IMPROVING STEREO 3D SOUND EVENT LOCALIZATION AND DETECTION: PERCEPTUAL FEATURES, STEREO-SPECIFIC DATA AUGMENTATION, AND DISTANCE NORMALIZATION

Technical Report

Jun-Wei Yeow, Ee-Leng Tan, Santi Peksi, Woon-Seng Gan*

Smart Nation TRANS Lab, Nanyang Technological University, Singapore
junwei004@e.ntu.edu.sg, {etanel, speksi, ewsgan}@ntu.edu.sg

ABSTRACT

This technical report presents our submission to Task 3 of the DCASE 2025 Challenge: Stereo Sound Event Localization and Detection (SELD) in Regular Video Content. We address the audio-only task in this report and introduce several key contributions. First, we design perceptually-motivated input features that improve event detection, sound source localization, and distance estimation. Second, we adapt augmentation strategies specifically for the intricacies of stereo audio, including channel swapping and time-frequency masking. We also incorporate the recently proposed FilterAugment technique that has yet to be explored for SELD work. Lastly, we apply a distance normalization approach during training to stabilize regression targets. Experiments on the stereo STARSS23 dataset demonstrate consistent performance gains across all SELD metrics. Code to replicate our work is available in this repository[1]

***Index Terms*—** Sound Event Localization and Detection, Sound Distance Estimation, Sound Source Localization, Sound Event Detection

1. INTRODUCTION

Sound Event Localization and Detection (SELD) is a form of machine-listening that enables systems to not only understand what sounds are happening, but also where they come from [1]. This form of spatial intelligence can be extended into three-dimensions by integrating Sound Distance Estimation (SDE), cumulating in 3D SELD. The transition of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge tasks from SELD to 3D SELD also signifies the growing interest in distance-aware systems [2].

Most existing SELD and 3D SELD systems use small microphone arrays, often configured in either the First Order Ambisonics (FOA) or Multi-channel Microphone Array (MIC) formats. In 2025, the DCASE Challenge Task 3 shifts the focus onto stereo-based 3D SELD, pivoting to a form of spatial awareness meant for consumer electronics. This reflects a greater trend towards more consumer-friendly environmental intelligence, such as for wearables [3] and online inference systems [4].

Compared to traditional FOA or MIC audio formats, stereo-based SELD has not yet been extensively explored [5]. In this paper, we outline our proposed system and general methodology for stereo-based 3D SELD. Our methods, including stereo-aware augmentation, perceptually-inspired features, and distance normalization, can apply to generic stereo 3D SELD pipelines to significantly improve performance.

*This research is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (MOE-T2EP20221-0014)

[1] https://github.com/itsjunwei/NTU_SNTL_Task3

2. INPUT FEATURES

Let $x_{\mathrm{L}}[n]$ and $x_{\mathrm{R}}[n]$ denote the left and right stereo input channels, respectively, with n being the discrete-time index. The Short-Time Fourier Transform (STFT) of the c-th channel at time frame t and frequency bin f is denoted as $X_c(t,f)$, for $c \in \{\mathrm{L}, \mathrm{R}\}$.

2.1. Mid-Side Conversion

Mid-Side (MS) conversion explicitly decomposes the stereo signal into Mid (M) and Side (S) components. This decomposition has been explored for acoustic analysis tasks using stereo audio, such as Acoustic Scene Classification [6]. The conversion process is performed in the time-domain as follows,

$$m[n] = \frac{x_{\mathrm{L}}[n] + x_{\mathrm{R}}[n]}{2}, \quad s[n] = \frac{x_{\mathrm{L}}[n] - x_{\mathrm{R}}[n]}{2}, \tag{1}$$

where $m[n]$ and $s[n]$ denote the discrete-time mid and side signals, respectively. Here, $m[n]$ represents the average pressure and $s[n]$ captures the horizontal pressure differential. The STFTs of $m[n]$ and $s[n]$ are therefore $M(t,f)$ and $S(t,f)$, respectively.

