THE NERC-SLIP SYSTEM FOR SOUND EVENT LOCALIZATION AND DETECTION WITH SOURCE DISTANCE ESTIMATION OF DCASE 2024 CHALLENGE

Technical Report

Qing Wang[1], *Yuxuan Dong*[1], *Hengyi Hong*[2], *Ruoyu Wei*[3], *Maocheng Hu*[4]
Shi Cheng[1], *Ya Jiang*[1], *Mingqi Cai*[3], *Xin Fang*[3], *Jun Du*[1]

[1] University of Science and Technology of China, Hefei, China
{qingwang2, jundu}@ustc.edu.cn, {yxdong0320, chengshi, yajiang}@mail.ustc.edu.cn
[2] Harbin Engineering University, Harbin, China, {hyhong}@hrbeu.edu.cn
[3] iFLYTEK, Hefei, China, {rywei, mqcai, xinfang}@iflytek.com
[4] National Intelligent Voice Innovation Center, Hefei, China, {mchu2}@nivic.cn

ABSTRACT

The technical report presents our submission system for Task 3 of the DCASE 2024 Challenge: Audio and Audiovisual Sound Event Localization and Detection (SELD) with Source Distance Estimation (SDE). In addition to direction of arrival estimation (DOAE) of the sound source, this challenge also requires predicting the source distance. We attempted three methods to enable the system to predict both the DOA and the distance of the sound source. First, we proposed two multi-task learning frameworks. One introduces an extra branch to the original SELD model with multi-task learning framework, resulting in a three-branch output to simultaneously predict the DOA and distance of the sound source. The other integrates the sound source distance into the DOA prediction, estimating the absolute position of the sound source. Second, we trained two models for DOAE and SDE respectively, and then used a joint prediction method based on the outputs of the two models. For the audiovisual SELD task with SDE, we used a ResNet-50 model pretrained on ImageNet as the visual feature extractor. Additionally, we simulated audio-visual data and used a teacher-student learning method to train our multi-modal system. We evaluated our methods on the dev-test set of the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset.

Index Terms— Sound event localization and detection, source distance estimation, model ensemble, Conformer, audiovisual fusion

1. TRACK A: AUDIO-ONLY INFERENCE

Sound event localization and detection (SELD) refers to the ability of a machine to automatically recognize the temporal activity trajectory of each sound category given a multi-channel audio input and to track the spatial location of the target sound source when a sound event is activate. In this technical report, we try to address the task with an additional source distance estimation (SDE), i.e., sound event detection, localization with distance estimation (3D SELD) [1]. We employ several effective audio data augmentation techniques to generate training samples. Subsequently, the Resnet-Conformer [2, 3], a robust deep neural network (DNN) architecture, was trained for 3D SELD. Previous works, such as [4] and [5], utilize a multi-task learning framework with two parallel branches for sound event detection (SED) and direction of arrival (DOA) estimation. Building on this structure, we investigated three ways of integrating distance estimation with the SELD task. The first method used a three-branch framework where, in addition to solving SED and DOA estimation, a separate branch is adopted to predict the source distance. The second method integrated the estimation of source distance into the DOA estimation. The third method involved training a separate SDE model to predict the source distance, which was later combined with the DOA estimation model to obtain the final 3D SELD result. Finally, model ensemble was employed to achieve robust prediction of sound categories, directions and distance. This technical report will provide a detailed description of the methodology's three main components: data augmentation, network training, and model ensemble.

