A HYBRID S5 SYSTEM BASED ON NEURAL BLIND SOURCE SEPARATION

Technical Report

Yuto Nozaki[*1], *Shun Sakurai*[*1,2], *Yoshiaki Bando*[*1], *Kohei Saijo*[1,3], *Keisuke Imoto*[1,4], *Masaki Onishi*[1]

[1] National Institute of Advanced Industrial Science and Technology, Tokyo, Japan,
[2] University of Tsukuba, Ibaraki, Japan, [3] Waseda University, Tokyo, Japan,
[4] Kyoto University, Kyoto, Japan
{yuto.nozaki, sakurai.shun, y.bando}@aist.go.jp

ABSTRACT

In this paper, we report our hybrid system for the DCASE 2025 Challenge Task 4 based on neural blind source separation (BSS). This task, called spatial semantic segmentation of sound scenes (S5), aims to detect and separate sound events from a multichannel mixture signal. To make the separation model robust against unseen audio environments, we leverage neural BSS to combine robust statistical signal processing and high-performing neural modeling. Specifically, our network architecture incorporates the iterative source steering algorithm to separate source signals using spatial statistics. The network is trained via multitask learning of source separation and classification with permutation invariant training. In addition, to improve the performance, we utilized an audio foundation model called BEATs and augmented the training data by curating AudioSet. The experimental results on the official development test set show that our best system (System 2) improved more than 2 dB in class-aware signal-to-distortion ratio improvement (CA-SDRi) from the official baseline system.

Index Terms— Spatial semantic segmentation of sound scenes (S5), neural blind source separation, BEATs

1. INTRODUCTION

Spatial semantic segmentation of sound scenes (S5) in DCASE 2025 Task 4 aims to detect and separate sound events from a mixture signal [1]. The input is provided in the first-order Ambisonics (FOA) format, consisting of four channels. Each mixture may contain up to three target sound events selected from eighteen predefined classes, along with background noise and up to two interference sounds. The baseline system adopts a cascading approach combining audio tagging and target source separation [2]. While the separation module is trained solely on the task dataset, the tagging module leverages an audio foundation model called Masked Modeling Duo (M2D) [3]. The performance is evaluated using class-aware signal-to-distortion ratio improvement (CA-SDRi) [1].

A key challenge in this task is making the separation model robust against unseen environments. While the train, validation, and test data in the development set are mainly generated from the same datasets, the eval set consists solely of new recordings by the organizers [1]. This domain shift significantly impacts performance. In fact, the baseline system achieved 11.09 dB in CA-SDRi on the in-domain test set but dropped to 6.60 dB on the out-of-domain eval set. One possible solution is blind source separation (BSS) based on spatial statistics [4–7]. However, classic BSS methods often deteriorate in the underdetermined scenarios of this challenge, where the number of sources may exceed the number of audio channels.

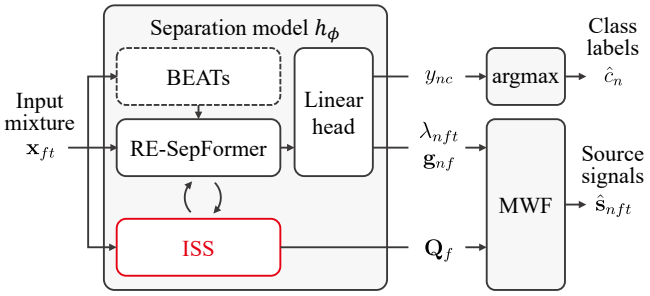


Figure 1: Overview of the proposed system based on neural BSS.

To address this, we employ neural BSS techniques that effectively combine neural networks and statistical BSS. A representative method is called independent vector analysis with a deep neural network (DNN-IVA) [8]. This method alternates between predicting time-frequency (TF) masks by a DNN and estimating the spatially demixing filters via IVA. This iterative process allows the DNN to access quasi-separated signals by IVA during inference, thereby facilitating the separation task. In addition, incorporating DNNs enables the system to be easily extended to tasks beyond separation (e.g., detection) by multi-task learning. The DNN-IVA has been extended with a jointly-diagonalizable (JD) spatial model [5, 9, 10] to handle diffuse noise and underdetermined conditions [11]. This method has shown promising results for separating conversational speech recordings [12].

