HYBRID FEATURES FOR MUSIC AND SPEECH DETECTION

Minsuk Choi
KAIST
minsukchoi@kaist.ac.kr

Jongpil Lee
KAIST
richter@kaist.ac.kr

Juhan Nam
KAIST
juhannam@kaist.ac.kr

ABSTRACT

This submission aims to show how learned music and speech features affect the performance on music and/or speech detection tasks. The learned features are extracted from a music model and a speech model in a transfer learning setting. The music model is trained to perform auto-tagging and the speech model is to recognize speech commands. The learning models are constructed using the SampleCNN architecture which showed powerful performances on both music auto-tagging and speech command recognition tasks recently. Engineered audio features are also utilized to add general audio characteristics. Finally, we concatenate them and put the hybrid features into a classifier to make final predictions for the music and speech detection tasks in MIREX 2018.

1. INTRODUCTION

Music and speech are often handled separately in the audio domain because of their different signal characteristics [3]. However, many of previous music and speech detection systems did not distinguish them and constructed the system using general hand-engineered audio features. In this work, we combine the general audio features with two types of learned features. The general audio features include spectral entropy, chromatic spectral entropy, and Mel-Frequency Cepstral Coefficient (MFCC). The learned features are extracted from a pre-trained music model and a speech model in a transfer learning setting. The speech model was trained for classifying speech commands to capture specific speech characteristics. On the other hand, the music model was trained for music auto-tagging. We expect that the combination of hand-engineered general audio features and data-driven learned features can help distinguishing between music and speech.

2. HYBRID FEATURES

2.1 General Audio Features (GAF)

Both spectral entropy and chromatic spectral entropy represent the unpredictability of the spectral behavior of the sound. Music is expected to have more predictable and regular spectrum, while speech is more unpredictable and irregular spectral distributions. Spectral entropy is calculated from the probability density function of a power spectrum over frequency bins for each frame. Chromatic spectral entropy is obtained from the probability density function of a mel-spectrogram. MFCC effectively depicts the overall spectral behavior in mel-frequency scale, which is a akin to the human auditory perception. 12 Mel-frequency bins are used for both chromatic spectral entropy and MFCC. Also, the delta features of spectral entropy, chromatic spectral entropy and MFCC are additionally used to represent their temporal dynamics.

2.2 Music Feature (MF)

The music feature is extracted from a pre-trained SampleCNN-SE-Multi model [3]. The model was trained for music auto-tagging task using the MagnaTagATune dataset [2]. The number of filters of each layer is set to be 32 and total 9 layers (with 3-sized 1D filters) are used for the 22050 input samples. Finally, the features from the last three hidden layer are utilized as a feature vector which has 96 dimensions.

2.3 Speech Feature (SF)

The same architecture was used for speech features. The model was trained for recognizing speech commands including 31 classes and *silence* [4]. The same feature extraction process as in the music model was also used for the speech model.

3. EXPERIMENTS

3.1 Dataset

We used the Muspeak dataset from MIREX 2015 to extract features given the pre-trained models [1]. The hybrid features are extracted from audio frame-wise. A single frame covers 220 samples, which corresponds to 10ms for the audio with 22050Hz sample rate. The label data is rearranged from the time unit to the frame unit.

3.2 Training

We used multilayer perceptron(MLP) as a primary classifier model. The MLP consists of 2 hidden layers with 50 dimensions each. The MLP takes the concatenated hybrid features of single frame as inputs, and makes class-wise predictions for each frame as outputs.

3.3 Smoothing

The output of frame-wise predictions are smoothed for stable estimation. The sequence of predictions is smoothed by a majority-voting filter in two steps. First, the new frame sequence is obtained by windowing the existing prediction sequence. Each frame of the new sequence is classified as the major class of 100 frames within the window on existing sequence. Then, the new sequence is smoothed again with the majority-voting filter. The class of the frames in the window is fixed to the major class of that window.

4. SUBMISSIONS

Here, we organize three combinations of GAF, MF and SF to compare the detection performance on different tasks.

- GAF+MF: MF from CNN trained with the MTAT dataset is expected to have higher activation for the frame with music. The performance on music detection, which is for task 1 and 3, can be advanced with MF features.
- GAF+SF: SF from CNN trained with the speech commands dataset is expected to have higher activation for the frame with speech. It may advance the performance on speech detection, which is for task 2 and 3.
- GAF+MF+SF: combination of MF and SF is expected to represent the overall activation on music and speech detection, which is for task 1, 2 and 3.

5. REFERENCES

[1] MuSpeak speech and music detection dataset. *http://mirg.city.ac.uk/datasets/muspeak/muspeak-mirex2015-detection-examples.zip*, 2015.

[2] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *International Society of Music Information Retrieval Conference (ISMIR)*, 2009.

[3] Jongpil Lee, Taejun Kim, Jiyoung Park, and Juhan Nam. Raw waveform-based audio classification using sample-level cnn architectures. In *NIPS, Machine Learning for Audio Signal Processing Workshop (ML4Audio)*, 2017.

[4] Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.