Similar to the Intensity Vector (IV) used in FOA-based SELD work [7], we derive a MS-based intensity feature for stereo audio. For each time-frequency (TF) bin, the real portion of the MS cross-spectrum is computed as follows,

$$I_x(t,f) \;=\; \Re\big\{M(t,f)\; S^*(t,f)\big\}, \tag{2}$$

before being normalized by the total MS power:

$$\tilde{I}_x(t,f) \;=\; \frac{I_x(t,f)}{|M(t,f)|^2 + |S(t,f)|^2 + \varepsilon}\,. \tag{3}$$

where ε is a small constant to prevent division by zero. Finally, $\tilde{I}_x(t,f)$ is typically projected onto a K-band Mel scale using the Mel filter bank matrix $\mathbf{W}_{\mathrm{mel}}$:

$$\mathrm{IV}(t,f) = \tilde{I}_x(t,f) \cdot \mathbf{W}_{\mathrm{mel}}(f,k). \tag{4}$$

This IV feature captures stereo intensity differences in a similar fashion to spatial features for FOA audio [8], providing important directionality cues to our stereo-based SELD system.

2.2. Spatial Coherence

Spatial coherence measures the similarity between channels as a function of frequency. This property has been investigated in many SDE frameworks due to its strong relationship with sound source distance [9]. In this work, we use the magnitude-squared coherence (MSC) between the LR channels.

Firstly, we define the cross-power spectral density between the two stereo channels as

$$\Phi_{\mathrm{L,R}}(t,f) = \mathbb{E}\Big[X_{\mathrm{L}}(t,f)\, X_{\mathrm{R}}^{*}(t,f)\Big]. \tag{5}$$

In practice, we estimate $\Phi_{\mathrm{L,R}}(t,f)$ using time-recursive averaging [10]:

$$\hat{\Phi}_{\mathrm{L,R}}(t,f) = \lambda\, \hat{\Phi}_{\mathrm{L,R}}(t-1,f) + (1-\lambda)\, X_{\mathrm{L}}(t,f)\, X_{\mathrm{R}}^{*}(t,f), \tag{6}$$

where $\lambda \in [0,1]$ is a smoothing coefficient, set as 0.8 in this work [11]. The MSC $\hat{\gamma}(t,f)$ is subsequently calculated as

$$\hat{\gamma}(t,f) = \frac{|\hat{\Phi}_{\mathrm{L,R}}(t,f)|^2}{\hat{\Phi}_{\mathrm{L,L}}(t,f)\, \hat{\Phi}_{\mathrm{R,R}}(t,f) + \varepsilon}, \tag{7}$$

where $0 \le \hat{\gamma}(t,f) \le 1$. Here, high MSC values typically signal direct, coherent sources (near/focused events), while lower values suggest diffuse or distant sources. Similarly, we project the MSC onto the same K-band Mel scale using $\mathbf{W}_{\mathrm{mel}}$:

$$\mathrm{MSC}(t,f) = \hat{\gamma}(t,f) \cdot \mathbf{W}_{\mathrm{mel}}(f,k). \tag{8}$$

3. DATA AUGMENTATION

We employ both waveform-level and spectrogram-level augmentation methods to generate meaningful variations in stereo spatial cues, thereby improving model robustness.

3.1. Waveform-level

Audio channel swapping (ACS) methods have been developed for both the FOA and MIC audio formats [12, 13]. ACS-based methods are extremely effective for the two-dimensional SELD task due to them being able to significantly increase the number of directional events, while preserving the natural reverberation conditions of the recording environments [14].

In the case of stereo audio, the ACS method becomes a simple swapping of the left and right channels. Accordingly, the azimuth labels are also inverted about the frontal axis. This can essentially double the amount of directional sound events available.

3.2. Spectrogram-level

In this work, we explore three different spectrogram-level data augmentation methods that are applied to the input features on-the-fly during training.

FilterAugment applies band-specific gains across the input spectrograms, simulating realistic distortions across frequency bands. Originally developed for Sound Event Detection [15], this method introduces variability in spectral coloration. Therefore, this can prevent models from relying on frequency-specific artifacts, making it attractive and applicable for SELD.

Frequency Shifting perturbs the input spectrograms by shifting frequencies within a controlled range, simulating pitch variation in the frequency domain. This has shown to improve generalization by encouraging the model to learn frequency-invariant representations of spatial cues [16, 17].

