DOI: 10.61782/fa.2025.0937

FORUM ACUSTICUM
EURONOISE 2025

UNIVERSAL AUDITORY SCENE ANALYSIS MODEL FOR SOURCE SEPARATION, EVENT LOCALIZATION, AND DETECTION

Dongheon Lee[1] **Jung-Woo Choi**[1*]
[1] School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Auditory scene analysis (ASA) seeks to address complex spatial audio tasks, including sound event detection (SED), direction of arrival estimation (DoAE), universal source separation (USS), and noise suppression (NS). This study introduces Universal DeFT-Mamba, a unified framework designed to tackle these challenges using a deep neural network trained on diverse multichannel audio mixtures. The proposed architecture integrates a transformer-based time-frequency attention network with the Mamba-feedforward network (Mamba-FFN), enabling it to simultaneously separate multichannel audio mixtures into unmixed signals and estimate acoustic parameters for SELD and DoAE. To this end, the Universal DeFT-Mamba adopts group-wise processing for individual sound objects, and features separated for each sound object are processed by the separation, SELD, and DoA decoders to accomplish the multitask objectives. In this way, permutation issues in aligning separated waveforms, event onsets/offsets, and DoAs can be naturally suppressed. Experimental results demonstrate that Universal DeFT-Mamba achieves superior multichannel separation and SELD performance, surpassing the traditional task-specific SELD network.

Keywords: *Universal sound separation, Auditory scene analysis, Sound event localization and detection*

**Corresponding author*: *jwoo@kaist.ac.kr.*

1. INTRODUCTION

Auditory scene analysis (ASA) refers to the process of organizing sound into perceptually meaningful elements, such as the classification of individual audio sources, the temporal localization of acoustic events, and the estimation of their direction-of-arrival (DoA). Recently, a variety of deep learning models have been proposed for audio classification, sound event detection, and DoA estimation [1–4]. However, analyzing acoustic scenes in which multiple sound sources overlap in the presence of background noise remains a challenging problem.

Sound event localization and detection (SELD) involves simultaneously performing sound event detection (SED) and direction-of-arrival estimation (DoAE) for individual sound sources in scenarios where multiple sources overlap. This task becomes particularly challenging under conditions with in-class polyphony, in which multiple sound sources of the same class overlap in time [5]. Existing approaches typically estimate sound events and their corresponding DoAs directly from the mixed audio input using neural networks, making it difficult to effectively handle polyphonic audio as distinct signals [6]. However, if individual sources can be successfully separated, the resulting simplified ASA can significantly improve overall performance.

On the other hand, universal sound separation (USS), which aims to extract individual source signals from an audio mixture, has recently attracted significant attention. Compared to conventional separation tasks, USS is more challenging due to variability in both the number and classes of sound sources. Nevertheless, successful source separation achieved through USS can considerably enhance SELD performance. The top-ranked model [7] of DCASE 2023 Task 3 achieved a high performance by utilizing the mixed audio features and concatenating the outputs of the separation model for each class as additional

features. Also, in [8], high performance was obtained by jointly addressing universal sound separation and polyphonic audio classification.

In this study, we propose a unified framework, Universal DeFT-Mamba, that can perform multichannel USS and SELD tasks concurrently. The proposed framework consists of the object separation network separating object-level features of individual sound sources and subdecoders performing USS, SED, and DoAE tasks using the separated object-level features. The object feature separation enables the consistent estimation of multiple targets without permutation ambiguity. We train and evaluate the proposed method using the auditory scene analysis (ASA) dataset generated through room simulation, which contains multiple moving sources and diffuse noise. Experimental results demonstrate that the proposed framework successfully accomplishes USS and SELD task objectives from a single model.

2. PROPOSED METHOD

The proposed research focuses on separating individual sound sources in noisy, reverberant environments and subsequently performing ASA using the separated sources. The audio mixture consists of S moving foreground source signals X and a single diffused background noise V. A microphone array with M channels is placed in the room to capture the audio mixture Y, and a deep learning model is then employed to isolate the individual sound sources. The multichannel USS task can be mathematically formulated as extracting individual source signals $X_{m,s}$ from the mixture given by

$$Y_m(t,f) = \sum_{s=1}^{S} X_{m,s}(t,f) + V_m(t,f), \qquad (1)$$

where $Y_m(t,f), X_{m,s}(t,f)$, and $V_m(t,f)$ are the multichannel spectrogram of the sound mixture, the reverberant sound of the s-th source, and the noise captured by the m-th microphone, respectively. Here, $t = 1, \cdots, T$ indicates the time frame index, and $f = 1, \cdots, F$ represents the frequency bin.

