2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) | 979-8-3503-0067-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/APSIPAASC58517.2023.10317385

Improving Sound Event Localization and Detection with Class-Dependent Sound Separation for Real-World Scenarios

Shi Cheng†, Jun Du†, Qing Wang† *, Ya Jiang†,
Zhaoxu Nian†, Shutong Niu†, Chin-Hui Lee‡, Yu Gao§ and Wenbin Zhang§
† University of Science and Technology of China, Hefei, Anhui, PR China
‡ Georgia Institute of Technology, Atlanta, GA, USA
§ AI Innovation Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China

***Abstract*—In this study, we propose a novel approach to sound event localization and detection (SELD) by using sound separation (SS) models to tackle key challenges of a high percentage of overlapped segments between sound events and imbalanced distributions of sound event classes in real-world scenarios. Specifically, we introduce class-dependent SS models to deal with overlapping mixtures and extract features from the SS model as prompts for SELD of a specific event class. The proposed SS-SELD method enhances the overall performance of the SELD system, resulting in improved accuracy and robustness in real-world scenarios. In contrast to many other classification methods that can be affected by the interference events, the proposed class-dependent SS framework enhances the overall performance of the SELD system, resulting in improved accuracies and robustness in real-world scenarios. When evaluated on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset, we demonstrate significant improvements in both sound event detection (SED) and direction-of-arrival (DOA) estimation. Our findings suggest that sound separation is a promising strategy to enhance the performance of SELD systems, particularly in scenarios with high overlaps between sound events and imbalanced distributions of event classes. In addition, our proposed framework had contributed building to our champion systems submitted to the Challenge of DCASE 2023 Task 3.**

I. Introduction

Sound event localization and detection (SELD) is an essential task in various audio processing applications, including surveillance systems [1]–[3], environmental monitoring [4], [5], and augmented reality [6], [7]. It was initially introduced as Task 3 in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge [8], [9]. The goal of SELD is to accurately estimate the temporal onsets and offsets, spatial locations, as well as categories of sound events simultaneously within an audio recording. Over the past few years, significant advancements have been made in SELD techniques, driven by the increasing demand for robust and efficient sound analysis algorithms.

SELDnet was proposed to map the input to the sound event detection (SED) and the 3-D Cartesian coordinates of direction-of-arrival (DOA) with two separated branches [10], [11]. In this paper, we use SEDDOA to replace the two-branch network. Shimada et al. introduced an output format called Activity-coupled Cartesian DOA (ACCDOA) for SELD [12]. ACCDOA associates event activity with the length of the Cartesian DOA vector, enabling the solution of SELD task without the requirement of separate branches [13], [14]. However, ACCDOA assumes that each event occurs only once within a frame. When multiple sources of the same event class exist in different directions at the same time, ACCDOA becomes unreliable, leading to angular errors. To address this issue, multi-ACCDOA was proposed [15], which adds a track for concurrence of sources in the prediction results, thereby yielding more accurate localization results. Wang et al. employed data augmentation techniques and performed model fusion with the previous SELD methods, achieving the state-of-the-art performance and ranking the first place in the DCASE 2022 Task 3 Challenge [16]. While SELD models have achieved impressive performance in real-world scenarios, obtaining excellent results in the DCASE 2019-2022 Task 3 Challenge series [8], [9], [13], [17], they face challenges due to the unbalanced distribution of sound events and severe overlapping in real-world scenes. As a result, pure SELD classification systems struggle to effectively model all sound event classes. Therefore, we aim to explore a frontend approach to enable the SELD system to better model each individual class. It has been proven that the separation method is effective in weakly labeled sound event detection (SED) tasks [18]–[20]. However, the effectiveness of sound separation in SELD tasks has been less extensively studied. In DCASE 2022, Jin-Young and his team attempted to use separation as a branch to handle SED information, ultimately merging it with DOA information for further processing [21]. However, such an approach led to the waste of spatial information contained in the 4-channel recordings within the SED branch and resulted in SED information relying entirely on the performance of separation, which could easily lead to unstable results.