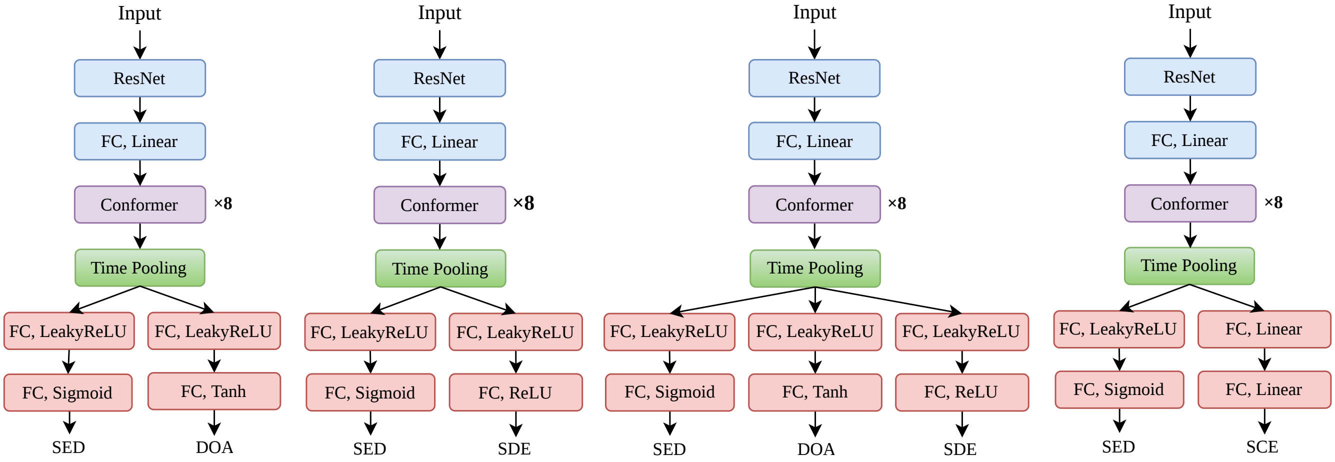
1.1. Audio Data Augmentation

The official dataset, named Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) [6, 7], contains only 7 hours and 22 minutes of real recordings. The dataset is split intro training data (90 clips) and testing data (78 clips). Therefore, data augmentation techniques are essential to improve the diversity of training samples. In this challenge, we use three data augmentation methods.

The first method is audio channel swapping (ACS) spatial augmentation, proposed in our previous work [5]. This method performs transformation on audio channels, which is based on the physical and rotational properties of the spherical microphone array, to augment the DOA representation. The second method involves simulating new multi-channel data using provided spatial room impulse responses (SRIRs) and sound samples selected from public dataset. Specifically, single-channel sound samples extracted from the FSD50K dataset [8] are convolved with the SRIRs to increase the amount of training data by using a recently released library [9]. With this method, we synthesized 40 hours of data. The third method is Manifold Mixup performed on randomly selected layers between input and hidden layers of the neural network [10].

1.2. Network Training

In this challenge, only FOA format data is used. A 1024-point discrete Fourier transform is applied to extract log-spectral features



(a) The SED-DOA model (b) The SED-SDE model (c) The SED-DOA-SDE model (d) The SED-SCE model

Figure 1: The network architecture of our proposed audio 3D SELD models.

from the 40 msec Hanning window and 20 msec hop-length multichannel audio sampled at 24 kHz. The 4-channel log-mel spectral features and 3-channel intensity vectors are then concatenated to obtain 7-channel features. The segment length input to the network is fixed at 10 seconds, which generates a feature shape of $7 \times 500 \times 64$. By applying the ACS strategy, the training data size can be increased up to 8 times, resulting in approximately 350 hours of data. We used Resnet-Conformer as the main network for 3D SELD.

In this technical report, we adopt four models with different output formats to address the 3D SELD as shown in Figure 1. The first is the SED-DOA model [5], where the network outputs two-branch for sound event detection and direction of arrival estimation. The second model is the SED-SDE, where the network outputs two-branch for sound event detection and source distance estimation. Mean square percent error (MSPE) [1] is used as the loss function for the SDE branch. The third model is SED-DOA-SDE, where the network outputs three-branch for sound event detection, direction of arrival and source distance estimation. The fourth model integrates the source distance into the DOA information, which uses two branches for SED and source coordinate estimation (SCE). We multiplies the normalized Cartesian coordinates of sound events with the source distance to obtain the absolute Cartesian coordinates, which serves as the source coordinate labels. This model aims to predict the absolute Cartesian coordinates of the sound source, with the direction of coordinate vector representing the DOA and the length of coordinate vector representing the source distance. Mean square error (MSE) is used as the loss function for the SCE branch.

Additionally, we found that initializing the parameters of Resnet-Conformer with the parameters of the best model trained by our team in DCASE 2023 Task 3 yielded good results. Therefore, we applied this technique in the final submission system as well.

