END-TO-END ITERATIVE S5 SYSTEM BASED ON TF-LOCOFORMER AND ATST-FRAME

Technical Report

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

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

ABSTRACT

This technical report describes our end-to-end (E2E) system based on TF-Locoformer and ATST-Frame. The best system in the previous challenge on spatial semantic segmentation of sound scenes (S5) leveraged an iterative architecture that alternately performs separation and classification to achieve excellent performance. Inspired by this architecture, we built an E2E system that iteratively applies TF-Locoformer and ATST-Frame. Specifically, the overall architecture is based on TF-Locoformer, which stacks Transformer-based blocks to process each time-frequency bin. We inserted signal and classification heads into the outputs of intermediate blocks, and applied ATST-Frame to the separated source signals. The source-wise embeddings extracted by ATST-Frame are then fed back to the Locoformer block to progressively improve the performance. The whole architecture is trained in an E2E manner with permutation invariant training. Our best model on the development set achieved a class-aware permutation invariant signal-to-distortion ratio improvement (CAPI-SDRi) of 16.3 dB and source-wise label accuracy of 73.6 % on the test subset of the development set.

Index Terms— Spatial semantic segmentation of sound scenes, neural source separation, TF-Locoformer, ATST-Frame

1. INTRODUCTION

Spatial semantic segmentation of sound scenes (S5) in DCASE 2026 Task 4 aims to detect and separate sound events from a mixture signal [1, 2]. In contrast to the previous challenge [3, 4], this year's task allows multiple sources with the same class label to be present in a mixture. This update emphasizes the importance of not only target source enhancement but also source separation. The best-performing system in the previous challenge demonstrated the effectiveness of an iterative architecture that alternately performs separation and classification [5].

Inspired by this iterative architecture, we develop an end-to-end (E2E) system that iteratively performs separation and classification. As summarized in Fig. 1, our architecture is based on the time-frequency-domain Transformer with local modeling by convolution (TF-Locoformer) [6]. We attach classification and signal reconstruction heads to intermediate block outputs. The separated signals are then fed into the frame-level Audio Teacher-Student Transformer (ATST-Frame) [7,8] to obtain frame-level embeddings. These embeddings are fed back into the next TF-Locoformer block to refine the results. The network is trained in an E2E manner using a model-parallel implementation of the TF-Locoformer.

[*]Equally contributed

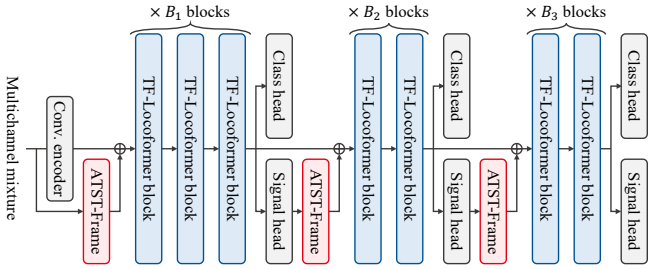


Figure 1: Overview of our network architecture.

2. MILLE-FEUILLE NET

This section first introduces the proposed architecture called mille-feuille net and then describe how the proposed model is trained.

2.1. Network architecture

The proposed model f takes an M-channel input signal with T samples $\mathbf{X} \in \mathbb{R}^{M \times T}$ and outputs N pairs of separated signals $\mathbf{s}_n \in \mathbb{R}^T$ and tagging probabilities $\mathbf{c}_n \in [0, 1]^C$:

$$\{\mathbf{s}_n, \mathbf{c}_n\}_{n=1}^N \leftarrow f_\theta(\mathbf{X}), \tag{1}$$