In this paper, we introduce a new framework that employs sound separation (SS) as the front-end of SELD model, named SS-SELD. By combining the separated sound with the original mixture, better results can be achieved compared to the traditional SELD methods which only use the mixture. We evaluate our approach on the Sony-TAu Realistic Spatial Soundscapes 2023 (STARSS23) dataset [22] and demonstrate significant

* Corresponding author: qingwang2@ustc.edu.cn

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.
979-8-3503-0067-3/23/$31.00 ©2023 IEEE

improvements in both SED and DOA estimation. Our findings suggest that sound separation is a promising strategy to enhance the performance of SELD systems. Besides, the newly proposed methods made a great contribution to our system ranking the first place in the DCASE 2023 Task3 Challenge.

II. Proposed Methods

In real-world scenarios, sound events often exhibit a high degree of overlap, as depicted in Table I. It can be observed that all sound events exhibit an overlap rate of over 50%, with certain classes such as clapping and laughter having overlap rates exceeding 80%. The Bell class, in particular, has a 100% overlap rate, indicating that all instances of this class overlap with other sound events in the dataset. Traditional SELD models focus on classification-based approaches, facing difficulties in adequately modeling these overlapping events. Building upon this foundation, we propose a novel SS-SELD method. Fig.1 illustrates this new framework.

A. Sound Separation

The newly proposed method consists of two main components. The first introduced component of SS-SELD is the SS method as depicted by the blue dashed box in Fig.1. SS illustrates the inference process of the separation model. As mentioned earlier, due to the uneven distribution of sound events, some events have short durations. Therefore, we sought to enhance the richness of the data by sourcing the required sound events from open-source datasets. Given a mixed audio signal $x(t)$, where t is the sample index in time domain. $x(t)$ contains target sound event $s(t)$ and multiple Interfering sound classes $x_1(t), \cdots, x_m(t)$:

$$x(t) = s(t) + \cdots + x_m(t), \quad m = 1, \ldots, n \tag{1}$$

where n represents the number of events in the mixed audio. The separation process is performed as follows:

$$f_\theta(x(t)) = \hat{s}(t) \tag{2}$$

Here, $f_\theta(\cdot)$ denotes the separation model, θ represents the model parameter set. $\hat{s}(t)$ is the extracted sound. In this process, we train a widely used time-domain model architecture, namely, Conv-TasNet [23] on the dataset simulated from both STARSS23 and Audioset. We first segment data into single-event segments based their labels, which are then added in the time domain to generate mix audios. Simulating training data for the sound separation model also considers the proportion of overlap in natural conditions. The separation model is capable of filtering out the mixed audios and providing separated event distributions, which can serve as guidance for further analysis and modeling. It is worth mentioning that due to significant differences between different event classes, we constructed a dedicated training set for each class with that class as the target. This allows us to achieve class-dependent sound separation. This approach effectively mitigates detection errors of low-intensity classes with significant overlap.

During the fusion process of SS-SELD, the mixture $x(t)$ will be first passed through Audio Channel Swapping (ACS) to obtain augmented signal $x'(t)$. Then for each class, $x'(t)$ will be sent into the separation model to obtain separated sound $\hat{s}(t)$ that only contains the specific class of interest. Afterward, audio signal $\hat{s}(t)$ and $x'(t)$ undergo feature extraction modules separately. $S_{\text{lm}}(k,l)$ denotes log Mel-spectrogram extracted from $\hat{s}(t)$, $X_{\text{lm}}(k,l)$ and $X_{\text{iv}}(k,l)$ denote log Mel-spectrogram and intensity vectors extracted from $x'(t)$, respectively. These features are concatenated to train the SS-SELD model, resulting in output O for detection and localization. We can represent the entire process using the following equations, corresponding representations in Fig.1:

$$x'(t) = \text{ACS}(x(t)) \tag{3}$$

$$\hat{s}(t) = \text{SS}(x'(t)) \tag{4}$$

$$S_{\text{lm}}(k,l) = \text{FE1}(\hat{s}(t)) \tag{5}$$

$$\{X_{\text{lm}}(k,l), X_{\text{iv}}(k,l)\} = \text{FE2}(x'(t)) \tag{6}$$

$$O = \text{SELD}(S_{\text{lm}}(k,l), X_{\text{lm}}(k,l), X_{\text{iv}}(k,l)) \tag{7}$$

where ACS$(\cdot)$, SS$(\cdot)$, FE1$(\cdot)$, FE2$(\cdot)$ and SELD$(\cdot)$ denote the audio channel swapping, sound separation, separated sound feature extraction, mixed sound features extraction and SELD predicting operations, respectively. The final output O contains both frame-level sound event labels and the angle information.

