Baseline Systems and Evaluation Metrics for Spatial Semantic Segmentation of Sound Scenes

Binh Thien Nguyen* Masahiro Yasuda* Daiki Takeuchi Daisuke Niizumi Yasunori Ohishi Noboru Harada
NTT Corporation *NTT Corporation* *NTT Corporation* *NTT Corporation* *NTT Corporation* *NTT Corporation*
Japan Japan Japan Japan Japan Japan

***Abstract*—Immersive communication has made significant advancements, especially with the release of the codec for Immersive Voice and Audio Services. Aiming at its further realization, the DCASE 2025 Challenge has recently introduced a task for spatial semantic segmentation of sound scenes (S5), which focuses on detecting and separating sound events in spatial sound scenes. In this paper, we explore methods for addressing the S5 task. Specifically, we present baseline S5 systems that combine audio tagging (AT) and label-queried source separation (LSS) models. We investigate two LSS approaches based on the ResUNet architecture: a) extracting a single source for each detected event and b) querying multiple sources concurrently. Since each separated source in S5 is identified by its sound event class label, we propose new class-aware metrics to evaluate both the sound sources and labels simultaneously. Experimental results on first-order ambisonics spatial audio demonstrate the effectiveness of the proposed systems and confirm the efficacy of the metrics.**

***Index Terms*—universal source separation, audio tagging, spatial semantic segmentation, immersive communication, label-aware signal-to-distortion ratio.**

I. Introduction

Recently, immersive communication has emerged as a promising field and gained attention [1]–[3], especially with the release of a codec for Immersive Voice and Audio Services (IVAS) [1]. Fundamental technologies for realizing immersive communication include decomposing a complex spatial sound scene into sound sources, accompanied by metadata describing their properties, such as the direction of arrival. These technologies are essential for spatial audio formats such as object-based audio [4] and Metadata-Assisted Spatial Audio (MASA) [5]. In these formats, spatial audio can be coded with fewer channels, reducing complexity and transmission delay, making them especially suitable for immersive communication on low computational capacity devices like smartphones.

Aiming for immersive communication, the DCASE 2025 Challenge has recently introduced a task on sound event detection and separation from spatial sound scenes, referred to as "Task 4: Spatial semantic segmentation of sound scenes (S5)[1]." An S5 system takes as input a reverberant multi-channel spatial audio mixture, consisting of multiple sound sources and diffuse background noise, and outputs the isolated single-channel dry sound sources along with their respective sound event class labels. This task adds complementary functions to other studies that estimate other meta-information, such as [6], contributing to the realization of immersive communication.

Separating sound mixtures into sources, known as source separation (SS), and predicting audio class labels, referred to as audio tagging (AT) or sound event detection (SED), are active areas of research, with their combination also being explored. SS has been used as a preprocessing step to improve the results of SED [7], and this approach served as the baseline for Task 4 of the DCASE Challenge 2020 and 2021. In contrast, [8] and [9] show that the performance of an SS network can be enhanced by incorporating the embedding information extracted by a sound classifier model. Further to this, [10] employs SED and AT to isolate individual sound sources from a weakly labeled signal, facilitating the training of a label-queried SS (LSS) model on a large-scale dataset. As another approach, [11] utilizes a multi-task learning framework for both source separation and label prediction. However, the main goal of these works is to improve either or both the SS and AT, while their performances are evaluated separately, or only the performance of the main task is considered. The connection between the separated sources and the predicted labels has not been thoroughly evaluated, which contrasts with the S5 task, where a separated source is identified by its label.

In this paper, we explore approaches for the S5 task. Specifically, we introduce baseline systems combining AT and LSS models, built upon a fine-tuned Masked Modeling Duo (M2D) model [12] and a modified version of ResUNet [10]. For LSS, in addition to extracting a single source as in [10], we also explore a model that queries multiple sound sources simultaneously. Furthermore, we propose new class-aware metrics that evaluate separated sources and their corresponding labels simultaneously. The proposed task setting, metrics, and systems will be reflected in the DCASE 2025 Challenge Task 4. The code is publicly available on GitHub[2].

