DCASE 2019 TASK 3: A TWO-STEP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Thi Ngoc Tho Nguyen[1*], *Douglas L. Jones*[2], *Rishabh Ranjan* [3], *Sathish Jayabalan*[3], *Woon Seng Gan*[1],

[1] Nanyang Technological University, Electrical and Electronic Engineering Dept., Singapore,
{nguyenth003, ewsgan}@ntu.edu.sg
[2] University of Illinois at Urbana-Champaign, Dept. of Electrical and Computer Engineering,
Illinois, USA, {dl-jones}@illinois.edu
[3] Nanyang Technological University, SCALE, Singapore,
{rishabh001, sathishj}@ntu.edu.sg

ABSTRACT

Sound event detection and sound event localization requires different features from audio input signals. While sound event detection mainly relies on time-frequency patterns to distinguish different event classes, sound event localization uses magnitude or phase differences between microphones to estimate source directions. Therefore, we propose a two-step system to do sound event localization and detection. In the first step, we detect the sound events and estimate the directions-of-arrival separately. In the second step, we combine the results of the event detector and direction-of-arrival estimator together. The obtained results show a significant improvement over the baseline solution for sound event localization and detection in DCASE 2019 task 3 challenge. Using the development dataset on 4-fold cross-validation, the proposed system achieves an F1 score of 86.9% for sound event detection and an error of 5.15 degrees for direction-of-arrival estimation while the baseline F1 score and error are 79.9% and 28.5 degrees respectively.

Index Terms— sound event detection, direction-of-arrival estimation

1. INTRODUCTION

Sound event localization and detection (SELD) has a wide application in acoustic monitoring, and context-aware devices [1, 2]. SELD can provide the information of sound classes and the corresponding directions-of-arrival (DOAs) of multiple sound sources. As a result, SELD is the core component of acoustic monitoring systems such as environmental noise monitoring, and surveillance system. For example, SELD can direct a surveillance camera to point toward the direction of some sounds of interest. In addition, SELD can also assist context-aware devices such as hearing aids, smart phones, autonomous cars, and robots to be aware of the surrounding environments.

DCASE 2019 task 3, sound event localization and detection, challenges participants to detect sound events and their corresponding directions-of-arrival [3]. There are a total of 11 sound classes taken from DCASE 2016 task 2 dataset. The clean data are convolved with real-file recorded room impulse responses from 5 different indoor locations. 50% of the synthesized clips has 2 temporal overlapping sound events. The room impulse responses are recorded using Eigenmike microphone array at every 10 degree azimuth angle between -180 and 180 degrees, and at every 10 degree elevation between -40 and 40 degree. The data are given in two formats: first-order ambisonics and tetrahedral microphone array.

In general, SELD consists of two components, which are sound source localization (SSL) and sound event detection (SED). A microphone array is required to do sound source localization. In the context of the DCASE 2019 task 3 challenge, SSL refers to DOA estimation. The main challenges of SED tasks are multiple sources, overlapping events, varying background noises, and lacking of labeled data. In the past decade, deep learning is the most common solution for SED tasks [4]. The state-of-the-art SED models are often learnt using convolutional neural networks (CNN), recurrent neural networks (RNN), and some combinations of these two networks such as recurrent convolutional neural networks (RCNN). The main challenges of DOA estimation are reverberation, multiple sources, and varying background noises. Traditionally, DOA tasks for small microphone arrays are solved by using signal processing algorithms such as minimum variance distortionless response (MVDR) beamformer, and multiple signal classification (MUSIC) [5]. Recently, several researches have applied deep learning to DOA estimations [6]. Compared to the signal processing methods, the deep learning methods require more data for training, and the models need to be retrained when another microphone configuration is used.

