A REPORT ON SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Yifeng Lin，Zhisheng Wang

ABSTRACT

In this paper, we make a little change to the baseline of Sound Event Localization and Detection. We add a Gaussian-noise on the input data to find if noise would help us improve the neutral network. Sound event detection is performed stacked convolutional and recurrent neural network and the evaluation is reported using standard metrics of error rate and F-score. The studied neutral network with noise on input data are seen to consistently perform equal to the origin baseline with respect to error rate metric.

Index Terms—Gaussian-noise, convolution, recurrent, neutral network

1. INTRODUCTION

Sound event detection (SED) is the task of recognizing the sound events and their respective temporal start and end time in a recording. Sound events in real life do not always occur in isolation, but tend to considerably overlap with each other. Recognizing such overlapping sound events is referred as polyphonic SED. Applications of such polyphonic SED are numerous. Recognizing sound events like alarm and glass breaking can be used for surveillance. Environmental sound event detection can be used for monitoring biodiversity studies. Further, SED can be used for automatically annotating audio datasets, and the sound events recognized can be used as a query for retrieval. Polyphonic SED using mono-channel audio has been studied extensively. Different approaches have been proposed using supervised classifiers like Gaussian mixture model - hidden Markov model, fully-connected networks, convolutional neural networks (CNN) and recurrent neural networks (RNN). More recently, the state of the art method for polyphonic SED was proposed in and evaluated on multiple private and publicly available datasets. They used log mel-band energies along with a convolutional recurrent neural network (CRNN) architecture as their method. Recognizing overlapping sound events using mono-channel audio is a difficult task. These overlapping sound events can potentially be recognized better with multichannel audio. We train the network method with features extracted from the publicly available TUT Sound Events 2019 dataset and present the results. The feature extraction and neural network used is described in section 2. The dataset creation, evaluation metrics and procedure are explained in section 3. Finally, the results and discussion are presented in section 4.

2. METHOD

The block diagram of the proposed DOAnet is presented in Figure 1. The DOAnet takes multichannel audio as the input and first extracts the spectrograms of all the channels. The phases and the magnitudes of the spectrograms are mapped using a CRNN to two outputs sequentially. The first output, spatial pseudo-spectrum (SPS) is generated as a regression task, followed by the DOA estimates as a classification task. The DOA is defined by the azimuth ϕ and elevation λ with respect to the microphone and the SPS is the intensity of sound along the DOA given by S(ϕ ; λ). In this paper, we use discrete ϕ and λ by uniformly sampling the 2-D polar coordinate space, with a resolution of 10 degrees in both azimuth and elevation, resulting in 614 sampled directions. The SPS is computed at each sampled direction, whereas, a subset of 432 directions is used for DOA, where the elevations are limited between -60 and 60 degrees.

A. Feature extraction

The spectrogram is calculated for each of the audio channels whose sampling frequencies are 44100 Hz. A 2048-point discrete Fourier transform (DFT) is calculated on Hamming windows of 40 ms with 50 % overlap. We keep 1024 values of the DFT corresponding to the positive frequencies, without the zeroth bin. L frames of features, each containing 1024 magnitude and phase values of the DFT extracted in all the C channels, are stacked in a L ×1024 × 2C 3-D tensor and used as the input to the proposed neural network. The 2C dimension results from ordering the magnitude component of all channels first, followed by the phase. We use a sequence length L of 100 (= 2 s) in this work.

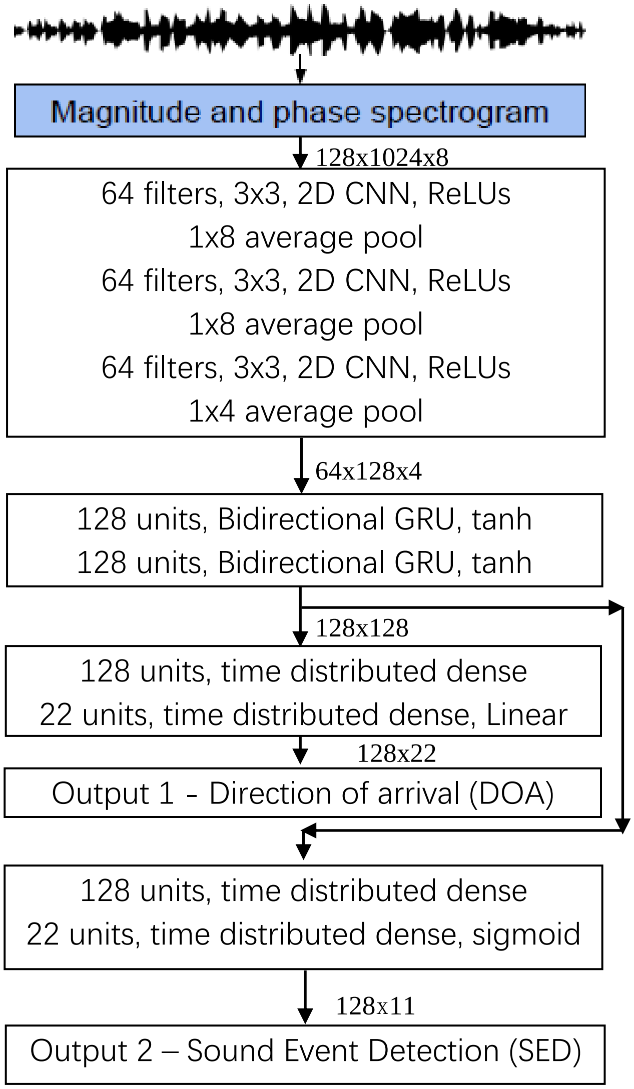
Magnitude and phase spectrogram
128x1024x8
64 filters, 3x3, 2D CNN, ReLUs
1x8 average pool
64 filters, 3x3, 2D CNN, ReLUs
1x8 average pool
64 filters, 3x3, 2D CNN, ReLUs
1x4 average pool
64x128x4
128 units, Bidirectional GRU, tanh
128 units, Bidirectional GRU, tanh
128x128
128 units, time distributed dense
22 units, time distributed dense, Linear
128x22
Output 1 - Direction of arrival (DOA)
128 units, time distributed dense
22 units, time distributed dense, sigmoid
128x11
Output 2 – Sound Event Detection (SED)

