DESCRIPTION AND DISCUSSION ON DCASE 2025 CHALLENGE TASK 4: SPATIAL SEMANTIC SEGMENTATION OF SOUND SCENES

Masahiro Yasuda[*1], *Binh Thien Nguyen*[*1], *Noboru Harada*[1], *Romain Serizel*[2], *Mayank Mishra*[2]
Marc Delcroix[1], *Shoko Araki*[1], *Daiki Takeuchi*[1], *Daisuke Niizumi*[1]
Yasunori Ohishi[1], *Tomohiro Nakatani*[1], *Takao Kawamura*[3], *Nobutaka Ono*[3]

[1] NTT, Inc., Japan, masahiro.yasuda@ntt.com
[2] University de Lorraine, CNRS, Inria, Loria, France
[3] Tokyo Metropolitan University, Japan

ABSTRACT

Spatial Semantic Segmentation of Sound Scenes (S5) aims to enhance technologies for sound event detection and separation from multi-channel input signals that mix multiple sound events with spatial information. This is a fundamental basis of immersive communication. The ultimate goal is to separate sound event signals with 6 Degrees of Freedom (6DoF) information into dry sound object signals and metadata about the object type (sound event class) and representing spatial information, including direction. However, because several existing challenge tasks already provide some of the subset functions, this task for this year focuses on detecting and separating sound events from multi-channel spatial input signals. This paper outlines the S5 task setting of the Detection and Classification of Acoustic Scenes and Events (DCASE) 2025 Challenge Task 4 and the DCASE2025 Task 4 Dataset, newly recorded and curated for this task. We also discuss the performance and characteristics of the S5 systems submitted to DCASE 2025 Challenge Task 4 based on experimental results.

Index Terms— Sound event detection and separation, Semantic segmentation of sound scenes, Spatial signal

1. INTRODUCTION

This paper summarizes a newly introduced task for the Detection and Classification of Acoustic Scenes and Events (DCASE) 2025 challenge, named Task 4: Spatial Semantic Segmentation of Sound Scenes (S5) [1], and discusses the experimental results of submitted systems for this challenge [2, 3, 4, 5, 6, 7, 8, 9].

As illustrated in Fig. 1, S5 consists of detecting and extracting sound events from multi-channel spatial input signals. The input signal contains multiple simultaneous sounds as well as background noise. Each output signal should contain one isolated sound event with a predicted label for the event class.

One promising application of S5 is XR services that capture a user's surrounding acoustic scene and transmit it to remote participants. To deliver a believable experience, the mixture must first be decomposed into individual sound objects. Each object is paired with its class label and 6DoF (three-dimensional position and rotation) spatial metadata. By using this, the rendering engine can then update direction and distance as the listener moves or edits the scene in real time. S5 technology can also be applicable for home-assisted living through sound monitoring of the room environment. As a first step for these applications, the present S5 task asks systems to detect the constituent sound events and separate their dry signals from multi-channel spatial recordings.

The S5 task relates to earlier DCASE challenges. DCASE 2021 Task 4 (Sound Event Detection and Separation in Domestic Environments) used single-channel recordings. Separation was optional and did not affect the score; it was only considered as a potential way to improve sound event detection when overlapping sound events are present [10, 11]. S5 instead supplies multi-channel input, allowing systems to exploit spatial cues, and it directly scores the quality of the separated sources. DCASE 2024 Task 3 (Audio and Audiovisual Sound Event Localization and Detection with Source Distance Estimation; SELD) targets direction-of-arrival (DoA) and distance metadata [12]. Although S5 does not require explicit estimation of geometric metadata like SELD, spatial information remains an important key.

2. TASK SETTING OF S5

2.1. Formulation and notation

The S5 task, originally proposed in our prior work [13], aims to detect and separate the sounds of each sound event from a mixture observed by a multi-channel microphone array at various locations in a real environment. This section introduces the notation and task settings.