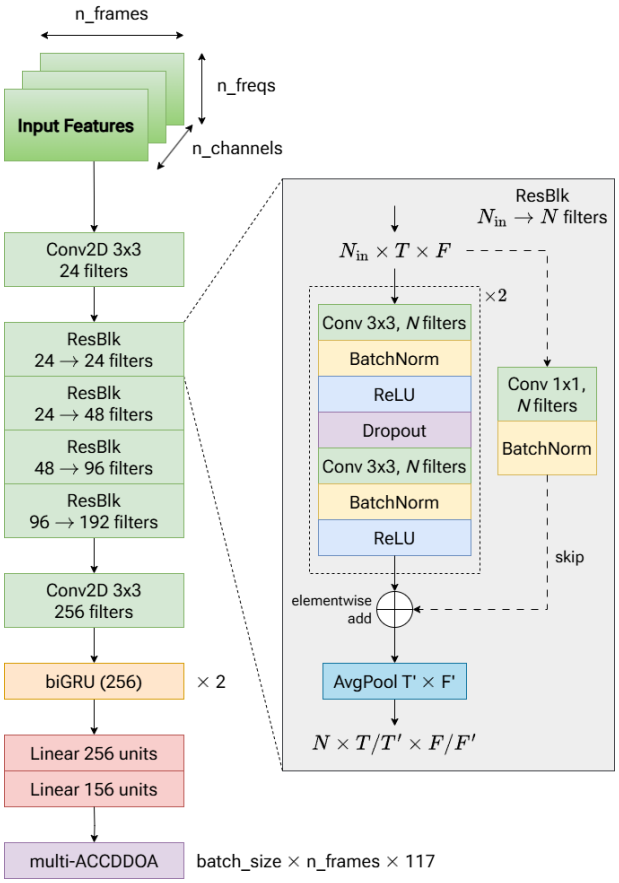

Figure 1: Block diagram of the ResNet-biGRU CRNN used in our DCASE 2025 submission.

Inter-Channel-Aware Time-Frequency Masking (ITFM) is our proposed adaptation of TF masking (TFM) for stereo audio. Traditional TFM methods, such as SpecAugment [18] or Cutout [19], typically apply independent or identical masks to each of the spectral channels. This risks distorting or erasing inter-channel differences, which are critical for robust localization and distance estimation in stereo-based tasks. Our ITFM method pre-computes and reapplies inter-channel differences post-masking. Therefore, this helps to preserve inter-channel differences, maintaining spatial information essential for robust 3D SELD.

4. NETWORK ARCHITECTURE

For this year's challenge, we employed a relatively lightweight convolutional recurrent neural network (CRNN) built on a ResNet backbone followed by bi-directional gated recurrent units (bi-GRUs). The architecture follows similarly to our previous DCASE submission in 2024 [20]. For the output, we use the multi-ACCDDOA output format proposed by Shimada et al. [21] and extended for distance estimation by Krause et al. [22].

Figure 1 showcases the CRNN system that we use for our sub-

missions. The Mean Squared Error (MSE) is used as the loss function. Notably, we opt not to use Conformer modules [13, 14] to reduce the substantial computational and environmental costs associated with training such large, complex models. Quantitatively, our full systems only uses around 4 million parameters, and requires 1.89G multiply-accumulate operations (MACs) per forward pass.

5. DISTANCE NORMALIZATION

The distance values in the STARSS23 dataset can range from $[0.04, 7.64]$ in meters. If we were to directly regress these values in ACCDOA-based output format variants, the MSE loss function can very easily be biased towards further or more distant sound events [22]. Therefore, to mitigate this problem, we apply the distance normalization method that was first proposed in our previous work on 3D SELD [20].

This distance normalization procedure scales the distribution of distances, d, to a uniform range of $[-1, 1]$ in two steps:

$$d' = \frac{d - \bar{d}}{\sigma_d}, \quad d_{\text{norm}} = \frac{d'}{\max(d')}, \tag{9}$$

where $\bar{d}$ and σ_d represent the mean and standard deviation of all distances, respectively. This normalization ensures that all elements in the multi-ACCDDOA vector lies within the same scale of $[-1, 1]$, preventing larger distances from disproportionately affecting the MSE loss, thereby yielding better overall 3D SELD performance.