II. Related Work

A. M2D-based Audio Tagging

An M2D-based AT model is a fine-tuned M2D [12] model on AudioSet [13] with 527 classes. M2D is a self-supervised learning (SSL) foundation model that is pre-trained with M2D's masked prediction-based objective using only audio

This work was partially supported by JST Strategic International Collaborative Research Program (SICORP), Grant Number JPMJSC2306, Japan.

* These authors contributed equally to this work.

[1]https://dcase.community/challenge2025/#task4

[2]https://github.com/nttcslab/dcase2025_task4_baseline

samples from AudioSet. It has been highlighted that the audio representations extracted using M2D achieve state-of-the-art performance in various downstream tasks in diverse domains such as speech, music, and environmental sound. Furthermore, M2D-X [12], an extension of M2D, accommodates additional training objectives for learning representations specialized to applications. M2D-X provides a variant, M2D-AS, tailored to AudioSet tagging using a combination of SSL objective and additional supervised loss for label prediction. We utilize an M2D-based AT model with an M2D-AS backbone for label prediction in our system.

B. Label-Queried Source Separation Using ResUNet

Kong et al. [10] proposed a single-input, single-output model for LSS. The model takes as input a single-channel sound mixture and a one-hot vector representing the sound class label, and it outputs the separated target sound source corresponding to the label query. The input signal is first converted into magnitude and phase spectrograms using the short-time Fourier transform (STFT). Then, the magnitude spectrogram is processed by an encoder-decoder separator network, built on a ResUNet backbone. The separator network consists of 6 symmetric encoder and decoder blocks with skip-connections, where each block consists of 2D convolutional/deconvolutional layers, batch normalization, and leaky ReLU activation layers. The label query information is incorporated into the separator network using Feature-wise Linear Modulation (FiLM) [14]. The separator network outputs the magnitude mask and phase residual, which are applied to the magnitude and phase spectrograms of the input mixture to obtain the spectrograms of the target sound source. Finally, the target sound source is reconstructed using the inverse STFT. Experimental results [10] have demonstrated the effectiveness of LSS in a wide range of separation tasks, including music and general sound separation.

III. Proposed Spatial Semantic Segmentation Systems

A. Task Setting

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

$$\boldsymbol{y}^{(m)} = \sum_{k=1}^{K} \boldsymbol{h}_k^{(m)} * \boldsymbol{s}_k + \boldsymbol{n}^{(m)} = \sum_{k=1}^{K} \boldsymbol{x}_k^{(m)} + \boldsymbol{n}^{(m)}, \quad (1)$$

where $\boldsymbol{s}_k \in \mathbb{R}^T$ is the single-channel dry source signal corresponding to the label c_k, $\boldsymbol{h}_k^{(m)} \in \mathbb{R}^H$ is the m-th channel of the length-H room impulse response (RIR) at the spatial position of $\boldsymbol{s}_k$, and $\boldsymbol{n}^{(m)} \in \mathbb{R}^T$ is the m-th channel of the multi-channel noise signal. The wet source $\boldsymbol{x}_k^{(m)} \in \mathbb{R}^T$ can be split into two components: the direct path and early reflections, $\boldsymbol{x}_k^{(m,\text{d})} \in \mathbb{R}^T$, and late reverberation, $\boldsymbol{x}_k^{(m,\text{r})} \in \mathbb{R}^T$, as

$$\boldsymbol{x}_k^{(m)} = \boldsymbol{x}_k^{(m,\text{d})} + \boldsymbol{x}_k^{(m,\text{r})} = \boldsymbol{h}_k^{(m,\text{d})} * \boldsymbol{s}_k + \boldsymbol{h}_k^{(m,\text{r})} * \boldsymbol{s}_k, \quad (2)$$

where $\boldsymbol{h}_k^{(m,\text{d})}, \boldsymbol{h}_k^{(m,\text{r})} \in \mathbb{R}^H$ are the corresponding early and late parts of the RIR, respectively. We denote by N the number of sound event classes.

The goal of S5 is to extract all the single-channel sources $\{\hat{\boldsymbol{x}}_1^{(m_{\text{ref}},\text{d})}, \dots, \hat{\boldsymbol{x}}_{\hat{K}}^{(m_{\text{ref}},\text{d})}\}$ at a reference microphone m_{ref} and their corresponding class labels $\hat{C} = \{\hat{c}_1, \dots, \hat{c}_{\hat{K}}\}$ from the multi-channel mixture $\boldsymbol{Y}$.