There are two main approaches to solve for SELD. One is the end-to-end approach [2] where one system learns to detect the sound events and estimate their DOAs simultaneously. The other approach is to solve for SED and DOA separately, and match the sound events with the DOA estimates later. The end-to-end approach is attractive since it does not have to match the SED and DOA. However, SED and DOA estimation require different types of information and thus a joint estimation can hurt the performance of the whole SELD system. The baseline solution in DCASE task 3 is a joint RCNN model for SED and DOA estimation. The baseline model inputs both the magnitude and phase spectrograms of all microphone channels, and do multi-label classification for SED

[*] This material is based on research work supported by the IAF-ICP: Singtel Cognitive and Artificial Intelligence Lab for Enterprises@NTU under the Research Theme on Edge Intelligence.

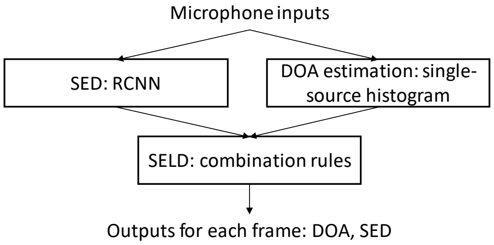
Microphone inputs
SED: RCNN
DOA estimation: single-source histogram
SELD: combination rules
Outputs for each frame: DOA, SED

Figure 1: A two-step SELD system

Table 1: SED network architecture

Stage	Output	Layers
1	128x64x32	32 5x5 conv, BN, ReLU, maxpool, stride = (1,2)
2	128x64x128	conv block, filters =[32,32,128], stride = (1,1)
		identity block, filters =[32,32,128]
		identity block, filters =[32,32,128]
3	128x16x128	Average pooling, stride = (1,4)
4	128x8x128	conv block, filters =[64,64,256], stride = (1,2)
		identity block, filters =[64,64,256]
		identity block, filters =[64,64,256]
5	128x2x256	Average pooling, stride = (1,4)
6	128x512	Reshape
7	128x128	Bidirectional, 128 GRU units, tanh
8	128x128	Bidirectional, 128 GRU units, tanh
9	128x128	128 fully connected
10	128x11	11 fully connected, sigmoid activation
Number of parameters		1008235

and regression for DOAs. The resulting DOA error on the development dataset is relatively large at 28.5 degrees for the first-order ambisonic format. In order to maximize the performance of the two subtasks, we propose a two-step SELD system, where we detect the sound events and estimate their DOAs separately and joint them later. The obtained results show a significant improvement over the joint estimation proposed in the baseline system. The main drawback of the two-step system is that mismatches of the sound events and their DOAs occur when there are overlapping events. We organize the paper as follows. Section II shows our proposed two-step system for SELD task. Section III presents our submissions to the DCASE task 3 challenge together with discussions. Finally, we conclude the paper in Section IV.

2. A TWO-STEP SELD SYSTEM

We use the first-order ambisonic (FOA) format for the SELD task. The theoretical DOA information is embedded in the magnitude differences between the four microphone channels up to 9 kHz [3]. The development set consists of 400 one-minute audio clips divided into 4 folds. Microphone input signals are transformed into the short-time Fourier transform domain with the following parameters: sampling rate of 48 kHz, window length of 2048 samples, hop length of 960 samples (0.02 second), and 2048 FFT points. This results in 3000 time frames for each one-minute audio clip. We are asked to predict the classes of sound events and their corresponding DOAs for each of the 3000 time frame. The block diagram of our two-step SELD system is shown in Fig 1. We use convolutional recurrent neural network (CRNN) for SED, and a single-source histogram algorithm [10] for DOA estimation. The results of SED and DOA estimation are fused together for each time frame using some rule-based logics.

2.1. Sound event detection

We extract 128 log mel-band energies of all the 4 microphones as input features for the SED block. The audio signals are divided into 128 frame segments. The size of input features into RCNN model is 128 frame x 128 mel x 4 channels. We replace the 3 convolutional layers in the baseline model [3] with the first three stages of Resnet-50 network [7]. Since the size of the SELD dataset is much smaller compared to the ImageNet [8], we also reduce the number of filters in each stage. Table. 1 shows the SED network architecture. We will refer to this network in the subsequent sections as the RCNN-Resnet. The log-mel input features are normalized along the mel-band by mean and standard deviation. We use the same normalization factors for all 4 channels. Throughout the network, we do max pooling only on the mel-band dimension, so that we can have the output predictions at each time frame.