6. EXPERIMENTAL METHOD

6.1. Dataset

The STARSS23 dataset consists of real-world, multi-room recordings with annotations of event activity, spatial trajectories, and distances [23]. The stereo version of STARSS23 comprises of 30,000 five-second audio recordings, with the training set containing 22.5h of audio data. To enrich the number of real-world directional examples, we apply ACS to the STARSS23 dataset to double the amount of real data to roughly 45h.

Manual annotation of 3D SELD data is costly, resulting in the class distribution in the STARSS23 dataset to be severely imbalanced. We mitigate this challenge by first generating additional FOA data using the SpatialScaper generator [24], before converting them into stereo audio using the provided conversion generator. In total, a total of 30,000 additional five-second synthetic stereo audio samples were generated. The combined dataset used for training therefore spans approximately 86.7h.

For feature extraction, we use a sampling rate of 24kHz, using a 1024-point FFT with the Hann window of length 1024 samples and a hop length of 300 samples, resulting in 400 time frames per audio clip. All features are mapped onto 96 Mel bands.

6.2. Training

The base feature stack consists of LR log-Mel spectrograms. We extend this by including the proposed perceptually-motivated features. In particular, the addition of MS log-Mel spectrograms and IV gives the *MSI* feature set. The subsequent addition of MSC yields the richer *MSIC* feature set.

Spatial diversity during training is further enhanced by two alternative spectrogram-level augmentation pipelines – either using the proposed stereo-based TFM method (ITFM) or frequency-domain perturbation that combines FilterAugment and Frequency Shifting (FAFS).

We train each of our systems for 100 epochs using the Adam optimizer, with a peak learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a batch size of 64. We follow the implementation of the baseline system and evaluate the model on the test split of the development set of the STARSS23 dataset, saving the model with the best validation location-dependent F-Score as our final model. Furthermore, we further fine-tune the models for another 20 epochs on only real audio recordings after the initial training.

Table 1: 3D SELD performance of the SELDNet baseline system when distance normalization is applied.

Experiment	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
Baseline	23.72	20.8°	0.347	0.409
+ DN	24.60	17.0°	0.287	0.379

Table 2: Performance of our submitted systems using different combinations of input features and data augmentation pipelines.

	Setup	$F_{20^\circ/1}\uparrow$	$LE_{CD}\downarrow$	$RDE_{CD}\downarrow$	$\mathcal{E}_{SELD}\downarrow$
A	MSI + ITFM	43.95	13.2°	0.271	0.302
B	MSIC + ITFM	43.12	**12.7**$^\circ$	**0.259**	0.300
C	MSI + FAFS	**45.32**	13.2°	0.262	**0.294**
D	MSIC + FAFS	43.44	13.2°	0.261	0.300

7. RESULTS

We employ the same validation metrics as used in DCASE 2025 Challenge Task 3. These include the location-dependent F-score ($F_{20^\circ/1}$), class-dependent localization error (LE_{CD}), and class-dependent relative distance error (RDE_{CD}). In addition, we also calculate an aggregated SELD error ($\mathcal{E}_{\text{SELD}}$) to provide an overview of the overall performance of the system as follows [25]:

$$\mathcal{E}_{\text{SELD}} = ((1 - F_{\leq 20^\circ/1}) + \frac{LE_{CD}}{180^\circ} + RDE_{CD})/3. \tag{10}$$

First, we demonstrate the effectiveness of using distance normalization (DN). Table 1 showcases the performance of the baseline SELDNet trained using the stereo version of the STARSS23 dataset. From the results, we can see that using DN yields consistent performance improvements across all metrics. In particular, $\mathcal{E}_{\text{SELD}}$ decreases by 7.33%, showing the benefits of distance normalization in improving overall 3D SELD performance.

For our submitted systems, we use a combination of improved feature sets with different augmentation methods. Table 2 showcases the 3D SELD performance of our submitted systems. All submitted systems apply DN to the ground truth labels.

Compared to the baseline system in Table 1, we can see that our proposed approach yields significant improvements in 3D SELD performance. We leave the optimal combination of spatial features and augmentation pipelines for future work. We theorize also that our methods, in combination with more complex and sophisticated model architectures, can yield even better performance.