B. Sound Event Localization and Detection

Fig.1 (b) illustrates the training process of the traditional SELD model. Mixture from STARSS23 undergo the data augmentation technique ACS [24] to increase the data size by approximately 8 times. The augmented mixture are then fed into the feature extraction module, which extracts log Mel-spectrogram containing event label information and intensity vector features containing spatial location information. These features are used to train the SELD which follows a similar principle to the traditional SELD approach, consistent with the SELD methodology employed in our winning solution for DCASE 2022 TASK 3. For more detailed information, please refer to reference [16].

III. Experiment Setup

We evaluate SELD on the official development set of the DCASE 2023 Task 3, which is collected in realistic spatial soundscapes [22]. The set totals 168 recording clips (about 9 hours), which can be split into a training part (dev-train, 90 clips) and a testing part (dev-test, 78 clips) according to the official set. We have utilized labeled segments of individual sound events from Audioset [25], in addition to the DCASE 2023 Task 3 dev-train data, to construct training data for sound separation. The total duration of single-event segments amounts to approximately 39.42 hours. Out of this duration, 6.21 hours are obtained from the STARSS23 dev-train dataset, while the remaining 33.21 hours are sourced from the Audioset dataset. For each sound event class, we utilize its corresponding single-event segments as targets and treat the remaining sound events as noise. During the data simulation

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.

TABLE I
OVERLAPPING PERCENTAGE (%).

	Woman speaking	Man speaking	Clap	Telephone	Laughter	Domestic sounds	Footsteps	Door	Music	Musical instrument	Water tap	Bell	Knock
Overlap	59.98	54.25	83.19	51.89	84.79	69.56	54.79	66.77	67.20	58.62	79.79	100.00	82.22