1.3. Model Ensemble

Model ensemble is utilized to improve the generalization ability and achieve better results. First, we propose a fusion strategy by combining the outputs of the SED-DOA and SED-SDE models. The SED-DOA model can predict the DOA of sound events, while the SED-SDE model can predict the source distance. By utilizing the outputs of these two models, we can simultaneously obtain the sound event categories, direction and distance estimation, which are required for the 3D SELD task. Considering that the SED-DOA model benefits from the ACS method and provides more robust SED predictions, we select its SED prediction as the fusion SED result in terms of posterior probability. And the DOA and SDE results are obtained from these two models, respectively.

Combining this system with the SED-DOA-SDE and SED-SCE models improves generalization ability and achieves better results. The final result is obtained from the model ensemble of this system with the SED-DOA-SDE and SED-SCE systems.

2. TRACK B: AUDIO-VISUAL INFERENCE

2.1. Video Data Augmentation

The STARSS23 dataset contains about 3.8 hours of audio-video training data [6], which is too small to train a robust audio-visual 3D SELD network. To obtain more video data, we took two data augmentation methods. The first method utilizes the audio-video simulation method proposed by Adrian S. Roman et al. [11]. Spatialized sound events are generated using room impulse responses (RIR) from the METU-SPARG RIR dataset [12]. And a spatial audio synthesizer extracts audio from YouTube videos and convolves it with RIR. This method provided us with approximately 6 hours of audio and video data.

Additionally, in the audio-only 3D SELD system, we perform ACS [5] 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 use an audio-visual pixel swapping (AVPS) approach to increase the audio and video data size [13]. Unlike our previous work [2], we generate completely new video frames by flipping and rotating the original frames. With these two methods, we obtained approximately 80 hours of audio and video data.

2.2. Audio-Visual Network Training

The audio-visual 3D SELD network takes both audio features and visual features as input. Audio features are extracted in the same manner as in audio-only 3D SELD networks. Visual features are extracted using a pre-trained ResNet-50 network [14] at a frame rate of 10 fps. Global average pooling is applied on the last layer of ResNet-50, resulting in a 7 × 7 feature map. As the segment length input to the network is fixed at 10 seconds, the visual feature shape is 100 × 7 × 7.

To align the visual features with the audio features along the temporal dimension, we repeat each visual feature map five times. We consider one dimension of the visual feature map as the channel dimension and concatenate the remaining dimension with the frequency dimension of the audio features. This results in fused audio-visual features with a shape of 7 × 500 × 71. This concatenated feature set is then fed into the Resnet-Conformer network for training. We trained the previously described SED-DOA, SED-SDE and SED-SCE models using audio-visual data. The SED-SCE model that we used in the challenge is shown in Figure 2. For simplicity, we have omitted the linear layer between the ResNet and Conformer, as well as the time pooling layer after the Conformer. Additionally, instead of training from scratch, we employed the parameters of the best audio model trained by our team in DCASE 2023 Task 3 for initialization. This strategy also proved effective.

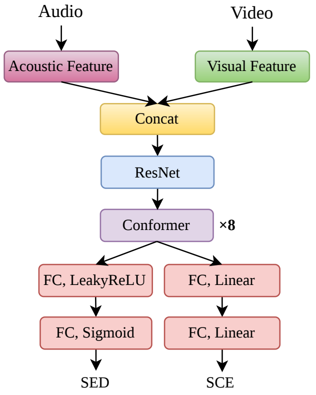


Figure 2: The network architecture of our proposed audio-visual SED-SCE model.

2.3. Model Ensemble and Post-processing

We trained two kinds of audio-visual 3D SELD systems as described in the previous subsection. The submission system is obtained by fusing these two single systems with the model trained on the audio-only track using posterior probability fusion. Additionally, we used a post-processing scheme, called video-guided decision fusion [15], to generate more accurate DOA results.

