A ROBUST FRAMEWORK FOR SOUND EVENT LOCALIZATION AND DETECTION ON REAL RECORDINGS

Technical Report

Jin Sob Kim*, Hyun Joon Park*, Wooseok Shin*, Sung Won Han**

School of Industrial and Management Engineering, Korea University, Seoul, Republic of Korea

ABSTRACT

This technical report describes the systems submitted to the DCASE2022 challenge task 3: sound event localization and detection (SELD). The task aims to detect occurrences of sound events and specify their class, furthermore estimate their position. Our system utilizes a ResNet-based model under a proposed robust framework for SELD. To guarantee the generalized performance on the real-world sound scenes, we design the total framework with augmentation techniques, a pipeline of mixing datasets from real-world sound scenes and emulations, and test time augmentation. Augmentation techniques and exploitation of external sound sources enable training diverse samples and keeping the opportunity to train the real-world context enough by maintaining the number of the real recording samples in the batch. In addition, we design a test time augmentation and a clustering-based model ensemble method to aggregate confident predictions. Experimental results show that the model under a proposed framework outperforms the baseline methods and achieves competitive performance in real-world sound recordings.

Index Terms— DCASE2022, Sound event localization and detection, Framework, Test time augmentation, Ensemble

1. INTRODUCTION

Sound event localization and detection (SELD) aims identification of both the sound event occurance (SED) and the direction of arrival from the sound source (DOA). Since the localization and detection joining task challenge launched in DCASE2019 [1, 2], the conceptual formulation of the SELD task has been established in DCASE2021 based on activity-coupled Cartesian direction of arrival (ACCDOA) [3]. Unlike the task has been trained and evaluated on the emulated sound scenes in control up to previous iterations, DCASE2022 engages the real sound scape recordings along with strong temporal and spatial annotations bringing the task to the real-world problem [4]. The challenge also inspires the exploitation of external data to solve the lack of real-world labeled data.

In this report, we introduce the total training framework achieving the generality of model performance on the real sound recordings through utilizing the emulations from the external data sources. Emulated sound data consist of the synthesis of the individual audio samples corresponding to the task-predefined classes, extracted from the external datasets [5, 6, 7, 8, 9, 10, 11], along with spatial room impulse responses (SRIR) and spatial ambient noises (SNoise) [12]. Considering the difference between the size of real

---

*Equal contribution.
**Corresponding author.

sound recordings dataset and of emulated sounds, the framework exploits the down-sized bootstrapping of emulations per every training epoch. The framework also employees the data augmentation techniques to train further generalized model.

Additionally, aiming at the increment of robustness in model prediction, the clustering-based test time rotation-augmentation ensemble method of ACCDOA is designed in this work. The method allows the effective aggregation of the predictions from confident majorities whilst selective rejection of the abnormal candidates.

2. PROPOSED METHOD

In this section, we describe a proposed framework for SELD, and it consists of feature extraction, data augmentation, external mix, network, and test time augmentation. The overall process of the framework is shown in Figure 1.

2.1. Features

We use first-order Ambisonic (FOA) format signals, and multichannel log mel-spectrograms and FOA intensity vectors are used as frame-wise features. The parameters $(sr, nfft, hop, window) = (24000, 2048, 600, 1200)$ are set for Short-time Fourier transformation.

2.2. Data augmentation

We adopt three types of augmentation including each wave and spectrogram augmentation for guaranteeing the generalized performance on the test set. Regarding wave augmentation on the audio inputs, we adjust the gain, shift the pitch, and apply the band-pass filter on the audio. Furthermore, we rotate the directions of events in a sound source as in [13]. The rotation function sets are pre-defined as 16 patterns to change azimuth and elevation angles of a sound source. The rotation function sets are the combinations of channel swapping and channel sign inversion on the X, Y, and Z axis of FOA inputs. Finally, we use spectrogram augmentation [14], which masks time and frequency information after log mel-spectrogram transformation on the audio inputs.

