TS-TFGRIDNET: EXTENDING TFGRIDNET FOR LABEL-QUERIED TARGET SOUND EXTRACTION VIA EMBEDDING CONCATENTAITON

Technical Report

Fulin Wu

Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, Guangdong, China
12110411@mail.sustech.edu.cn

Zhong-Qiu Wang

Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, Guangdong, China
wang.zhongqiu41@gmail.com

ABSTRACT

The DCASE2025 Challenge Task 4 - Spatial Semantic Segmentation of Sound Scenes (S5) challenges participants to separate a set of mixed sound events (sampled from 18 targeted sound events) to individual sound-event signals. The baseline system provided by the challenge organizors first performs audio tagging to identify the sound events existed in the mixture, and then conducts label-queried target sound extraction (TSE) to extract the signal of each identified sound event. Building on the baseline system, we propose to improve the label-queried TSE component by using a novel model named *Target Sound Extraction TF-GridNet* (TS-TFGridNet), leveraging the strong capability of TF-GridNet at speech separation for TSE. TS-TFGridNet concatenates audio and sound-class embeddings along the frequency or feature dimension, thereby conditioning TF-GridNet to perform TSE. Clear improvement is observed over the baseline system.

***Index Terms*—** Target sound extraction, TF-GridNet.

1. INTRODUCTION

In recent years, simultaneously separating the distinct sound sources from their mixtures and tagging their corresponding sound classes have gained significant attention. Systems like the universal sound separation (USS) model [1] has demonstrated strong performance and potential to solve the two tasks together. Building upon such advances, the DCASE2025 Challenge introduces the S5 task [2], focusing on separating reverberant multi-channel spatial audio mixtures into monaural sound sources, and at the same time predicting their associated class labels. The S5 baseline system [3] combines two key components: the M2D model [4] for audio tagging, and a ResUNet model, adapted based on [5], for label-queried TSE.

Considering that TF-GridNet [6, 7] has shown strong performance and potential in speech separation and that ResUNet-style models are known to have much weaker separation capability than dual-path models such as TF-GridNet [7], we propose to replace the ResUNet module with the more advanced TF-GridNet model and adapt TF-GridNet for label-queried TSE. We realize the adaption by concatenating the audio embedding in TF-GridNet with trainable sound-class embeddings along the frequency or feature dimension. The resulting model, named *Target Sound Extraction TF-GridNet* (TS-TFGridNet), exhibits clearly better TSE performance over the baseline model.

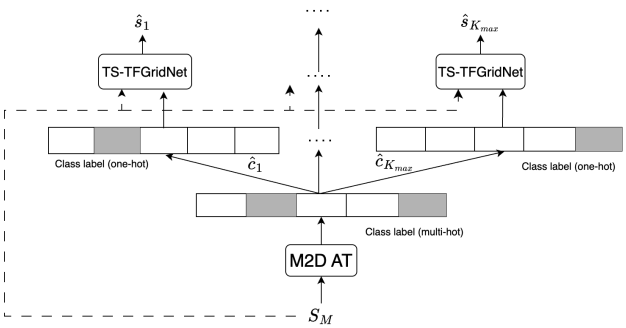
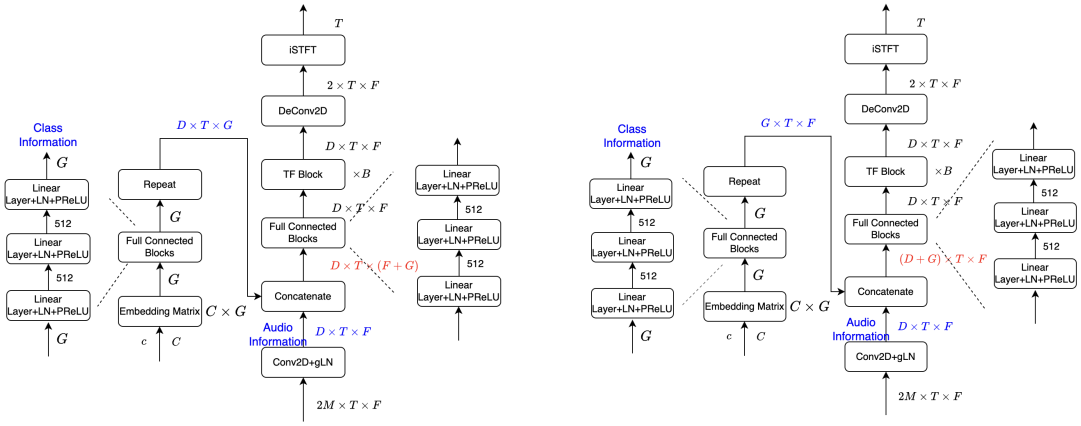


Figure 1: Overview of proposed system.

2. PROPOSED METHOD

Building on the baseline system, our proposed system replaces the default ResUNet model in the baseline with our proposed TS-TFGridNet model. See Fig. 1 for an illustration. Given an input mixture $S_M \in \mathbb{R}^{M \times N}$ with M denoting the number of microphones and N the number of time-domain samples, the M2D-based audio tagging (AT) system [4] first identifies all the sound events exisitng in the mixture, outputing a multi-hot vector denoting whether each sound event exists in the mixture. The multi-hot vector is then converted into a set of one-hot vectors $(\hat{c}_1, \ldots, \hat{c}_{K_{max}})$, with K_{max} denoting the number of idenfied sound events. Each one-hot vector is then utilized as an extra input (besides the input mixture signal) to condition TS-TFGridNet for TSE.

