# Polyphonic Sound Event Localization and Detection

Abhinav Pavithran-171EC202[1], Nadeem Roshan-171EC226[2], Kushal Jaju-181EC221[3], Rahul Grover-171EC125[4]

***Abstract*— Polyphonic Sound event localization and detection (SELD) is the combined task of identifying the temporal onset and offset of a sound event, tracking the spatial location when active, and further associating a textual label describing the sound event. TAU Spatial Sound Events 2019 - Microphone Array dataset is used which consists of one minute long audio recordings. Sound event detection (SED) is performed as a multi-label multi-class classification task, allowing the network to simultaneously estimate the presence of multiple sound events for each frame. At the second output, direction of arrival (DOA) estimates in the continuous 3D space are obtained as a multi-output regression task, where each sound event class is associated with two regressors that estimate the spherical coordinates azimuth (azi) and elevation (ele) of the DOA on a unit sphere around the microphone.**

***Index Terms*— Polyphonic Acoustic Event Detection(AED), Deep Neural Network (DNN), Sound event localization and detection (SELD), Convolutional Neural Networks (CNN), Directio of Arrival(DOA)**

## I. INTRODUCTION

Sound event detection is a quickly developing research area that aims to analyze and recognize a variety of sounds in urban and natural environments. Compared to sound tagging, event detection also involves estimating the time of occurrence of sounds. Automatic recognition of sound events would have a major impact in a number of applications. For instance, sound indexing and sharing, bio-acoustic scene analysis for animal ecology, smart home automatic audio event recognition (baby cry detection, window break alarm),security surveillance in smart cities.

The task of AED is broadly categorized into two perspectives based on event composition: monophonic AED and polyphonic AED. Detection of at most one simultaneous acoustic event in a given time instance is known as monophonic AED. In a real-time scenario, multiple acoustic events occur at the same time in a overlapped manner unlike the monophonic (isolated) acoustic events. Detection of such multiple overlapped acoustic events is known as polyphonic AED.

Sound source localization, which focuses on identifying the locations of sound sources, on the other hand, has been an active research topic for decades. It plays an important role in applications such as robotic listening, speech enhancement, source separation, and acoustic visualization. Unlike the dominance of neural network-based techniques in SED, DOA is mainly studied using two methods: parametric-based methods and learning-based methods.

Learning-based DOA methods have the advantages of good generalization under different levels of reverberation and noise. They are designed to enable the system to learn the connections between input features and the DOA. There has already been a series of research addressing DOA using deep neural networks. Results show that they are promising and comparable to parametric methods. In this paper we will be exploring the neural networks method for DOA.

In real-world applications, a sound event is always transmitted in one or several directions. Given this fact, it is reasonable to combine sound event detection and localization with not only estimating their respective associated spatial location, but also identifying the type and temporal information of sound. Therefore, it is worthwhile to study them together and investigate the effects and potential connections between them.

A recently-developed system known as SELDnet was used as the baseline system. SELDnet uses magnitude and phase spectrograms as input features and trains the SED and DOA objectives jointly. However, phase spectrograms are hard for neural networks to learn from, and further relationships between SED and DOAE have not been revealed.

The SELD output can automatically describe social and human activities. This description can be used by machines to be context-aware. For instance, robots and humanoids can use SELD for navigation and natural interaction with surroundings. Smart meeting rooms can recognize the active speaker among other sound events, and track their motion with respect to time. This tracked location of the speaker can be employed to enhance speech using beam forming for teleconferencing or automatic speech recognition applications. According to the World Health Organization, 5% of the world's population suffer from hearing disability. With the help of SELDT, we can build assistants that will help these hearing-impaired people to visualize sounds and enable them to interact with the world naturally.

The data association problem occurs especially when the two sub-tasks of sound event detection and sound event localization and tracking are done separately, for a real-life sound scene with overlapping sound events. One of the solutions to overcome this problem is to jointly perform the two sub-tasks of sound event detection and sound event localization and tracking which will be shown in this paper.

## II. LEARNING METHOD

### A. SELDnet architecture

The SELDnet architecture is as shown below. The input is the multichannel audio, from which the phase and magnitude components are extracted and used as separate features. The proposed method takes a sequence of consecutive spectrogram frames as input and predicts all the sound event classes active for each of the input frame along with their respective spatial location, producing the temporal activity and DOA trajectory for each sound event class. In particular, a convolutional recurrent neural network (CRNN) is used to map the frame sequence to the two outputs in parallel. At the first output, sound event detection (SED) is performed as a multi-label multi-class classification task, allowing the network to simultaneously estimate the presence of multiple sound events for each frame. At the second output, direction of arrival (DOA) estimates in the continuous 3D space are obtained as a multi-output regression task, where each sound event class is associated with two regressors that estimate the spherical coordinates azimuth (azi) and elevation (ele) of the DOA on a unit sphere around the microphone.

