SOUND EVENT DETECTION AND LOCALIZATION USING CRNN MODELS

Technical Report

Arunodhayan Sampathkumar, Danny Kowerko

Technische Universität Chemnitz, Juniorprofessur Media Computing, Chemnitz, Germany
{Arunodhayan Sampathkumar, Danny Kowerko}@informatik.tu-chemnitz.de

ABSTRACT

Sound Event Localization and Detection (SELD) requires both spatial and temporal information of sound events that appears in an acoustic event. The sound event localization and detection DCASE2020 task3 developed a strongly labelled dataset consisting of 14 classes. In this research work the existing method from DCASE2019 is used with significant modifications, where this method utilizes logmel features for sound event detection, and uses intensity vector and generalized cross-correlation (GCC) GCC-PHAT features for sound source localization. The Convolutional Recurrent Neural Network (CRNN) is developed that jointly predicts the Sound Event Detection (SED) and Degree of Arrival (DOA) hence minimizing the overlapping problems. The developed model significantly outperformed the baseline system.

Index Terms— Convolutional Recurrent Neural Network (CRNN), Sound Event Localization and Detection (SELD), logmel, intensity vector, GCC-PHAT

1. INTRODUCTION

Sound Event Localization and detection is a combined task of estimating the spatial location of trajectories and further syndicating the textual labels with sounds. Sound Event Localization and Detection (SELD) is a complex task that covers a wide area of research and its application in acoustic monitoring, robots which can employ this method for interaction with the surroundings, smart homes, virtual reality can assist users in visualizing sound events [1].

SELD can be divided into two subtasks named as Sound Event Detection (SED) and Sound Source Localization (SSL). The SED aims at detecting sounds and further syndicating the sound with text labels. The state of the art SED uses different supervised learning methods such as Hidden Markov Model (HMM), Gaussian Mixture Models(GMM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN) and Fully Connected(FC) [1]. The state of the art DCASE 2018 and 2019 achieved better results when stacking CNN, RNN and FC layers consecutively.

The SSL aim is to determine the direction or position of sound sources with respect to the microphone (array). This research paper only deals with the estimation of sound event direction, basically called as Degree of Arrival (DOA) [2]. Most state of the art DOA estimation algorithms were based on parametric methods such as Time of Difference of Arrival (TDOA), Steered Response Power (SRP), Multiple Signal Classification (MUSIC) and signal parameters estimation via rotational invariance technique (ESPRIT) [3]. The state of the art method states that DNN methods perform well when compared to parametric methods [4].

The DOA estimation tracks the single sound source events correctly. Meanwhile in multiple sound source, DOA estimation tracks the respective sound source events without correctly identifying the source. The real-time detection of multiple overlapping sound is called as polyphonic SED [5]. In most state of the art, performing SED and DOA separately results in data syndicating problem between the recognized sound events and DOA estimation. The solution for data syndicating problem is to jointly predict SED and DOA. The author performed by using DNN based method and DCASE2019 baseline by using CRNN method [6].

The DCASE2020 dataset consists of 600 one-minute long sound scene recordings for development and 200 one-minute long sound scene recordings for evaluation. The dataset comprises of 14 classes. Each scene recording is delivered in two spatial recording formats, a microphone array one (MIC) and first-order Ambisonics one (FOA) [7].

The paper is organized as follows, the feature extraction is explained in section 3, followed by the network architecture in section 4 and the results in section 6 which compares our results with the baseline.

2. EXPERIMENTAL SETUP

Our hardware and software configuration is described in the following bullet points.

- Hardware Configuration:
  - Intel Core i7
  - 8 GB RAM
  - Operating system: Ubuntu 18.04 LTS
  - CPU (4 cores @ 2.40 GHz)
  - NVIDIA Titan X with 12 GB graphics memory
- Software:
  1. *Programming Language:*
     - Python 3.7
  2. *Libraries:*
     - CUDA 10.0 tool kit for GPU acceleration
     - CUDnn 7.5.4
     - Tensorflow=1.15.2
     - Keras=2.2.4
  3. *Dependencies:*
     - numpy 1.14.2

- scipy 1.0.0
- python speech features 0.6
- pydub 0.21.0
- python-openCV=4.2.0
- librosa=0.7.2
- matplotlib 2.2.0
- cmake 3.5.1
- cython 0.29.2
- libblas-dev liblapack-dev

3. FEATURE EXTRACTION

The DCASE 2020 task3 provides two formats of TAU Spatial sound events is of 4-channels, 3-dimensional recordings namely First Order Ambisonic (FOA) and tetrahedral microphone array. Each recording is 1 minute long approx. with a sampling rate of 24kHz. We used log-mel space for SED, intensity vector in log-mel space and a GCC with phase transform is used for DOA estimation. The Short Time Fourier Transform (STFT) is set to a sampling rate of 24kHz with window length and hop length 0.04 and 0.02, respectively. The raw audio syndicated is transformed to FOA/MIC channel and then transormed into spectrogram of size 3000×400 as presented in the Figure 1. The log mel space and intensity vectors utilizes FOA input data, whereas GCC-PHAT utilizes the microphone array as input data. This feature exraction method is inspired from the work of two stage sound event localization and detection.