In this report, we describe our hybrid S5 system submitted to DCASE 2025 Task 4 (Fig. 1). Specifically, we utilized the JD-extended version of DNN-IVA [11, 12] to address underdetermined conditions. Our model architecture is built upon a resource-efficient SepFormer (RE-SepFormer) [12–14]. The network is designed to classify separated signals by attaching a classification head trained through multitask learning. To further improve generalizability, we incorporated two additional techniques. First, we utilized an audio foundation model called BEATs [15]. We obtained embeddings of BEATs for the input mixture and fed them as input features for the RE-SepFormer. Second, we augmented the training data by curating the audio samples from AudioSet [16]. We collected more than 60K source signals of the target events by automatically filtering the audio samples. We report the impact of each technique on CA-SDRi for the development validation and test sets.

[*] Equally contributed

2. A HYBRID S5 SYSTEM

Our system extracts N source images $\hat{\mathbf{s}}_{nft} \in \mathbb{C}^M$ with their class labels $\hat{c}_n \in \{0, \ldots, C\}$ from an input M-channel mixture $\mathbf{x}_{ft} \in \mathbb{C}^M$ in the short-time Fourier transform (STFT) domain (Fig. 1). Here, $n = 1, \ldots, N$, $f = 1, \ldots, F$, and $t = 1, \ldots, T$ denote source, frequency, and time-frame indices, respectively. Since the number of targets in a mixture is unknown, we assign $\hat{c}_n = 1, \ldots, C$ to target events and $\hat{c}_n = 0$ to silence (or noise).

2.1. Signal model based on local Gaussian model

Assuming N source tracks, our method utilizes a local Gaussian model (LGM) [5, 7, 17]. In this model, the mixture $\mathbf{x}_{ft}$ is decomposed into N pairs of the power spectral densities $\lambda_{nft} \in \mathbb{R}_+$ and spatial covariance matrices (SCMs) $\mathbf{H}_{nf} \in \mathbb{S}_+^{M \times M}$ as follows:

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n=1}^{N} \lambda_{nft} \mathbf{H}_{nf} \right). \tag{1}$$

The SCMs $\mathbf{H}_{nf}$ are further decomposed with a diagonalizer $\mathbf{Q}_f \in \mathbb{C}^{M \times M}$ and diagonal elements $\mathbf{g}_{nf} \in \mathbb{R}_+^M$ as follows:

$$\mathbf{H}_{nf} \triangleq \mathbf{Q}_f^{-1} \mathrm{diag}\left(\mathbf{g}_{nf}\right) \mathbf{Q}_f^{-\mathsf{H}}. \tag{2}$$

Given the model parameters λ_{nft}, $\mathbf{Q}_f$, and $\mathbf{g}_{nf}$, the source image $\hat{\mathbf{s}}_{nft} \in \mathbb{C}^M$ is obtained by multichannel Wiener filtering (MWF) as:

$$\hat{\mathbf{s}}_{nft} \leftarrow \mathbf{Q}_f^{-1} \mathrm{diag} \left(\lambda_{nft} \mathbf{g}_{nf} \Big/ \sum_{n=1}^{N} \lambda_{nft} \mathbf{g}_{nf} \right) \mathbf{Q}_f \mathbf{x}_{ft}, \tag{3}$$

where $\cdot/\cdot$ denotes an element-wise division operator. Since this model assumes (JD) full-rank SCMs, it can handle underdetermined conditions and diffuse noise [5, 7, 10].

2.2. Network architecture based on RE-SepFormer and BEATs

The network is designed to predict the LGM parameters $\boldsymbol{\lambda}$, $\mathbf{G}$, and $\mathbf{Q}$ and the label log-probabilities $\mathbf{Y} \triangleq \{y_{nc} \in \mathbb{R}\}_{n,c=1}^{N,C}$ as follows:

$$\{\boldsymbol{\lambda}, \mathbf{G}, \mathbf{Q}, \mathbf{Y}\} \leftarrow h_\phi(\mathbf{X}), \tag{4}$$

