MUSIC AND/OR SPEECH DETECTION METHODS FOR MIREX 2018

Byeong-Yong Jang[1] **Woon-Haeng Heo**[1] **Jung_Hyun Kim**[2] **Oh-Wook Kwon**[1]
[1]Chungbuk National University, South Korea
[2]Electronics Telecommunications Research Institute, South Korea
[1]{byjang, whheo, owkwon}@cbnu.ac.kr, [2]bonobono@etri.re.kr

ABSTRACT

In this submission we propose and introduce methods for the MIREX 2018 Music and/or Speech Detection. For music detection, we propose to use CNN with mel-scale kernels. The parameters of the mel-scale kernels in CNN are learned from spectrograms of a mixed data set. For speech detection, we introduce RNN with bidirectional GRU whose model is trained by mel-scale spectrograms of a broadcast data set. Finally, frame-based classification results are smoothed with a median filter to produce event-level segments.

1. INTRODUCTION

The music and/or speech task suggested in MIREX 2018 consists of four sub-tasks: Music detection, speech detection, music and speech detection, and music relative loudness estimation. Among these sub-tasks, we implement music detection, speech detection, and music and speech detection using deep neural networks. Music signals have less temporal variation than speech signals, and have an energy distribution mostly concentrated in the specific frequency region. Considering this fact, we use Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to detect music and speech segments, respectively.

2. PREPROCESSING

First, we downsample the input wave of 22,050 Hz sampling rate to 16 kHz because we want to utilize the existing abundant music and speech data of 16 kHz sampling rate. Then, we compute log power coefficients (spectrogram) of short time Fourier transform (STFT) with the 25ms window size, 10ms shift size, and 512-point FFT (fast Fourier transform). The spectrogram is used as CNN input for music detection. In addition, the mel-scale spectrogram is obtained by multiplying the spectrogram by 64 mel-filters [1] and is used as input of RNN for speech detection. The final dimension of each feature vector of CNN and RNN is 257×101 and 64×101, respectively, by splicing 50 frames on either side.

3. DEEP NEURAL NETWORK STRUCTURES

3.1 CNN for music detection

We use a CNN with a mel-scale convolutional layer and 3 convolutional layers appended with 2 fully connected feed-forward layers and a softmax layer for class output. The mel-scale convolutional layer uses a mel-scale kernel. That is, the column size of kernel is equal to the number of FFT points contained in each mel-scale bin. Figure 1 compares the convolutional layers with fixed-size kernels (a) and mel-scale kernels (b). We expect that the mel-scale convolutional layer can extract more robust feature from the spectrogram for music detection. The row size of a mel-scale kernel is 5 with stride 1, the number of filters in the mel-scale convolutional layer is 3, and the activation function is hyperbolic tangent. The subsequent three conventional convolutional layers of the CNN have 32, 64, and 128 filters, respectively. Each convolutional layer has a 3x3 kernel with stride 1, ReLU (rectified linear unit) [2] activation function, and 2x2 max pooling with stride 2. The detailed structure of the CNN is shown in Figure 2. The CNN was trained for 50 epochs with cross entropy loss function, Adam optimizer, mini-batch size of 300, learning rate of 0.001, and dropout probability of 0.4.

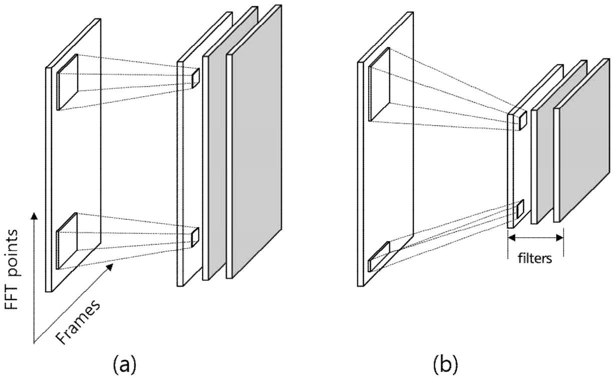

Figure 1. Convolutional layers with fixed-size kernels (a) and mel-scale kernels (b)

3.2 RNN for speech detection

We use the RNN which is composed of bidirectional GRU (gated recurrent unit) [3] for speech detection. The RNN consists of three bidirectional GRU layers, and the number of hidden nodes of each GRU layer is 1024. The detailed structure of the RNN is shown in Figure 3. The RNN was trained for 5 epochs with cross entropy loss function, Adam optimizer, mini-batch size of 300, learning rate of 0.0001, and dropout probability of 0.4.