2.3. External mix

To keep the model fit in real-world scenario contexts while taking advantage of various audio samples from external datasets, dataset mixing technique is adopted to consist model training dataset. The technique balances the size of each dataset on the model training phase, between the small real recording set and the large emulated scenarios. The balancing of two different datasets is conducted as
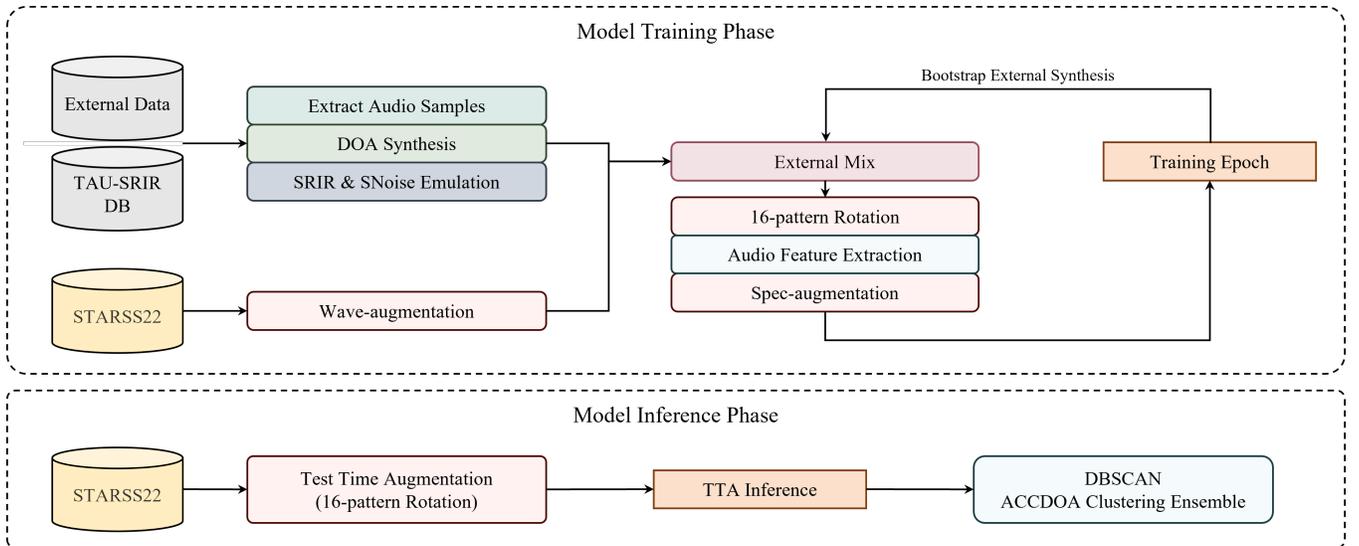
Figure 1: Proposed framework overview.

follows: the half of the model training dataset consists of the real-world sound scenarios, where it is maintained to the end of the training phase. The other half comprises with the scenarios down sampled from the emulated sounds, and the sampling occurs dynamically at the start of the every training epoch.

2.4. Network architecture

Figure 2 presents the process of our SELD architecture consisting of an encoder, a self-attention pooling, and a decoder used in this study. First, we adopt the squeeze-and-excitation residual networks [15] (SE-ResNet), which have recently been applied to audio classification [16, 17], as the SELD encoder. Our SE-ResNet-34 is the same original SE-ResNet-34, except we use average pooling instead of the stride convolution in the second and third blocks. In addition, to preserve the frame-wise localization information of the input audio, we aggregate frequency dimensions into channel dimensions using the variant of self-attention pooling [18, 19]. For the SELD decoder, we employ two bidirectional GRU layers followed by layer normalization and tangent hyperbolic activation. Then, the SELD output is obtained by applying two fully-connected layers followed by tangent hyperbolic activation.

2.5. Test time augmentation

To maximize the generalization ability of models in the real recording test phase, we adopt a test time augmentation (TTA). TTA is widely used in computer vision to increase the robustness and performance of models [20, 21, 22]. Although they proposed better augmentation policies and aggregation methods, they are not appropriate to apply on SELD due to the different predicted outputs. In the specific, the unknown number of events and the presence of coordinates information make challenging to apply TTA on SELD.

