THE NERC-SLIP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION OF DCASE2023 CHALLENGE

Technical Report

Qing Wang[1], *Ya Jiang*[1,*], *Shi Cheng*[1,*] *Maocheng Hu*[2], *Zhaoxu Nian*[1], *Pengfei Hu*[1], *Zeyan Liu*[1], *Yuxuan Dong*[1], *Mingqi Cai*[3], *Jun Du*[1], *Chin-Hui Lee*[4]

[1] University of Science and Technology of China, Hefei, China
{qingwang2, jundu}@ustc.edu.cn,
{yajiang, chengshi, zxnian, hudeyouxiang, xy671231, anonymous}@mail.ustc.edu.cn
[2] National Intelligent Voice Innovation Center, Hefei, China, {mchu2}@nivic.cn
[3] iFLYTEK, Hefei, China, {mqcai}@iflytek.com
[4] Georgia Institute of Technology, Atlanta, USA, {chl}@ece.gatech.edu

ABSTRACT

The technical report details our submission system for Task 3 of the DCASE2023 Challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. To address the audio-only SELD task, we apply the audio channel swapping (ACS) technique to generate augmented data, upon which a ResNet-Conformer architecture is employed as the acoustic model. Additionally, we introduce a class-dependent sound separation (SS) model to tackle overlapping mixtures and extract features from the SS model as prompts to perform SELD for a specific event class. In the case of audio-visual SELD task, we leverage object detection and human body key point detection algorithms to identify potential sound events and extract Gaussian-like vectors, which are subsequently concatenated with acoustic features as the input. Moreover, we propose a video data augmentation method based on the ACS method of audio data. Finally, we present a post-processing strategy to enhance the results of audio-only SELD models with the location information predicted by video data. We evaluate our approach on the dev-test set of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset.

Index Terms— Sound event localization and detection, speech separation, data augmentation, model ensemble, Conformer, human keypoint detection, object detection

1. TRACK A: AUDIO-ONLY INFERENCE

The proposed approach employs several effective audio data augmentation techniques to generate training data for sound event localization and detection (SELD). Subsequently, Resnet-Conformer [1], a robust deep neural network (DNN) architecture, is trained for SELD systems of different target representations. [2,3] adopted two parallel branches for sound event detection and direction-of-arrival (SEDDOA) estimation with a multi-task learning framework. Shimada *et. al.* [4, 5] proposed an activity-coupled Cartesian DOA (ACCDOA) vector with the length indicating event activity, which was later extended to a multi-ACCDOA version. To enhance the SELD performance, Conv-TasNet [6], a sound separation (SS) network, is implemented. Finally, to obtain robust SED and DOA estimation, a combination of model ensemble and post-processing is employed. This technical report will provide a detailed description about the five parts of the approach: data augmentation, sound separation, network training, model ensemble, and post-processing.

1.1. Audio Data Augmentation

The official training dataset contains 24 hours of data, among which only about 4 hours are recorded in real sound scenes. Therefore, data augmentation techniques are necessary, and we adopted three such methods for SELD and SS.

One of the methods is ACS spatial augmentation, which was proposed in our previous work [3]. The method utilizes the rotational properties of the recorded data set to increase DOA representations. The other method is to simulate new multi-channel data by using provided spatial room impulse responses (SRIRs) and sound samples selected from several public datasets. Specifically, single-channel sound samples extracted from FSD50K dataset [7] are convoluted with SRIRs to generate 1-minute long multi-channel scene recordings with a maximum polyphony of 3. This results in 20 hours of data. Moreover, sound event clips from AudioSet [8] and the DCASE2023 Task 3 dev-train split are used to build a training data set for separation model. We generated a total of 39 hours of sound event clips, which is much higher than the development dataset of the DCASE2023 Task 3.

1.2. Fusing SS and SELD: SS-SELD

We have utilized labeled segments of individual sound events from Audioset [8], in addition to the DCASE2023 Task 3 dev-train data, to construct training data for sound separation. The simulated data was used to train SS models for specific classes based on Conv-TasNet architecture. As for the SELD model, we used SEDDOA output format for fusion.

Subsequently, we used the trained SS model as the front-end and separated the augmented data with ACS, aiming at extracting specific class sounds while removing the sounds of other classes. This approach effectively mitigates detection errors of low-intensity classes with significant overlap. We extracted log-mel spectra for

* These authors contributed equally to this work and should be considered co-second authors.

separated data and concatenated them with original features for SELD training. The framework of the method is shown in Figure 1.