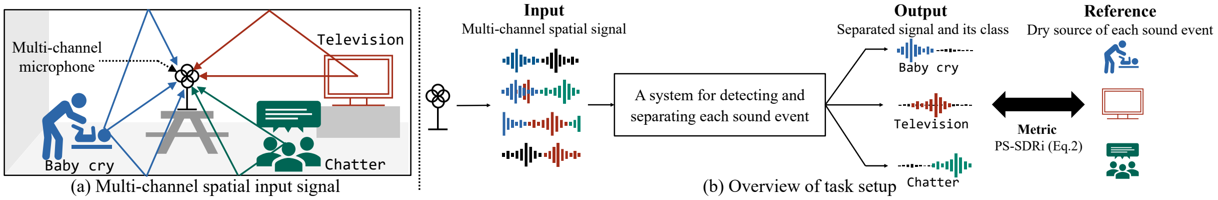
Let $\boldsymbol{Y} = [\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(M)}]^\top \in \mathbb{R}^{M \times T}$ be the multi-channel time-domain mixture signal of length T, recorded with an array of M microphones, where $\{\cdot\}^\top$ is the matrix transposition. We denote $C = \{c_1, ..., c_K\}$ the set of source labels in the mixture, where the source count K can vary from 1 to K_{max}. The m-th channel of $\boldsymbol{Y}$ can be modeled as

$$\boldsymbol{y}^{(m)} = \sum_{k=1}^{K} \boldsymbol{h}_k^{(m)} * \boldsymbol{s}_{c_k} + \left[\sum_{j=1}^{J} \boldsymbol{h}_j^{(m)} * \boldsymbol{s}_{c_j} + \boldsymbol{n}^{(m)} \right]_{\textit{optional}} \tag{1}$$

where $\boldsymbol{s}_{c_k}, \boldsymbol{s}_{c_j} \in \mathbb{R}^T$ are the single-channel dry source signal corresponding to the labels of the target event c_k and interference event c_j, respectively. $\boldsymbol{h}_k^{(m)}, \boldsymbol{h}_j^{(m)} \in \mathbb{R}^H$ are the m-th channel of the length-H room impulse response (RIR) at the spatial position of $\boldsymbol{s}_{c_k}$ and $\boldsymbol{s}_{c_j}$. $\boldsymbol{n}^{(m)} \in \mathbb{R}^T$ is the m-th channel of the multi-channel noise signal. In the S5 task, the desired outputs are the estimates

This work was partially supported by JST Strategic International Collaborative Research Program (SICORP), Grant Number JPMJSC2306, Japan. This work was partially supported by the Agence Nationale de la Recherche (Project Confluence, grant number ANR-23-EDIA-0003).

[*] These authors contributed equally to this work.


Multi-channel microphone
Television
Baby cry
Chatter
Input
Multi-channel spatial signal
A system for detecting and separating each sound event
Output
Separated signal and its class
Baby cry
Television
Chatter
Reference
Dry source of each sound event
Metric
PS-SDRi (Eq.2)

(a) Multi-channel spatial input signal

(b) Overview of task setup

Figure 1: Overview of spatial semantic segmentation of sound scenes.

of individual target sound events $\hat{S} = \{\hat{\boldsymbol{s}}_{\hat{c}_1}, \dots, \hat{\boldsymbol{s}}_{\hat{c}_{\hat{K}}}\}$ and their corresponding class label $\hat{C} = \{\hat{c}_1, \dots, \hat{c}_{\hat{K}}\}$.