To utilize TTA on SELD, we propose a clustering-based aggregation method to obtain confident predicted outputs and aggregate them. We take 16 pattern rotation augmentation for test time augmentation, making 16 predicted outputs, that is candidates. To obtain confident aggregated outputs, we use DBSCAN [23] for clus-
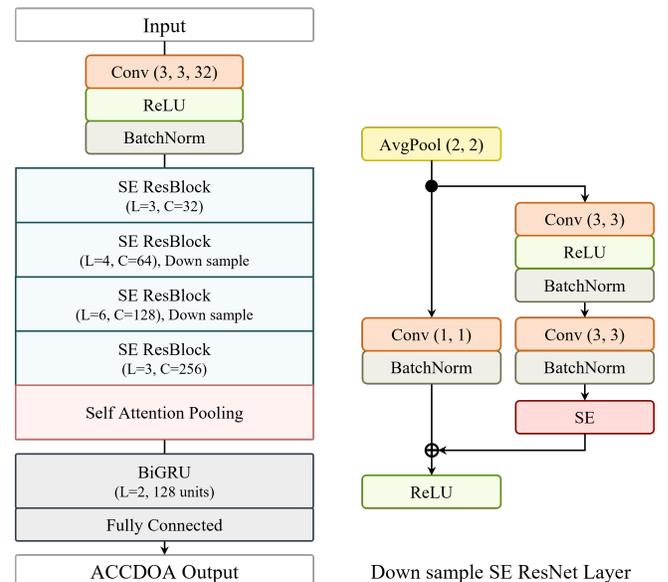

Figure 2: Model architecture.

tering candidates, only when the candidates have the same class and the number of them is over the threshold. The epsilon of DB-SCAN is the distance calculated by the threshold of unification for inference in degrees. By exploiting DBSCAN, the outliers of the candidates can be excluded and it is possible to divide the events which have the same class but are differently located. After acquiring clusters of candidates, we take an average to aggregate them. When the number of the aggregated outputs is over three tracks, we select top-3 outputs depending on the weight which is calculated by multiplication between the number of candidates per cluster and their norm. Furthermore, we fit the model on rotated validation set to maximize TTA effect. Finally, we apply this strategy to cross-

validation models for the submission of the evaluation dataset.

3. EXPERIMENTS

3.1. Dataset

The proposed method is evaluated on STARSS22 [4]. To synthesize emulated sound scenarios from the external data, we use class-wise audio samples extracted from six external datasets, which are AudioSet [5], FSD50K [6], DCASE2020 and 2021 SELD datasets [7, 8], ESC-50 [9], IRMAS [10], and Wearable SELD [11]. W axis of the ambisonic audio format is extracted for the external data that originally consisted in multichannel recording format aiming the SELD task. As the same way in former SELD task challenges, extracted audio samples are synthesized through SRIR and SNoise from TAU-SRIR DB [12] emulating the spatial sound environment.

Before synthesizing the external data, we split the external audio samples from each dataset into four folds for the cross validation. Then, we synthesized two sets for each fold, one keeps the original class distribution of external samples and the other balances the distribution via down sampling the larger class population.

At the developmental stage, the model is trained on the real recording training set and the last external emulation fold while it is validated and also tested on the dev-test set of the given STARSS22 dataset. On the other hand, two types of cross validating divisions are additionally set for the evaluation stage, the first is based on class event-wise stratified split and the other is based on the room-wise split. Both of the splits consist to have four folds each, and the corresponding external emulation folds are considered as external data to be fed the model in training framework.

3.2. Experimental setup

We used the official metrics [2] to evaluate our SELD system. Similar to previous DCASE SELD challenges, DCASE 2022 SELD challenge includes four evaluation metrics: location-dependent error rate ($ER_{20°}$), location-dependent F-score ($F_{20°}$), class-dependent localization error (LE_{CD}), and localization recall metric (LR_{CD}). In contrast to the previous challenges, macro-average is used as a class averaging method for $ER_{20°}$.

During the training, we splitted the audio inputs into 5-second segments with 4-second overlapped. An AdamW optimizer with a learning rate of 1×10^{-3} was employed. We calibrated the learning rate in half if the validation score did not reduce after 25 epochs. We set a batch size to 128 and total epochs to 200.

Table 1: Experimental results of our systems for the dev-test set.