where h_ϕ is a neural network with model parameters ϕ. This network is built by alternately stacking B RE-SepFormer blocks [13] and $B - 1$ iterative source steering (ISS) [4, 8] blocks. Specifically, the RE-SepFormer block is applied in a channel-wise manner to predict $\sum_n \lambda_{nft}^{-1} g_{nfm}^{-1}$ as M-channel TF masks from $\mathbf{Q}_f \mathbf{x}_{ft} \in \mathbb{C}^M$. The ISS block, on the other hand, optimizes the diagonalizer $\mathbf{Q}_f$ from the TF mask to maximize the likelihood of Eq. (1). After iterating this process $B - 1$ times, the final (B-th) RE-SepFormer block predicts network outputs λ_{nft}, g_{nfm}, and y_{nc} by an M-to-N attention mechanism followed by linear layers. The RE-SepFormer blocks are enhanced by transform-concatenate-average (TAC) modules [14] for inter-channel communication, which was originally proposed in [12]. In addition, we improve the separation and detection performance by utilizing an audio foundation model called BEATs [15]. The BEATs is applied to the reference (0-th) channel of the mixture $\mathbf{x}_{ft}$, and the obtained embeddings are concatenated with $\mathbf{Q}_f \mathbf{x}_{ft}$ to feed the RE-SepFormer blocks.

2.3. Permutation-invariant training

The network is trained via multitask learning of source separation and detection. Specifically, the cost function $\mathcal{L}_\phi$ consists of the separation loss $\mathcal{L}_\phi^{(\mathrm{sep})}$ and classification loss $\mathcal{L}_\phi^{(\mathrm{cls})}$ as follows:

$$\mathcal{L}_\phi = \mathcal{L}_\phi^{(\mathrm{sep})} + \alpha \mathcal{L}_\phi^{(\mathrm{cls})}. \tag{5}$$

where $\alpha \in \mathbb{R}_+$ is a scaling hyperparameter. We utilized the SDR loss[1] on the time-domain signals of $\hat{\mathbf{s}}_{nft}$ for $\mathcal{L}_\phi^{(\mathrm{sep})}$, where the time-domain signals are obtained via inverse STFT. The SDRs are computed and averaged only over source tracks with target labels ($c_n \neq 0$). The classification loss $\mathcal{L}_\phi^{(\mathrm{cls})}$ is, on the other hand, defined as the negative cross entropy for all the source tracks. The track indices n are aligned between estimated and reference tracks using permutation invariant training [18] to minimize $\mathcal{L}_\phi$.

2.4. Dataset augmentation utilizing AudioSet

To compensate for the lack of target source signals in the challenge datasets (including FSD50K [19] and EARS [20]), we augment the training data using AudioSet [16]. Since AudioSet consists of web videos, each sample often includes multiple sound events, making it difficult to leverage as a source signal for simulated mixtures. We first identified AudioSet clips that had only a single weak label corresponding to the target event classes in FSD50K. These clips were then further curated using BEATs (AudioSet Fine-tuned Model 1) [15] to ensure each contained only one event. Through this filtering process, we obtained 64,125 clips containing only a single sound event, which were used as additional target signals.

2.5. Inference

Once the network is trained, the source signals $\hat{\mathbf{s}}_{nft}$ are separated by Eq. (3), and the labels $\hat{c}_n$ are predicted as:

$$\hat{c}_n \leftarrow \operatorname*{argmax}_c y_{nc}. \tag{6}$$

Since each class appears at most once in the mixture, we summed $\hat{\mathbf{s}}_{nft}$ having the same $\hat{c}_n$. Additionally, as each mixture has at least one target, we applied post-processing to select the track with the highest non-silent probability when all tracks are predicted as silent.

3. EXPERIMENTAL RESULTS

This section reports the experimental results on the development validation and test sets of the challenge dataset.