In the benchmark method, the variables in the image below have the following values, T = 128, M = 2048, C = 4, P = 64, MP1 = MP2 = 8, MP3 = 4, Q = R = 128, N = 11.

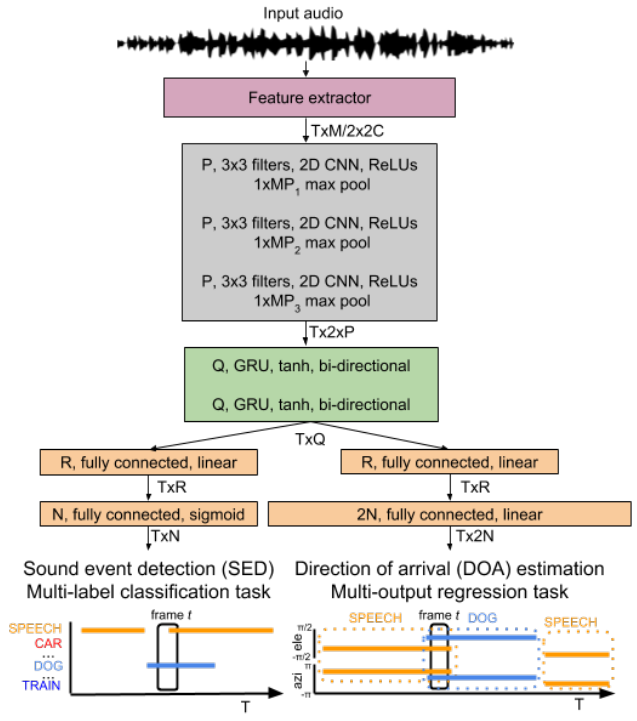


Fig. 1. SELDnet Architecture

### B. Learning Process

The batch feature extraction is a script, that extracts the features, labels, and normalizes the training and test split features for a given dataset.

The parameter script consists of all the training, model, and feature parameters.

The feature class script has routines for labels creation, features extraction and normalization.

The data generator script provides feature/label data in generator mode for training.

The keras model script implements the SELDnet architecture.

The evaluation metrics script, implements the core metrics from sound event detection evaluation and the DOA metrics.

The seld is a script that trains the SELDnet. The training stops when the SELD error stops improving.

## III. EVALUATION

Polyphonic sound event detection and localization are evaluated with individual metrics for SED and DOA. SED eval is an open source Python toolbox which provides a standardized, and transparent way to evaluate sound event detection systems (see Sound Event Detection). In addition to this, it provides tools for evaluating acoustic scene classification systems, as the fields are closely related (see Acoustic Scene Classification).

### A. SED

For SED F-Score and Error Rate are used to evaluate performance. Higher the F-Score the better the performance and conversely lower the Error Rate the better the performance.

$$F = \frac{precision \times recall}{precision + recall} \tag{1}$$

$$ER = \frac{S + D + I}{N} \tag{2}$$

Substitutions(S)-events in system output with correct temporal position but incorrect class label

Insertions(I)-events in system output unaccounted for as correct.

Deletions(D)-events in ground truth unaccounted for as correct.

N-Total number of events in ground truth

### B. DOA

DOA error and frame recall are used. A lower DOA error or a higher frame recall indicates better performance.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{3}$$

$$DOAerror = \frac{1}{\sum_{t=1}^{T} D_E^t} \times \sum_{t=1}^{T} \mathcal{H}(DOA_R^t, DOA_E^t) \tag{4}$$

DOA error is the average angular error in degrees between the predicted and reference DOAs. For a recording of length T time frames,let $DOA_R^t$ be the list of all reference DOAs at time-frame t and $DOA_E^t$ be the list of all estimated DOAs. where $D_E^t$ is the number of DOAs in $DOA_E^t$ at t-th frame, and $\mathcal{H}$ is the Hungarian algorithm which solves this by estimating the pair-wise costs between individual predicted and reference DOA using the spherical distance between them.

### C. Dataset

TAU Spatial Sound Events 2019 - Microphone Array provides four-channel directional microphone recordings from a tetrahedral array configuration. Both formats are extracted from the same microphone array, and additional information on the spatial characteristics of each format can be found below. The participants can choose one of the two, or both the datasets based on the audio format they prefer. Both the datasets, consists of a development and evaluation set. The development set consists of 400, one minute long recordings sampled at 48000 Hz, divided into four cross-validation splits of 100 recordings each. The evaluation set consists of 100, one-minute recordings. These recordings were synthesized using spatial room impulse response (IRs) collected from five indoor locations, at 504 unique combinations of azimuth-elevation-distance. Furthermore, in order to synthesize the recordings the collected IRs were convolved with isolated sound events dataset from DCASE 2016 task 2. Finally, to create a realistic sound scene recording, natural ambient noise collected in the IR recording locations was added to the synthesized recordings such that the average SNR of the sound events was 30 dB.

