END-TO-END SPATIAL SEMANTIC SEPARATOR WITH DOA MODULE

Technical Report

Yoohan Park

KUBIG, Korea University
Seoul, South Korea
youan1024@korea.ac.kr

Haejin Cho

KUBIG, Korea University
Seoul, South Korea
indicigahby@korea.ac.kr

ABSTRACT

This technical report describes our submission to DCASE 2026 Challenge Task 4, Spatial Semantic Segmentation of Sound Scenes. The task requires a system to separate sound events from 4-channel First-Order Ambisonics (FOA) recordings while simultaneously predicting their sound classes and directions of arrival (DoA). To address this challenge, we propose an end-to-end framework that jointly performs source separation, sound classification, and DoA estimation. The model generates a fixed set of source representations, each associated with a separated waveform, a class label, and a DoA estimate. A frozen BEATs encoder is used to provide robust acoustic representations, while lightweight task-specific modules are trained for spatial modeling and prediction. The entire system is optimized using a joint permutation-invariant training objective that encourages consistent source assignment across all outputs.

1. INTRODUCTION

Spatial Semantic Segmentation of Sound Scenes (S5), introduced in DCASE 2026 Challenge Task 4[1], aims to decompose a multi-channel acoustic scene into individual sound events together with their semantic labels and spatial information. Unlike conventional source separation, which focuses on waveform extraction, or sound event localization and detection (SELD), which predicts event classes and directions without generating separated audio, this task requires all three outputs simultaneously.

The challenge becomes more difficult when multiple sources of the same class appear at the same time or when no target event is present in the scene. In such cases, the system must correctly associate separated signals with their corresponding class labels and directions while avoiding false detections in silent mixtures.

The official baseline addresses the task using a two-stage pipeline composed of audio tagging and label-queried source separation. While this approach provides a strong starting point, the classification and separation modules are optimized independently, which may limit the consistency between predicted labels and separated sources. Moreover, information learned during source separation cannot be fully utilized to improve classification performance.

To address these limitations, we propose an end-to-end framework that jointly performs source separation, sound classification, and direction-of-arrival estimation. The model generates a fixed number of source slots, where each slot contains a separated waveform, a DoA estimate, and a class prediction. A frozen BEATs encoder is employed to provide robust acoustic representations, while lightweight task-specific modules are trained for spatial modeling and prediction[2]. In contrast to mixture-level classification, class prediction is performed on separated source representations, allowing the model to exploit source-specific information. Finally, a joint permutation-invariant training objective is used to align separation, classification, and localization outputs under a shared source assignment.

2. DATASET

2.1. DCASE 2026 Task 4 Development Set

We use the official DCASE 2026 Task 4 development dataset (zenodo.org/records/19328046), which provides isolated sound event sources, First-Order Ambisonics Room Impulse Responses (FOA-RIRs), background noise recordings, and interference sounds. Mixtures are synthesized on-the-fly during training using SpAudSyn, the spatial audio synthesis tool provided by the organizers.

The development set consists of three splits: training (on-the-fly synthesis), validation (1,800 fixed mixtures), and test (1,512 fixed mixtures). Each mixture is 10 seconds long at 32 kHz. The number of target sound events per mixture ranges from 0 to 3, and interference events from 0 to 2. The 18 target sound event classes are: AlarmClock, BicycleBell, Blender, Buzzer, Clapping, Cough, CupboardOpenClose, Dishes, Doorbell, FootSteps, HairDryer, MechanicalFans, MusicalKeyboard, Percussion, Pour, Speech, Typing, and VacuumCleaner.

2.2. External Data: FSD50K

To augment the interference sound pool, we additionally incorporate the FSD50K development set (zenodo.org/record/4060432), which contains 40,966 audio clips totaling 80.4 hours of audio across 200 sound classes drawn from the AudioSet ontology[3]. As required by the task rules, we use only the dev split of FSD50K (the eval split is excluded). The FSD50K clips are processed into the scaper-compatible format and added to the interference pool using the provided `add_interference.py` script.