Ver.	# Params.	$ER_{20°}$	$F_{20°}$	LE_{CD}	LR_{CD}
Baseline	0.60M	0.71	21%	29.30°	46%
- w/o Ext. Data	-	0.84	16%	43.15°	31%
+ w/ Larger Ext. Data	-	0.67	35%	18.86°	45%
SE-ResNet+GRU	6.04M	0.59	40%	17.52°	50%
- w/o Ext. Data	-	0.83	17%	36.86°	30%
+ w/ Larger Ext. Data	-	0.64	44%	17.53°	65%
SE-ResNet+GRU	6.04M	0.59	40%	17.52°	50%
+ Augmentation	-	0.55	46%	15.78°	54%
+ External Mix	-	0.45	54%	14.53°	69%
+ TTA	-	0.43	58%	12.86°	69%

3.3. Experimental results

In Table 1, we conducted an ablation study to validate the influence of the proposed framework. The results of the first (Baseline model) and second (SE-ResNet) blocks show the effect of the external dataset on performance. The first row of each block, which uses dev-train data and FSD50K (Baseline synthesized data), presented that SE-ResNet significantly outperforms the baseline model; however, the second row of each block using dev-train only showed poor performance for both models. In addition, both models showed additional performance gains when using the seven external datasets mentioned in 3.1. In the last block, we found that significant improvements were obtained from three components (Augmentation, External Mix, and TTA). Among them, the external mix method contributed more to the performance improvement than the other methods.

3.4. Submission

Table 2 shows the setups of our submitted systems. Submission #1 and #2 are the system trained on development training set with validated and tested solely via dev-test set. The first system is validated without considering the TTA while the second system is validated for all 16-patterns of rotating augmentation. System #3 and #4 are trained and validated from each cross validation division aforementioned in Section 3.1, class-wise stratified split and room-wise split respectively. TTA was considered for validating models from both systems.

Table 2: Experimental results of our submitted systems.

System	# Params.	Description
Submission #1	6.04M	SE-ResNet+GRU (Aug & Ext. mix)
Submission #2	6.04M	SE-ResNet+GRU (Aug & Ext. mix & TTA)
Submission #3	24.16M	SE-ResNet+GRU (CV & Aug & Ext. mix & TTA)
Submission #4	24.16M	SE-ResNet+GRU (CV & Aug & Ext. mix & TTA)

4. CONCLUSION

In this report, we presented our method for DCASE2022 task 3: sound event localization and detection. We proposed a framework for SELD which composes of data augmentation, external mix, and test time augmentation. The augmentations including the transformation of wave feature, directional rotation, and spectrogram augmentation were effective to increase the diversity of samples. The external mixing technique, which balances the size of the dataset between real-world and synthetic sounds during the training phase, contributed to the improvement of performance on the real-world recordings. Regarding test time augmentation, we designed a clustering-based test time augmentation, which is appropriate for SELD task, and it showed a performance gain. Experimental results demonstrated that the proposed robust framework for SELD with our SE-ResNet-34 can guarantee the generalized performance on the real-world recordings even when the train set of real-world recordings was not enough.

5. ACKNOWLEDGMENT

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, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[2] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9306885

[3] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Ontario, Canada, June 2021.

[4] 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," 2022. [Online]. Available: https://arxiv.org/abs/2206.01948

[5] J. F. Gemmeke, D. P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[6] 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, 2022.

[7] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 165–169. [Online]. Available: https://dcase.community/workshop2020/proceedings

[8] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 125–129. [Online]. Available: https://dcase.community/workshop2021/proceedings

[9] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. ACM Press, pp. 1015–1018. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[10] J. J. Bosch, F. Fuhrmann, and P. Herrera, "IRMAS: a dataset for instrument recognition in musical audio signals," Sept. 2014. [Online]. Available: https://doi.org/10.5281/zenodo.1290750

[11] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 156–160.

[12] A. Politis, S. Adavanne, and T. Virtanen, "TAU Spatial Room Impulse Response Database (TAU- SRIR DB)," Apr. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6408611

[13] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] J. H. Yang, N. K. Kim, and H. K. Kim, "Se-resnet with gan-based data augmentation applied to acoustic scene classification," in *DCASE 2018 workshop*, 2018.

[17] H. Shim, J. Kim, J. Jung, and H.-j. Yu, "Audio tagging and deep architectures for acoustic scene classification: Uos submission for the dcase 2020 challenge," *Proceedings of the DCASE2020 Challenge, Virtually*, pp. 2–4, 2020.

[18] W. Cai, J. Chen, and M. Li, "Exploring the encoding layer and loss function in end-to-end speaker and language recognition system," *arXiv preprint arXiv:1804.05160*, 2018.

[19] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.

[20] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[22] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1214–1223.

[23] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.