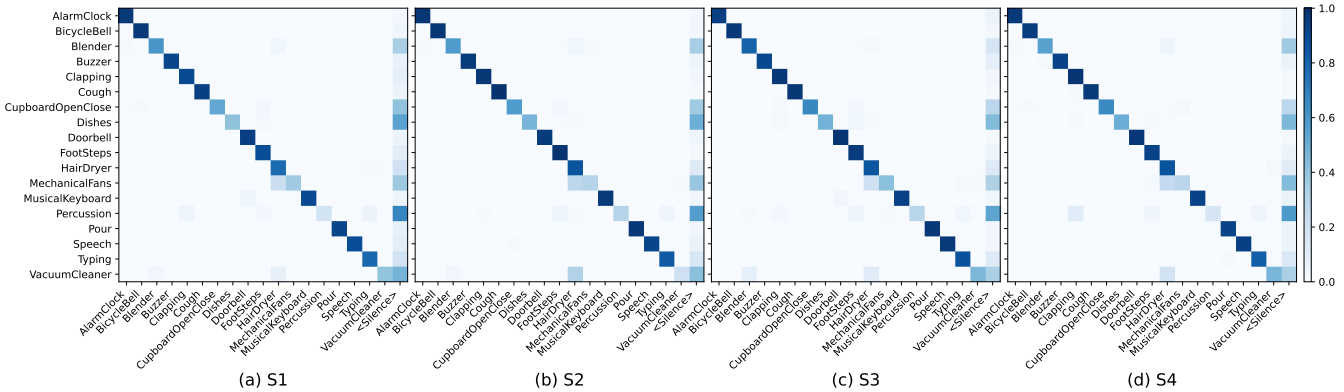
where θ denotes the model parameters, and $C = 18$ is the number of class labels. As in Fig. 1, the mixture is first processed by ATST-Frame in a channel-wise manner and by the convolutional encoder of TF-Locoformer. The resulting representations are fed into the first B_1 TF-Locoformer blocks. The processed embeddings are then fed into the classification and signal reconstruction heads to obtain tagging probabilities and separated signals, respectively. Each separated signal is further fed into ATST-Frame to obtain frame-wise embeddings. We selected ATST-Frame for its low memory footprint, which was essential for E2E training [9–12]. The obtained embeddings are fed back into the subsequent B_2 TF-Locoformer blocks to improve both separation and classification. This process is repeated once more, and the resulting embeddings are fed into the final B_3 TF-Locoformer blocks to further refine the results.

2.2. Permutation invariant training

The model is trained using permutation-invariant training (PIT) [13, 14]. Specifically, we calculate the loss for each output as a weighted sum of the signal-to-noise ratio (SNR) loss and the binary cross-entropy (BCE) loss. We used the BCE instead of the CE loss, inspired by image detection [15–17]. The permutation of the source index n is determined for each output by minimizing the SNR loss.

Table 1: Separation and detection performance on the development set of DCASE 2026 Challenge Task 4. $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}$ are the precision, recall and F1-scores of the predicted labels. `S1`–`S4` corresponds to System 1–4 submited to the challenge, respectively.

ID	System	Epoch	Validation set			Test set									
			CAPI-SDRi	Acc. (mix)	Acc. (src)	CAPI-SDRi	Acc. (mix)	Acc. (src)	TP-SDRi Avg.	TP-SDRi $K=1$	TP-SDRi $K=2$	TP-SDRi $K=3$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
`B1`	M2DAT_4c + ResUNetK [2]	–	–	–	–	8.5	60.7	70.4	–	–	–	–	–	–	–
`P1`	TF-Locoformer (Medium)	200	14.4	52.7	61.3	13.9	54.0	62.4	21.3	18.7	21.8	21.8	89.3	67.5	76.8
`P2`	TF-Locoformer (Large)	200	16.1	59.4	67.4	15.3	57.7	65.9	22.1	19.5	22.4	22.8	87.4	72.9	79.5
`P3`	`P1` + ATST-Frame	200	16.3	62.4	69.9	15.3	61.0	69.3	21.2	18.5	21.4	22.1	88.7	76.0	81.9
`P4`	`P2` + ATST-Frame	200	16.5	60.5	67.9	15.9	60.3	68.8	22.1	19.5	22.8	22.6	88.8	75.4	81.5
`S1`	TF-Locoformer (Medium)	320	16.2	59.2	66.4	15.8	57.7	68.2	22.2	19.5	22.9	22.7	89.8	74.0	81.1
`S2`	TF-Locoformer (Large)	251	16.9	62.7	69.1	16.5	62.3	70.3	22.4	19.6	22.9	23.1	89.4	76.7	82.6
`S3`	`S1` + ATST-Frame	257	16.9	64.3	71.4	16.3	64.9	73.6	21.5	18.7	21.8	22.3	90.0	80.1	84.8
`S4`	`S2` + ATST-Frame	216	16.7	60.2	68.7	16.4	61.5	70.5	22.4	19.9	23.1	22.7	89.5	76.9	82.7



(a) S1 (b) S2 (c) S3 (d) S4

Figure 2: Confusion matrices of detection results for the test set. `<Silence>` denotes the label predicted as silence.

3. EXPERIMENTAL EVALUATION

Our models were trained using only the dataset provided for the baseline system, without additional training data [1].

3.1. Experimental configuration

The model was trained using AdamW [18] with a learning rate of 1.0×10^{-3}, a weight decay of 0.01, and linear warmup [19]. Spectrograms were obtained using the short-time Fourier transform with a window size of 1024 samples and a hop length of 512 samples. We followed the configurations of TF-Locoformer described in [6]. Specifically, we trained a medium model with B_1, B_2, and B_3 set to 4, 1, and 1, respectively, and a large model with B_1, B_2, and B_3 set to 3, 3, and 3, respectively. The batch size was set to 4. N was set to 5. To minimize the domain mismatch, we fed the entire 10-second mixture to the model. These hyperparameters were determined using the validation set. To maximize training speed, we implemented the model in a model-parallel manner [20, 21]. Specifically, each sample is processed using four graphics processing units by parallelizing the time- or frequency-wise processing in the Locoformer blocks and the source-wise processing in ATST-Frame.

