SELF-GUIDED TARGET SOUND EXTRACTION AND CLASSIFICATION THROUGH UNIVERSAL SOUND SEPARATION MODEL AND MULTIPLE CLUES

Technical Report

Younghoo Kwon[1*], *Dongheon Lee*[1*], *Dohwan Kim*[1], *Jung-Woo Choi*[1†]

[1] KAIST, School of Electrical Engineering, Daejeon, South Korea,
{k0hoo, donghen0115, rlaehghks5, jwoo}@kaist.ac.kr

ABSTRACT

This paper presents a multi-stage framework that integrates Universal Sound Separation (USS) and Target Sound Extraction (TSE) for sound separation and classification. In the first stage, USS is applied to decompose the input audio into multiple source components. Each separated source is then individually classified to generate two types of clues: enrollment and class clues. These clues are subsequently utilized in the second stage to guide the TSE process. By generating multiple clues in a self-guided manner, the proposed method enhances the performance of target sound extraction. The final output of the TSE module is then re-classified to improve the classification accuracy further.

***Index Terms*—** Self-guided clue, multi-stage training, universal sound separation

1. INTRODUCTION

Sound separation and classification in environments with overlapping sound sources have been actively studied in recent years. Universal Sound Separation (USS) aims to separate all types of sound sources in an auditory scene, regardless of their class or number. This task offers the advantage of extracting all source objects without requiring prior information about their identity. In contrast, Target Sound Extraction (TSE) focuses on isolating sound sources aligned with specific clues, such as class labels, direction-of-arrival (DoA), or enrollment samples. By leveraging such target-specific clues, TSE generally achieves superior performance compared to USS.

In DCASE 2025 Task 4, the objective is to extract foreground sources, ranging from one to three per mixture, by removing reverberation and isolating them from interfering sources (up to two) and a background noise component (always present). Each extracted foreground source must be classified into one of 18 predefined sound event classes. This task is particularly challenging because it requires the selective separation of only the foreground sources that belong to the target classes, while simultaneously suppressing interfering sources. In the baseline framework, audio tagging (AT) is first performed on the mixed audio to identify active classes, and the resulting class labels are used as clues for TSE. However, the AT model operates on a single-channel input and thus does not leverage spatial information when classifying source signals. This leads to limitations in the quality of guidance. Furthermore, since only class clues are used during training, the TSE model becomes highly dependent on the accuracy of the AT output, making it difficult to recover when the predicted class clues are incorrect.

The proposed framework leverages the state-of-the-art USS model, DeFT-Mamba, which has recently demonstrated strong performance on source separation and classification tasks in polyphonic scenarios. In this work, waveforms separated by DeFT-Mamba are utilized as clues to guide the TSE process. In the first stage, USS is applied to multi-channel input, which enables the model to exploit spatial information for improved separation and classification. The separated audio signals are then used as enrollment clues, while their corresponding classification outputs serve as class clues for the second-stage TSE. By incorporating both enrollment and class clues estimated in a self-guided manner, the proposed method mitigates the dependency on a single type of clue. In cases where one clue is inaccurate, the complementary clue may still provide effective guidance. Finally, the outputs from the TSE process are re-classified to obtain the final predicted sound event classes.

2. PROPOSED METHOD

The overall framework is illustrated in Figure 1. Given a mixed audio spectrogram as input, the first-stage model, DeFT-Mamba-USS, separates the audio into foreground object features, interference object features, and a noise object feature. Each object feature is passed through a separation decoder and a class decoder, allowing the model to jointly separate and classify sources without pairing ambiguity. The DeFT-Mamba architecture is detailed in [1], and to reduce model complexity, a modified version is employed in this work by removing the unfold operation and excluding the Mamba module from the F-Hybrid Mamba blocks. The object separator of DeFT-Mamba is designed to separate three foreground object features, as well as two interference source features and one background noise feature. The separated foreground features are subsequently decoded to object waveforms by audio decoders and then classified through the Masked Modeling Duo for Single-labeled Classification (M2D-SC). Since each separated foreground source is assumed to correspond to a single class, M2D-SC is designed as a single-label classifier based on the M2D [2]. The last two layers of the pretrained M2D model are fine-tuned to identify the individual labels of separated waveforms, just as the M2D-AT was fine-tuned in the baseline [3]. However, the separated waveforms can include silence signals corresponding to non-existent objects. To classify these silence signals, energy-based learning [4] is incorporated in the training, and the energy function is utilized as a metric to discriminate silence signals from active foreground

*Equal contribution.
†Corresponding author.