In the first stage, the input features are fetched into a convolutional layer with 32 filters of size 5x5, batch normalization, ReLU activation, and max pooling of size 1 x 4. The second and fourth stages are the modified stage 2 and 3 of the Resnet-50 network respectively. We reduce the number of filters to [32,32,128] for stage 2, and [64,64,256] for stage 4. We inserted two average pooling layers of stride (1, 4) to reduce the number of parameters, and avoid overfitting. The subsequent recurrent layers and fully connected layers are similar to the baseline [3]. The output layer use sigmoid activation to do multi-label classification.

We train the network using Adam optimizer with learning rate of 0.0001 for 100 epochs. Cross validation is used to select the best parameters to train the final model. The competition uses individual evaluation metrics for SED and DOA estimation. For SED task, evaluation metrics are error rate and F1 score calculated in one-second segment [9]. To make our predictions more robust against the random segmentation of the audio clip, for the validation and testing data, we shift each of the audio clip 0, 32, and 64 samples before dividing them into 128-frame segments, and input them into the SED network. After that we combine the 3 predictions of the SED network at each frame using geometric mean. On the 4-fold cross validation, this shifting scheme reduces our error rate about 0.06 and improves the F1 score of 1%. In the final prediction, a sound event is consider active if the prediction probability is greater than 0.5.

2.2. DOA estimation

We use a single-source histogram algorithm proposed in [10] to estimate DOAs. The evaluation metrics used for DOA task are angle error and frame recall that are defined in [11]. Compared to the MUSIC algorithm, on the 4-fold cross validation, the single-source histogram reduces the DOA error by 2 degrees and increase the frame recall by 2%. The single-source histogram finds all the time-frequency (TF) bins that contains energy from mostly one source. A TF bin is considered to be a single-source TF bin when it passes all three tests: magnitude, onset, and coherence test. Magnitude test finds TF bins that are above a noise floor to mitigate the effect of background noise. Onset test finds TF bins that belong to direct-path signals to reduce the effect of reverberation in the DOA

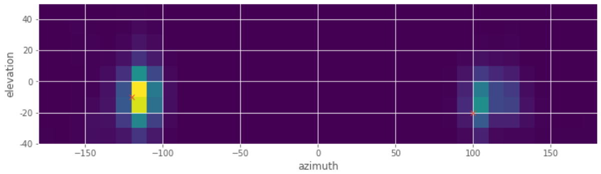

Figure 2: Smoothed single-source histogram for a 2-source frame

estimation. Coherence test finds TF bins of which the covariance matrices are approximately rank-1. After all the single-source TF bins are found, the DOA at each bin is computed using the theoretical steering vector of the microphone array [10]. These DOAs are discretized using the required resolution of azimuth and elevation angles. After that, these DOAs are populated into a histogram. The histogram is smoothed to reduce the estimation errors. The final DOA estimates are the peaks of this histogram.

For this task 3, we compute one histogram per time frame. Since the SELD dataset have only maximum two overlapping sources and moderate level of reverberation, we do not use the onset test. The theoretical steering vector for the first-order Ambisonics format is approximately true for up to 9 kHz. Therefore we only search for single-source TF bins between bin 2 and bin 384, which correspond to 50 Hz and 9000 Hz respectively.