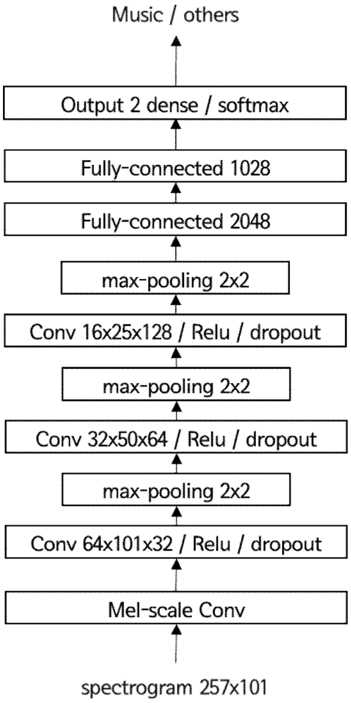

Figure 2. The network structure of CNN

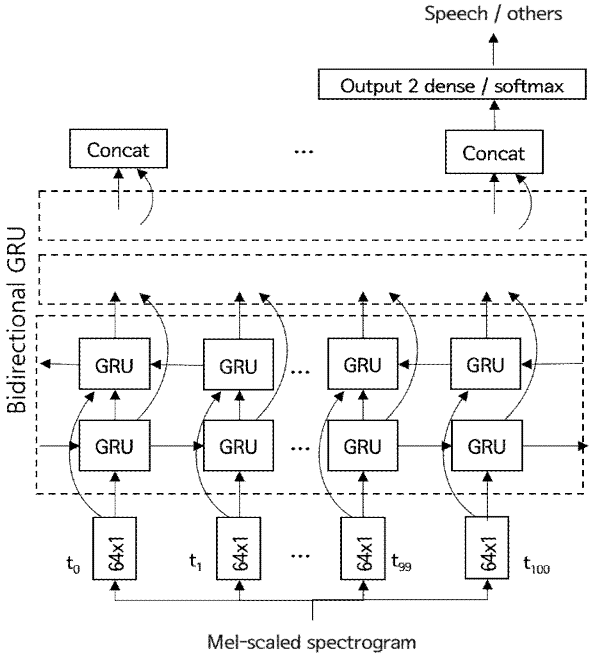

Figure 3. The network structure of RNN

3.3 Music and speech detection

The results for the music and speech detection sub-task are obtained by integrating the classification results from CNN and RNN instead of implementing a new classifier.

4. POST PROCESSING

We apply a median filter to the frame-by-frame detection results in order to obtain the smoothed segmentation results. The size of median filter is 5 second and 1 second for music and speech detection, respectively.

5. TRAINING DATA SETS

We used two sets of data to train the model for music and/or speech detection. The first data set is the mixed data set. This data set was created by mixing music, speech and noise; we used 25h of library music (song, classic, instrument, etc.), 25h of the librivox (speech) in MUSAN database [4], and 2h 46m of the ESC-50 database (noise) [5]. The second data set is 35h 47m of broadcast data (drama, documentary, news, kids, and so on) in the Spanish, British English, and German languages. The broadcast data were manually tagged for music and speech sections. We used the mixed data set to train CNN models for music detection, and used the broadcast data set to train RNN models for speech detection.

6. SUBMITTED PROGRAMS

We submitted three programs. The first one is a music detection program whose parameters are learned by using mel-scale spectrogram without a mel-scale convolutional layer. The second is a music detection program with the proposed mel-scale convolutional layer. Both the first and second programs were trained by using only the mixed data set. The third is a speech detection program with bidirectional GRU, which was trained by using only the broadcast data set.

7. ACKNOWLEDGMENTS

This research project was supported by Ministry of Culture, Sports and Tourism (MCST) and from Korea Copyright Commission in 2018. [2018-micro-9500, Intelligent Micro-Identification Technology for Music and Video Monitoring]

8. REFERENCES

[1] S. Umesh, L. Cohen, and D. Nelson: "Fitting the mel scale," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 217-220, 1999.

[2] V. Nair and G. E. Hinton: "Rectified linear units improve restricted Boltzmann machines," *Proc. International Conference on Machine Learning (ICML-10),* pp. 807-814, 2010.

[3] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio: "Learning phrase

representations using RNN encoder-decoder for statistical machine translation," *Arxiv preprint arXiv:1406.1078*, 2014.

[4] D. Snyder, G. Chen, and D. Povey: "Musan: A music, speech, and noise corpus," *Arxiv preprint arXiv:1510.08484*, 2015.

[5] K. J. Piczak: "ESC: Dataset for environmental sound classification," *Proc. ACM International Conference on Multimedia*, pp. 1015-1018, 2015.