IMPROVING LOW-RESOURCE SOUND EVENT LOCALIZATION AND DETECTION VIA ACTIVE LEARNING WITH DOMAIN ADAPTATION

Technical Report

Yuhao Wang[*1], *Yuxin Duan*[*1], *Pingjie Wang*[1], *Yu Wang*[†1,2], *Wei Xue*[†3],

[1] Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai, China
[2] Shanghai AI Lab, Shanghai, China
[3] Dept. Computer Science, Hong Kong Baptist University, Hong Kong, China
{colane, duanyuxin}@sjtu.edu.cn, applewpj@gmail.com,
yuwangsjtu@sjtu.edu.cn, weixue@comp.hkbu.edu.hk

ABSTRACT

This report describes our systems submitted to DCASE2022 challenge task3: sound event localization and detection (SELD) evaluated in real spatial sound scenes. We present two approaches to improve the performance of this task. The first one is to leverage active learning to bring in and filter the AudioSet dataset based on the pre-trained audio neural networks (PANNs). The second one is to adapt the generic models to different sound event categories, thereby improving the performance on classes with scarce data. We have also explored various model structures incorporating attention machanisms. Finally, we combine models trained on different input recording formats. Experimental results on the validation set show that the proposed systems can greatly improve all the metrics when compared to the baseline systems.

Index Terms— DCASE2022, Sound event localization and detection, Active learning, AudioSet, Adaptation

1. INTRODUCTION

Sound event localization and detection (SELD) plays an important role in robots sensing [1], audio surveillance for animals [2], navigation for the hearing impaired [1] and so on. Its goal is to recognize the onset and offset of sound events when active and the corresponding temporal trajectory. SELD involves two subtasks, sound event detection (SED) and directional of arrival estimation (DOAE).

Unlike previous years of SELD tasks, DCASE2022 makes some changes to this task. The most significant difference is that all the audios are recorded in real sound scenes and manually annotated instead of computationally generated. This transition brings several impacts on nature of this task. First, as the recording procedure is complex and strict , creating a large dataset is very expensive. Second, the sound making difficulty varies across different classes, making the data distribution severely imbalanced. For example, the sounds easy to perform contribute a great part, like male speech, female speech and music, while the sounds hard to control are fairly few, like knock and door.

To solve the above two problems, we explore two strategies. Firstly, to settle the lack of data, we introduce the temporally-strong labeled release of AudioSet [3] (AudioSet-strong)[1] as one of the external data sources to this challenge. To further augment data with reliable labels, we apply active learning to our models, where the audios are scored using the pre-trained audio neural networks (PANNs) [4]. Secondly, to mitigate the data imbalance problem, we train a number of task-specific models. As each trained model is domain adapted, the event classes with scarce data are able to be recognized more accurately. In addition to modifications from the data perspective, we also explore better model structures based on convolutional recurrent neural network (CRNN) used in the baseline [1, 5] for this SELD task, including increasing the number of convolutional layers and adding an attention layer or transformer encoder [6]. The submitted models combine a number of models with different architectures and input recording formats to improve the performance. Experiments conducted on the development dataset show that our systems improve greatly over the baseline.

2. PROPOSED APPROACH

In this section, we will first explain how to apply PANNs for filtering external data and the adaptive learning approach using the filtered data. Then we will explain the proposed domain adaptation approach where models are adapted to different categories to suit imbalanced data distribution. We will also describe our architecture modifications for the baseline model. Finally, our training setup is illustrated.

2.1. PANNs-based Active Learning

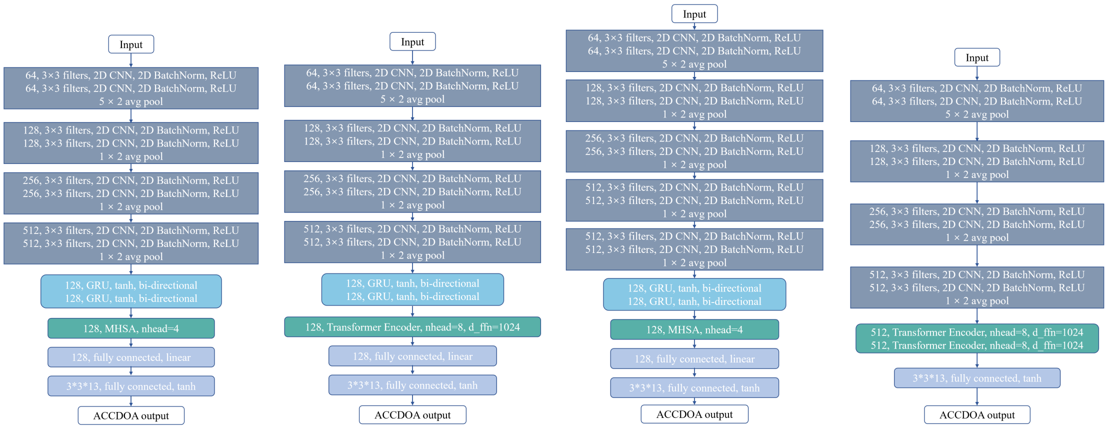
We use the AudioSet dataset as one of the external data source. Because the original release (AudioSet-original)[2] is weakly labeled with a resolution of 10s along with serious overlapping, which brings in multiple inference classes, for this task we use the AudioSet-strong release. To select data from AudioSet-strong, we exploit PANNs to identify the non-overlapping and reliable segments, which are then be used to synthesize spatial audio for active learning. Specifically, we first pick the sound classes identical to those in the original release and clean the dataset by filtering out audios shorter than 0.05s. Next, audios shorter than 0.5s are concatenated until they are longer than 0.5s. Then the PANNs outputs the probabilities $p_i^{(c)}$ of 527 classes defined by AudioSet ontology