The proposed framework is illustrated in Fig. 1. In this framework, the multichannel audio mixture is encoded into audio features, and the DeFT-Mamba [8] is employed as an object separation network to separate foreground signals and suppress background noise (Fig. 2).

The object separation network is largely a combination of Hybrid Mamba blocks designed for the analysis along the frequency and time dimensions. Each Hybrid Mamba block first analyzes features extracted from two 1D convolution layers through the gating mechanism in the Gated Convolution Block (GCB). The output from GCB is processed by the transformer with Flash Attention-2 [9], followed by the Mamba-FFN. This combination of transformer with position-aware feedforward network has demonstrated its outstanding performance in the USS task [8]. The last Conv2d layer of the object separation network separates features corresponding to individual objects.

Subsequently, the individual object features are passed through an audio decoder, a class decoder, and a DoA decoder to yield separated audio, predicted class labels, and estimated DoA, respectively. The DoA is estimated in form of a vector in cartesian coordinates, where a vector length is used as an object presence indicator. That is, the object with the DoA vector length greater than 0.5 is considered as an active source to enable the source event detection simultaneously. Compared to the previous work [8], the primary extensions in this study include generating a multichannel output with spatial information from the audio decoder and employing a DoA decoder to estimate DoA and onsets/offsets of individual sound sources in every time frame. To achieve these two challenging goals, the model is trained to focus more on spatial information essential in both audio separation and DoA estimation tasks.

To train the model using data captured in diverse acoustic environments, we employ the ASA dataset [8] recently introduced in our previous work. The ASA dataset includes spatialized multichannel audio signals simulated for moving sources, whose source signals consist of 13 foreground source classes drawn from a publicly accessible sound source database. The multichannel signals are mixed with background noises adopted from the TAU-SNoise dataset [1], such that the signal-to-noise ratio (SNR) ranges from 6 dB to 30 dB. The width and length of rooms vary between 5 m and 8 m, with a height ranging between 3–4 m. The reverberation time of the rooms is between 0.2 s and 0.6 s. All of these parameters were sampled from uniform distributions. The multichannel signals were simulated for a 4-channel tetrahedral microphone array with a radius of 4.2 cm. The dataset is publicly available for

[1] https://zenodo.org/records/6408611

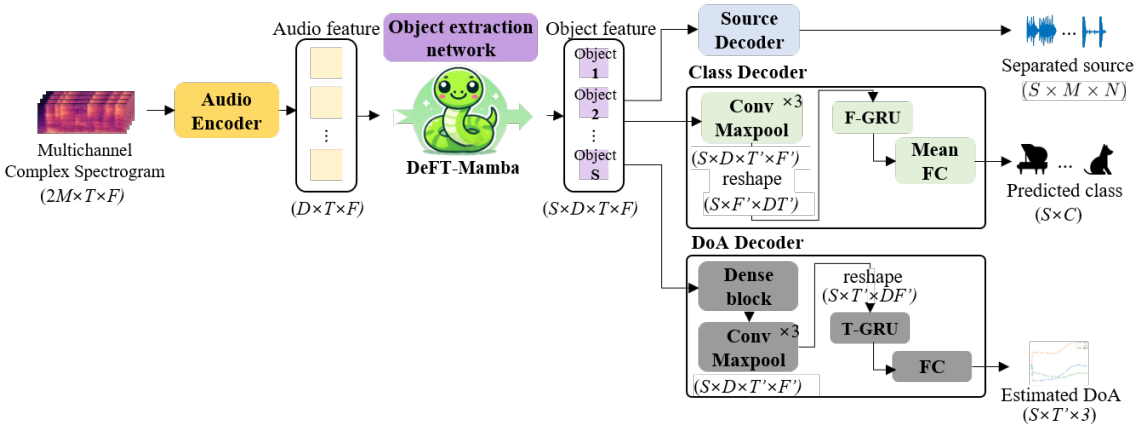

Figure 1: Overall framework of universal auditory scene analysis model for multichannel source separation, event localization, and detection