2.2. Evaluation method, metric

As evaluation metrics for the S5 task, we used class-aware signal-to-distortion ratio improvement (CA-SDRi) proposed for the S5 task [13]. The metric is defined as

$$\text{CA-SDRi}\left(\hat{S}, S, \hat{C}, C, \boldsymbol{y}^{(m_{\text{ref}})}\right) = \frac{1}{|C \cup \hat{C}|} \sum_{\bar{c} \in C \cup \hat{C}} P_{\bar{c}}, \quad (2)$$

where $|C \cup \hat{C}|$ is the length of the set union. The metric component $P_{\bar{c}}$ is calculated as

$$P_{\bar{c} \in C \cup \hat{C}} = \begin{cases} \text{SDRi}(\hat{\boldsymbol{s}}_{\bar{c}}, \boldsymbol{s}_{\bar{c}}, \boldsymbol{y}^{(m_{\text{ref}})}), & \text{if } \bar{c} \in C \cap \hat{C} \\ \mathcal{P}^{\text{FN}}_{\bar{c}}, & \text{if } \bar{c} \in C \wedge \bar{c} \notin \hat{C} \\ \mathcal{P}^{\text{FP}}_{\bar{c}}, & \text{if } \bar{c} \notin C \wedge \bar{c} \in \hat{C} \end{cases} \quad (3)$$

where the SDRi is calculated as

$$\text{SDRi}(\hat{\boldsymbol{s}}_{\bar{c}}, \boldsymbol{s}_{\bar{c}}, \boldsymbol{y}^{(m_{\text{ref}})}) = \text{SDR}(\hat{\boldsymbol{s}}_{\bar{c}}, \boldsymbol{s}_{\bar{c}}) - \text{SDR}(\boldsymbol{y}^{(m_{\text{ref}})}, \boldsymbol{s}_{\bar{c}}), \quad (4)$$

$$\text{SDR}(\hat{\boldsymbol{s}}, \boldsymbol{s}) = 10 \log_{10}\left(\frac{\|\boldsymbol{s}\|^2}{\|\boldsymbol{s} - \hat{\boldsymbol{s}}\|^2}\right). \quad (5)$$

The key concept of CA-SDRi is that estimated and reference sources are aligned based on their labels. The waveform metric, SDRi, is calculated only when the label is correctly predicted, i.e., in the first case of (3). For incorrect label predictions, including false negatives (FN) and false positives (FP), which correspond to the second and third cases in (3), the penalty values $\mathcal{P}^{\text{FN}}_{\bar{c}}$ and $\mathcal{P}^{\text{FP}}_{\bar{c}}$ are accumulated. In this study's evaluation, both $\mathcal{P}^{\text{FN}}_{\bar{c}}$ and $\mathcal{P}^{\text{FP}}_{\bar{c}}$ were set to 0, indicating that incorrect predictions do not contribute to the metric. For the reference channel m_{ref}, we use the omnidirectional channel i.e., $m_{\text{ref}} = 0$.

The CA-SDRi described above sets the dry source signal $\boldsymbol{s}_k$ as the target signal. However, direct comparison with the dry source can be challenging. It requires sample-accurate compensation for the distance-dependent delays between each source and microphone. Such precise alignment lies outside S5's main focus on detection and separation for now. To relax this requirement, we replace each dry source with the signal obtained by convolving it with the direct-path component of the impulse response, forming

$$\boldsymbol{s}^{(d)}_{c_k} = \boldsymbol{h}^{(m_{\text{ref}},d)}_k * \boldsymbol{s}_{c_k}. \quad (6)$$

Here, $\boldsymbol{h}^{(m_{\text{ref}},d)}_k$ is the direct path component of the impulse response at the reference microphone m_{ref}. CA-SDRi, in this study, is computed by using $\boldsymbol{s}^{(d)}_{c_k}$ as reference. Note that, although our notation differs, the metric is identical to the one adopted in the previous study [13].

Finally, the ranking score is the average CA-SDRi across all clips. In addition to the primary ranking score function CA-SDRi, we will also provide PESQ [14] and STOI [15] for speech, and PEAQ [16] scores for signals other than speech as informative metrics representing perceptual quality.