3.2. Experimental results

The performance is summarized in Table 1. First, increasing the model size (`P1`→`P2`) improves the class-aware permutation-invariant signal-to-distortion ratio improvement (CAPI-SDRi). Introducing ATST-Frame also improves classification accuracy (`P1`→`P3` and `P2`→`P4`). While ATST-Frame does not clearly improve the precision of class labels, it improves their recall.

While these results were obtained after 200 epochs of training, we continued training the model until shortly before the challenge deadline. `S1`–`S4` show the results of models trained as long as possble before the challenge deadline. We observed that performance improved significantly simply by continuing model training (e.g., `P1`→`S1`). The resulting system (`S2`) improved CAPI-SDRi by approximately 8.0 dB over the challenge baseline.

Fig. 2 shows the confusion matrices for `S1`–`S4`, where rows indicate reference labels and columns indicate estimated labels including silence. While our systems exhibit strong diagonal patterns, some classes are still confused with silence. This suggests that class-dependent thresholds at inference could be promising post-processing, although we did not investigate in this challenge.

4. CONCLUSION

We developed an E2E system for S5 based on TF-Locoformer and ATST-Frame. The proposed system outperformed the baseline system by 8.0 dB in CAPI-SDRi. We also found that the number of training iterations is the most important factor for improving separation performance. Future work includes improving classification performance through better integration strategies.

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, 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," *arXiv preprint arXiv:2604.00776*, 2026.

[2] 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 *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026, pp. 1–5.

[3] M. Yasuda, N. Binh Thien, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, T. Nakatani, T. Kawamura, and N. Ono, "Description and discussion on dcase 2025 challenge task 4: Spatial semantic segmentation of sound scenes," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2025, pp. 170–174.

[4] 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," in *Proc. of European Signal Processing Conference (EUSIPCO)*, 2025, pp. 266–270.

[5] Y. Kwon, D. Lee, D. Kim, and J.-W. Choi, "Self-guided target sound extraction and classification through universal sound separation model and multiple clues," in *Proc. of Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, 2025, pp. 235–239.

[6] K. Saijo, G. Wichern, F. G. Germain, Z. Pan, and J. Le Roux, "TF-Locoformer: Transformer with local modeling by convolution for speech separation and enhancement," in *Proc. of International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2024, pp. 205–209.

[7] X. LI and X. Li, "ATST: Audio representation learning with teacher-student transformer," 2022, pp. 4172–4176.

[8] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 32, pp. 1336–1351, 2024.

[9] 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 International Conference on Machine Learning (ICML)*, 2023, pp. 5178–5193.

[10] D. Bolya, P.-Y. Huang, P. Sun, J. H. Cho, A. Madotto, C. Wei, T. Ma, J. Zhi, J. Rajasegaran, H. Bangalath, *et al.*, "Perception encoder: The best visual embeddings are not at the output of the network," *Proc. of Neural Information Processing Systems (NeurIPS)*, vol. 38, pp. 60 884–60 937, 2026.

[11] S. Bharadwaj, S. Cornell, K. Choi, S. Fukayama, H.-j. Shim, S. Deshmukh, and S. Watanabe, "OpenBEATs: A fully open-source general-purpose audio encoder," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2025, pp. 1–5.

[12] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," 2024, pp. 547–551.

[13] 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 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 241–245.

[14] S. Maiti, Y. Ueda, S. Watanabe, C. Zhang, M. Yu, S.-X. Zhang, and Y. Xu, "EEND-SS: Joint end-to-end neural speaker diarization and speech separation for flexible number of speakers," in *Proc. of IEEE Workshop on Spoken Language Technology (SLT)*, 2023, pp. 480–487.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. of European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[16] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. of International Conference on Learning Representations (ICLR)*, 2021, pp. 1–16.

[17] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *Proc. of International Conference on Learning Representations (ICLR)*, 2023, pp. 1–19.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of International Conference on Learning Representations (ICLR)*, 2019, pp. 1–19.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. of Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[20] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[21] N. Shazeer, Y. Cheng, N. Parmar, D. Tran, A. Vaswani, P. Koanantakool, P. Hawkins, H. Lee, M. Hong, C. Young, *et al.*, "Mesh-tensorflow: Deep learning for supercomputers," *Proc. of Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.