Fig. 1. DOAnet - neural network architecture for direction of arrivalestimation of multiple sound sources. SEDnet - neural network architecture for sound event direction

B. *Direction of arrival estimation network (DOAnet)*

Before inputting the features into Neutral Network, we add a Gaussian-noise with the stddev of 0.02. However, we found that the network converged slowly and badly. Local shift-invariant features are extracted from the input spectrogram tensor (L × 1024 × 2C dimension) using CNN layers. In every CNN layer, the intra-channel time-frequency features are processed using a receptive field of 3 × 3, rectified linear unit (ReLU) activation and pad zeros to the resulting activation map to keep the output dimension equal to input. Batch normalization and max-pooling operation along frequency axis are performed after every CNN layer to reduce the final dimension to L × 2 × N_C, where N_C is the number of CNN filters in the last CNN layer. The CNN activations are reshaped to L×2N_C keeping the time axis length unchanged and fed to RNN layers in order to learn temporal structure. Specifically, the bi-directional gated recurrent units (GRU) with tanh activation are used. Further, the RNN output is mapped to the first output, the SPS, in regression manner using FC layers with linear activation.

C. *Sound Event Detection network(SEDnet)*

The output activation from CNN is further reshaped to a T frame sequence of length 2P feature vectors and fed to bidirectional RNN layers which are used to learn the temporal context information from the CNN output activations. Specifically, Q nodes of gated recurrent units (GRU) are used in each layer with tanh activations. This is followed by two branches of FC layers in parallel, one each for SED and DOA estimation. The FC layers share weights across time steps. The first FC layer in both the branches contains R nodes each with linear activation. The last FC layer in the SED branch consists of N nodes with sigmoid activation, each corresponding to one of the N sound event classes to be detected. The use of sigmoid activation enables multiple classes to be active simultaneously. The last FC layer in the DOA branch consists of 3N nodes with tanh activation, where each of the N sound event classes is represented by 3 nodes corresponding to the sound event location in x, y, and z, respectively. For a DOA estimate on a unit sphere centered at the origin, the range of location along each axes is [-1; 1], thus we use the tanh activation for these regressors to keep the output of the network in a similar range. We refer to the above architecture as SELDnet. The SED output of the SELDnet is in the continuous range of [0; 1] for each class, while the DOA output is in the continuous range of [-1; 1] for each axes of the sound class location. A sound event is said to be active, and its respective DOA estimate is chosen if the SED output exceeds the threshold of 0.5 as shown in Figure 1.

3. RESULT

The result of our work on task3 Sound Event Localization and Detection is shown in the table below:

SED ER	1.025
F1-SCORE	2.49
DOA ER	15.43
FRAME RECALL	30.7

Maybe the dropout rate can explain the poor result. We add dropout layers with a dropout rate of 0.2 to 0.3 after every CNN layer and the dropout rate of GRU layer also set to 0.2.

4. REFERENCES

[1] http://dcase.community/workshop2019/.

[2] http://www.ieee.org/web/publications/rights/copyrightmain.html

[3] E. G. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustic Holography,* London, UK: Academic Press, 1999.

[4] C. D. Jones, A. B. Smith, and E. F. Roberts, “A sample paper in conference proceedings,” in *Proc. IEEE ICASSP*, 2003, vol. II, pp. 803-806.

[5] A. B. Smith, C. D. Jones, and E. F. Roberts, “A sample paper in journals,” *IEEE Trans. Signal Process.*, vol. 62, pp. 291-294, Jan. 2000.

[6] Sharath Adavanne, Archontis Politis, Tuomas Virtanen, Direction of Arrival Estimation for Multiple Sound Sources Using Convolutional Recurrent Neural Network, Tampere University of Technology and Aalto University, Finland

[7] A. B. Smith, C. D. Jones, and E. F. Roberts, “A sample paper in journals,” *IEEE Trans. Signal Process.*, vol. 62, pp. 291-294, Jan. 2000.

[8] Sharath Adavanne, Tuomas Virtanen, A Report on Sound Event Detection with Different Binaural Features, Department of Signal Processing , Tampere University of Technology.

[9] Sharath Adavanne, Archontis Politis, Joonas Nikunen, Tuomas Virtanen, Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks, IEEE