3. DCASE2025 TASK4 DATASET

3.1. General overview

For the S5 task, we designed and recorded a new dataset named **DCASE2025 Task4 Dataset** [17, 18]. Because of the challenges in evaluating S5 for real recordings, we opted for a simulated dataset, making considerations to make it as realistic as possible. The resources needed to build such a dataset consist of the following.

- **Isolated target sound events $\boldsymbol{s}_{c_k}$:** isolated recordings of diverse sound event classes. In light of Eq. (1), these signals are preferably captured in anechoic conditions.
- **Room-impulse responses (RIRs) $\boldsymbol{h}^{(m)}_k$:** multichannel RIRs measured in various rooms.
- **Environmental noise $\boldsymbol{n}^{(m)}$:** environmental noise recorded with the same multichannel microphone array used for the RIRs in indoor and outdoor environments.

In addition to stationary, direction-independent environmental noise, it is often useful to include sporadic sound events that do not belong to the target classes. We refer to these as *interference sounds* and handle them as follows:

- **Interference sounds:** Recording of sound events of classes not included in the isolated target sound events mentioned above. When forming mixtures with Eq. 1, the interference signals are processed similarly to the target sound events.

The DCASE2025 Task4 Dataset consists of new recordings and curated data from publicly available datasets [19, 20, 21]. Eighteen classes listed in Table. 1 are selected for the target sound events. Among the materials that make up the DCASE2025 Task4 Dataset, our newly recorded material and curated materials from FOA-MEIR [21] are released on Zenodo [17, 18]. The remaining materials are obtained and filtered from their respective public datasets [19, 22, 20] via a download script that we provide on GitHub [23]. The training and evaluation mixtures $\boldsymbol{y}^{(m)}$ were synthesized from the above materials using a modified version of a spatial-audio simulator named SpatialScaper [24]. All mixtures were synthesized at 32kHz/16bit with a 10 second length.

3.2. Development dataset

The development dataset of the DCASE2025 Task4 Dataset [17] was constructed from various datasets, including both existing and newly recorded data specifically for this task. The isolated target sound events for the development set consist of our newly recorded data and curated samples from FSD50K [19] and EARS [22]. RIR

Table 1: The amount of recorded samples of isolated target sound events used to synthesize the DCASE 2025 Task 4 dataset. The numbers in parentheses indicate the number of samples curated from publicly available datasets [19, 22, 20], not newly recorded.

		Alarm Clock	Blender	Buzzer	Clapping	Cough	Cupboard OpenClose	Dishes	Doorbell	Foot Steps	Hair Dryer	Mechanical Fans	Musical Keyboard	Percus sion	Pour	Speech	Typing	Vacuum Cleaner	Bicycle Bell
Dev	duration [s]	699	1337	359	1022	1501	1434	1174	526	2143	988	1912	2959	5321	601	4047	3304	1414	548
	# of samples	72 (42)	95 (52)	73 (21)	376 (312)	288 (211)	262 (184)	269 (208)	68 (63)	336 (298)	29 (25)	82 (59)	434 (397)	1881 (1796)	71 (37)	1203 (1195)	308 (228)	53 (46)	110 (54)
Eval	duration [s]	157	260	164	147	59	108	120	147	243	78	202	250	69	163	109	258	156	179
	# of samples	12	19	20	23	24	21	16	19	19	2	4	21	13	24	28	35	3	21

dataset was constructed by merging newly recorded RIRs for this task with curated material from the publicly released FOA-MEIR dataset [21]. Newly recorded RIRs were captured at six microphone positions—two positions in each of three rooms. At each position we captured 108 RIRs, giving 648 responses in total. The three rooms represent different acoustic conditions: a small conference room, a music-listening room, and a room with reflective walls. All background noise were curated from the publicly available FOA-MEIR dataset [21]. All interference was curated by removing items related to our target 18 classes from the background set in the dataset for Semantic Hearing [20]. The procedures for the new recordings and the curated material drawn from public datasets are detailed in Sec. 3.4.

