Sound Event Detection and Localization Using ResNet RNN and Time-Delay DOA

Technical Report

Ee-Leng Tan

Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Ave, 639798, Singapore
etanel@ntu.edu.sg

Rishabh Ranjan

Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Ave, 639798, Singapore
rishabh001@ntu.edu.sg

Sathish s/o Jayabalan

Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Ave, 639798, Singapore
sathishj@ntu.edu.sg

Woon-Seng Gan

Nanyang Technological University
School of Electrical and Electronic Engineering
50 Nanyang Ave, 639798, Singapore
ewsgan@ntu.edu.sg

ABSTRACT

This paper presents a deep learning approach for sound events detection and time-delay direction-of-arrival (TDOA) for localization, which is also a part of detection and classification of acoustic scenes and events (DCASE) challenge 2019 Task 3. Deep residual nets originally used for image classification are adapted and combined with recurrent neural networks (RNN) to estimate the onset-offset of sound events, sound events class. Data augmentation and postprocessing techniques are applied to generalize the system performance to unseen data. Direction of sound events in a reverberant environment is estimated using a time-delay direction-of-arrival TDOA algorithm. Using our best model on validation dataset, sound events detection achieves F1-score of 0.84 and error rate of 0.25, whereas sound source localization task achieves angular error of 16.56 degree and 0.82 frame recall.

Index Terms— Sound events detection, source localization, ResNet RNN, TDOA

1. INTRODUCTION

Sound events detection (SED) and localization system allows one to have automated annotation of a scene in spatial dimension and can assist stakeholders to make informed decisions. It is an important tool for various applications as it can be used to identify critical events like gunshots, accidents, noisy vehicles, mixed reality audio where spatial scene information enhanced the augmented listening, and can be used to develop robots that listens and tracks sound source of interest just like humans.

In this paper, we propose a residual net (ResNet) combined with recurrent neural networks (RNN) for the estimation of respective labels for sound events detection (SED), and the direction of arrival (DoA) for sound events (except ambience) in a reverberant scene with one or two active sound sources is estimated using time-delay direction-of-sound (TDOA) algorithm (see Fig. 1). In the next section, an overview of the system will be described and followed by the discussion of results in Section 3. Section 4 details the DCASE task 3 submission and this report is then concluded in Section 5.

2. MODEL CONFIGURATION

For the classification of sound classes, a modified version of ResNet architecture combined with RNN is used. The ResNet model is adapted from residual net model originally designed for image recognition and described in [1].

2.1. Development Dataset

The development dataset consists of 4 splits and each split contains 100 audio files of length 60 sec, containing overlapping and non-overlapping sound events. Audio files are synthesized using 11 isolated sound labels taken from [2] and convolved with impulse response (IR) measured from 5 different rooms at 504 unique combinations of azimuth-elevation-distance and finally, mixed with natural ambient noise collected at IR recording locations.

2.2. Feature Extraction (SED)

Each of the audio file is sampled at 48kHz and short-time Fourier transform (STFT) is applied with hop size of 20 msec. Next, the magnitude STFT spectrogram is converted to log-mel spectrogram. After converting the STFT spectrogram into mel spectrogram features, low and high frequency components are removed and finally, resized to match the input shape of the neural network before training.

2.3. Model Training (SED)

For the development phase, 4 cross-fold sets from detection and classification of acoustic scenes and events (DCASE) challenge 2019 task 3 [3] is used as recommended by DCASE organizers.

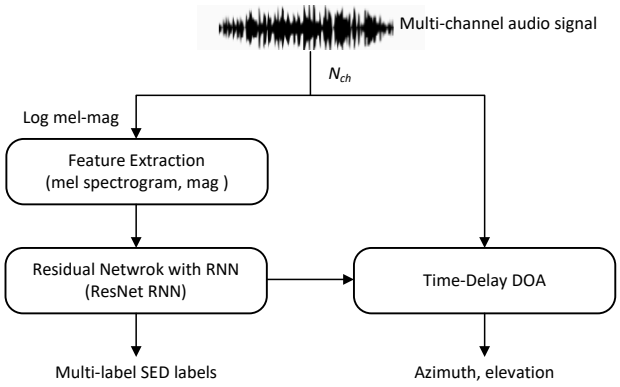

Figure 1: Block diagram of proposed system.

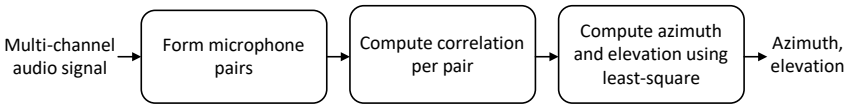

Figure 2: Block diagram of TDOA.

Each cross-fold consists of 2 training split, 1 validation split, and 1 test split as shown in Table I. For evaluation and DCASE submission, we used 3-splits for training and one split for validation as shown in Table II. During training, each processed audio feature file is split into sequence length of 128 frames and resized with fixed batch size of 96. The batch-feature dimension is $Batch\_size \times N_{ch} \times Seq\_length \times N_{mel}$, where $N_{ch} = 4$ corresponding to 4 channels for magnitude; N_{mel} is the number of filter banks and is varied between 64 and 128.