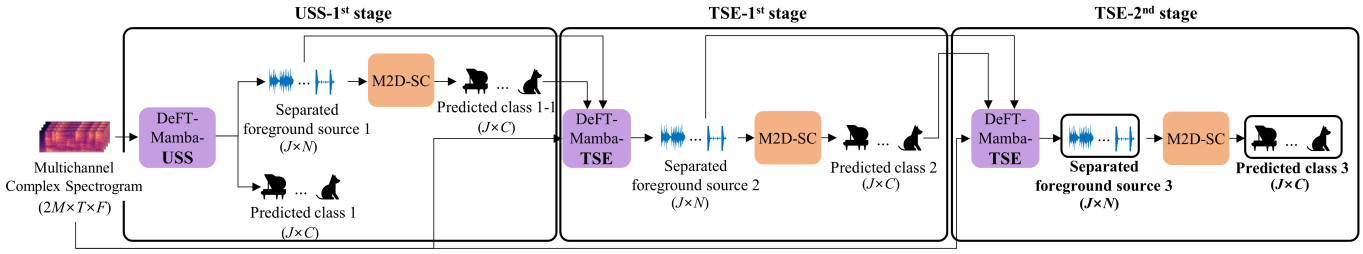

Figure 1: Overall framework of the proposed method

sources. The inference process of M2D-SC is illustrated in Figure 2. During inference, M2D-SC performs both classification over 18 classes and silence detection. The model first outputs unnormalized logits for the 18 classes, from which the most likely class is determined. Then, an energy score is computed from the same logits and compared to a threshold to decide whether the input corresponds to silence. We apply class-specific thresholds for silence detection, meaning that the energy threshold varies depending on the class predicted by M2D-SC.

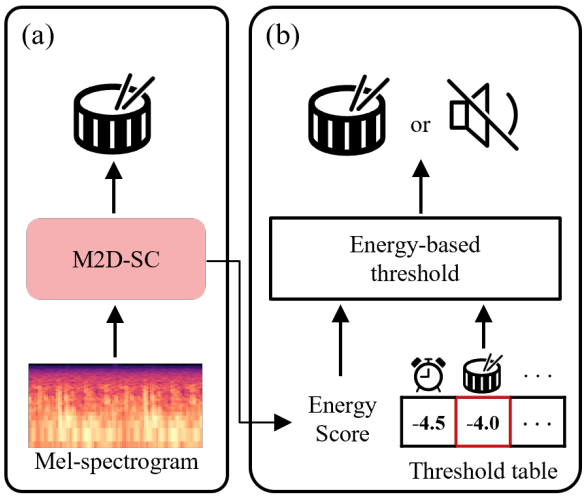

Figure 2: Inference procedure of M2D-SC (a) The model predicts class and calculates the energy score from unnormalized logits. (b) Silence is determined by comparing the energy score with a class-specific threshold.

The separated foreground source waveforms and their corresponding M2D-SC classification outputs are then used as enrollment and class clues, respectively, to guide the second-stage TSE. The TSE module, referred to as DeFT-Mamba-TSE, also adopts the modified DeFT-Mamba architecture. As described in [5], the enrollment clues are concatenated with the mixed audio features before being processed by the up-convolutional layers of DeFT-Mamba-TSE. Meanwhile, the class clues are injected into each DeFT-Mamba block via Res-FiLM conditioning, as reported in [6]. Finally, the signals extracted from the DeFT-Mamba-TSE are passed once again through M2D-SC for classification. The TSE process is repeated once more using the same approach. The final output of the proposed framework consists of the separated foreground sources and their predicted class labels, as shown in Fig.1 (e.g., separated foreground source 3, predicted class 3).