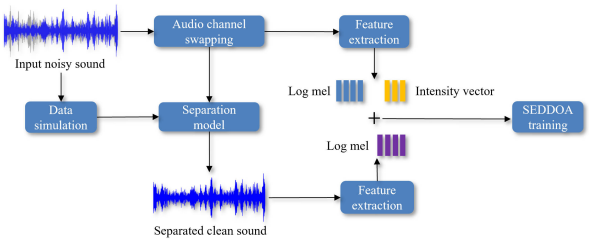

Figure 1: Framework of our proposed SS-SELD method. Log mel represents log-mel spectra features.

1.3. Network Training

In the proposed system, only FOA format data is adopted as the training data. Log-mel spectra features are extracted from multi-channel audio of 24 kHz sampling rate using 1024-point discrete Fourier transform from a 40 msec Hanning window and 20 msec hop length. Then 4-channel log-mel spectra features and 3-channel intensity vectors are concatenated together to get the 7-channel feature. And 4-channel log-mel spectra features extracted from the separated sound are concatenated to get the 11-channel feature input. The data size can be augmented to 8 times by applying the ACS strategy, which results in about 192 hours of data. As for sound separation, about 40 hours of single event clips are obtained. To train the SS model, we have generated a 40-hour dataset for each class, where sound events of all other classes are considered inferences. We utilized Conv-TasNet [6] as the separation network and Resnet-Conformer as the main network for SELD [9].

1.4. Model Ensemble

Two model ensemble strategies are utilized to improve the generalization ability and achieve better results. The ACCDOA and multi-ACCDOA fusion strategy is proposed to fully utilize the advantages of these two modeling methods. ACCDOA-based modeling method can provide accurate boundary information. Multi-ACCDOA-based method can process the overlap segments of the same event class but may introduce false alarms. Here the boundary information from ACCDOA is combined with the SED and DOA estimation of multi-ACCDOA. Assume class c happened at frame t in ACCDOA estimation. If the SED estimation of multi-ACCDOA at frame t is the same as ACCDOA, we will calculate the angle difference between the two estimated events. If the difference of DOA estimation is higher than a specific threshold, we think the DOA estimated by multi-ACCDOA model is more accurate. Otherwise we think these two events are exactly the same event, and the final DOA estimation will be the mean value of these two methods.

1.5. Post-processing

Two post-processing strategies are adopted to further improve the system performance. First, when testing the input data is cut into 20-second long segments with a 1 second hop length. Then the result of each frame is the mean value of the time-overlapped segments. Tested on time-overlapped segments can decrease the variance of the results. Second, dynamic threshold is adopted to improve the SED performance. We also explore the fusion of SS-SELD results and single SELD results. For a specific class c, we compare the scores between SS-SELD and SLED models. If SS-SELD is better than SELD, we believe the proposed SS-SELD model has a better discernment of class c and employ a class-level fusion between SS-SELD and SELD predictions.

2. TRACK B: AUDIO-VISUAL INFERENCE

In the proposed audio-visual SELD model, we design a video data augmentation method to match the audio data after performing audio channel swapping. To leverage the video information, we adopt a feature-level fusion approach, constructing an audio-visual Resnet-Conformer SELD network that utilizes both audio features and Gaussian features extracted from video frames. Additionally, we develop a decision-level fusion scheme to complement the prediction results of the audio modality with video data. Model ensemble and post-processing in Track A are also adopted to get the final SED and DOA estimation.

2.1. Video Data Augmentation

The size of Sony-Tau Realistic Spatial Soundscapes 2023 (STARSS23) audio-video dataset [10] is 3.8 hours, which is too small to train a robust audio-visual SELD network. In audio-only SELD system, we perform ACS [3] to expand the audio data by a factor of seven. The STARSS23 dataset includes simultaneous 360° video recordings with a resolution of 1920×960, corresponding to an azimuth angle range of $[180^\circ, -180^\circ]$ and an elevation angle range of $[-90^\circ, 90^\circ]$. We propose a video pixel swapping (VPS) approach as shown in Table 1 to expand the video data and match the augmented audio data. Take one transformation, $\phi = \phi + \pi, \theta = \theta$ for example, which means the azimuth angle is rotated by 180°, while the elevation angle remains the same. Based on such a transformation, the horizontal pixel points are panned by 960 pixel points in the negative direction while the vertical pixel points remain unchanged in the corresponding video image.

