ENHANCING STEREO SOUND EVENT LOCALIZATION AND DETECTION THROUGH PRETRAINED AUDIO REPRESENTATIONS AND HYBRID ARCHITECTURES

Technical Report

Tianbo Zhao, Zerui Han, Mengmei Liu

Xiaomi Corporation, MITC-Multimodal generation, Beijing, China,
{zhaotianbo, hanzerui, liumengmei}@xiaomi.com

ABSTRACT

The technical report presents our submission system for Task 3 of the DCASE 2025 Challenge: Stereo sound event localization and detection (SELD) in regular video content. This year we participate in the audio-only track. We propose a method that decomposes the SELD task into two sub-tasks. For the detection task, we employ the pre-trained Dasheng model [1], which is a high-performing audio encoder. For the localization task, we utilize the ResNet-Conformer [2] [3] architecture, which has demonstrated excellent performance in recent years' DCASE tasks. We evaluated our method on the dev-test set of the development dataset. The results show that our approach outperforms the baseline.

Index Terms— Sound event localization and detection, log-mel spectrogram, Conformer

1. INTRODUCTION

The Sound Event Localization and Detection (SELD) task aims to simultaneously identify the classes of sound events present in an audio scene and estimate their corresponding spatial locations. This task is of great significance in applications such as intelligent surveillance, robotic auditory perception, and augmented reality.

In recent years, the DCASE challenges have greatly advanced the development of SELD models. Since 2019 [4] , each edition of the DCASE challenge has included a related task, with increasing complexity year by year. New challenges introduced include the use of real sound sources, overlapping events, and distance estimation. In this year's challenge, the input to the model has been changed to stereo format for the first time, marking a significant step toward real-world applications. In this year's Task 3 [5] , participants are required to estimate sound event classes, azimuth angles, and distance information using only stereo signals. In fact, stereo content is much more prevalent and widely available in real-world scenarios compared to FOA-format audio. Therefore, investigating methods for achieving SELD using stereo signals holds significant research value and practical application potential.

To address the SELD task, researchers have mainly proposed two types of approaches. The first type employs a single-branch model that jointly outputs sound event detection (SED), direction of arrival (DOA), and source distance estimation (SDE) information. Among them, the ACCDOA method [6] proposed in accdoa estimates a directional vector, with its magnitude indicating the presence probability of the corresponding sound source. Building upon this, [7] introduced the Multi-ACCDOA algorithm, which incorporates Auxiliary Duplicating Permutation Invariant Training (ADPIT) [8] to better handle overlapping events from the same sound class.

Another popular approach is the multi-branch architecture, where the initial layers of the network share parameters to extract general audio features, and the final layer branches into multiple fully connected heads that independently output SED, SDE and DOA information. This method achieved the first place in DCASE 2024 Task 3 [9], demonstrating its strong capability in handling complex acoustic scenes with effective task decoupling and joint optimization.

Based on the two mainstream approaches mentioned above, we note that sound event detection and classification is a relatively more mature task compared to localization, with larger available datasets and a more established research community. Therefore, we believe it is beneficial to leverage pre-trained models trained on tens of thousands of hours of audio data, and fine-tune them on the competition dataset for the detection sub-task. Specifically, we adopt the Dasheng model, a high-performance audio encoder based on the Transformer architecture. For the localization task, considering the limited availability of well-annotated data, we choose the ResNet-Conformer architecture, which has demonstrated strong performance in recent DCASE SELD tasks.

2. PROPOSED METHOD

2.1. Data Augmentation

This year's Task 3 dataset is a 5-second clip version derived from the STARSS23 dataset [10] [11]. STARSS23 contains approximately 7.5 hours of real-world recordings covering 13 sound classes, with annotations for sound source class, azimuth angle, and distance provided every 0.1 seconds. In comparison, this year's dataset includes a development set consisting of 30,000 5-second audio clips, amounting to a total duration of approximately 41.7 hours. Furthermore, the dataset converts the original four-channel FOA-format audio into stereo format through a defined transformation.

Although the dataset has seen a noticeable increase in scale compared to previous editions, we observe that the number of samples for certain sound classes remains limited, which is insufficient to support the training of a high-performing model. Therefore, data augmentation is essential to improve model generalization and overall performance.