From now on, we drop some indices for the clarity of the formulation. The notations of $\boldsymbol{x}_k^{(m_{\text{ref}},\text{d})}$, $\hat{\boldsymbol{x}}_k^{(m_{\text{ref}},\text{d})}$, and $\boldsymbol{y}^{(m_{\text{ref}})}$ become $\boldsymbol{x}_k$, $\hat{\boldsymbol{x}}_k$, and $\boldsymbol{y}$, respectively.

B. Proposed Evaluation Metrics

In this section, we introduce class-aware metrics that evaluate both the sound sources and their class labels simultaneously for the S5 task. The main idea is that the estimated and reference sources are aligned by their labels, with the waveform metric being calculated when the label is correctly predicted, i.e., $c_k \in C \cap \hat{C}$. In cases of incorrect label prediction, penalty values are accumulated. We define a class-aware signal-to-distortion ratio improvement (CA-SDRi) metric as

$$\begin{aligned}&\text{CA-SDRi}\left(\{\hat{\boldsymbol{x}}_1, \dots, \hat{\boldsymbol{x}}_{\hat{K}}\}, \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_K\}, \hat{C}, C, \boldsymbol{y}\right) \\ &\qquad = \frac{1}{|C \cup \hat{C}|} \sum_{c_k \in C \cup \hat{C}} P_{c_k},\end{aligned} \quad (3)$$

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

$$P_{c_k \in C \cup \hat{C}} = \begin{cases} \text{SDRi}(\hat{\boldsymbol{x}}_k, \boldsymbol{x}_k, \boldsymbol{y}), & \text{if } c_k \in C \cap \hat{C} \\ \mathcal{P}_{c_k}^{\text{FN}}, & \text{if } c_k \in C \text{ and } c_k \notin \hat{C} \\ \mathcal{P}_{c_k}^{\text{FP}}, & \text{if } c_k \notin C \text{ and } c_k \in \hat{C} \end{cases}, \quad (4)$$

where the SDRi is calculated as

$$\text{SDRi}(\hat{\boldsymbol{x}}_k, \boldsymbol{x}_k, \boldsymbol{y}) = \text{SDR}(\hat{\boldsymbol{x}}_k, \boldsymbol{x}_k) - \text{SDR}(\boldsymbol{y}, \boldsymbol{x}_k), \quad (5)$$

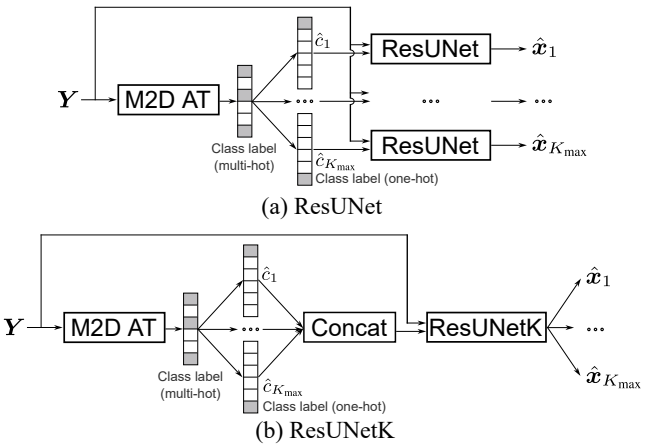
$$\text{SDR}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10 \log_{10} \left(\frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2} \right). \quad (6)$$

$\mathcal{P}_{c_k}^{\text{FN}}$ and $\mathcal{P}_{c_k}^{\text{FP}}$ are the penalty values for false negative (FN) and false positive (FP), both set to 0, indicating that incorrect predictions do not contribute to any improvement in the metric.

In a similar calculation method, the class-aware scale-invariant SDRi (CA-SI-SDRi), can be calculated by replacing the SDR in (6) with the SI-SDR [15] defined as

$$\text{SI-SDR}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = 10 \log_{10} \left(\frac{\|\alpha \boldsymbol{x}\|^2}{\|\alpha \boldsymbol{x} - \hat{\boldsymbol{x}}\|^2} \right), \quad (7)$$

where $\alpha = \hat{\boldsymbol{x}}^\top \boldsymbol{x} / \|\boldsymbol{x}\|^2$ is a scaling factor.