The development dataset was further divided into three subsets: training, validation, and test. Mixtures for the test subset were pre-synthesized, while for the training and validation subsets, participants may generate mixtures using the provided data. Each clip contains between one and three target sound events, with at most three target sound events active simultaneously. The SNR for each target sound event ranges from 5 to 20 dB, with respect to the background environmental noise. Each mixture also includes one or two interfering sound events, with SNRs ranging from 0 to 15 dB.

3.3. Evaluation dataset

To ensure fairness in the challenge, isolated sources of the target sound events, RIRs, interference sounds, and environmental noise were all newly recorded for the evaluation data. In other words, the evaluation data does not include any publicly available data.

The main subset (files 0000–1619) contains mixtures with 1, 2 or 3 target sound events; 0, 1 or 2 interference events; and RIRs recorded at six microphone positions. The six positions are distributed across three rooms, with two positions recorded in each room: a small conference room, a large conference room and a room with reflective walls. These rooms are not included in the development set. These factors yield $3 \times 3 \times 6 = 54$ acoustic settings, and 30 mixtures were synthesized for each setting. The SNR settings for mixing are the same as for the development set.

Files 1620–2033 constitute the ‘partially known conditions’ subset. Mixtures in this subset were synthesized so that one element—RIRs, target sound event, background noise, or interference sound—was drawn from the training partition of the development set. This design enables a factor-specific evaluation of how well the systems generalize beyond their training conditions. The following subsets exceed the scope of this year’s S5 task. Files 2034–2141 contain no target events, while files 2142–2249 include multiple same-class target events arriving from different directions. Files 2250–2289 are recordings made in real indoor and outdoor environments; consequently, oracle target sources are unavailable in this split.

3.4. Recording details

Isolated target sound events were recorded in an anechoic chamber. Fig. 2 shows the configuration of microphones for this recording.

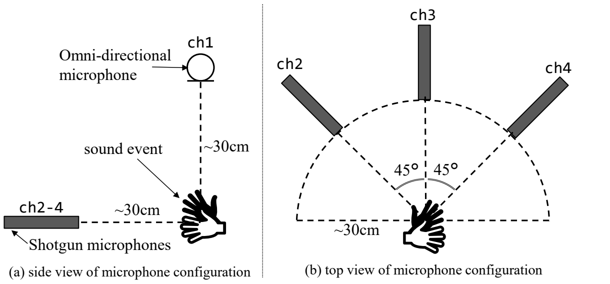

Omni-directional microphone
ch1
sound event
~30cm
ch2-4
~30cm
Shotgun microphones
(a) side view of microphone configuration
ch3
ch2
ch4
45°
45°
~30cm
(b) top view of microphone configuration

Figure 2: The Configuration of microphones used to record isolated sound events in an anechoic chamber

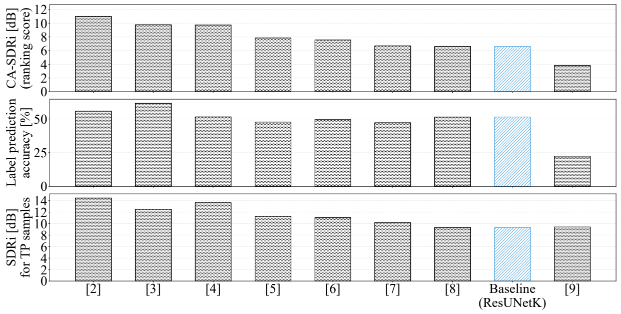


Figure 3: This plot shows performance of the submitted systems and the baseline (ResUNetK) on the evaluation set. Only the systems with the best CA-SDRi performance among those submitted by each team are selected.

The recording was made using three cardioid microphones to capture the sound events from the left, front, and right, and one omnidirectional microphone to capture the sound from above. The 4-channel configuration was not intended to be used as a microphone array, but simply to increase the variety of data for the target sound event by recording from different directions. During mixture synthesis for S5, one of these four channels is chosen at random and used as the monaural target sound event.

