SPATIAL SEMANTIC SEGMENTATION OF SOUND SCENES BASED ON ADAPTER FINE-TUNING

Technical Report

Zehao Wang, Sen Wang
Zhicheng Zhang, Jianqin Yin,

Beijing University of Posts and Telecommunications, China
{wzhao, senwang, zczhang, jqyin}@bupt.edu.cn

ABSTRACT

In this work, we present our submission system for DCASE 2025 Task4 on Spatial semantic segmentation of sound scenes (S5).Among them, we introduce the audio tagging (AT) and label-query source separation (LSS) systems built on the fine-tuned M2D and the modified version of ResUnet. By introducing the bidirectional recurrent neural network (DPRNN) module into ResUNet and improving the Feature-wise Linear Modulation (FiLM) mechanism, the model's ability to capture long-term dependent features in spatial audio and the flexibility of dynamic feature adjustment are enhanced. Experimental results show that the improved system outperforms the baseline system in class-aware evaluation metrics (CA-SDRi, CA-SI-SDRi), verifying the effectiveness of the method.

Index Terms— M2D, ResUNet, DPRNN

1. INTRODUCTION

The development of immersive communication technologies has driven the upgrading of requirements for spatial sound scene analysis[1][2][3]. The rapid advancement of deep learning has also provided support for improving audio processing accuracy, prompting people to pursue precise recognition and separation of sound events in spatial sound scenes to understand complex acoustic environments. Sound event detection largely relies on neural network architectures such as convolutional neural networks (CNN), convolutional recurrent neural networks (CRNN), and Transformer, while spatial sound scene separation mostly depends on CRNN and Transformer architectures, with some studies also using U-net and its variants. DCASE 2025 Task4 requires separating single-channel dry sound signals from multi-channel mixed audio and predicting their category labels, which is crucial for achieving object-based audio coding (e.g., MASA, object audio). The baseline system implements the task through a two-stage framework: The first stage uses a fine-tuned M2D model for audio tagging; the second stage achieves label-query source separation through the ResUNet family of models.

Separating sound mixtures into sources, known as source separation (SS), and predicting audio class labels, referred to as audio tagging (AT) or sound event detection (SED), are active areas of research, with their combination also being explored. In the baselines of DCASE Challenge Task 4 in 2020 and 2021, SS was used as a preprocessing step to improve the results of SED[4]. Most of the current related tasks mainly aim to improve one or both of SS and AT, and evaluate their performances separately at the same time. Therefore, the connection between the separated sources and the predicted labels has not been thoroughly evaluated. However, in Task S5, the separated sources are identified by their labels. That is, Task S5 requires meeting both the separation accuracy and label consistency.

An M2D-based AT model is a fine-tuned M2D[5] model on AudioSet[6] with 527 classes. M2D is a self-supervised learning (SSL) foundation model that is pre-trained with M2D's masked prediction-based objective using only audio samples from AudioSet. In this paper, for the M2D model, we introduce a novel Adapter fine-tuning method by inserting an Adapter into the feed-forward network of M2D. By modifying the pre-trained structure, this approach enhances the model's ability to capture comprehensive feature information and enables effective fine-tuning.

The second stage of the original system takes two forms: The first is a single-input single-output ResUNet, where the input layer is adjusted to accept an M-channel mixed spectrogram, and the output layer is modified to predict an M-channel magnitude mask and phase residual applied to the input spectrogram. A 1×1 convolutional layer is then used to generate the single-channel spectrogram of the target source. The other form is a single-input multi-output variant named ResUNetk, which differs from the first form in that it outputs $K_{max}N$ channels of magnitude masks and phase residuals applied to the input mixed spectrogram. Since the convolutional layers in the original model cannot effectively extract dependencies between frequency bins or time frames, we integrate a bidirectional recurrent neural network (DPRNN) into ResUNet and improve the Film conditioning.

2. METHODS

2.1. Adapter

The baseline method of DCASE 2025 Task 4, for the AT and LSS tasks, uses the fine-tuned Masked Modeling Duo (M2D) and a modified version of ResUnet. The audio representations extracted by M2D already reach the state-of-the-art level. The M2D model used in the Baseline is fine-tuned on AudioSet, and we use a more advanced fine-tuning method on this basis.