We extend the 2D DOA estimation in [10] to 3D DOA estimation for DCASE 2019 task 3. The ranges of azimuth and elevation angles of the room impulse responses in task 3 are $[-180, 180)$, and [-40, 40] respectively. The azimuth and elevation resolutions are 10 degrees. After discretizing, the dimensions of the 2D DOA histogram are 16x9. From our observations, the estimated elevation angles have higher variation than the estimated azimuth angles. Therefore, we smooth the 2D histogram using a 2D Gaussian kernel with higher variance for elevation. Validation sets are used to find the best threshold to select the peaks on the DOA histograms. Fig. 2 shows a smoothed single-source histogram for a 2 source case. The active source classes are *drawer* and *laughter* at $(-120, -10)$ and $(100, -20)$ respectively. The cross markers show the estimated DOAs which coincide with the ground truths.

2.3. Combine SED and DOA estimations

We use a set of rules to combine SED and DOA estimation at each frame. The SED results have higher precedence than the DOA results. Let denote n_{sed} and n_{doa} as the number of sound events detected by SED and DOA estimator respectively. Since the SELD dataset has maximum of 2 overlapping events for each time frame, we limit n_{sed} and n_{doa} to 2. Algorithm 1 describes the combination rules. The limitation of this approach is that when SED and DOA estimator return 2 sources, the DOAs and classes of the sound events have 50% chance of being mismatched. This mismatched issue will be studied in our future research.

3. SYSTEM SUBMISSION AND DISCUSSION

Table 2 shows our four submitted systems to the SELD challenge. The first system consists of a single RCNN-Resnet model for SED as described in Section 2.1, and the single-source histogram algorithm for individual time frame for DOA estimation as described in Section 2.2. The second system uses the same RCNN-Resnet model as the first system, but the single-source histogram is computed for all the frames that belong to each sound event detected by the RCNN-Resnet model. The third and fourth systems are ensembles of different variations of the RCNN-Resnet model, and the single-source histogram algorithm for individual time frame. We trained 4 variations of the RCNN-Resnet: the first variation use 128 log-mel energy of background-normalized magnitude spectrograms [12]; the second variation uses LSTM instead of GRU; the third variance uses additional inputs which is the largest eigenvector of the covariance matrix of each TF bin; the fourth variance has additional output which indicates if a time frame has an active signal. System 3 ensembles the original RCNN-Resnet model and all the variances. System 4 excludes variance 2 since the validation results does not show improvement when variation 2 is added. All of the submitted system used the same combination logics to combine SED and DOA results.

Algorithm 1 Combine SED and DOA estimation

```
For each time frame
if $n_{SED} == 0$ then
    Return None
else if $n_{SED} < n_{DOA}$ then
    Assign all the DOAs to the sound event
else if $n_{SED} == n_{DOA}$ then
    Randomly assign one DOA for each sound event
else
    Look for additional DOAs in the neighbourhood frames
    if find addition DOAs so that $n_{SED} == n_{DOA}$ then
        Randomly assign one DOA for each sound event
    else
        Randomly ignore the extra sound event
    end if
end if
Return pairs of (sound event, DOA)
```

Table 2: Submission systems

Submission	SED (# of params)	DOA histogram
NGUYEN_NTU_task3_1	single model (1M)	individual frame
NGUYEN_NTU_task3_1	single model (1M)	signal support
NGUYEN_NTU_task3_1	5-model ensemble (5.6M)	individual frame
NGUYEN_NTU_task3_1	4-model ensemble (4.4M)	individual frame

Table 3 shows the 4-fold cross-validation results on the development dataset of 4 submission systems against the baseline system. The development dataset is divided into 4 folds. For each fold cross-validation, 2 folds are used for training, 1 fold is used for validation, and 1 fold is used for test [3]. The baseline system learns a CRNN model that jointly estimate sound events and DOAs from both magnitude and phase spectrogram. Because SED and DOA estimation requires different types of information from the microphone inputs, we do the SED and DOA estimation separately to maximize the performance of both tasks. The downside of our approach is the mismatch between sound classes and their corresponding DOAs when there are more than one sound event. However, because the evaluation metrics do not penalize this mismatch, we could not quantify this mismatch in both the baseline and our proposed algorithms.