[*] Equal contribution.

[†] Corresponding author.

[1] https://research.google.com/audioset/download_strong.html

[2] https://research.google.com/audioset/download.html


Input
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
5 × 2 avg pool
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
128, GRU, tanh, bi-directional
128, GRU, tanh, bi-directional
128, MHSA, nhead=4
128, fully connected, linear
3*3*13, fully connected, tanh
ACCDOA output
Input
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
5 × 2 avg pool
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
128, GRU, tanh, bi-directional
128, GRU, tanh, bi-directional
128, Transformer Encoder, nhead=8, d_ffn=1024
128, fully connected, linear
3*3*13, fully connected, tanh
ACCDOA output
Input
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
5 × 2 avg pool
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
128, GRU, tanh, bi-directional
128, GRU, tanh, bi-directional
128, MHSA, nhead=4
128, fully connected, linear
3*3*13, fully connected, tanh
ACCDOA output
Input
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
64, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
5 × 2 avg pool
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
128, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
256, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
512, 3×3 filters, 2D CNN, 2D BatchNorm, ReLU
1 × 2 avg pool
512, Transformer Encoder, nhead=8, d_ffn=1024
512, Transformer Encoder, nhead=8, d_ffn=1024
3*3*13, fully connected, tanh
ACCDOA output

(a) model #1 (b) model #2 (c) model #3 (d) model #4

Figure 1: Illustration of proposed model architecture. (a) model #1, (b) model #2, (c) model #3, (4)model #4. $nhead$, d_ffn denote the number of heads in attention layer, the intermediate dimension in feed forward network.

for audio i, in which $c \in \{1, 2, ..., 527\}$. Suppose that the corresponding label class indexes are C, audios whose output probabilities not satisfying

$$\begin{cases} p_i^c > thd1, c \in C, \\ p_i^c < thd2, c \notin C \end{cases} \tag{1}$$

will be filtered out, where thresholds $thd1$ and $thd2$ determine whether audio i is correctly labeled, and therefore need careful tuning for different target classes. The remaining audios are finally used to simulate real data with spatial room impulse responses using provided data generator[3].

2.2. Domain Adaption

As there are 13 sound event target classes, we train single model with the same structure for each class, and each model can only classify the specific class. During inference on unseen validation set or evaluation set, all the 13 models' output will be combined to form the final predicted output.

2.3. Model Architecture

Figure 1 shows our model architectures. The models are based on the CRNN structure. The input is acoustic features for tetrahedral microphone array (MIC) format or first-order ambisonics (FOA) format. For MIC format, we extract logmel and GCC-PHAT features, while for FOA format, logmel and intensity vectors are computed. The output is activity-coupled Cartesian direction of arrival (ACCDOA) representation. In order to increase the diversity and robustness of the model ensembles, four model variants are introduced, that is model #1∼#4. In general, all the four variants double the number of convolutional blocks compared to the baseline. Our model architectures also incorporate single multi-head self-attention (MHSA) layer or transformer encoder to the original network.

2.4. Hyper-parameters and Training settings

Our hyper-parameters and training setup mostly follows the baseline. The training data include Sony-TAU Realistic Spatial Soundscapes 2022 (STARSS22)[4] [7] training set (∼5 hours), and synthesized data (55 hours in total) with FSD50K[5] (which is provided) [8] and AudioSet-strong. The audios are synthesized with 10 room impulse responses and maximum 2 events are overlapped, and each audio is 60s length. The model output format is the multi-ACCDOA format introduced in [5].