Specifically, we use Adapter fine-tuning, which adopts an inverted bottleneck structure to project the input features into a high-dimensional space, and uses the ReLU function to process the projection to extract high-richness information, and then restores it to the original dimension. This high-richness information is a mixture that not only contains the time-frequency patterns of sound events but also contains the adjacent information of different events. The structure of the Adapter can be represented as follows:

$$x'_{\ell} = LN\left(\text{attn} + x_{\ell-1}\right) \tag{1}$$

$$x''_{\ell} = \text{ReLU}\left(x'_{\ell} \cdot W_{\text{up}}\right) \cdot W_{\text{down}} \tag{2}$$

$$x_{\ell} = LN\left(x''_{\ell} + x'_{\ell}\right) \tag{3}$$

where, $x_{\ell-1} \in R^{T \times D}$ is the input of the BEATs block from the $\ell - 1$ layer, x'_{ℓ} is the input of the M3A-FFN in the ℓ layer, x''_{ℓ} are the outputs of the Adapter, attn $\in R^{T \times D}$ is the output of the Multi-Head Self-Attention, W_{down} and W_{up} are the parameters of the down-projection layer and up-projection layer respectively, x_{ℓ} is the output from the ℓ layer.

By inserting this Adapter into M2D and modifying the pre-trained structure, we enhance its ability to capture comprehensive feature information and address the multi-scene nature of the heterogeneous dataset SED. Effective fine-tuning of the model can improve its robustness in multi-scene applications.

2.2. FiLM conditioning DPRNN

The FiLM used in the baseline only performs multiplication with β. On this basis, we introduce element-wise multiplication γ, enabling the model to adjust feature distributions more flexibly[7].To bridge the text encoder and the separation model, use a FiLm layer after each ConvBlock deployed in the ResUNet. We use $H \in R^{m \times h \times w}$ to denote the output feature map produced by ConvBlock l with m channels, here h and w are the height and width of the feature map $H^{(l)}$, respectively. The modulation parameters are applied per feature map H_i with the FiLm layer as follows:

$$\text{FiLM}(H \mid \gamma_i, \beta_i) = \gamma_i H_i + \beta \tag{4}$$

where $H_i \in R^{h \times w}$, and $\gamma, \beta \in R^{m}$ are the modulation parameters from $g(.)$, i.e., $(\gamma, \beta) = g(e_q)$, such that $g(.)$ is a neural network and e_q is the text embedding obtained from the text encoder. In this work, we model $g(\cdot)$ with two fully connected layers followed by ReLU activation, which is jointly trained with the ResUNet separation model.

To achieve dynamic feature extraction, we insert a DPRNN module into the encoder-decoder intermediate layer (bottleneck layer) of the baseline ResUNet[8]. The specific structure is as follows: The DPRNN module consists of a bidirectional LSTM in the time dimension and a bidirectional LSTM in the frequency dimension. After the input feature map is processed in the time dimension, it is recursively processed along the frequency dimension and finally projected back to the original number of channels through a 1×1 convolution. The input x has the shape B: Batch Size, C: Channels, T: Time Steps, F: Frequency Bins.

3. EXPERIMENT

3.1. Dataset

This task is based on DCASE2025Task4Dataset. This dataset was recorded and designed for DCASE 2025 Challenge Task 4: Spatial Semantic Segmentation of Sound Scenes (S5). The dataset contains 18 classes of sound events recorded in an anechoic chamber (ASE1K) and room impulse responses (RIRs) recorded using first-order ambisonics (FOA) microphones. All of the acoustic data and RIR formats included in this dataset are 32kHz/16bit. In the following part of this description, we will briefly summarize the recording of sound events and RIR. ASE1K is a one-shot recording of 18 classes of sound events recorded in an anechoic chamber. The recording was made using three cardioid microphones to capture the sound events from the left, front and right, and one omnidirectional microphone to capture the sound from above. In the S5 task, it is assumed that you will simply select a single channel (e.g. ch=3) from these and use it as a monaural sound event. For each class, 50 to 80 events were recorded, and a total of over 1K samples were recorded.