Table 1: The VPS augmentation approach for video data corresponding to the ACS approach, where x and y denote the horizontal pixel point and vertical pixel point in video image frames.

DOA transformation	Pixel point swapping
$\phi = \phi - \pi/2, \theta = -\theta$	x = (x + 1440) mod 1920, y = 960 - y
$\phi = -\phi - \pi/2, \theta = \theta$	x = (-x + 1440) mod 1920, y = y
$\phi = \phi, \theta = \theta$	x = x, y = y
$\phi = -\phi, \theta = -\theta$	x = 1920 - x, y = 960 - y
$\phi = \phi + \pi/2, \theta = -\theta$	x = (x + 480) mod 1920, y = 960 - y
$\phi = -\phi + \pi/2, \theta = \theta$	x = (-x + 480) mod 1920, y = y
$\phi = \phi + \pi, \theta = \theta$	x = (x + 960) mod 1920, y = y
$\phi = -\phi + \pi, \theta = -\theta$	x = (-x + 960) mod 1920, y = 960 - y

2.2. Video Keypoint Detection and Object Detection

We perform human keypoint detection and object detection on every frame of the video. For human keypoint detection, we adopt a top-down strategy [11]. Firstly, based on the mmdetection [12]

framework, we employ a faster-rcnn-based two-stage target detection model [13] to predict human bounding boxes and corresponding confidences. The model is pre-trained on the COCO object detection dataset [14]. By setting a threshold of 0.3, we filter out human bounding boxes with lower confidence scores. Next, we scale the filtered human bounding boxes to 384×288 and feed them into the keypoint detection model HRNet [15] based on the mmpose framework [16]. The model is pre-trained on the COCO-WholeBody dataset [17]. Based on the cropped human bounding boxes, we choose the coordinates for 5 keypoints of the human body: mouth, left hand, right hand, left foot and right foot.

Regarding object detection, we resize the video images to 640×640, and adopt the PP-YOLOE model based on the PaddleDetection framework [18] to detect the bounding boxes of the target event classes. The confidence threshold is set to 0.7.

2.3. Audio-Visual SELD Network Training

The audio-visual SELD network takes audio features and visual features as inputs. The audio features are extracted in the same way as the audio-only SELD network. We transform the bounding boxes of the target object or human keypoint, which are generated by selecting two video frames in one second, into two Gaussian-like vectors of 64 dimensions [19]. These vectors represent the likelihood of objects being present along the image's horizontal and vertical axes. We add the Gaussian-like vectors of all detected objects and perform a normalization operation to generate location-based features. We repeat the 2-channel video features several times along the time dimension and concatenate them with the 7-channel 64-dimensional audio features. This concatenated feature set is then fed into the Resnet-Conformer network for training. Additionally, we have implemented the outputs of three different modeling approaches in the audio-visual SELD network: ACCDOA, multi-ACCDOA and SEDDOA.

2.4. Decision-level Fusion and Post-processing

We adopt two schemes to make better use of the video information. The first scheme is based on the approximate conversion relationship between the pixel coordinates in image and the DOA coordinates in real-world. We design a matching rule to generate more accurate DOA estimation by utilizing the keypoint detection results and the object detection results.

The details of the decision-level fusion scheme are as follows. The DOA predictions may be inaccurate. Intuitively, visual-based localization methods tends to yield more accurate results in comparison to audio-based ones. Therefore, for certain categories, we employ object detection techniques to enhance the accuracy of DOA predictions by audio model. Taking Figure 2(a) as an example, it displays the DOA prediction d of the category *Male Speech* at frame t that generated by the audio model. To refine the DOA prediction, firstly, we detect all the mouths present in the visual modality at the same frame, and depict the detection results $d_1, \ldots, d_5$ in Figure 2(b). Secondly, we select a candidate from $d_1, \ldots, d_5$ that is closest to d, and denote it as $\hat{d}$. Thirdly, if the angular distance between $\hat{d}$ and d is less than a pre-defined threshold value $\theta = 30°$, we replace d with $\hat{d}$ as the final DOA estimation, which is shown in Figure 2(c). Figure 2(d) lists the corresponding relationships between audio categories and visual objects. We exclude several audio categories due to the lack of corresponding visual objects (e.g., *Knock*) or poor visual detection performance (e.g., *Door*).