2.3. External Data: EARS

To strengthen the Speech class, we use the EARS (Expressive Anechoic Recordings of Speech) dataset (github.com/facebookresearch/ears_dataset), which provides approximately 100 hours of clean, anechoic speech recordings from 107 speakers in diverse recording conditions and speaking styles[4]. We use all 107 speakers (p001 through p107), applying a speaker-level train/validation split to prevent data leakage: 83 speakers are assigned to the training set and the remaining 9

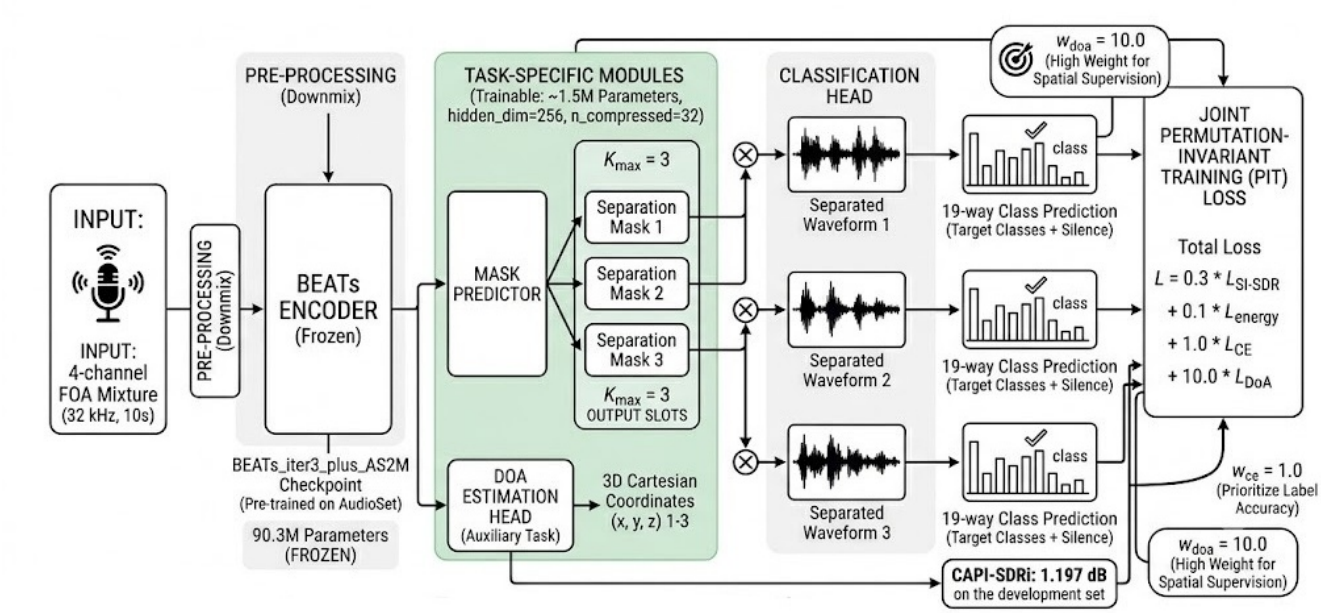

Figure 1: Overall pipeline.

speakers (p090, p091, p094, p095, p097, p101, p102, p105, p107) are exclusively reserved for the validation set. Speech clips are processed and added to the sound_event pool using the provided `add_sound_event.py` script.

3. SYSTEM DESCRIPTION

3.1. Overview

Our system, SpatialSeparatorModel, is a single-model architecture that jointly performs source separation, classification, and DoA estimation from the four-channel FOA mixture input. This contrasts with the two-stage baseline, which first runs an audio tagging model to predict class labels and then queries a separate ResUNet-based separation model conditioned on those labels.

The primary motivation for incorporating DoA estimation as an auxiliary task is to provide the model with explicit spatial supervision. In the DCASE 2026 Task 4 setting, same-class sources appear at directions separated by at least 60 degrees. By training the model to predict the spatial position of each source, we hypothesize that the internal representations may better encode directional information that is useful for disambiguating same-class sources during separation and classification.

3.2. BEATs Encoder

We use BEATs (BEATs_iter3_plus_AS2M checkpoint, github.com/microsoft/unilm/tree/master/beats) as a frozen audio feature encoder. BEATs is a self-supervised audio pre-training model trained on AudioSet, and provides rich semantic representations. The encoder weights are kept frozen throughout training to preserve the pre-trained representations and reduce the number of trainable parameters. The FOA mixture (4 channels) is downmixed to a single channel before being passed to the BEATs encoder, as BEATs expects single-channel input.

3.3. Separation Module

The separation module applies mask-based waveform separation. Given the BEATs features, a mask predictor generates $K_{max} = 3$ soft masks (one per output slot), which are applied to the mixture waveform to produce K_{max} separated single-channel waveforms. The maximum number of output slots corresponds to the maximum number of target sources per mixture in the task ($K_{max} = 3$).

3.4. Classification Head

Each separated waveform slot is independently classified into one of 18 target classes or a silence class, resulting in 19-way classification per slot. The silence class allows the model to indicate that a given output slot contains no target sound event, which is essential for handling zero-target mixtures and mixtures with fewer than K_{max} sources. The model architecture uses $hidden\_dim = 256$ and $n\_compressed = 32$ for the intermediate feature dimensions.