All RIRs were captured with the same first-order Ambisonics microphone (Sennheiser AMBEO VR Mic) and are provided in B-format (AmbiX). In both the development set and evaluation set, 108 RIRs were measured at each microphone location. These are composed of the relative positions of the following speakers and microphones: (i) the azimuth was swept in 20° steps to cover the full 360°; (ii) the elevation was set to $-20°$, $0°$, or $20°$; and (iii) the source distance was chosen from the range 0.75–1.50 m.

4. CHALLENGE RESULTS

4.1. Overview of submitted system’s performance

We received 24 submissions from 8 teams. Figure 3 shows the best-performing system from each team in terms of CA-SDRi (ranking score). As Fig. 3 (top) shows, seven of the eight teams surpassed our baseline, some of them by a large margin of more than 4 dB, indicating that significant progress toward solving S5 task has been

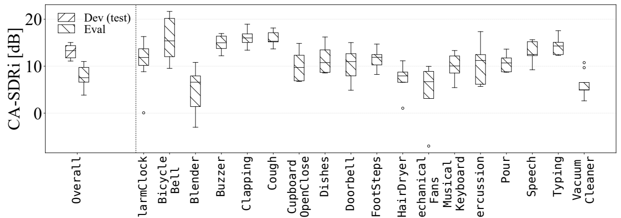


Figure 4: Class-wise box-plot for CA-SDRi score of submitted systems. The leftmost pair shows average score for development set (Dev (test)) and evaluation set (Eval) scores. To its right, eighteen plots present CA-SDRi for each target sound event class in evaluation set.

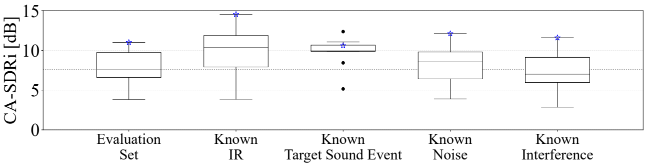


Figure 5: Box plot of CA-SDRi for each team's best system under four partially known conditions. 'Known IR': evaluation RIRs seen in training; 'Known Target': target sound event samples seen; 'Known Noise': background noise seen; 'Known Interference': interference samples seen. The blue star indicates the performance of the top-ranked system [2].

achieved.

As described in Section 2.2, CA-SDRi combines label-prediction accuracy and source-separation quality. To examine these factors separately, we also report label-prediction accuracy (Fig. 3, middle) and SDRi computed only on clips where every sound event was correctly labeled (Fig. 3, bottom). The gains achieved by most systems primarily stem from improved separation. Notably, the top-ranked system places second in label-prediction accuracy but achieves the highest CA-SDRi due to its superior separation performance. This indicate that further improvement may be achieved by combining the separation and label prediction approaches of the two best teams.

Figure 4 compares CA-SDRi on the development-set test split and the evaluation set for best systems per team highlighted in Fig. 3. Across the board, CA-SDRi drops on the evaluation set, suggesting overfitting. In the class-wise CA-SDRi in the evaluation set, particularly poor performance was observed for "Blender" and "VaccumCleaner," and these confusing classes may be lowering the overall performance.

4.2. Characteristic of submitted systems

This section outlines the characteristics of the submitted systems. The teams coupled source separation and its label prediction in several ways. Sequential pipelining, where tagging precedes separation, similar to the Baseline system [13], is employed in systems [4] and [7]. Systems [3] and [9] conditioned the separator with tag information through Feature-wise Linear Modulation (FiLM). An iterative sequential strategy—tagging, separation, and refined tagging—was adopted by [2], [5], and [8]. Systems [2] and [6] performed joint estimation of sound-event labels and separated signals. The system [6] stands out in that it employs a hybrid approach that combines a separation technique based on multichannel signal processing (multichannel Wiener Filter; MWF) and data-driven modeling based on DNN.