Table I: Cross-fold configuration for model development

Fold	Training sets	Validation sets	Test sets
1	Split 3, 4	Split 2	Split 1
2	Split 4, 1	Split 3	Split 2
3	Split 1, 2	Split 4	Split 3
4	Split 2, 3	Split 1	Split 4

Table II: Cross-fold configuration for model evaluation

Fold	Training sets	Validation sets
1	Split 2, 3, 4	Split 1
2	Split 3, 4, 1	Split 2
3	Split 1, 2, 4	Split 3
4	Split 1, 2, 3	Split 4

For SED, binary cross-entropy loss function with sigmoid activation function in output layer is used for multi-label classification of 11 classes.

2.4. Data Augmentation

To improve model generalization capability on unseen test data, data augmentation using frame shifting is applied to each of the processed audio file. Each audio file with 3000 frames is shifted by 32, 64 and 96 frames in temporal dimension before splitting into sequence of 128 frames. Therefore, total data after augmentation is 4 times larger than the original dataset size and is selected randomly for training in each epoch.

2.5. Evaluation Metrics

Model performance is evaluated using 4 metrics, 2 each for SED and DoA. SED is evaluated using error rate (ER) and F-score. ER is the total error based on total number of insertions (I), deletions (D) and substitutions (S) [4]:

$$ER = \frac{S + D + I}{N}, \quad (1)$$

where N is total number of frames. F-score is calculated as harmonic mean of precision (P) and recall (R) [5]:

$$F - Score = \frac{2PR}{P + R}. \quad (2)$$

DoA is evaluated using average angular error and frame recall. DoA error is defined as average angular error in degrees between estimated and ground truth directions and computed using the Hungarian algorithm [5] to account for the assignment problem of matching the individual estimated direction with respective reference direction. DoA frame recall (FR) is defined as percentage of frames where number of estimated and reference directions are equal.

2.6. Microphone pairs used in TDOA

To improve the performance of the azimuth and elevation estimation, a total of 6 microphone pairs are used, namely, mic1-mic2, mic1-mic3, mic1-4, mic2-mic3, mic2-mic4, and mic3-mic4. The block diagram of the TDOA algorithm is shown in Fig. 2. Cross-correlation of these microphone pairs are computed, and then the time-delay of each microphone pair is estimated from the cross-correlation. After combining all the time-delays computed from each microphone pair, the computation of azimuth and elevation is performed using the least-square method [6] based on the cross-correlation of the 6 microphone pairs.

3. RESULTS

Table III shows the model performance of proposed ResNet RNN with TDOA for 3-split model training configurations as described in Table II.

Table III: Proposed ResNet RNN model & TDOA estimation

Fold	ER	F-Score (%)	DoA Error (°)	FR (%)
1	0.1711	89.10	18.63	83.86
2	0.2147	86.28	15.63	82.48
3	0.1749	89.34	17.47	84.83
4	0.2308	86.07	16.88	84.42
Overall	0.1979	87.68	17.16	83.90

4. DCASE SUBMISSION

Based on the 3-split model results, the best model was selected for submission, and the details of our submission is as follows:

Model corresponding to *fold 3* in Table III:

- SED: Model trained using 128 mel bands; magnitude only
- DoA: TDOA; time-domain signal

5. CONCLUSION

In this report, ResNet model combined RNN architecture is used for sound events classification and TDOA for the localization task. With data augmentation and post-processing techniques, the proposed model is significantly improved overall, and especially for the source localization with DoA error improvement of more than 10° for both single and two overlapping sources.

6. ACKNOWLEDGMENT

This research was conducted in collaboration with Singapore Telecommunications Limited and supported by the Singapore Government through the Industry Alignment Fund - Industry Collaboration Projects Grant.

The authors are also thankful to Maggie Leong at Amazon Web Services (AWS) Singapore to generously provide the resources for model development on the cloud.

7. REFERENCES

[1] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "*Deep residual learning for image recognition.*" In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

[2] TUT audio dataset: http://www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-synthetic-audio#audio-dataset.

[3] DCASE Challenge Task 3: http://dcase.community/challenge2019/task-sound-event-localization-and-detection

[4] Mesaros, Annamaria, Toni Heittola, and Tuomas Virtanen. "*Metrics for polyphonic sound event detection.*" Applied Sciences 6, no. 6 (2016): 162.

[5] H. W. Kuhn, "The hungarian method for the assignment problem," *in Naval Research Logistics Quarterly*, no. 2, 1955, p. 8397

[6] J. Smith, J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," *IEEE Trans. Acoustics, speech, and signal process.*, vol. ASSP-35, no. 12, pp. 1661-1669, dec 1987.