTRACT

We describe our methods that we have submitted for the MIREX 2020 task of Lyrics Transcription (Lyrics-to-Audio Alignment). Given lyric and audio of a song, lyrics-to-audio alignment's goal is to automatically detect the time boundaries of each word in lyrics. The main idea of our submission is to use a trained SVS (singing voice separation) model to obtain the vocal of given song to mitigate the influence of background music. After the SVS preprocess step, we apply forced alignment to align lyrics using DNN-HMM hybrid model. This abstract describes the details of both preprocessing and acoustic model training.

datasets	# of songs	# of line boundaries
DALI [1]	3919	180,034

Table 1. dataset description

1. DATASETS

The Dataset we use to train our acoustic model is shown as table 1. The DALI dataset consists of 3919 English songs, total 180034 lines with line-level boundary. We trimmed all songs into sentence-wise, then training our model using trimmed songs.

2. SINGING VOICE SEPARATION

We trained a CNN model to perform singing voice separation. Our model architecture is shown as Figure 1. In testing phase, the input audio is fed into the network, and followed by a series of convolution and transpose convolution. The output of the model is the vocal of the input song without background instrumental music.

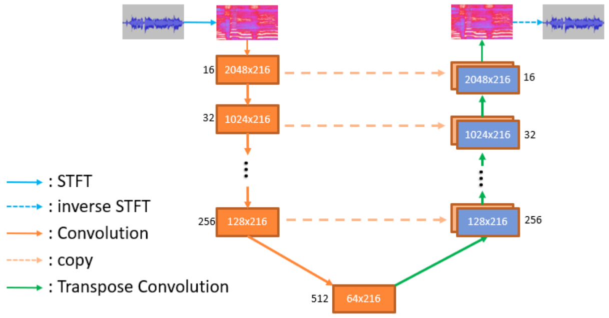

Figure 1. SVS model architecture

3. TRAINING OF ACOUSTIC MODEL

We follow the Kaldi [2] training procedure to train our acoustic model, the training process is shown as Figure 2.

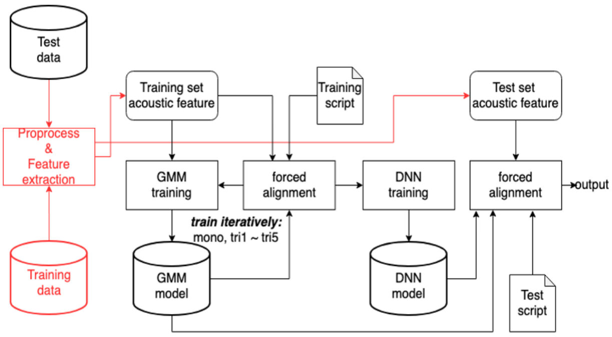

Figure 2. acoustic model training

Training data are first fed into SVS network to obtain pure vocal audio. In the training procedure, we first train the traditional GMM model, from mono phone model training to speaker adaptive training. Then we use the tri5 model to align training data, and make the alignment result as DNN model's ground truth. Finally, we train the DNN model iteratively until it convergence.

In testing phase, we also extract vocal from testing data using SVS model, and use the trained DNN model to align testing data to obtain the alignment result.

4. REFERENCES

[1] 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.

[2] 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.