Table 2: PESQ, STOI, and PEAQ for the top-three submitted systems abd the baseline sysmte (ResUNetK). PESQ and STOI are evaluated for clips in which speech classes are detected. PEAQ is evaluated for the rest. Errors are indicated by standard deviation.

System	# of detected speech	PESQ↑ (1 – 4.5)	STOI↑ (0 – 1)	PEAQ↑ (-4 – 0)
Baseline [13]	246/251	2.39 ± 0.63	0.84 ± 0.11	-3.60 ± 0.43
Rank 1 [2]	241/251	2.88 ± 0.58	0.91 ± 0.08	-3.43 ± 0.48
Rank 2 [3]	249/251	2.97 ± 0.60	0.90 ± 0.07	-3.39 ± 0.51
Rank 3 [4]	246/251	2.77 ± 0.58	0.90 ± 0.10	-3.43 ± 0.50

4.3. Performance under partially known conditions

As described in Sec. 3, the evaluation set of the DCASE 2025 Task 4 Dataset contains **RIR**, **target sound event**, **background noise**, and **interference sound** that do not appear in the development set. This section examines how CA-SDRi changes when exactly one of these four components is made known, that is, drawn from the training data. Comparing these conditions reveals which unknown factors limit system performance. Figure 5 shows CA-SDRi for the original evaluation set and for the four partially known conditions. The highest median scores occur when either the **RIR** or the **target sound event** are known, indicating that unfamiliar acoustic environments and unseen target sound events are the primary sources of degradation. In contrast, making the **background noise** or the **interference sound** known yields little change relative to the original evaluation set.

4.4. Quality of separated sound events

Table 2 summarizes the perceptual quality of the separated signals for the three top-ranked systems and the baseline. PESQ and STOI were computed on clips containing speech, while PEAQ was computed on the remaining non-speech clips. The top submissions [2, 3, 4] outperform the baseline on every metric. For speech, PESQ scores near 3 ("fair/toll") and STOI values around 0.90 indicate natural and intelligible quality, though additional refinement is needed for high-fidelity applications. For non-speech content, all systems obtain comparatively low PEAQ scores, revealing that perceptual fidelity for other sound events still requires improvement.

5. CONCLUSTION AND FUTURE VIEWS

This paper introduced the Spatial Semantic Segmentation of Sound Scenes (S5) task of DCASE 2025 Challenge Task 4, which targets joint detection and separation of spatial sound events from multichannel mixtures. For this task, we released the DCASE2025 Task 4 Dataset [17, 18], which comprises isolated target events, multichannel RIRs, environmental noise, and interference sounds. We evaluated 24 systems submitted for this challenge by eight teams and analyzed their performances and characteristics. The collective results confirmed that many approaches already surpass the naïve baseline, indicating steady progress in both tagging and separation modules. Future work should address generalization to unseen acoustic environments and novel sound events. A further challenge is separating and labeling multiple instances of the same class within a single clip. This scenario is common in practice but lies outside the current task, and all submissions performed poorly on it. Finally, the current systems evaluated on RIR-based simulation; extending evaluation to real recordings is essential to validate robustness in real-world deployments [1].

[1] Samples of real-recording inference in the top3 system is shown on the challenge results page.

6. REFERENCES

[1] M. Yasuda, N. Harada, B. T. Nguyen, D. Takeuchi, D. Niizumi, M. Delcroix, S. Araki, T. Nakatani, Y. Ohishi, R. Serizel, M. Mishra, N. Ono, and T. Kawamura, "DCASE2025 challenge task 4: Spatial semantic segmentation of sound scenes," 2025. [Online]. Available: https://dcase.community/challenge2025/task-spatial-semantic-segmentation-of-sound-scenes

[2] Y. Kwon, D. Lee, D. Kim, and J.-W. Choi, "Self-guided target sound extraction and classification through universal sound separation model and multiple clues," DCASE2025 Challenge, Tech. Rep., June 2025.