Our proposed TS-TFGridNet is illustrated in Fig. 2a and 2b. Given a total of C considered sound-event classes, we start with initializing an embedding matrix with size $C \times G$ to derive a class embedding of dimension G for each class $c \in \{1, \ldots, C\}$. This class embedding then undergoes processing through fully-connected blocks, which consist of multiple linear layers interleaved with non-linear activations. To facilitate diverse concatenation strategies with audio embeddings, we perform dimension matching, which involves replicating the processed class embedding to align its dimension with that of the audio embedding. The audio embedding has a dimension of $D \times T \times F$ and is obtained by applying 2D convolution and global layer normalization to the input mixture, which, after we apply short-time Fourier transform



(a) Architecture of TS-TFGridNet v1.

(b) Architecture of TS-TFGridNet v2.

Figure 2: Various variants of proposed TS-TFGridNet architectures.

(STFT), has a tensor shape of $2M \times T \times F$, with T denoting the number of frames, F the number of frequency bins, and 2 meaning stacking the real and imaginary (RI) components.

In TS-TFGridNet v1 (shown in Fig. 2a), we concatenate the processed class embedding with the audio embedding along the frequency dimension. This combined representation is then fed into secondary fully-connected blocks, which mirror the architecture of the initial embedding processing blocks. In TS-TFGridNet v2 (shown in Fig. 2b), the processed class embedding is concatenated with the audio embedding along the feature dimension. Similarly to TS-TFGridNet v1, this combined representation is subsequently passed into secondary fully-connected blocks that replicate the architecture of the initial embedding processing blocks. The integrated features then proceed through the dual-path block in the standard TF-GridNet [7] to predict the RI components of the target signal via complex spectral mapping [8]. Finally, the time-domain signal is obtained by applying inverse STFT (iSTFT) to the predicted RI components, and the loss function is the same as the time-domain CA-SDRi loss [2] proposed in the challenge baseline.

3. EXPERIMENTAL SETUP

We valudate TS-TFGridNet based on the DCASE2025 S5 dataset [2] provided by the challenge organizors. This section describes the dataset and system configurations.

Each input mixture contains up to three target sound events, and multiple non-target sound events and non-directional background noises. All the mixtures are sampled at 32 kHz. The dry sound source signals are sourced from [9], which contains 20 types of sound events. However, the "music" and "singing" classes are excluded, and therefore there are 18 target sound events to extract. Room impulse responses and noise data are obtained from [10].

The hyper-parameters of TS-TFGridNet v1 and v2 are exactly the same and are listed in Table 1. The STFT window and hop sizes are respectively 16 and 8 ms. All the training configurations are the same as the baseline model. Following the challenge setup [2], we use CA-SDRi as the evaluation metric.

4. EVALUATION RESULTS

Based on the M2D audio tagging model [3], which has an accuracy of 59.8% for audio tagging, we report the TSE performance of TS-TFGridNet in Table 2. The results demonstrate a substantial performance gain over the baseline ResUNet model. TS-TFGridNet v1 reaches a CA-SDRi of 14.2 dB, representing a 28.4% improvement over ResUNet, while TS-TFGridNet v2 obtaines a slightly-worse CA-SDRi of 14.0 dB, on the validation set. These results confirm the effectiveness of the proposed TS-TFGridNet model.

Table 1: Model configurations. Except M, G and C, the other hyper-parameters are defined in the same way as [7].

Symbol	Description	Value
M	Number of input microphones	4
G	Dimension of class embedding	512
C	Total number of classes	18
D	Dimension for each T-F unit embedding	48
B	Number of TF blocks	2
I	Kernel size for Unfold and Deconv1D	4
J	Stride size for Unfold and Deconv1D	1
H	Total number of hidden units of BLSTMs in each direction	128
L	Total number of heads in self-attention	4
E	Output channels in point-wise Conv2D to obtain key and query tensors in self-attention	256

Table 2: Comparison with challenge baseline ResUNet [3].

TSE Model	CA-SDRi (dB)	Improvement
ResUNet [3]	11.0	-
TS-TFGridNet v1	14.2	+28.4%
TS-TFGridNet v2	14.0	+27.0%

5. CONCLUSION

We have proposed TS-TFGridNet, which adapts TF-GridNet for label-queried TSE by concatenating class embedding with audio embedding along the frequency or feature dimension. Evaluation results on the development dataset of the DCASE2025 Challenge Task 4 show the effectiveness of TS-TFGridNet.

6. REFERENCES

[1] 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.

[2] M. Yasuda, B. T. Nguyen, N. Harada, R. Serizel, M. Mishra, M. Delcroix, S. Araki, D. Takeuchi, D. Niizumi, Y. Ohishi, *et al.*, "Description and discussion on DCASE 2025 Challenge Task 4: Spatial semantic segmentation of sound scenes," *arXiv preprint arXiv:2506.10676*, 2025.

[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," *arXiv preprint arXiv:2503.22088*, 2025.

[4] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked Modeling Duo: Towards a universal audio pre-training framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2391–2406, 2024.

[5] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94–114, 2020.

[6] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[7] ——, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[8] Z.-Q. Wang, P. Wang, and D. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speaker separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2001–2014, 2021.

[9] B. Veluri, M. Itani, J. Chan, T. Yoshioka, and S. Gollakota, "Semantic hearing: Programming acoustic scenes with binaural hearables," in *Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–15.

[10] M. Yasuda, Y. Ohishi, and S. Saito, "Echo-aware adaptation of sound event localization and detection in unknown environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 226–230.