3. MODEL COMBINATION

Table 3 shows the performance of our submitted systems. We combine proposed models trained on MIC and FOA formats. In addition, we also combine each system's models by applying a simple average strategy to their results. Specifically, the four systems use 0.4 as the threshold, which we use to compare with the class output's ACCDOA vector length, and if it exceeds the threshold, the corresponding class is supposed to be active. To convert multi-ACCDOA with three tracks to single-ACCDOA output format, the result of the most active track is saved as the final output of each model.

4. EXPERIMENTAL EVALUATION

In this section, we show our experimental settings, and compare the experimental results with the baseline.

4.1. Experimental settings

We evaluate our approach using the STARSS22 evaluation set. The four metrics used for evaluation measure the accuracy of detection (ER_{20°, F_{20°) and localization (LE_{CD}, LR_{CD}). ER_{20° and F_{20° compute the classification error rate and F-score, which is based on whether the prediction is further or closer to the true DOA than

[3] https://dcase.community/challenge2022/task-sound-event-localization-and-detection-evaluated-in-real-spatial-sound-scenes#example-external-data-use-with-baseline

[4] https://zenodo.org/record/6387880

[5] https://zenodo.org/record/4060432.YqicNqhByl4

Table 1: System configuration. Format denotes recording format.

System	Format	Base model
System #1	FOA	model #1, #2, #4
	MIC	model #1, #2, #4
System #2	FOA	model #1, #2, #4
	MIC	model #1, #2, #3, #4
System #3	FOA	model #1, #2, #3, #4
	MIC	model #1, #2, #4
System #4	FOA	model #1, #2, #3, #4
	MIC	model #1, #2, #3, #4

20°. LE_{CD} and LR_{CD} depend on correct sound event classification, which represent the class-based localization error and localization recall respectively. Contrary to previous challenges, macro-averaging is performed in this challenge, in which F_{20°, LE_{CD} and LR_{CD} are first calculated for separate classes, and are then averaged across all the classes. The aggregated SELD score can be calculated as

$$score = \frac{1}{4}\left[ER_{20^\circ} + (1 - F_{20^\circ}) + \frac{LE_{CD}}{180^\circ + (1 - LR_{CD})}\right]. \quad (2)$$

4.2. Experimental Results

To explore the improvement of our approaches for the dataset and model, we conduct a series of experiments.

We evaluate the performance of different dataset ensembles with CRNN as used model and MIC as input recording format. As shown in table 2, additional datasets improve the performance of all metrics, especially for LR_{CD} with external AudioSet dataset compared to only adding FSD50K.

Table 2: SELD performace of our models with external data and baseline. * means the results are reported officially.

Dataset	ER_{20°	F_{20°	LE_{CD}	LR_{CD}
STARSS22*	0.71	18.0	32.2°	47.0
+FSD50K	0.63	32.0	24.6°	42.5
+FSD50K&AudioSet	**0.59**	**34.0**	**21.4°**	**52.4**

We also compare model #1∼#4 with CRNN trained on STARSS22, FSD50K and AudioSet, and the results show that all the proposed models outperform the original CRNN. Finally, we conduct an experiment on model #1 trained on all mentioned datasets, and the four metrics respectively improve by 3.9%, 35.2%, 37.8%, 11.4% relatively, which validates the effectiveness of the domain adaptation.

Table 3 compares the performance of our systems with the baseline on validation set, where the baseline model is trained on STARSS22 dataset with multi-ACCDOA output format. As shown in the table, our systems outperform the baseline on all metrics by a large margin, especially for F_{20°.

5. CONCLUSION

The focus of DCASE2022 SELD task is for the real spatial sound scenes. The main challenges arises from the small size and the imbalance of the data. To solve these two challenges, we propose a PANN-based active learning strategy to agument the dataset, and propose a domain adaptation approach to improve the performance on the low-resource categories. Experimental results show that our systems can improve all the metrics significantly compared to the baseline systems for the DCASE2022 SELD task.

Table 3: SELD performance of our systems and baseline. * means the results are reported officially.

System	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	$score$
Baseline-FOA*	0.71	21.0	29.3°	46.0	0.551
Baseline-MIC*	0.71	18.0	32.2°	47.0	0.560
System #1	0.47	**62.2**	**11.3°**	**69.0**	**0.305**
System #2	**0.46**	61.8	11.4°	68.4	0.306
System #3	0.48	61.4	11.5°	**69.0**	0.309
System #4	0.47	61.6	11.4°	68.7	0.307

6. REFERENCES

[1] 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*, vol. 13, no. 1, pp. 34–48, 2018.

[2] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: a survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[3] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[4] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[5] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, 2022.

[8] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM*

Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 829–852, 2021.