In the construction of the training dataset which we call MixedAudioDataset, to obtain 5-second audio clips containing sound events (referred to as "mix" hereafter), we first randomly select a signal-to-noise ratio (SNR) and generate a background noise mix. Then, using the event dataset from FSD50K-FMA [12] [13], we randomly sample 4 to 6 event instances (with replacement) from each class. During dataset initialization, additional augmentation event types can be introduced to adjust the occurrence ratios of certain events. Each event instance is randomly cropped to a specific duration and undergoes various augmentations at different time intervals, including gain control, polarity inversion, multi-band EQ, time-domain masking, audio channel swapping(ACS) [14] [15] and convolution with room impulse responses (RIRs). The start time for inserting each event into the mix is also selected randomly. This completes the data generation pipeline. Finally, the dataloader returns the synthesized 5-second audio clip along with its corresponding labels.

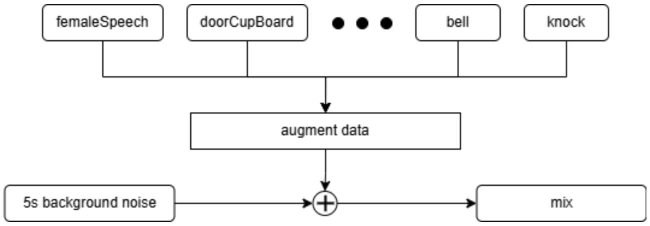

Figure 1: Data Augmentation Pipeline

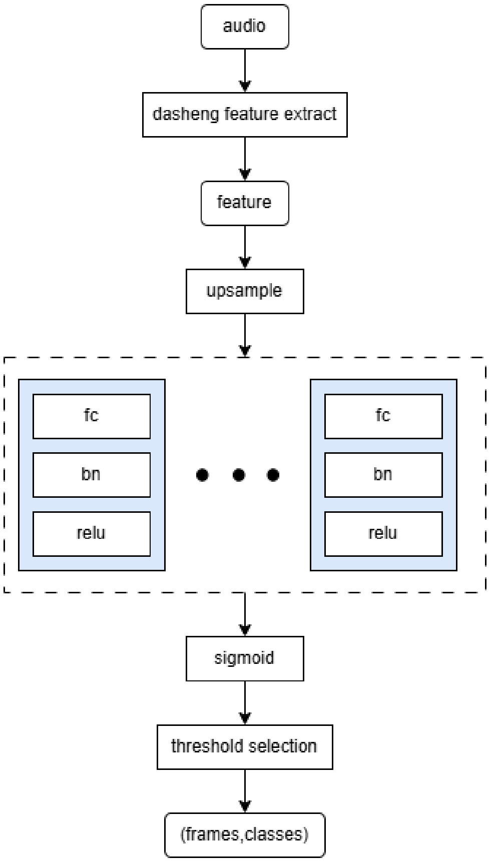

Figure 2: Detection Model Architecture

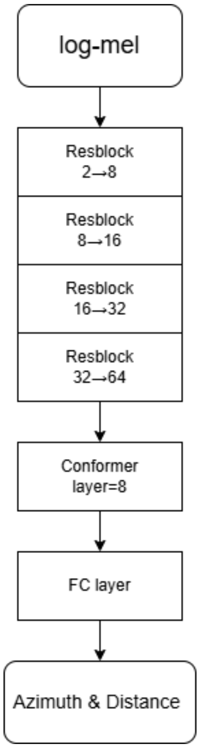

Figure 3: Localization Model Architecture

2.2. Detection Model Architecture

In this part, we employ the Dasheng-Base pre-trained model, which has 86 million parameters, to encode the input audio into 768-dimensional embedding vectors. Dasheng is a large-scale self-supervised audio encoder, based on the Masked Autoencoder (MAE) framework, designed to learn general-purpose audio representations. Trained on over 272,000 hours of audio data—including speech, music, and environmental sounds—Dasheng features up to 1.2 billion parameters , making it one of the largest self-supervised audio models to date.

The model adopts an asymmetric Transformer-based encoder-decoder architecture, in which only unmasked segments of the input are processed, significantly reducing computational overhead while enabling scalable training. Experimental results demonstrate that Dasheng shows strong capabilities across music and environmental sound classification. Furthermore, Dasheng's learned embeddings exhibit strong zero-shot transferability, enabling direct application to downstream tasks without fine-tuning

In the detection model, audio features are extracted through Dasheng. After upsampling the embedding output, it is mapped to the (frames, classes) dimension using multiple fully connected (FC) layers combined with batch normalization (BN) and ReLU activation. Finally, multi-label detection is performed using the Sigmoid function.