8. CONCLUSION

This technical report details our proposed methods for the stereo-based 3D SELD task. We use perceptually-motivated input features to improve both localization and distance estimation performance. We introduce FilterAugment for the 3D SELD task, and propose a stereo-specific form of spectrogram masking augmentation. Overall, our proposed approach yields consistent and extensive improvements across all 3D SELD metrics, and can be applied to generic stereo-based 3D SELD methodologies.

9. REFERENCES

[1] 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, no. 1, pp. 34–48, 2018.

[2] D. Diaz-Guerra, A. Politis, P. Sudarsanam, K. Shimada, D. A. Krause, K. Uchida, Y. Koyama, N. Takahashi, S. Takahashi, T. Shibuya, Y. Mitsufuji, and T. Virtanen, "Baseline models and evaluation of sound event localization and detection with distance estimation in dcase2024 challenge," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, Tokyo, Japan, 2024, pp. 41–45.

[3] K. Nagatomo, M. Yasuda, K. Yatabe, S. Saito, and Y. Oikawa, "Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 156–160.

[4] J. W. Yeow, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Real-time sound event localization and detection: Deployment challenges on edge devices," *arXiv preprint arXiv:2409.11700*, 2024.

[5] J. Wilkins, M. Fuentes, L. Bondi, S. Ghaffarzadegan, A. Abavisani, and J. P. Bello, "Two vs. four-channel sound event localization and detection," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, 2023, pp. 216–220.

[6] Y. Han, J. Park, and K. Lee, "Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification." in *DCASE*, 2017, pp. 46–50.

[7] Y. Cao, T. Iqbal, Q. Kong, M. Galindo, W. Wang, and M. Plumbley, "Two-stage sound event localization and detection using intensity vector and generalized cross-correlation," *DCASE2019 Challenge, Tech. Rep*, 2019.

[8] L. Perotin, R. Serizel, E. Vincent, and A. Guérin, "Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 22–33, 2019.

[9] K. Zhagyparova, R. Zhagypar, A. Zollanvari, and M. T. Akhtar, "Supervised learning-based sound source distance estimation using multivariate features," in *2021 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2021, pp. 1–5.

[10] A. Schwarz and W. Kellermann, "Coherent-to-diffuse power ratio estimation for dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 1006–1018, 2015.

[11] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, "Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals," in *2011 19th European Signal Processing Conference*. IEEE, 2011, pp. 1347–1351.

[12] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for dnn-based direction of arrival estimation," *arXiv preprint arXiv:1910.04388*, 2019.

[13] 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.

[14] S. Niu, J. Du, Q. Wang, L. Chai, H. Wu, Z. Nian, L. Sun, Y. Fang, J. Pan, and C.-H. Lee, "An experimental study on sound event localization and detection under realistic testing conditions," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] H. Nam, S.-H. Kim, and Y.-H. Park, "Filteraugment: An acoustic environmental data augmentation method," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4308–4312.

[16] T. N. T. Nguyen, D. L. Jones, K. N. Watcharasupat, H. Phan, and W.-S. Gan, "Salsa-lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 716–720.

[17] T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, and W.-S. Gan, "Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1749–1762, 2022.

[18] 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.

[19] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[20] J. W. Yeow, E.-L. Tan, J. Bai, S. Peksi, and W.-S. Gan, "Squeeze-and-excite resnet-conformers for sound event localization, detection, and distance estimation for dcase2024 challenge," DCASE2024 Challenge, Tech. Rep., 2024.

[21] 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.

[22] D. A. Krause, A. Politis, and A. Mesaros, "Sound event detection and localization with distance estimation," *arXiv preprint arXiv:2403.11827*, 2024.

[23] K. Shimada, A. Politis, P. Sudarsanam, D. A. Krause, K. Uchida, S. Adavanne, A. Hakala, Y. Koyama, N. Takahashi, S. Takahashi, *et al.*, "Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] I. R. Roman, C. Ick, S. Ding, A. S. Roman, B. McFee, and J. P. Bello, "Spatial scaper: A library to simulate and augment soundscapes for sound event localization and detection in realistic rooms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, South Korea, 2024.

[25] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.