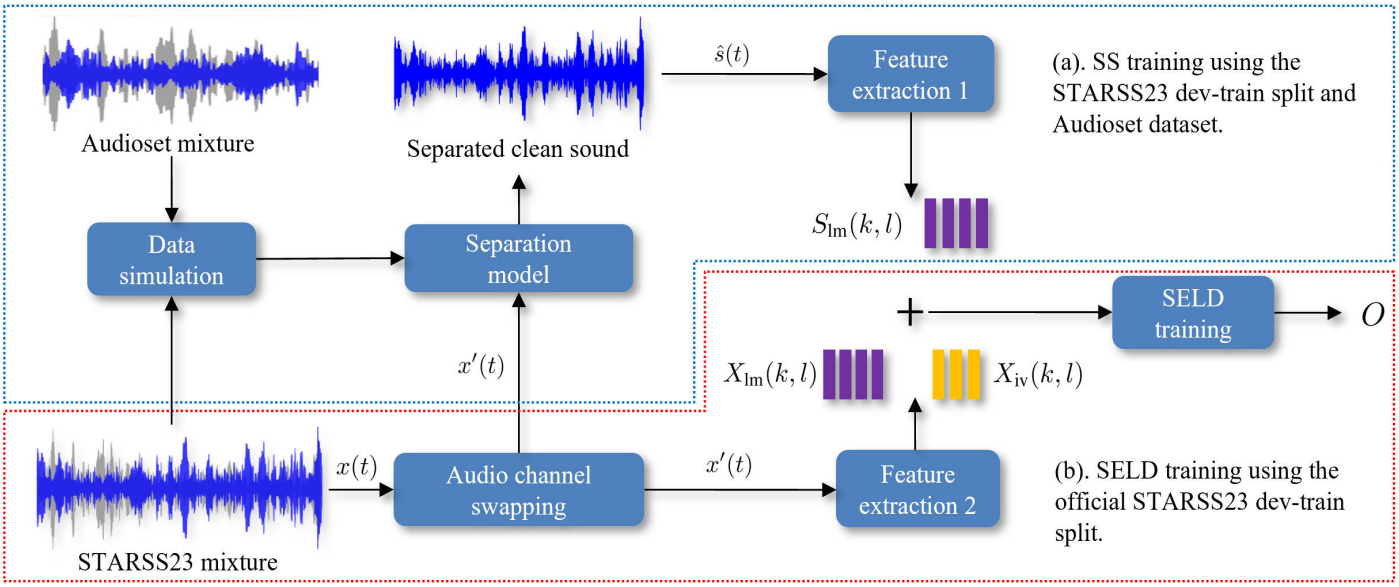

Fig. 1. Framework of the proposed SS-SELD system. Note that (a) in the blue dashed box represents the inference process of SS. while (b) in the red dashed box represents the inference process of the SELD model. In the data simulation part, the data from the dev-train set and open-source dataset are split into single-event segments based on their labels. Once the target class is determined, other classes are considered as interference, and simulation is carried out accordingly.

phase, we overlay the target sound event segments with other sound event segments in the time domain to create mixed audio. This allows us to simulate approximately 40 hours of data, which is used for training the corresponding separation model. All recordings from STARSS23 are 4-channel with a 24 kHz sampling rate. Similarly, single-channel speech from Audioset are also downsampled to 24 kHz. There are 13 sound classes in total. We apply the short-term Fourier transform (STFT) with 40 ms frame length and 20 ms frame hop on 4-channel first-order Ambisonics (FOA) [26] audios to extract log Mel-spectrogram. We concatenate them to get the 11-channel feature at each frame. Considering its outstanding performance in real-world scenarios in recent years, the separation models is trained on ConvTasNet architecture and the model parameter settings can be referenced in [23].

For the SS-SELD training, the data size can be augmented to 8 times by applying the ACS strategy, which results in about 192 hours of data. We use the ResNet-Conformer as our main model architecture [27]. The model employs 8 attention heads, with input, key, and value vectors having dimensions of 256, 32, and 32, respectively. It consists of 8 Conformer layers. The optimizer used is Adam, and a warm-up learning schedule is employed with a maximum learning rate of 0.001. Batch-size is set to 32. The training process is conducted for a maximum of 180,000 steps (which is around). All experiments are conducted on a NVIDIA Tesla V100 Graphic Card-32GB GPU.

All methods are evaluated using SELD_{score} [28], which is calculated as follows:

$$\text{SELD}_{score} = \frac{\text{ER}_{20^\circ} + (1 - \text{F}_{20^\circ}) + \text{LE}'_{\text{CD}} + (1 - \text{LR}_{\text{CD}})}{4} \tag{8}$$

$$\text{LE}'_{\text{CD}} = \frac{\text{LE}_{\text{CD}}}{\pi} \tag{9}$$

where ER_{20° and F_{20° are location-dependent error rate and F-score when the spatial error is within 20°. Note that the F_{20°, LE_{CD} and LR_{CD} are calculated through macro-averaging follow the official definition.

IV. RESULTS AND ANALYSIS

The dev-train split of the official development set of the DCASE 2023 Task 3 is used for training the SELD model as

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.

TABLE II
EXPERIMENTAL RESULTS FOR DEVELOPMENT SET OF DCASE 2023 TASK 3 BASELINE AND SOME SELD METHODS, AS WELL AS THE PROPOSED SS-SELD METHOD. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

	$ER_{20^\circ} \downarrow$	$F_{20^\circ} \uparrow$	$LE_{CD} \downarrow$	$LR_{CD} \uparrow$	$SELD_{score} \downarrow$
Baseline-FOA [22]	0.57	0.30	21.60	0.48	0.478
SEDDOA	0.41	0.59	14.05	0.70	0.304
Multi-ACCDOA	0.44	0.58	13.75	0.74	0.303
ACCDOA	0.42	0.59	13.72	0.72	0.300
SS-SELD	**0.40**	**0.64**	**13.40**	**0.74**	**0.279**

well as the newly proposed SS-SELD framework. The dev-test split is used to validate the performance of the models.