[3] T. Morocutti, F. Schmid, J. Greif, P. Primus, and G. Widmer, "Transformer-aided audio source separation with temporal guidance and iterative refinement," DCASE2025 Challenge, Tech. Rep., June 2025.

[4] F. Wu and Z.-Q. Wang, "TS-TFGRIDNET: Extending tfgridnet for label-queried target sound extraction via embedding concatentaiton," DCASE2025 Challenge, Tech. Rep., June 2025.

[5] X. Zhou, H. Wang, C. Li, B. Han, X. Zheng, and Y. Qian, "Sjtu-audiocc system for dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes," DCASE2025 Challenge, Tech. Rep., June 2025.

[6] Y. Nozaki, S. Sakurai, Y. Bando, K. Saijo, K. Imoto, and M. Onishi, "A hybrid s5 system based on neural blind source separation," DCASE2025 Challenge, Tech. Rep., June 2025.

[7] J. Park, J. Lee, D.-H. Lim, H. K. Kim, H. Geum, and J. E. Lim, "Performance improvement of spatial semantic segmentation with enriched audio features and agent-based error correction for dcase 2025 challenge task 4," DCASE2025 Challenge, Tech. Rep., June 2025.

[8] V. Stergioulis and G. Potamianos, "Redux: An iterative strategy for semantic source separation," DCASE2025 Challenge, Tech. Rep., June 2025.

[9] Z. Wang, S. Wang, Z. Zhang, and J. Yin, "Spatial semantic segmentation of sound scenes based on adapter fine-tuning," DCASE2025 Challenge, Tech. Rep., June 2025.

[10] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[11] R. Serizel, N. Turpault, F. Ronchini, S. Wisdom, H. Erdogan, J. Hershey, J. Salamon, P. Seetharaman, E. Fonseca, S. Cornell, and D. P. W. Ellis, "DCASE2021 challenge task 4: Sound event detection and separation in domestic environments," 2021. [Online]. Available: https://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments

[12] A. Politis, K. Shimada, Y. Mitsufuji, T. Virtanen, P. Sudarsanam, D. Krause, K. Uchida, D. Diaz-Guerra, Y. Koyama, N. Takahashi, T. Shibuya, and S. Takahashi, "DCASE2024 challenge task 3: Audio and audiovisual sound event localization and detection with source distance estimation," 2024. [Online]. Available: https://dcase.community/challenge2024/task-audio-and-audiovisual-sound-event-localization-and-detection-with-source-distance-estimation

[13] B. T. Nguyen, M. Yasuda, D. Takeuchi, D. Niizumi, Y. Ohishi, and N. Harada, "Baseline systems and evaluation metrics for spatial semantic segmentation of sound scenes," in *2025 33rd European Signal Processing Conference (EUSIPCO)*, 2025. [Online]. Available: https://arxiv.org/abs/2503.22088

[14] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[16] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ – the ITU standard for objective measurement of perceived audio quality," *Journal of Audio Engineering Society, Volume 48, Issue 1/2, pp. 3-29, February 1, 2000*, 2000.

[17] M. Yasuda, B. T. Nguyen, N. Harada, and D. Takeuchi, "DCASE2025 Task 4 S5 Development set," 2025. [Online]. Available: https://zenodo.org/records/15117227

[18] ——, "DCASE2025 Task 4 S5 Evaluation set," 2025. [Online]. Available: https://zenodo.org/records/15553984

[19] 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.

[20] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–15.

[21] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2022, pp. 226–230.

[22] J. Richter, Y.-C. Wu, S. Krenn, S. Welker, B. Lay, S. Watanabe, A. Richard, and T. Gerkmann, "Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Interspeech 2024*, 2024, pp. 4873–4877.

[23] "DCASE2025 Task 4 Baseline," 2025. [Online]. Available: https://github.com/nttcslab/dcase2025_task4_baseline

[24] 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 Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2024, pp. 1221–1225.