(a) ResUNet

(b) ResUNetK

Fig. 1. Proposed spatial semantic segmentation system flow.

1) Relation to Conventional Metrics: Our metrics share some similarities with conventional metrics [16]–[18] for source separation with an unknown number of sources, where the mixture contains K sources and the model outputs $\hat{K}$ sources. When $\hat{K} = K$, these metrics are typically computed using a permutation-invariant objective, where the output sources are permuted to best match the reference sources in order to maximize the metrics. When $\hat{K} \neq K$, one approach [16], [17] is to pad all-zero signals to the estimated sources when the source count is underestimated ($\hat{K} < K$) and to retain only a subset that best matches the reference sources in the case of overestimation ($\hat{K} > K$). Alternatively, [18] proposed a penalized metric composed of two components: the waveform metrics calculated on the best-matching subsets of size $\min(K, \hat{K})$ from the estimated and reference sources, and the penalties for source count deviation, $|K - \hat{K}| \cdot \mathcal{P}_{\text{ref}}$, where $\mathcal{P}_{\text{ref}}$ is a fixed penalty value.

The primary distinction is that our metric identifies sources using their labels, while these conventional metrics align sources using a permutation-invariant objective. In addition, unlike [18], which calculates the penalty based on source count deviation, our metric explicitly penalizes the FP and FN cases, avoiding situations where the source count is correct but the label prediction is incorrect.

C. Proposed Systems

To address the S5 task, we introduce two two-stage systems, (a) ResUNet and (b) ResUNetK, as shown in Fig. 1. The first stage predicts the source labels from the mixture, while the second stage extracts the corresponding sound sources. The details of these systems are as follows.

The first stage is identical for both systems, where we fine-tune an AT model consisting of a feature extractor backbone (i.e., M2D) and a head layer. The input of the model is the first channel of $\boldsymbol{Y}$. We replace the head of the original M2D model with a linear layer, which outputs the probabilities for N sound event classes in our dataset. To fine-tune the M2D-based AT, we initially trained only the head while keeping the rest of the model frozen. Next, we unfroze the last two layers and fine-tuned them along with the head for several epochs. Binary cross-entropy was used as the loss function. After fine-tuning, $\hat{C}$ is obtained by applying a threshold γ, with the estimated number of sources, $\hat{K}$, limited to the range $[1, K_{\text{max}}]$. If $\hat{K} < 1$, the highest probability is selected. If $\hat{K} > K_{\text{max}}$, the top K_{max} probabilities are chosen. It is worth noting that the output of the first stage in Fig. 1 consists of K_{max} label vectors, and if $\hat{K} < K_{\text{max}}$, the remaining $(K_{\text{max}} - \hat{K})$ vectors are all zeros. These zero labels, however, are not considered in the set $\hat{C}$ during the evaluation phase.

For the second stage of the first system in Fig. 1(a), we modify the single-input, single-output ResUNet from [10]. We adjust the input layer to accept the M-channel mixture spectrogram and modify the output layer to predict the M-channel magnitude mask and phase residual, which are applied to the input spectrogram. Following this, we add a 1×1 convolutional layer to generate the single-channel spectrogram for the target source. During inference, ResUNet is executed $\hat{K}$ times to extract $\hat{K}$ sources from the mixture.

Furthermore, we introduce a single-input, multiple-output variant of ResUNet, named ResUNetK, as illustrated in Fig. 1(b). ResUNetK has a similar structure to ResUNet, except that it outputs $K_{\text{max}} \times M$-channel magnitude mask and phase residual. These mask and residual are applied to the input mixture spectrogram, followed by a 1×1 convolutional layer that generates K_{max} single-channel spectrograms, corresponding to K_{max} output sound sources. Another difference is that the query of ResUNetK is a vector of length $K_{\text{max}} \times N$, which is a concatenation of K_{max} one-hot label vectors. The output sound sources are generated in the same order as the label vectors in the query. In addition, we add a connection so that if the label contains only zeros, the corresponding output waveform will also be all zeros.