Table II shows the overall experimental results of the proposed method for development dataset. "ACCDOA", "SEDDOA" represent the ACCDOA- and SEDDOA-based modeling method, where SEDDOA adopts a two-branch (SED and DOA) approach for SELD modeling. "Multi-ACCDOA" represents the multi-ACCDOA-based method. SS-SELD denotes the proposed SELD model with SS front-end, with the same output format as SEDDOA. As shown in the table, each proposed single model outperforms the two baseline systems by a large margin. It can also be observed that compared to the baseline, traditional SELD models (SEDDOA, ACCDOA, Multi-ACCDOA) exhibit a decrease of approximately 0.18 in $SELD_{score}$, dropping from 0.478 to around 0.300. Despite attempting to use more data and iterative optimization, there was no significant improvement in the performance of the SELD models. However, with the assistance of the separation method, our SS-SELD system further reduces $SELD_{score}$ to 0.279, breaking through the bottleneck around 0.30 for SELD performance. Additionally, the results of the ER_{20° and F_{20° metrics demonstrate that the inclusion of the separation method significantly improves the model's performance in the detection aspect. As the localization results rely on the accuracy of the detection, the introduction of separation indirectly enhances the localization performance, leading to improvements in LE_{CD} and LR_{CD} as well.

Furthermore, we present the results for each individual class as shown in Table III. It can be observed that the separation method achieves significant improvements in the majority of classes, especially for classes like 'class11' (i.e., *Bell*). These classes have short durations within the entire test set but poss distinct features. The separation method effectively distinguishes them from other classes. However, for the class Water tap, the separation method is ineffective. Through specific analysis, we discovered that Water tap sounds are often low in volume, distant from the microphones (the official dataset provides auxiliary video information, but it is not allowed to be used during the testing process in the Audio-only track. Therefore, we only utilized the video information for analysis purposes.) More importantly, there exists interference noise in the test set that closely resembles the sound of Water tap. This makes it challenging for the separation model to accurately isolate the target Faucet sounds, thereby misleading

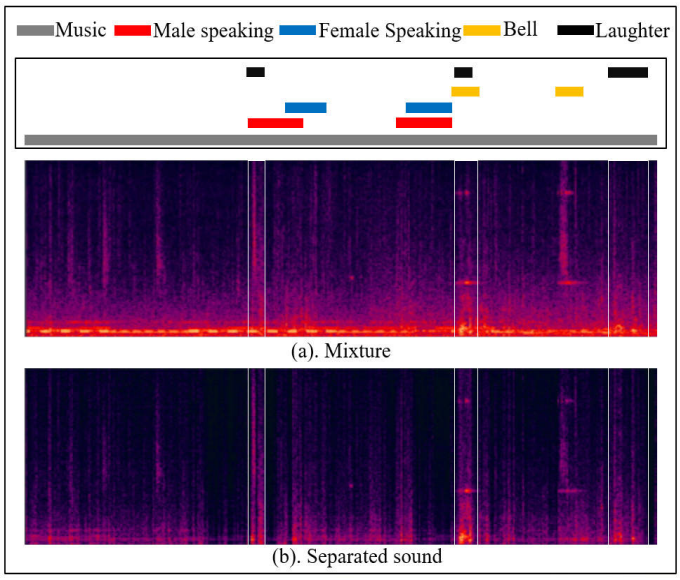

Fig. 2. Comparison of spectrograms before and after sound separation, where the colored rectangular bars at the top represent temporal event labels. The target class in this case is laughter. The model is able to preserve the laughter while effectively removing other sounds such as music.

the training of the subsequent SS-SELD model.

Fig.2 illustrates the spectrogram comparison of the input mixture before and after separation. The colored rectangular tags at the top represent different sound events, with laughter being the target class in this example. The white regions represent the target sound event, which is *Laughter*. It can be observed that the input mixture contains background 'class 8' (*Music*) with high energy in the low-frequency range throughout the entire audio segment, as well as overlapping conversations between different speakers ('class 0' and 'class 1', i.e., *Female speaking* and *Male speaking*), and occasional instances of bell sounds. After passing through the separation model, the target class, laughter, is preserved completely while effectively removing the background music and other interfering sound events. Additionally, in the first white rectangular region, 'class 4' (*Laughter*) in the original mixed audio is almost completely masked by 'class 8' (*Music*). This would cause significant interference for the laughter class in the SELD methods. However, after applying the separation process, it can be observed that the music sound is almost completely removed, allowing the laughter to be clearly revealed. This ensures 'class 4' can be detected more accurately. All of the above indicate that the application of the separation model can eliminate non-target classes with high energy, allowing the extraction of target classes that may have been masked by other sounds with lower energy. This provides a cleaner reference for the SELD model.