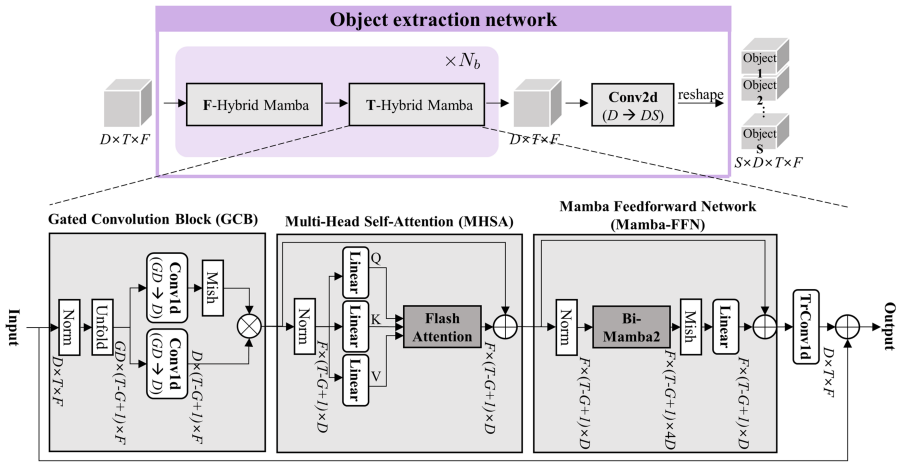

Figure 2: Architecture of Object Separation Network (DeFT-Mamba) consisting of a gated convolution block (GCB), multi-head self-attention (MHSA), and Mamba feedforward network (Mamba-FFN)

Table 1: Experimental result of universal auditory scene analysis (ASA dataset)

Model	SI-SDRi (dB) ↑	SDRi (dB) ↑	ER (%) ↓	F1 (%) ↑	LE (°) ↓	LR (%) ↑	SELD ↓
SELD-Net	-	-	51.5	48.1	31.2	54.2	0.416
Universal DeFT-Mamba	**11.0**	**12.1**	**33.8**	**67.9**	**21.8**	**68.2**	**0.275**

download at Zenodo [2] .

3. EXPERIMENT

The model was trained using a multi-task loss for audio, classification, and DoA estimation. For source separation, the source-aggregated signal-to-distortion ratio (SA-SDR) loss [10] was employed, enabling the separation of an arbitrary number of sources. Cross entropy was used as the classification loss, and mean squared error (MSE) was utilized as the DoA loss. These losses were summed as a joint loss function using the weights of 1, 0.1, and 0.01 for the MSEs of a reference microphone channel (mic 1),

[2] https://zenodo.org/records/13749621

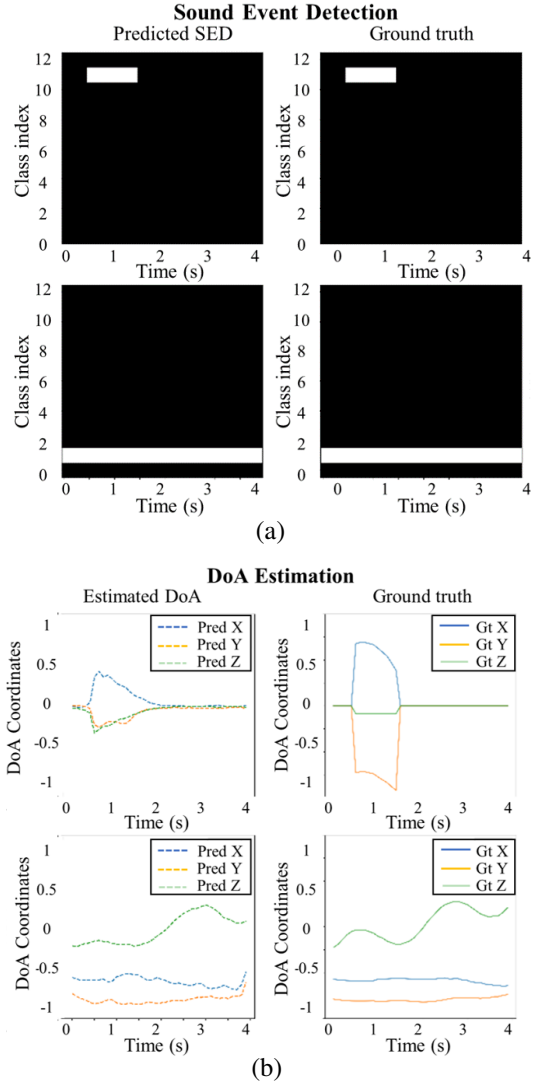

Figure 3: An example of universal auditory scene analysis using the proposed architecture. (a) Sound event detection and classification (b) DoA estimation

the MSEs of non-reference channels, and for the cross entropy loss, respectively. Apart from these adjustments, all other parameter settings were kept identical to those in the previous study [8].