Both the ResUNet and ResUNetK are trained with reference labels and sound sources using the SDR loss function. For ResUNetK, the loss is the mean SDR of all output sources. We randomly shuffled the labels to ensure that the model could extract the sources regardless of the label order.

IV. Experiments and Results

A. Experimental Setting

For the test data, we generated 1296 mixtures from the test split of sound sources and RIR datasets, with 432 mixtures containing 1, 2, and 3 sources, respectively. For validation, a similar number of mixtures was generated from the validation split, while for training, mixtures were synthesized on-the-fly.

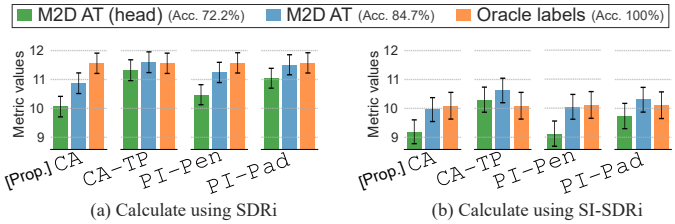
All sources were sampled at 32 kHz with the duration T of 10 seconds. Each mixture may contain 1 to $K_{\text{max}} = 3$ sources.

For the sound sources $\boldsymbol{s}_k$, we used the same train, validation, and test splits of the dataset in [19], which is derived from FSD50K [20] and ESC-50 [21]. We selected only 18 out of the 20 available sound classes, excluding "music" and "singing".

For the RIR and noise, we used the FOA-MEIR dataset [22], which contains first-order ambisonics RIR recorded in various environments. The Reverb-C and Test subsets of FOA-MEIR, consisting of 7 environments and 216 IR recordings

TABLE I
ACCURACY (%) OF M2D-BASED AUDIO TAGGING MODELS

Number of sources	1	2	3	Average
M2D AT (head)	91.2	71.1	54.4	72.2
M2D AT	93.1	81.7	79.4	84.7



(a) Calculate using SDRi (b) Calculate using SI-SDRi

Fig. 2. Comparison of various metrics for assessing S5 systems. The bar colors indicate first-stage methods. The second-stage methods are all ResUNet.

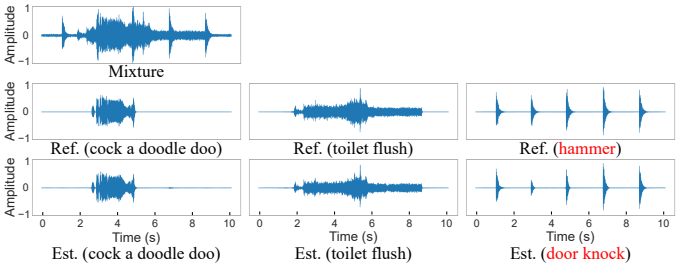


Fig. 3. Example of an S5 system when the labels are incorrectly predicted.

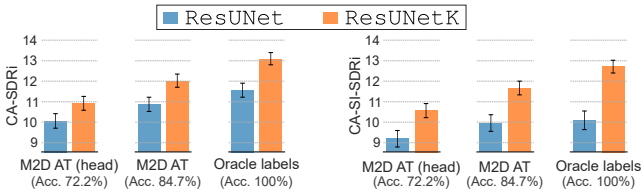


Fig. 4. Class-aware metrics on various S5 systems. The x-axis shows first-stage methods, and bar colors indicate second-stage methods.

each, were used for training. The Reverb-S subset, comprising 96 environments with 3 IR recordings each, was randomly split into 48 environments for testing and the rest for validation.

The multi-channel mixtures were synthesized using a modified version of the SpatialScaper toolkit [23]. For each mixture, we first selected a room from the RIR dataset and a 10-second segment from the corresponding noise signal. Then, we randomly generated a number of sources $K \in [1, K_{\max}]$, after which K positions in the room and K sound events were selected. In this setting, we assumed that the sound events in a mixture are mutually exclusive. The mixture was then synthesized using (1), with the SNR of each sound event ranging from 5 to 20 dB. The direct path IR, $\boldsymbol{h}_k^{(m,\mathrm{d})}$, was calculated by retaining the range of $[-6, 50]$ ms around the first peak of $\boldsymbol{h}_k^{(m)}$ and setting the rest to zero. The reference channel, m_{ref}, was set to 1, corresponding to the omnidirectional channel.