In order to analyze the performance of the model more comprehensively, we visualized the SED and DOA estimation, presented in Fig.3. The horizontal axis in all subplots

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.

TABLE III
SELD$_{score}$ COMPARISON BETWEEN SELD SYSTEM AND THE PROPOSED SS-SELD METHOD OF ALL 13 CLASSES FOR DEVELOPMENT DATASET. THE BETTER RESULTS ARE HIGHLIGHTED IN BOLD.

	Woman speaking	Man speaking	Clap	Telephone	Laughter	Domestic sounds	Footsteps	Door	Music	Musical instrument	Water tap	Bell	Knock
SELD	0.243	0.231	0.254	0.345	0.377	0.341	0.458	0.318	0.241	0.257	**0.276**	0.338	0.278
SS-SELD	**0.227**	**0.217**	**0.231**	**0.313**	**0.359**	**0.303**	**0.421**	**0.282**	**0.236**	**0.254**	0.314	**0.248**	**0.244**

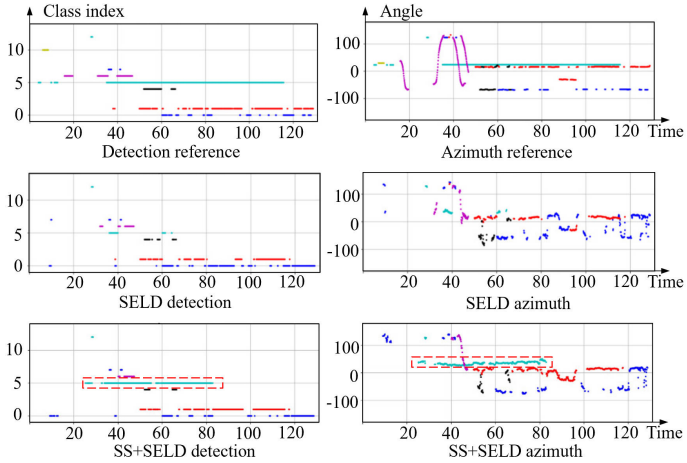

Fig. 3. The visualized results comparing SELD and SS-SELD, with the two columns representing the detection results and azimuth angle results, respectively. Different colors represent different events. The horizontal axis represents time (s) for all subplots. The vertical axis in the left column represents event class index, while the vertical axis in the right column represents angles in degrees (°).

represents time in seconds, while different colors represent different sound events. The first column represents the SED results, where the vertical axis represents 13 sound event classes. The occurrence of a color indicates the presence of the corresponding sound event at that time. The second column represents the DOA results. It is worth mentioning that the presented information pertains to the azimuth angle. The trend in the elevation angle is similar to that of the azimuth angle. To avoid clutter and congestion in the figure, the elevation angle information is not shown. The azimuth angle ranges from -180° to 180°, illustrating the relative angles of sound events with respect to the origin of the microphone array. The first row represents the ground truth results, while the second and third rows represent the results of SELD and SS-SELD, respectively. From the figure, it can be observed that SS-SELD exhibits a significant improvement over SELD in the 'class 5' (light blue line, *Domestic sounds*), as shown in the dashed boxes. SELD almost entirely missed the home-related sounds, thus failing to provide angle information for this class. In contrast, SS-SELD can detect the majority of these sounds and accurately provide angle information. This further demonstrates that our SS-SELD not only enhances the model's detection capability but also brings improvements in localization performance.

V. CONCLUSIONS

In this study, we explored the effectiveness of the sound separation method in sound event localization and detection task. By using a separation model as a front-end to obtain separated audio for each event class, we were able to extract the features of both the sound event mixtures and the separated sound. The combined SS-SELD model trained on these features effectively improved the performance of single SELD model. In future work, we will continue to explore more effective and universally applicable separation strategies to address the existing issues and enhance the performance of the SELD model.