3.1. Experimental conditions

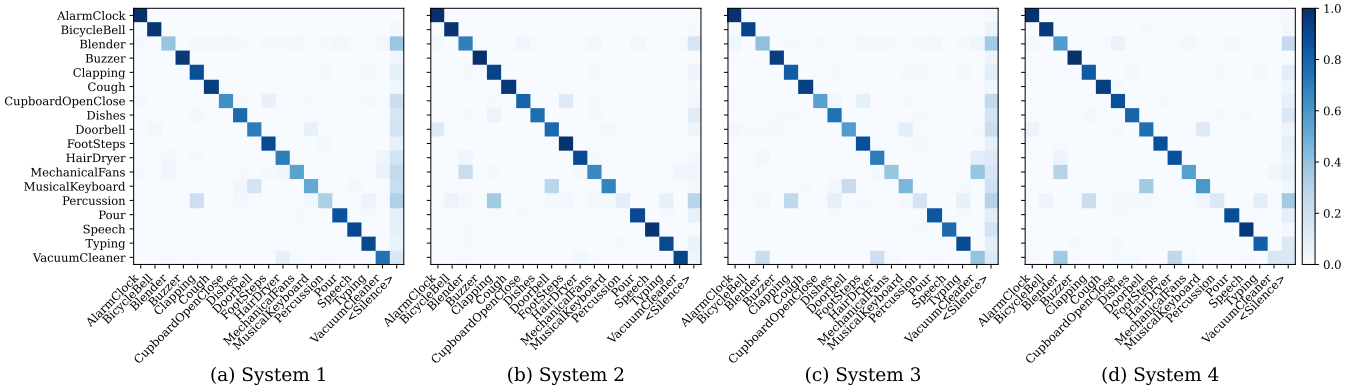
The separation model consisted of $B = 8$ RE-SepFormer blocks, each with 256-dimensional 8-head attentions and 1024-dimensional feedforward layers. Spectrograms were obtained from 32-kHz input signals using STFT with a 1024-sample window and a 320-sample hop length. The intra-chunk processing of RE-SepFormer was performed with 100-frame (1-second) chunks without overlap. The ISS blocks with two iterations were inserted between RE-SepFormer blocks. The number of source tracks was set to $N = 4$, assuming three target sources and one noise track. The noise track is assumed to include both background and interference sounds.

The separation model was trained using the AdamW optimizer [21] having a learning rate of 1.0×10^{-4} and weight decay of 1.0×10^{-5}. Full 10-second clips were fed to the model during training with a batch size of 64. One epoch was defined as 1250 updates,

[1] We refer to the signal-to-noise ratio (SNR) as the SDR, following [1].

Table 1: Separation and detection performance on the development set of DCASE 2025 Challenge Task 4. SDRi$^{(K)}$ denotes the average SDRi for mixtures having K target sources. $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ are the micro precision, recall and F1-scores of the predicted labels.

System	BEATs	AS	Validation set		Test set							
			CA-SDRi	Acc.	CA-SDRi	Acc.	SDRi$^{(1)}$	SDRi$^{(2)}$	SDRi$^{(3)}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Baseline (ResUNetK) [2]	N/A	N/A	11.28	59.26	11.09	59.80	–	–	–	0.84	0.80	0.82
System 1			11.74	45.56	12.38	57.13	14.53	16.43	16.89	0.87	0.76	0.81
w/o post-processing			11.50	44.07	12.26	56.00	14.53	16.43	16.89	**0.89**	0.75	0.81
w/o model averaging			11.23	45.56	11.85	55.47	14.09	16.23	16.47	0.85	0.75	0.80
System 2	✓		**14.07**	**64.82**	**13.31**	**64.07**	**15.15**	16.92	**17.43**	0.84	**0.83**	**0.84**
System 3		✓	10.88	42.59	11.23	48.80	14.51	16.43	16.54	0.80	0.69	0.74
System 4	✓	✓	13.30	56.67	12.46	55.93	15.02	**17.00**	17.38	0.81	0.76	0.79



(a) System 1 (b) System 2 (c) System 3 (d) System 4

Figure 2: Confusion matrices of detection results for the test set. `<Silence>` denotes the label predicted as silence ($c = 0$).

and training continued for up to 150 epochs. Following the baseline system, the dynamic mixing (synthesizing) [2] was applied. The scaling factor α was set to 1 or 2, selected based on validation CA-SDRi for each training. To address initialization sensitivity, we trained models with two seeds and selected the best. The final system was obtained by averaging 10 sets of the model weights with the highest CA-SDRi scores on the validation set. These hyperparameters were experimentally determined by using the validation set.

We submitted four systems (System 1–4), differing in the use of the BEATs and AudioSet (AS) augmentation. In addition to the CA-SDRi and label prediction accuracy, we evaluated two additional metrics. One was the separation performance by SDRi$^{(K)}$, which is the average SDRi for mixtures having K target sources, ignoring the predicted labels. The other is the label prediction performance by the micro-precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F1-scores ($\mathcal{F}$).

3.2. Experimental results

As summarized in Table 1, System 2, which used BEATs but didn't used the AudioSet augmentation, achieved the best performance among all the four systems. This system achieved 13.31 dB in CA-SDRi for the test set, which was more than 2 dB better than the official baseline system based on ResUNet. In addition, all the systems had better CA-SDRi than the baseline system, which indicates the effectiveness of the proposed architecture based on neural BSS. We can also see the consistent improvement of CA-SDRi by using BEATs (System 1→2 and 3→4). While the improvement of SDRi are limited, the micro recall ($\mathcal{R}$) was significantly improved. This improvement is also shown in the confusion matrices (Fig. 2), where the number of mistakenly predicted silence labels was significantly reduced. The post-processing and model averaging also slightly improved CA-SDRi. However, using AudioSet didn't improved the performance. While the SDRi was not affected by the augmentation, it significantly degraded the prediction performance for the `VacuumCleaner` as shown in Fig. 2-(b) and (d).