For the first stage of the S5 systems, we used the M2D-based AT model available online[3]. The probability threshold γ was set to 0.5. For the second stage, both ResUNet and ResUNetK were trained with a batch size of 5 on 8 RTX 3090 GPUs using the Adam optimizer with a learning rate of 10^{-4} for 250K steps. The training conditions were identical for both models, except that, for ResUNet, we randomly selected one source as the target for each mixture.

B. Audio Tagging Results

Table I presents the accuracy of the M2D AT models on the test dataset. "M2D AT (head)" refers to the checkpoint from the first step of fine-tuning, where only the head layer is trained. Since the separation model is LSS, the accuracy of label prediction in the first stage will directly affect the separated sources in the second stage.

C. Assessment of Evaluation Metrics

To assess the suitability of metrics for S5 systems, we compared our class-aware metrics (`CA`) with conventional metrics using permutation-invariant objectives described in Section III-B1, including the penalized metric (`PI-Pen`) [18] and the zero-padded metric (`PI-Pad`) [17]. We set $\mathcal{P}_{\mathrm{ref}}$ in `PI-Pen` to 0 to be consistent with our metric. For reference, we also included our metric without penalty (`CA-TP`), calculated using true positive predictions, with FN and FP ignored.

The results are shown in Fig. 2. We can see that, with the oracle labels in the first stage, all the metrics yield similar values. Since the separation model is LSS, given correct labels, if the separated sources are of sufficient quality, aligning the sources by permutations is nearly equivalent to aligning them by labels. The values vary across the metrics when the labels are predicted. It appears that `CA-TP` and `PI-Pad` may not be appropriate metrics, as, when calculated with SI-SDRi, the predicted labels using M2D AT yield even higher metric values than the oracle labels. `CA` and `PI-Pen` are more effective at reflecting the performance of the label prediction, as their metric values increase with higher label prediction accuracy in both SDRi and SI-SDRi cases. However, for the S5 task, `PI-Pen` may fail to evaluate in certain cases when the labels are incorrectly predicted, an example of which is shown in Fig. 3. In this figure, although the label prediction misclassifies the "hammer", the source separation model still provides a reasonable estimate of the sound source using the "door knock" label, which is most likely due to similarities between these two events. In this case, `CA` considers "hammer" as FN and "door knock" as FP, while `PI-Pen`, by only evaluating the estimated waveform without considering the label, imposes no penalty. These observations demonstrate the efficiency of `CA` in evaluating the S5 task, compared to conventional metrics.

D. Evaluation of Proposed Systems using Proposed Metrics

We evaluate the proposed S5 systems: (a) ResUNet (`ResUNet`) and (b) ResUNetK (`ResUNetK`), with the proposed evaluation metrics, the results of which are depicted in

[3]"M2D-AS fine-tuned on AS2M" on https://github.com/nttcslab/m2d

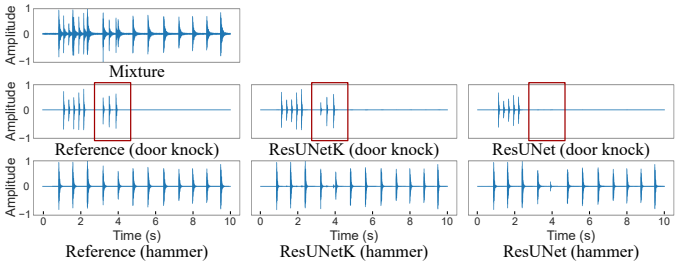


Fig. 5. Example of ResUNet and ResUNetK with given labels.

Fig. 4. It can be seen that `ResUNetK` outperforms `ResUNet` in all metrics when the same label predictions are used in the first stage. A possible reason is that, while being trained under nearly identical conditions, `ResUNetK` reconstructs all the sources simultaneously, which may facilitate mutual assistance among the reconstructions of these sources. We observed that in many cases, the sum of the output sources from `ResUNetK` is close to the mixture, even though we did not impose a loss function for mixture reconstruction. `ResUNet` may not benefit from this property since it extracts each source separately. Fig. 5 shows an example of the superior performance of `ResUNetK` compared to `ResUNet`, where the mixture contains two similar sound events, "door knock" and "hammer". As shown, `ResUNetK` successfully extracted the "door knock" sound at 4 seconds, while `ResUNet` did not. These findings illustrate the advantage of `ResUNetK` over `ResUNet` in the S5 task, with `ResUNetK` not only achieving higher performance but also being faster due to parallel source extractions.