3. RESULTS ON DEVELOPMENT DATASET

3.1. Results on Track A

We evaluated our proposed method using the STARSS23 development dataset. For Track A, we generated a larger training set using the data augmentation methods described above. Table 1 shows the experimental results of the proposed 3D SELD methods on the development dataset of audio-track. In the table, "SED-DOA" denotes the modeling method based on SED-DOA output format and "SED-SDE" denotes the modeling method based on SED-SDE output format. "SED-DOA+SED-SDE" denotes the fusion system of these two models. "SED-DOA-SDE" denotes the modeling method based on SED-DOA-SDE output format and "SED-SCE" denotes the modeling method based on SED-SCE output format. "Model Ensemble" represents employing model ensemble for joint prediction of these four models. Our proposed 3D SELD systems outperform the Baseline by a large margin.

Table 1: Experimental results of the audio-only 3D SELD systems on the development dataset using FOA format data.

System	F_{20° ↑	DOAE↓	RDE↓
Baseline-A	0.13	36.90°	0.33
SED-DOA	0.59	12.90°	-
SED-SDE	0.57	-	0.22
SED-DOA+SED-SDE	0.59	12.93°	0.23
SED-DOA-SDE	0.52	13.85°	0.24
SED-SCE	0.52	12.85°	0.24
Model Ensemble	0.59	12.42°	0.21

3.2. Results on Track B

For Track B, we utilized approximately 80 hours of audio-visual training data. We fine-tune the audio-visual 3D SELD models based on the audio pre-trained parameters. Table 2 presents the experimental results of the proposed AV 3D SELD methods on the development dataset. Systems in Table 2 share the same network architectures as those in 1 expect they are trained using audio-visual data. Similar observations can be made for the audio track and the audio-visual track. Using two separate models, namely SED-DOA and SED-SDE, to perform joint prediction, yields slightly better results compared to using a single model to solve the 3D SELD task. In the table, "AV SED-DOA+SED-SDE" refer to the AV fusion system of the DOA and distance estimation models. "AV Model Ensemble" represents the fusion among several AV 3D SELD systems with a single audio system, denoted as "SED-DOA+SED-SDE" in Table 1. "+PP" indicates the use of a video-guided decision fusion. Our proposed audio-visual system demonstrates significant improvement over the baseline system. Through model ensemble and post-processing methods, all three metrics are improved.

Table 2: Experimental results of the audio-visual 3D SELD systems on the evelopment dataset using FOA format data.

System	F_{20° ↑	DOAE↓	RDE↓
Baseline-AV	0.11	38.40°	0.36
AV SED-DOA	0.58	13.10°	-
AV SED-SDE	0.63	-	0.24
AV SED-DOA+SED-SDE	0.56	12.77°	0.24
AV SED-SCE	0.55	13.16°	0.25
AV Model Ensemble	0.59	12.39°	0.22
AV Model Ensemble +PP	0.61	10.94°	0.22

4. REFERENCES

[1] D. A. Krause, A. Politis, and A. Mesaros, “Sound event detection and localization with distance estimation,” *arXiv*, 2024.

[2] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C.-H. Lee, “The NERC-SLIP system for sound event localization and detection of DCASE2023 challenge,” DCASE2023 Challenge, Tech. Rep., June 2023.

[3] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] 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.

[5] 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.

[6] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, *et al.*, “STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[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.

[9] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, “Spatial scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, April 2024.

[10] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, “Manifold mixup: Better representations by interpolating hidden states,” in *International conference on machine learning*. PMLR, 2019, pp. 6438–6447.

[11] A. S. Roman, B. Balamurugan, and R. Pothuganti, “Enhanced sound event localization and detection in real 360-degree audio-visual soundscapes,” *arXiv preprint arXiv:2401.17129*, 2024.

[12] O. Olgun and H. Hacihabiboglu, “METU SPARG eigenmike em32 acoustic impulse response dataset v0. 1.0,” *Graduate School Inform., Middle East Tech. Univ., Ankara, Turkey, Tech. Rep*, 2019.

[13] D. Berghi, P. Wu, J. Zhao, W. Wang, and P. J. Jackson, “Fusion of audio and visual embeddings for sound event localization and detection,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8816–8820.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[15] Y. Jiang, Q. Wang, J. Du, M. Hu, P. Hu, Z. Liu, S. Cheng, Z. Nian, Y. Dong, M. Cai, X. Fang, and C.-H. Lee, “Exploring audio-visual information fusion for sound event localization and detection in low-resource realistic scenarios,” *Accepted by International Conference on Multimedia and Expo (ICME)*, 2024.