2.3. Localization Model Architecture

Our localization model is based on the ResNet-Conformer architecture. The inputs are log-mel spectrograms with shape B×2×251×64, where the four dimensions represent batch size, channel number, time steps, and the number of mel filters, respectively. This is followed by four ResBlocks, which replace the standard convolutional blocks used in the baseline. Subsequently, Conformer blocks are introduced to better capture both local and global features. Finally, a fully connected layer maps the output to a tensor of shape B×50×26, where the second dimension aligns with the number of labels, and the last dimension represents the azimuth angle and distance predictions for each of the 13 classes.

2.4. Network Training

For the detection task, we use the MixedAudioDataset with full event class random augmentation for sound event classification. For validation and testing, we employ the AudioFrameDataset, deriving from the STARSSS23 dataset. In each validation epoch, the optimal threshold for each event class is computed based on performance, and these thresholds are then applied during the test epoch to calculate the F-score. The batch size is set to 128, and the learning rate starts from 0, warms up to 0.001, and then follows a cosine decay schedule over 200 epochs back to 0. We keep the top five models corresponding to the highest F-scores on the validation set. For each of these models, we also evaluate the F-score on the test set, and the final model for inference is selected based on both validation and test performance.

For the localization task, The batch size is set to 64, and the learning rate starts from 0, warms up to 0.001, and then follows a cosine decay schedule over 300 epochs back to 0. The loss function used was mean squared error (MSE).

3. RESULTS ON DEVELOPMENT DATASET

We trained the final model using the two training folders provided by DCASE and validated it on two separate test folders. The training folders contain a total of 16,214 samples from various rooms, while the test folders include 13,786 samples. We made a total of four submissions, with each submission differing slightly in data augmentation strategies. Experimental results demonstrate that our model outperforms the baseline approach across all evaluation metrics.

Table 1: Comparison of model performance with the baseline for the development dataset

Model	F-score (20°/1)	DOAE	RDE
Baseline	22.8 %	24.5°	41 %
Sub1	35.2 %	17.4°	38 %
Sub2	36.3 %	17.2°	37 %
Sub3	37.0 %	16.9°	39 %
Sub4	35.1 %	18.0°	37 %

4. CONCLUSION AND FUTURE WORK

In this year's DCASE Challenge Task3, we employed a method that combines a detection model with a localization model to achieve SELD. For the detection part, we utilized the pre-trained Dasheng model, while for localization, we adopted the ResNet-Conformer architecture, which has previously demonstrated strong performance. With the inclusion of necessary data augmentation techniques, experimental results show that our approach achieves promising results.

There are still several issues worth further investigation in the future. For instance, how to predict the distance of sound sources more accurately remains an open challenge. Additionally, identifying individual source instances within the same event class is also an interesting and meaningful research direction.

5. REFERENCES

[1] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification."

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[3] Q. Wang, L. Chai, H. Wu, Z. Nian, S. Niu, S. Zheng, Y. Wang, L. Sun, Y. Fang, J. Pan, J. Du, and C.-H. Lee, "The nerc-slip system for sound event localization and detection of dcase2022 challenge," DCASE2022 Challenge, Tech. Rep., June 2022.

[4] http://dcase.community/challenge2019/.

[5] http://dcase.community/challenge2025/.

[6] K. Shimada, Y. Koyama, N. Takahashi, S. Takahashi, and Y. Mitsufuji, "Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection," in *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2021, pp. 915–919.

[7] K. Shimada, Y. Koyama, S. Takahashi, N. Takahashi, E. Tsunoo, and Y. Mitsufuji, "Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 316–320.

[8] Y. Liu and D. Wang, "Permutation invariant training for speaker-independent multi-pitch tracking," in *ICASSP 2018-2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5594–5598.

[9] Q. Wang, Y. Dong, H. Hong, R. Wei, M. Hu, S. Cheng, Y. Jiang, M. Cai, X. Fang, and J. Du, "The nerc-slip system for sound event localization and detection with source distance estimation of dcase 2024 challenge," DCASE2024 Challenge, Tech. Rep., June 2024.

[10] https://zenodo.org/records/15559774.

[11] https://zenodo.org/records/7880637.

[12] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[13] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv preprint arXiv:1904.08779*, 2019.

[15] 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.