## IV. RESULTS

### A. Training

The model was trained with a batch size of 16 and maximum number of epochs as 20 however the training stops once the error stops improving. The model uses Adam optimizer, binary cross-entropy and MSE loss functions. The model was trained for 8 epochs before the error stopped improving. The image below shows the training.

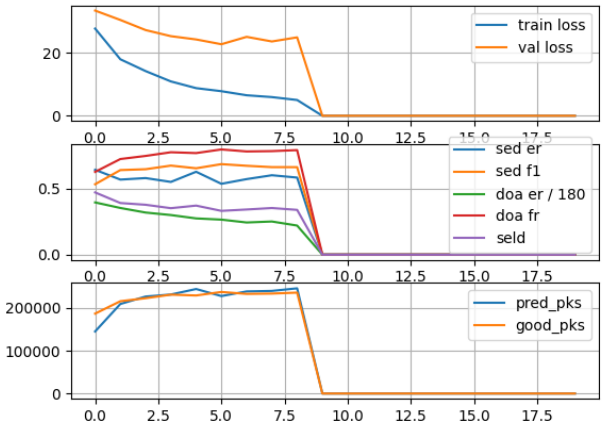


Fig. 2. Training Graph

The table below shows the metric values after training. SELD Score = 0.3400

| Metrics | DOA Error Overall | Frame Recall Overall |
|---|---|---|
| Values | 39.5210 | 0.7900 |
| Metrics | ER Overall | F1 Overall |
| Values | 0.5801 | 0.6693 |

TABLE I
TRAINING METRICS

### B. Testing

The trained model was saved and was tested on the test data which was split using train/test split and the output is shown below.

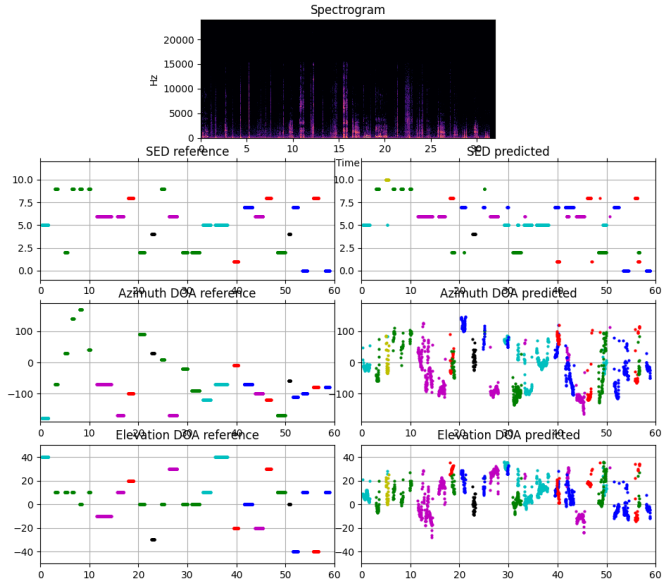


Fig. 3. Sample Output

The model was tested on the test dataset and the metrics are as follows.

SELD Score = 0.2461

| Metrics | DOA Error Overall | Frame Recall Overall |
|---|---|---|
| Values | 47.3634 | 0.8838 |
| Metrics | ER Overall | F1 Overall |
| Values | 0.3999 | 0.7941 |

TABLE II
TESTING METRICS

## V. CONCLUSIONS

In this paper, a SELDnet-driven feature learning method is proposed for polyphonic Acoustic Event Detection and Localization. The proposed DNN learns from spectral features and extracts high-level features for polyphonic AED. It is seen that SELDnet performs well despite incorporating both AED and DOA tasks in a single neural network approach and can be seen that the incorporation helps the model perform better in both aspects of SED and DOA.

## VI. FUTURE WORK

Future work involves training the network using all folds and improving the SELDnet architecture to better incorporate DOA task which will produce a better output for DOA prediction. Testing out different LSTM methods which will utilize memory better and produce better and more accurate results.

## REFERENCES


[1] Manjunath Mulimani, Akash B. Kademani, Shashidhar G. Koolagudi "A DEEP NEURAL NETWORK-DRIVEN FEATURE LEARNING METHOD FOR POLYPHONIC ACOUSTIC EVENT DETECTION FROM REAL-LIFE RECORDINGS"

[2] Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, Mark D. Plumbley "POLYPHONIC SOUND EVENT DETECTION AND LOCALIZATION USING A TWO-STAGE STRATEGY"

[3] Sharath Adavanne, Archontis Politis, Tuomas Virtanen "A MULTI-ROOM REVERBERANT DATASET FOR SOUND EVENT LOCALIZATION AND DETECTION"