SOUND EVENT DETECTION IN MULTICHANNEL AUDIO LSTM NETWORK

Jianchao Zhou

Peking University
Institute of Computer Science and Technology
Beijing China

ABSTRACT

In this paper, a polyphonic sound event detection system is proposed. This system uses log mel-band energy features with long short term memory (LSTM) recurrent neural network. Human listeners have been successfully recognizing overlapping sound events by two ears. Motivated by that we propose to extend the system to use multichannel audio data. The original stereo (multichannel) audio signal has two channels, we construct three different channel data and use different fusion strategies to extend our system. Experiments show that our system achieved superior performances compared with the baselines.

Index Terms— polyphonic sound event detection, log mel-band energy, LSTM, multichannel

1. INTRODUCTION

Polyphonic sound event detection is the task of detecting overlapped sound events from audio stream. Deep neural networks have shown promising results for polyphonic sound event detection tasks [1, 2, 3]. In this project, we propose a polyphonic sound event detection system by employing LSTM network with log mel-band energy features as input. Motivated by human we propose to extend the system to use multichannel audio data [4]. Since the original audio files are stereo signals, we construct three different channel data called right channel, mean channel and diff channel to extend our system. Sound event detection task is carried out for all three channels and different fusion strategies are used.

2. SYSTEM OVERVIEW

We have two subsystems called MC-LSTM-1 and MC-LSTM-2. MC-LSTM-1 subsystem do fusion work in feature stage while MC-LSTM-1 subsystem do fusion work in prediction stage. Detailed explanations are as follow.

2.1. Feature Extraction

As a pre-processing step for the feature extraction, the recordings are divided into frames with 40 ms duration and 50% overlap. The 40 log mel-filter bank coefficients are then computed. The feature vectors of each recording is normalized to zero mean and unit variance, and the scaling parameters are saved for normalizing the test feature vectors.

For each time frame a 6 dimensional target output vector is obtained. Each element of the vector is a binary variable encoding whether the event is present in a given time frame.

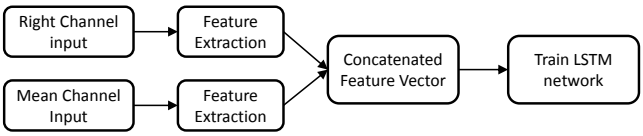

Figure 1: MC-LSTM-1 subsystem overview

2.2. LSTM Networks Configurations

We use a multi-label LSTM with a single hidden layer having 32 units. The output layer has one neuron for each class. The network is trained by back propagation through time (BPTT) [5] using binary cross-entropy as loss function and Adam optimizer [6]. Early stopping is used to reduce over-fitting, the training is halted if the validation loss does not decrease for 5 epochs. The total training epochs are 20 and the best model is saved during training.

To train the LSTM network, the feature vector sequence of each recording is further split into smaller sequences of length 50 with hop length of 10 frames. Correspondingly we do the same work on target output vector sequences. While at test time stage, feature vector sequence of each recording is split into nonoverlapping sequences of 50 frames, and threshold the outputs with a fixed threshold of 0.5, i.e., we mark an event is active if the posterior in the output layer of network is greater than 0.5 and otherwise inactive.

2.3. Multichannel Data

The original recordings are stereo signals. Besides left channel and right channel, we construct two more channels. Mean channel is obtained by averaging the two channels and diff channel is obtained by calculating the difference between the two channels. Since the left channel data performs poor, it is abandoned in our system. Feature extraction and LSTM network training are carried out for all the other three channels.

2.4. Subsystems

The MC-LSTM-1 subsystem shown in Figure 1 extracted log mel-filter bank feature vectors for both right channel and diff channel. The two 40 dimensional feature vectors at each time are concatenated together to form an 80 dimensional feature vector. Then the concatenated feature vectors is used for training LSTM network. Obviously the input layer of LSTM has 80 units. The trained LSTM is used to predict sound event.

In the MC-LSTM-2 subsystem shown in Figure 2, feature extraction and LSTM training are carried out for right channel, mean

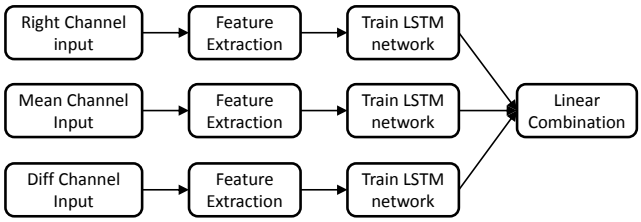

Figure 2: MC-LSTM-2 subsystem overview

	ER	F-score
overall	0.66	54.5
brakes squeaking	1.01	0.0
car	0.55	72.0
children	1.17	10.1
large vehicle	1.08	43.7
people speaking	1.12	17.5
people walking	0.74	58.7

Table 1: performance of MC-LSTM-1 subsystem on development set

channel and diff channel respectively. We obtain the final prediction result by doing linear combination of the three prediction results. The coefficients of right channel LSTM result, mean channel LSTM result and diff channel LSTM result are 0.3, 0.3, 0.4. The three LSTM networks in this subsystem all have an input layer of 40 units.

2.5. Post Processing

In order to smoothen the outputs in the testing stage, a median filter is applied to the results of the system output which is same as the baseline DNN model does.

3. EXPERIMENT RESULT

The performances of MC-LSTM-1 and MC-LSTM-2 subsystems on development set are shown in Table 1 and Table 2.

4. CONCLUSION

In this paper, a polyphonic sound event detection system is proposed, which employs log mel-band energy features with LSTM network. We construct three different channel data and use different fusion strategies to extend our system. Experiments show that our system achieved superior performances compared with the baselines.

	ER	F-score
overall	0.64	54.4
brakes squeaking	1.01	0.0
car	0.56	71.7
children	1.08	0.0
large vehicle	0.95	42.8
people speaking	1.04	10.8
people walking	0.71	58.8

Table 2: performance of MC-LSTM-2 subsystem on development set

5. REFERENCES

[1] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Multi-label vs. combined single-label sound event detection with deep neural networks," in *Signal Processing Conference*, 2015, pp. 2551–2555.

[2] O. Gencoglu, T. Virtanen, and H. Huttunen, "Recognition of acoustic events using deep neural networks," in *Signal Processing Conference*, 2010, pp. 506–510.

[3] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *International Joint Conference on Neural Networks*, 2015, pp. 1–7.

[4] S. Adavanne, G. Parascandolo, P. Pertil, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," 2017.

[5] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[6] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.