Table 1: Performance of the improved system

system	model	CA-SDRi	mean Accuracy
ResUNet	Baselin	11.032	59.80%
ResUNet	Ours	**11.073**	**62.67%**
ResUNetk	Baseline	11.088	59.80%
ResUNetk	Ours	**11.783**	**65.47%**

The RIR dataset is made up of RIRs recorded in six environments for DCASE2025 Task4, combined with RIRs that have already been released as the FOA-MEIR dataset. All recordings were made using the same FOA Microphone (Sennheiser Ambeo VR Mic). RIR recordings were made from multiple locations in each environment, and these are compiled in sofa file format.

This dataset also includes noise recordings in the FOA format. All of noise recordings are all the same as those included in FOA-MEIR. The recordings were made using the same FOA microphones as those used in RIR.

During evaluation, systems will undergo assessment using labels of varying granularity to gain a comprehensive understanding of their performance and assess their adaptability across diverse applications. Given that different datasets feature distinct target classes, it is possible that sound labels present in one dataset may not be annotated in another. As a result, systems must be capable of handling potential missing target labels during training. Furthermore, SED system is required to operate without knowledge of the origin of the audio clips during evaluation, emphasizing the need for robust and generalized performance across varied scenarios.

3.2. Experiment setup

We train the M2D model for 800 epochs and the ResUNet for 80 epoch. The batch size is set to 4. Each training session is deployed on the NVIDIA RTX 4090 and lasts 200 hours.

3.3. Results and submissions

Table 1 shows the performance of the commit system. The baseline model initially used both the M2D and ResUNet models.Our system improves the fine-tuning method by using a self-designed Adapter for the M2D module, and modifies the FiLM conditioning and adds the DPRNN module for the ResUNet part. Eventually, it shows significant performance improvements. For ResUNet, we increase the CA-SDRi from 11.032 to 11.073 and the mean Accuracy from 59.80% to 62.67%; for ResUNetK, we train with a smaller number of epochs, raising the CA-SDRi from 11.088 to 11.783 and the mean Accuracy from 59.80% to 65.47%.

4. CONCLUSION

In this study, we significantly improved the performance of feature extraction and spatial acoustic scene semantic segmentation by adding a special Adapter to M2D to modify the fine-tuning method, as well as introducing the DPRNN module and improving the FiLM conditioning mechanism in ResUNet. Our system outperforms the baseline in terms of both the CA-SDRi and Accuracy metrics, achieving 11.073 and 62.67% on the ResUNet system, and 11.783 and 65.47% on ResUNetK.

5. REFERENCES

[1] M. Multrus, S. Bruhn, J. Torres, E. Fotopoulou, T. Toftgård, E. Norvell, S. Döhla, Y. Gao, H.-y. Su, L. Laaksonen, *et al.*, “Immersive voice and audio services (ivas) codec-the new 3gpp standard for immersive communication,” in *157th AES Convention*, 2024.

[2] E. Fotopoulou, K. Sagnowski, K. Prebeck, M. Chakraborty, S. Medicherla, and S. Döhla, “Use-cases of the new 3gpp immersive voice and audio services (ivas) codec and a web demo implementation,” in *2024 IEEE 5th International Symposium on the Internet of Sounds (IS2)*. IEEE, 2024, pp. 1–6.

[3] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Østergaard, F. Font, T. Bäckström, and C. Fischione, “The internet of sounds: Convergent trends, insights, and future directions,” *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, 2023.

[4] N. Turpault, S. Wisdom, H. Erdogan, J. Hershey, R. Serizel, E. Fonseca, P. Seetharaman, and J. Salamon, “Improving sound event detection in domestic environments using sound separation,” *arXiv preprint arXiv:2007.03932*, 2020.

[5] 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*, 2024.

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[7] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, “Separate anything you describe,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[8] H. Yin, J. Bai, M. Wang, and J. Chen, “Language-queried audio source separation via resunet with dprnn,” *DCASE2024 Challenge, Tech. Rep*, 2024.