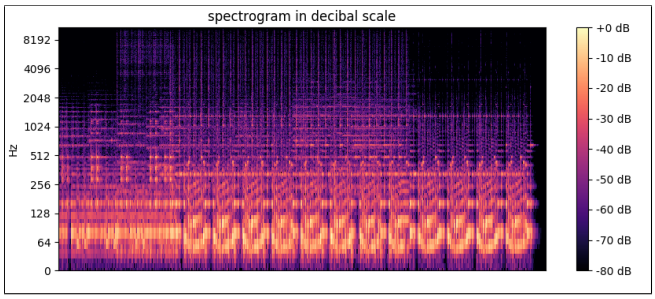

Figure 1: Extracted spectrogram of 60 seconds length consisting of different audio events. For technical details, see text.

4. NETWORK ARCHITECTURES

From state of the art, Convolutional Recurrent Neural Network (CRNN) performs well for SELD when compared with other state of the art literature networks. Here, SELD network consist of two branches, the SED branch and DOA branch, respectively, as shown in the Figure 2 . The shape of the input to the network is $A \times T \times M$ where A represents the feature map, T represents the time of bins and M represents the size of mel bands. This network was inspired from this research work [2]

The network architecture consists of CNN layers with four groups followed by RNN and FC layers. Each CNN groups consists of filters in the increasing order of 64 to 512, the kernel size is 3×3, batch normalization and relu activation function. After 4 CNN groups, the network is fed to a dropout layer of 0.2 followed by bidirectional Gated Recurrent Unit (GRU). The output size of the

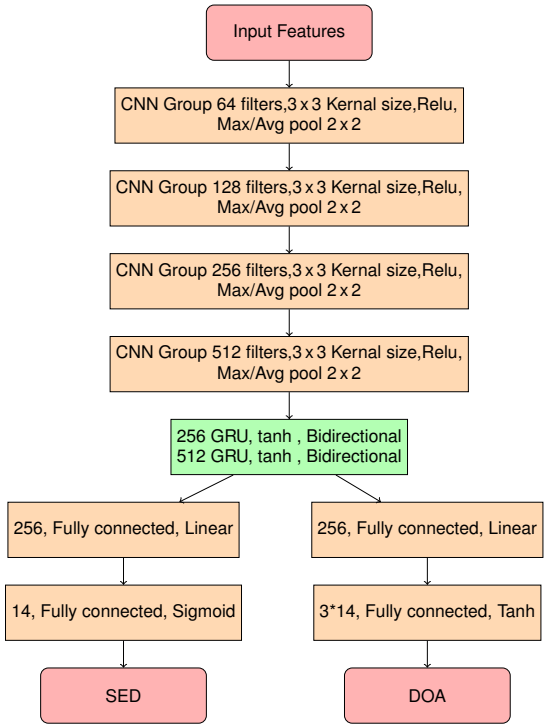

Figure 2: Network Architecture for SELD. For details, see text

GRU is maintained as constant S and fed to a two fully connected layer. The sigmoid activation function is used in one of the FC layer with binary cross entropy to determine the SED output. The tanh activation function is used in another FC layer to determine the DOA, where DOA output is a vector of $c \times 2$ which contains azimuth and elevation angles for c events. The DOA output are masked by SED ground truth during training to determine if the corresponding angles are active. First SED is predicted, fine-tuned and used SED masked with DOA to perform the joint prediction.

5. HYPERPARAMETERS

The input feature is extracted using STFT with a sampling rate of 24kHz, hop length of 0.02 and window length of 0.04. For 4 channels of FOA and MIC, there are 8 channels of log mel features, 3 channels of intensity vector features, and 6 channels of GCC-PHAT features summing up to 17 input channels are fed to the network. The network is trained for 500 epochs with a learning rate of 0.0001 till 120 epochs and degrading slowly until it reaches 500 epochs. The dropout is set to 0.2 after 4 CNN layers and followed by RNN with 0.2 as the recurrent dropout.

6. RESULTS

The results are evaluated based on the evaluation dataset provided. For SED, the segment based error rate (ER) and F-Score is calculated. For DOA, localization error and localization recall is calculated. Lower error rates and higher F-Scores indicate that the model performs better. The figure 3 represents the ground truth and the predicted results. The X axis represents the audio signal which is 60 seconds long. The Y axis for SED represents the number of