VI. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grants No. 62171427.

REFERENCES

[1] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, pp. 1–46, 2016.

[2] A. Greco, N. Petkov, A. Saggese, and M. Vento, "Aren: A deep learning approach for sound event recognition using a brain inspired representation," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3610–3624, 2020. DOI: 10.1109/TIFS.2020.2994740.

[3] V. Suruthhi, V. Smita, R. Gini J, and K. Ramachandran, "Detection and localization of audio event for home surveillance using crnn," *International Journal of Electronics and Telecommunications*, vol. 67, no. 4, 2021.

[4] F. Angulo, S. Essid, G. Peeters, and C. Mietlicki, "Cosmopolite sound monitoring (cosmo): A study of urban sound event detection systems generalizing to multiple cities," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095833.

[5] W. Xiong, X. Xu, L. Chen, and J. Yang, "Sound-based construction activity monitoring with deep learning," *Buildings*, vol. 12, no. 11, p. 1947, 2022.

[6] A. Härmä, J. Jakka, M. Tikander, *et al.*, "Augmented reality audio for mobile and wearable appliances," *Journal of the Audio Engineering Society*, vol. 52, no. 6, pp. 618–639, 2004.

[7] B. S. Liang, A. S. Liang, I. Roman, *et al.*, "Reconstructing room scales with a single sound for augmented reality displays," *Journal of Information Display*, vol. 24, no. 1, pp. 1–12, 2023.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.

[8] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," *arXiv preprint arXiv:1905.08546*, 2019.

[9] 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.

[10] 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, 2019. DOI: 10.1109/JSTSP.2018.2885636.

[11] S. Kapka and M. Lewandowski, "Sound source detection, localization and classification using consecutive ensemble of crnn models," *arXiv preprint arXiv:1908.00766*, 2019.

[12] 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 *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 915–919. DOI: 10.1109/ICASSP39728.2021.9413609.

[13] 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," *arXiv preprint arXiv:2106.06999*, 2021.

[14] O. Slizovskaia, G. Wichern, Z.-Q. Wang, and J. Le Roux, "Locate this, not that: Class-conditioned sound event doa estimation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 711–715. DOI: 10.1109/ICASSP43922.2022.9747604.

[15] 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)*, 2022, pp. 316–320. DOI: 10.1109/ICASSP43922.2022.9746384.

[16] Q. Wang, L. Chai, H. Wu, *et al.*, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," *DCASE2022 Challenge, Tech. Rep.*, 2022.

[17] A. Politis, K. Shimada, P. Sudarsanam, *et al.*, "Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *arXiv preprint arXiv:2206.01948*, 2022.

[18] N. Turpault, S. Wisdom, H. Erdogan, *et al.*, *Improving sound event detection in domestic environments using sound separation*, 2020. arXiv: 2007.03932 [cs.SD].

[19] Q. Zhou, Z. Feng, and E. Benetos, "Adaptive noise reduction for sound event detection using subband-weighted nmf," *Sensors*, vol. 19, no. 14, p. 3206, 2019.

[20] S. Liu, F. Yang, F. Kang, and J. Yang, "A multi-task learning method for weakly supervised sound event detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8802–8806. DOI: 10.1109/ICASSP43922.2022.9746947.

[21] J.-Y. Park, D.-H. Kim, B. H. Ku, *et al.*, "Sound event localization and detection based on cross-modal attention and source separation,"

[22] A. Politis, K. Shimada, P. Sudarsanam, *et al.*, "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, Nov. 2022, pp. 125–129. [Online]. Available: https://dcase.community/workshop2022/proceedings.

[23] 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. DOI: 10.1109/TASLP.2019.2915167.

[24] 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. DOI: 10.1109/TASLP.2023.3256088.

[25] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "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.

[26] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[27] S. Niu, J. Du, Q. Wang, *et al.*, "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.

[28] 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. DOI: 10.1109/WASPAA.2019.8937220.

Authorized licensed use limited to: University of Science & Technology of China. Downloaded on May 08,2024 at 14:05:09 UTC from IEEE Xplore. Restrictions apply.