3.5. DoA Estimation Head

A separate DoA regression head predicts the 3D Cartesian coordinates (x, y, z) of each source from the separated feature representation. The DoA head is trained as an auxiliary task with ground-truth positions extracted from the SpAudSyn synthesis metadata. For silence slots, the DoA target is set to $[0, 0, 0]$. The DoA threshold parameter ($doa\_threshold = 0.1$) is used at inference to filter low-confidence DoA predictions.

4. TRAINING

4.1. Loss Function

We use FinalJointPITLoss, a joint permutation-invariant training loss that simultaneously optimizes separation quality, classification

accuracy, and spatial accuracy. The loss is defined as:

$$L = w_{si\_sdr} L_{SI\text{-}SDR} + w_{energy} L_{energy} + w_{ce} L_{CE} + w_{doa} L_{DoA} \quad (1)$$

with weights: $w_{si\_sdr} = 0.3$, $w_{energy} = 0.1$, $w_{ce} = 1.0$, $w_{doa} = 10.0$. The classification loss (Cross-Entropy) is assigned the highest weight among the primary task losses ($w_{ce} = 1.0$) to prioritize label accuracy, which directly impacts the CAPI-SDRi metric[5] through correct source alignment. The DoA loss weight ($w_{doa} = 10.0$) is set relatively high to ensure meaningful spatial supervision despite DoA being an auxiliary task, since DoA targets are unit-scale 3D coordinates that would otherwise produce small gradient magnitudes. The loss weight of si_sdr ($w_{si\_sdr} = 0.3$) is kept lower than classification to avoid overfitting to waveform-level reconstruction at the expense of semantic accuracy. Permutation-invariant training (PIT) is applied across all output slots to resolve the assignment ambiguity introduced by same-class sources.

4.2. Training Setup

The model is trained using PyTorch Lightning with AdamW optimizer ($lr = 1\mathrm{e}{-4}$) and bf16-mixed precision. Early stopping is applied with patience = 10 epochs, monitoring validation loss. Training is performed on a single NVIDIA L40S GPU on Lightning AI (AWS).

Table 1: Training hyperparameters.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	1e-4
Precision	bf16-mixed
Batch size	4
Num workers	12
Max epochs	100
Early stopping patience	10
Gradient clip	0.5
GPU	NVIDIA L40S (Lightning AI AWS)

5. RESULTS AND LIMITATIONS

The proposed system achieved a CAPI-SDRi score of 1.197 dB on the DCASE 2026 Task 4 development set. In addition, the model obtained a source-level classification accuracy of 55.02% and a mixture-level accuracy of 40.28%. All results were produced using a single model without ensembling.

The positive CAPI-SDRi score indicates that the model successfully performs source separation and provides improved source estimates compared to the original mixture. However, the relatively modest improvement suggests that there is still room for improvement in both separation quality and source detection performance.

Several limitations remain in the current system. Several factors may have contributed to this performance gap. First, only a small portion of the model parameters (approximately 1.5M out of 91.8M total parameters) were trainable, while the BEATs encoder remained frozen throughout training. Although this design reduces computational cost and preserves pretrained representations, it may limit the model's ability to adapt to the specific characteristics of the S5 task. Second, the FOA input was downmixed to a single channel before being processed by the BEATs encoder, which may result in the loss of spatial information that could otherwise be useful for source separation and localization.

Another limitation is that the available training data was not fully utilized during development because of computational constraints. As a result, the reported results should be viewed as a baseline implementation of the proposed architecture rather than the best achievable performance. Future work will focus on improving the utilization of multichannel spatial information, increasing trainable model capacity, and exploring more effective training strategies for the available dataset.

6. REFERENCES

[1] B. T. Nguyen, M. Yasuda, N. Harada, R. Serizel, M. Mishra, M. Delcroix, C. Hernandez-Olivan, S. Araki, D. Takeuchi, T. Nakatani, and N. Ono, "Description and discussion on DCASE 2026 challenge task 4: Spatial semantic segmentation of sound scenes," 2026. [Online]. Available: https://arxiv.org/abs/2604.00776

[2] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[3] 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, 2022.

[4] J. Richter, R. O. Aichinger, R. Bliefnick, A. Brueggemann, J. Brunner, M. Hafezi, B. Haider, and T. Habigt, "Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation," in *Proceedings of Interspeech*, 2023.

[5] B. T. Nguyen, M. Yasuda, D. Takeuchi, D. Niizumi, and N. Harada, "Class-aware permutation-invariant signal-to-distortion ratio for semantic segmentation of sound scene with same-class sources," in *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.