The 4-fold cross-validation test results on the development set show a significant improvement of the proposed algorithms over the joint SED and DOA estimation model in the baseline. Our single RCNN model and DOA histogram achieved an error rate of 0.21

Table 3: Development results of four submissions

Submission	SED ER	SED F1	DOA Er	DOA FR
NGUYEN_NTU_task3_1	0.21	86.9	5.15	**88.9**
NGUYEN_NTU_task3_2	0.24	84.9	5.86	80.5
NGUYEN_NTU_task3_3	**0.17**	**89.3**	**5.12**	87.5
NGUYEN_NTU_task3_4	**0.17**	**89.3**	**5.12**	87.6
baseline	0.34	79.9	28.5	85.4

Table 4: SED network architecture

	ID	SED ER	SED F1	DOA Er	DOA FR
Fold	1	**0.17**	**89.7**	5.17	**89.6**
	2	0.26	84.2	5.15	88.1
	3	0.18	**89.7**	5.22	89.2
	4	0.25	0.84	**5.05**	88.7
Overlap	1	**0.16**	**90.4**	**4.66**	**96.0**
	2	0.24	85.0	5.41	81.8
Impulse Response	1	**0.21**	87.2	**3.74**	89.3
	2	**0.21**	**87.3**	3.99	**89.9**
	3	0.22	86.4	7.17	88.2
	4	**0.21**	87.2	4.69	89.4
	5	0.22	86.4	6.03	87.7
Total		0.21	86.9	5.15	88.9

and a F1 score of 86.9% for SED task, a DOA error of 5.15 degrees and a frame recall of 88.9% for DOA task. The DOA error has the largest improvement from 28.5 degrees in the baseline. The ensembles obtain the best performance for SED tasks. The DOA errors are relatively similar for all submitted system. The DOA frame recalls slightly reduce for ensembles. System 2 computes DOA histograms using all the frames that belong to an event detected by the SED model. This approach improves the DOA performance when there is no overlapping event. When there is one source, the DOA error and the frame recall of system 1 are 4.66 degrees and 96.0%; and those of system 2 are 4.36 degrees and 96.3%. However when there are two sources, the DOA error and the frame recall of system 1 are 5.41 degrees and 81.8%; and those of system 2 are 6.56 degrees and 64.8%. System 2 makes more mistakes when there are overlapping events, the DOA algorithm fails to pick up all the present DOAs.

Table 4 shows the performance of submission 1 across all folds, overlaps, and room impulse responses. Across 4 folds, the error rate and the F1 score for SED task vary the most, while the DOA error and the frame recall are quite similar. The system performance degrades when there are overlapping events as expected. The SED error increases from 0.16 to 0.24 and the DOA frame recall drops from 96.0% to 81.8%. Across different rooms, the performance of SED is quite stable, while the DOA error has more fluctuation. These results show that overlapping sound event and the different room acoustics are the main challenges for the SELD task.

4. CONCLUSION

SELD is an interesting task with many real-life applications. Our experiments show that a joint model for SED and DOA might be suboptimal. The separate SED and DOA estimation models achieve better performance on the DCASE task 3 dataset compared to the joint model in the baseline system. The drawback of our proposed system is the mismatch of the sound classes and the DOAs. We nominate submission 1 (NGUYEN_NTU_task3_1) for the Judge Award.

5. REFERENCES

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 405–409.

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2018.

[3] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and uetection," in *Submitted to Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019. [Online]. Available: https://arxiv.org/abs/1905.08546

[4] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[5] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[6] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2814–2818.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: http://www.mdpi.com/2076-3417/6/6/162

[10] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.

[11] S. Adavanne, A. Politis, and T. Virtanen, "Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1462–1466.

[12] T. N. T. Nguyen, N. K. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2018 task 2: iterative training, label smoothing, and background noise normalization for audio event tagging," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, November 2018, pp. 54–58.