3.3. Limitations and future directions

We briefly outline here several limitations and future directions:

1. **Training epochs**: Due to the time constraints, System 3 and 4 were trained with only 108 and 105 epochs, respectively, while System 1 and 2 were trained with 150 epochs. Therefore, care should be taken when comparing their performance.
2. **Data augmentation**: We augmented only the target source signals with AudioSet, leaving the interference signals unchanged. Along with the target augmentation, we will investigate improved curation methods to better align with the original domain.
3. **Modeling assumption**: Our system is tailored to the challenge conditions and assumes stationary sound sources. While this matches the current challenge setup, handling moving sources remains an important future direction. We plan to extend the system using a time-varying spatial model [22–24].

4. CONCLUSION

We developed a hybrid S5 system based on neural BSS for the DCASE 2025 Challenge Task 4. The proposed system improved more than 2 dB in CA-SDRi from the baseline system. Our future work includes handling moving sources and diverse audio events.

5. ACKNOWLEDGMENT

We used ABCI 3.0 provided by AIST and AIST Solutions with support from "ABCI 3.0 Development Acceleration Use".

6. REFERENCES

[1] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, *et al.*, "Description and discussion on DCASE 2025 Challenge Task 4: Spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2506.10676*, 2025.

[2] 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," *arXiv preprint arXiv:2503.22088*, 2025.

[3] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Modeling Duo: Towards a universal audio pre-training framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 32, pp. 2391–2406, 2024.

[4] R. Scheibler and N. Ono, "Fast and stable blind source separation with rank-1 updates," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 236–240.

[5] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 28, pp. 2610–2625, 2020.

[6] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 5, pp. 971–982, 2013.

[7] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, no. 7, pp. 1830–1840, 2010.

[8] R. Scheibler and M. Togami, "Surrogate source model learning for determined source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 176–180.

[9] K. Sekiguchi, A. A. Nugraha, Y. Bando, and K. Yoshii, "Fast multichannel source separation based on jointly diagonalizable spatial covariance matrices," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2019, pp. 1–5.

[10] N. Ito and T. Nakatani, "FastMNMF: Joint diagonalization based accelerated algorithms for multichannel nonnegative matrix factorization," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 371–375.

[11] Y. Bando, Y. Masuyama, A. A. Nugraha, and K. Yoshii, "Neural fast full-rank spatial covariance analysis for blind source separation," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2023, pp. 51–55.

[12] Y. Bando, T. Nakamura, and S. Watanabe, "Neural blind source separation and diarization for distant speech recognition," in *Proc. of Interspeech*, 2024, pp. 722–726.

[13] L. Della Libera, C. Subakan, M. Ravanelli, S. Cornell, F. Lepoutre, and F. Grondin, "Resource-efficient separation transformer," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 761–765.

[14] Y. Luo, Z. Chen, N. Mesgarani, and T. Yoshioka, "End-to-end microphone permutation and number invariant multi-channel speech separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6394–6398.

[15] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: audio pre-training with acoustic tokenizers," in *Proc. of the International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[16] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[17] M. Togami, "Multi-channel itakura saito distance minimization with deep neural network," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 536–540.

[18] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[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 (TASLP)*, vol. 30, pp. 829–852, 2021.

[20] 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 *Proc. of Interspeech*, 2024, pp. 4873–4877.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of the International Conference on Learning Representations (ICLR)*, pp. 1–8.

[22] T. Fujimura and R. Scheibler, "Multi-channel separation of dynamic speech and sound events," in *Proc. of Interspeech*, 2023, pp. 3749–3753.

[23] H. Munakata, Y. Bando, R. Takeda, K. Komatani, and M. Onishi, "Joint separation and localization of moving sound sources based on neural full-rank spatial covariance analysis," *IEEE Signal Processing Letters*, vol. 30, pp. 384–388, 2023.

[24] Y. Nozaki, Y. Bando, and M. Onishi, "Source-aware spatial self-supervision for sound event localization and detection," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.