Tab. 1 compares the results of the baseline model and the proposed architecture on the ASA dataset. Here, we only consider the ASA dataset because other datasets, e.g., STARSS23 [11], do not provide the ground truth signals for the multichannel separation task. Most of the conventional models have reported the performance on

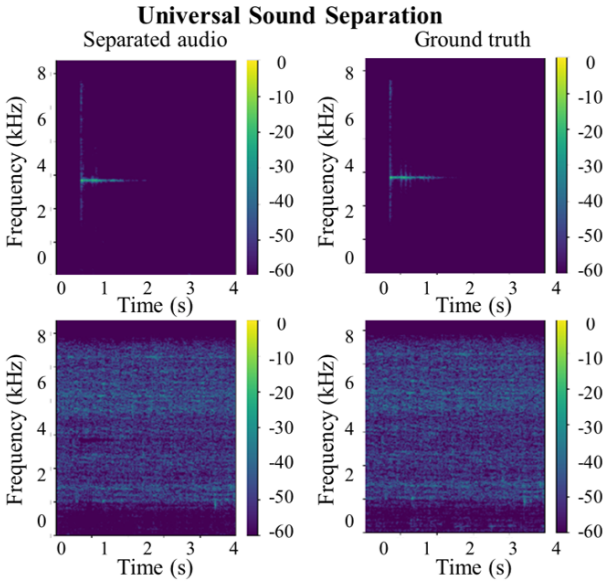

Figure 4: Spectrograms of separated source signals. (top row) Source 1 (bottom row) Source 2.

STARSS23 datasets, so we only utilize the open-source model, SELD-Net [12], as the comparison baseline. The comparison was made in terms of SELD performance metrics, such as error rate (ER), F1 score (F1), localization error (LE), localization recall (LR), and the SELD score adopted from the evaluation criteria of DCASE 2023 Task 3 [11]. The experimental results show that the proposed method achieves significantly higher performance than the baseline model. Both models employ the same decoders for classification and DoA estimation, but incorporating the multichannel object separation network with the source separation objective makes a substantial difference in performance. The baseline model cannot separate objects in polyphonic audio scenarios, so it often misclassifies classes and events, resulting in high error rates. In contrast, the proposed method with Universal DeFT-Mamba trained jointly with the multichannel separation task shows robust SELD performance. Fig. 3a and Fig. 3b present an example of universal auditory scene analysis using the proposed architecture. In this example including two overlapping sound objects, accurate event detection is accomplished without interference between objects. The DoA estimation, however, shows room for improvement, displaying the DoA vector difference to the ground truth DoA in the source-overlapping regions. The spectrograms of separated source signals (Fig. 4) are also close to the

ground truth spectrograms, except the slightly underestimated spectral energy of the second source in the T-F bins overlapping with the spectrogram of the first source.

4. CONCLUSION

This study proposes the Universal DeFT-Mamba framework, which separates individual sound objects in multichannel audio and concurrently performs ASA tasks such as sound event localization and detection. Experimental results demonstrate superior separation and event detection performance compared with a conventional SELDNet, particularly in polyphonic audio scenarios where the source separation contributes greatly to overall SELD performance.

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT of Korea government (MSIT) (No. RS-2024-00337945, RS-2024-00464269), the BK21 FOUR program through the NRF grant funded by the Ministry of Education of Korea government (MOE).

6. REFERENCES

[1] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1251–1264, 2023.

[2] Y. Shul and J.-W. Choi, "Cst-former: Transformer with channel-spectro-temporal attention for sound event localization and detection," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8686–8690, IEEE, 2024.

[3] D. Mu, Z. Zhang, and H. Yue, "Mff-einv2: Multi-scale feature fusion across spectral-spatial-temporal domains for sound event localization and detection," in *Proc. Interspeech 2024*, pp. 92–96, 2024.

[4] O. L. D. Santos, K. Rosero, B. Masiero, and R. de Alencar Lotufo, "w2v-seld: A sound event localization and detection framework for self-supervised spatial audio pre-training," *IEEE Access*, vol. 12, pp. 181553–181569, 2024.

[5] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[6] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[7] Q. Wang, Y. Jiang, S. Cheng, M. Hu, Z. Nian, P. Hu, Z. Liu, Y. Dong, M. Cai, J. Du, and C.-H. Lee, "The nerc-slip system for sound event localization and detection of dcase2023 challenge," tech. rep., DCASE2023 Challenge, June 2023.

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

[9] T. Dao, "Flashattention-2: Faster attention with better parallelism and work partitioning," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[10] 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. Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, pp. 6022–6026, Singapore, 2022.

[11] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, T. Virtanen, and Y. Mitsufuji, "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. Advances in Neural Information Processing Systems*, vol. 36, pp. 72931–72957, Curran Associates, Inc., 2023.

[12] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 34–48, March 2018.