3. EXPERIMENTAL SETTINGS

We used the training set provided in DCASE 2025 Task 4 to train our models. To enhance generalization across diverse audio conditions, we replaced the speech class subset from FSD50K with samples from the VCTK corpus and excluded the percussion class samples originating from FSD50K. A total of 50,000 training mixtures were utilized. Audio signals were sampled at 32 kHz and converted into spectrograms using a 40 ms Hann window and a 20 ms hop size. The DeFT-Mamba architecture consists of six blocks, each with a channel dimension of 64. The input channel dimension for the up-convolution layers was set to 8 in the first stage and 14 in the second stage. All models were trained using a learning rate of 0.0004 for 100 epochs. The DeFT-Mamba-USS, M2D-SC, and DeFT-Mamba-TSE models were trained independently. For energy-based training of M2D-SC, we include the silence data in the training dataset by making 5% of the training data to exclude foreground signals. The interference and noise signals are included in the training data with a probability of 70% each. For training DeFT-Mamba-TSE, we used the separated foreground sources from DeFT-Mamba-USS as enrollment clues, and the ground-truth class labels as class clues. For inference, enrollment clues were generated from the DeFT-Mamba-USS output (e.g., separated foreground source 1), and class clues were provided by the M2D-SC predictions (e.g., predicted class 1-1).

The loss functions and weighting strategies used for training each module are as follows. For DeFT-Mamba-USS, the audio decoder for the foreground signal was trained using the negative Source-Aggregated Signal-to-Distortion Ratio (SA-SDR) loss [7]. The same loss function was used to estimate the interference signal. Given M estimated signals ($\hat{s}_m$) and target signals (s_m), the negative SA-SDR loss is calculated as shown in equation 1. In accordance with the challenge data configuration, M is fixed at 3 for foreground signals and 2 for interference signals. The noise source was trained with the negative Scale-Invariant Signal-to-Noise Ratio (SI-SNR) loss. Since the DeFT-Mamba-USS estimates one noise signal ($\hat{n}$), the negative SI-SNR loss can be calculated as equation 2 with the target noise (n). The scaling factor $\alpha =< \hat{n}, n > /\|n\|^2$ normalizes the scale of the target noise. Both non-foreground losses were weighted with a factor of 0.01.

$$\mathcal{L}_{\text{SA-SDR}} = -10\log_{10}\frac{\sum_{m=1}^{M}\|s_m\|_2^2}{\sum_{m=1}^{M}\|s_m - \hat{s}_m\|_2^2} \quad (1)$$

$$\mathcal{L}_{\text{SI-SNR}} = -10\log_{10}\frac{\|\alpha\cdot n\|^2}{\|\hat{n} - \alpha\cdot n\|^2},\ \alpha = \frac{<\hat{n}, n>}{\|n\|^2} \quad (2)$$

$$\mathcal{L}_{\text{USS}} = \mathcal{L}_{\text{foreground}} + 0.01\cdot(\mathcal{L}_{\text{interference}} + \mathcal{L}_{\text{noise}}) \quad (3)$$

The class decoder within DeFT-Mamba-USS was trained using cross-entropy (CE) loss for foreground sources. For silence seg-

ments, outlier exposure was applied to encourage uniform class distributions, using Kullback–Leibler (KL) divergence loss. The silence decision was trained as a binary classification task using binary cross-entropy (BCE) loss, and the sigmoid output was thresholded at 0.5 to determine the presence of foreground signals (sigmoid ≥ 0.5) or silence (≤ 0.5). For M2D-SC, a two-stage training strategy was employed. In the first stage, the model was trained to classify foreground sources using the ArcFace loss [8], which introduces an additive angular margin to enhance inter-class separability. The ArcFace loss is defined as:

$$\mathcal{L}_{\text{ArcFace}} = -\log \frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cdot\cos(\theta_j)}}, \quad (4)$$

where θ_j is the angle between the input feature and the weight vector of class j, $s = 32$ is the scale factor, and $m = 0.5$ is the angular margin. For silence segments, the same KL divergence loss used in the DeFT-Mamba-USS class decoder was applied to ensure uniform probability distributions across classes. In the second stage, energy-based learning was applied to detect silence. The energy score is defined as:

$$E(x) = -\log \sum_{k=1}^{C} e^{l_k} \quad (5)$$

where l_k is the unnormalized logit for class k, and C is the number of classes. The loss encourages inlier samples (non-silence foreground samples) to have low energy, and outlier samples (silence samples) to have high energy. To this end, the hinge-based loss with individual margins is applied:

$$\mathcal{L}_{\text{energy}} = \mathbb{E}_{x_{\text{in}}\sim\mathcal{D}_{\text{in}}^{\text{train}}}(\max(0, E(x_{\text{in}}) - m_{\text{in}}))^2 + \mathbb{E}_{x_{\text{out}}\sim\mathcal{D}_{\text{out}}^{\text{train}}}(\max(0, m_{\text{out}} - E(x_{\text{out}})))^2 \quad (6)$$

where $m_{\text{in}} = -6.0$ and $m_{\text{out}} = -1.0$ are the energy thresholds for inlier and outlier samples, respectively. This hinge-based loss function penalizes non-silence samples whose energy exceeds m_{in}, and silence samples whose energy falls below m_{out}. This hinge loss for energy-based learning was weighted with a factor of 0.001.

$$\mathcal{L}_{\text{SC}}^{1st} = \mathcal{L}_{\text{ArcFace}} + \mathcal{L}_{\text{KL}} \quad (7)$$

$$\mathcal{L}_{\text{SC}}^{2nd} = \mathcal{L}_{\text{ArcFace}} + \mathcal{L}_{\text{KL}} + 0.001 \cdot \mathcal{L}_{\text{energy}} \quad (8)$$

The target signal extractor, DeFT-Mamba-TSE, was trained using the masked SNR loss adopted from the baseline framework [3]. In addition, for non-foreground signals, the SI-SNR loss was applied with a weighting factor of 0.01 to penalize non-foreground signals underestimated from their corresponding object features.

4. RESULTS

The experimental results are summarized in Table.1. We evaluated four configurations based on different combinations of Foreground Source Separation (FSS) and Class Prediction (CP) available at various stages of the proposed framework. The configurations include (1) **FSS 1 + CP 1** using the separated waveforms and estimated classes from DeFT-Mamba-USS, (2) **FSS 1 + CP 1-1** using the waveforms from DeFT-Mamba-USS but processing them by M2D-SC to estimate classes, (3) **FSS 2 + CP 1-1** performing the second stage processing using DeFT-Mamba-TSE but using the classification results from the first stage M2D-SC, (4) **FSS 2 + CP 2** using the waveforms separated by DeFT-Mamba-TSE and classes predicted by feeding them into the second-stage M2D-SC, (5) **FSS 3 + CP 3** two-stage TSE model. Among all configurations, the FSS 3 + CP 3 model achieved the best performance, demonstrating the effectiveness of the proposed two-stage multi-clue framework. These results demonstrate the effectiveness of using USS-derived outputs as multi-clues to perform self-guided target sound extraction.

Table 1: Experimental result of the proposed framework

	SNRi	Accuracy	CA-SNRi
FSS 1 + CP 1	15.1 dB	73.2 %	10.8 dB
FSS 1 + CP 1-1	15.1 dB	81.8 %	12.7 dB
FSS 2 + CP 1-1	18.3 dB	81.8 %	14.6 dB
FSS 2 + CP 2	18.3 dB	83.4 %	14.7 dB
FSS 3 + CP 3	18.4 dB	84.5 %	14.9 dB

5. REFERENCES

[1] D. Lee and J.-W. Choi, "DeFT-Mamba: Universal multichannel sound separation and polyphonic audio classification," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[2] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *ICASSP 2023-2023 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] 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," 2025. [Online]. Available: https://arxiv.org/abs/2503.22088

[4] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 21 464–21 475, 2020.

[5] D. Wu, X. Wu, and T. Qu, "Leveraging sound source trajectories for universal sound separation," *IEEE Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 33, pp. 2337–2348, 2025.

[6] Q. Kong, K. Chen, H. Liu, X. Du, T. Berg-Kirkpatrick, S. Dubnov, and M. D. Plumbley, "Universal source separation with weakly labelled data," 2023. [Online]. Available: https://arxiv.org/pdf/2305.07447

[7] T. von Neumann, K. Kinoshita, C. Boeddeker, M. Delcroix, and R. Haeb-Umbach, "Sa-sdr: A novel loss function for separation of meeting style data," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6022–6026.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.