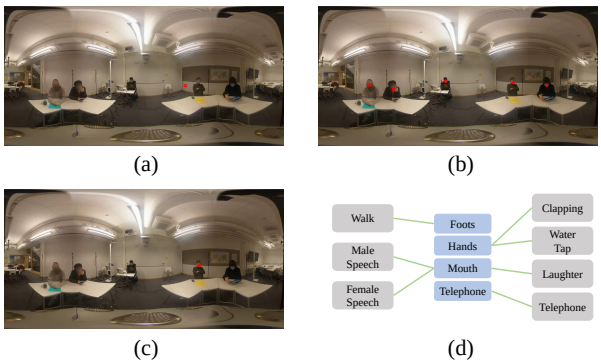

Figure 2: The illustration of the decision-level fusion. (a) The DOA predictions of the category *Male Speech*. (b) The detection results of mouths. (c) The final result with the decision-level fusion. (d) Corresponding relationships between audio categories (Grey) and visual objects (Blue).

The second scheme is implemented by model ensemble and post-processing, as mentioned in Section 1.4 and 1.5. We employ a posteriori fusion of the audio-only model and the audio-visual model on 10-second long segments with a 1-second hop length. We also apply dynamic threshold on the SED output.

3. RESULTS ON DEVELOPMENT DATASET

We evaluate our proposed method on the development dataset of STARSS23 with joint localization and detection metrics [20].

For Track A, we generate a larger training set with the above mentioned data augmentation approaches, namely audio channel swapping and multi-channel data simulation. Table 2 shows the experimental results of the proposed method for development dataset of Track A. "ACCDOA" represents the ACCDOA-based modeling method and "Multi-ACCDOA" represents the multi-ACCDOA-based method, both of which used post-processing strategies. "SS-SEDDOA" denotes the fusion method of SS and SELD. As shown in the table, each proposed single model outperforms the two baseline systems by a large margin. By fusing the SELD results predicted by ACCDOA and multi-ACCDOA methods, further improvements are achieved as shown in the last row of Table 2. "Model Ensemble+PP" denotes employing model ensemble and post-processing strategies upon the above systems.

Table 2: Experimental results of the audio-only SELD systems for development dataset using FOA format data.

	$ER_{20°} \downarrow$	$F_{20°} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
Baseline-FOA	0.57	0.30	21.60	0.48
SEDDOA	0.41	0.59	14.05	0.70
ACCDOA	0.42	0.59	13.72	0.72
Multi-ACCDOA	0.44	0.58	13.75	0.74
SS-SEDDOA	0.40	0.64	13.40	0.74
Model Ensemble+PP	0.38	0.66	12.81	0.75

For Track B, we fine-tune the audio-visual SELD model based on the audio pre-trained parameters, utilizing about 30 hours of audio-video data. Table 3 shows the experimental results of the pro-

posed audio-visual methods for development dataset. As shown in Table 3, the proposed audio-visual systems outperforms the baseline system by a large margin. Performance gain is achieved by employing model ensemble and post-processing strategies. With the decision-level fusion scheme, the SELD metrics are improved, especially for localization error. A small localization error can help to improve detection metrics, as shown in the last row of Table 3.

Table 3: Experimental results of the audio-visual (AV) SELD systems for development dataset using FOA format data.

	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$
Baseline-AV Model	1.07	0.14	60.40	0.33
AV SEDDOA	0.41	0.56	14.47	0.64
AV ACCDOA	0.42	0.59	14.06	0.70
AV Multi-ACCDOA	0.45	0.58	14.62	0.71
Model Ensemble+PP	0.39	0.63	13.12	0.73
Decision-level Fusion	0.37	0.68	10.76	0.73

4. REFERENCES

[1] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng, Y. Wang, L. Sun, Y. Fang, J. Pan, J. Du, and C.-H. Lee, "The NERC-SLIP system for sound event localization and detection of dcase2022 challenge," DCASE2022 Challenge, Tech. Rep., June 2022.

[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.

[3] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[4] 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)*, 2021, pp. 915–919.

[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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 316–320.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] 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.

[8] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[9] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[10] 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," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022, pp. 125–129.

[11] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4903–4911.

[12] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[16] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[17] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 196–214.

[18] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, *et al.*, "PP-YOLOE: An evolved version of YOLO," *arXiv preprint arXiv:2203.16250*, 2022.

[19] X. Qian, Z. Wang, J. Wang, G. Guan, and H. Li, "Audio-visual cross-attention network for robotic speaker tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 550–562, 2023.

[20] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 333–337.