V. Conclusions

In this paper, we introduce two-stage systems to address the S5 task, aiming at simultaneously detecting and separating sound events from multi-channel mixture signals. The first stage is an AT model that predicts source labels in the mixture, which is obtained by fine-tuning the pre-trained SSL M2D model. The second stage is an LSS model that extracts sound sources using the labels predicted in the first stage. Two variants for the LSS model have been explored: ResUNet, which extracts a single source, and ResUNetK, which simultaneously separates all sources from a mixture. To evaluate the performance of the proposed systems, we also proposed the evaluation metric dedicated to the S5 task. Experimental results highlight the effectiveness of the ResUNetK and confirm the efficacy of the proposed metrics in evaluating both source separation and class label prediction simultaneously.

The proposed task setting, metrics, and systems will be reflected in those of the DCASE 2025 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes. We expect to have enhanced solutions for the S5 task with the challenge.

References

[1] M. Multrus, S. Bruhn, J. Torres, E. Fotopoulou, T. Toftgård, E. Norvell, S. Döhla, Y. Gao, H.-y. Su, L. Laaksonen *et al.*, "Immersive voice and audio services (IVAS) codec-the new 3GPP standard for immersive communication," in *157th AES Convention*, 2024.

[2] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, "The internet of sounds: Convergent trends, insights, and future directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[3] E. Fotopoulou, K. Sagnowski, K. Prebeck, M. Chakraborty, S. Medicherla, and S. Döhla, "Use-cases of the new 3GPP immersive voice and audio services (IVAS) codec and a web demo implementation," in *IEEE 5th Int. Symp. on the IoS (IS2)*, 2024, pp. 1–6.

[4] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke, "Platform independent audio," *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, pp. 130–165, 2013.

[5] J. Paulus, L. Laaksonen, T. Pihlajakuja, M.-V. Laitinen, J. Vilkamo, and A. Vasilache, "Metadata-assisted spatial audio (MASA)–an overview," in *IEEE 5th Int. Symp. on the IoS (IS2)*, 2024, pp. 1–10.

[6] M. Yasuda, S. Saito, A. Nakayama, and N. Harada, "6DoF SELD: Sound event localization and detection using microphones and motion tracking sensors on self-motioning human," in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2024, pp. 1411–1415.

[7] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, "Improving sound event detection in domestic environments using sound separation," *arXiv preprint arXiv:2007.03932*, 2020.

[8] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2020, pp. 96–100.

[9] C. Hernandez-Olivan, M. Delcroix, T. Ochiai, D. Niizumi, N. Tawara, T. Nakatani, and S. Araki, "Soundbeam meets M2D: Target sound extraction with audio foundation model," *arXiv:2409.12528*, 2024.

[10] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," *arXiv preprint arXiv:2305.07447*, 2023.

[11] D. Lee and J.-W. Choi, "DeFT-mamba: Universal multichannel sound separation and polyphonic audio classification," *arXiv preprint arXiv:2409.12413*, 2024.

[12] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Towards a universal audio pre-training framework," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, 2024.

[13] 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 *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2017, pp. 776–780.

[14] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[15] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2019, pp. 626–630.

[16] S. R. Chetupalli and E. A. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 31, pp. 1681–1692, 2023.

[17] Y. Lee, S. Choi, B.-Y. Kim, Z.-Q. Wang, and S. Watanabe, "Boosting unknown-number speaker separation with transformer decoder-based attractor," in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2024, pp. 446–450.

[18] J. Zhu, R. A. Yeh, and M. Hasegawa-Johnson, "Multi-decoder DPRNN: Source separation for variable number of speakers," in *IEEE Intl. Conf. on Acoust., Speech & Sig. Proc. (ICASSP)*, 2021, pp. 3420–3424.

[19] 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.

[20] 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.

[21] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. of the 23rd ACM Intl. Conf. on Multi.*, 2015, pp. 1015–1018.

[22] 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.

[23] 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.