A DATA GENERATION METHOD FOR SOUND EVENT LOCALIZATION AND DETECTION IN REAL SPATIAL SOUND SCENES

Technical Report

Jinbo Hu[1,2]*, Yin Cao*[3]*, Ming Wu*[1]*, Feiran Yang*[1]*, Wenwu Wang*[4]*, Mark D. Plumbley*[4]*, Jun Yang*[1,2]

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics,
Chinese Academy of Sciences, Beijing, China, {hujinbo, mingwu, feiran, jyang}@mail.ioa.ac.cn
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Department of Intelligent Science, Xi'an Jiaotong Liverpool University, China, yin.k.cao@gmail.com
[4]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

This technical report describes our submission systems for Task 3 of the DCASE 2023 Challenge: Sound Event Localization and Detection (SELD) Evaluated in Real Spatial Sound Scenes. Our proposed solution includes data synthesis, data augmentation, and track-wise model training. We focus on data generation and synthesize multi-channel spatial recordings by convolving monophonic sound event examples with multi-channel spatial room impulse responses (SRIRs) to overcome the problem of lacking real-scene recordings. The sound event samples are sourced from FSD50K and AudioSet. On the other hand, the SRIRs are extracted from the TAU Spatial Room Impulse Response Database (TAU-SRIR DB) dataset and computationally generated using the image source method (ISM). Furthermore, we utilize our previously proposed data augmentation chains, which randomly combine several data augmentation operations. Finally, based on the manually synthesized and augmented data, we employ the Event-Independent Network V2 (EINV2) with a track-wise output format to detect and localize up to three different sound events. These different sound events can be of the same type from different locations. Our proposed solution significantly outperforms the baseline method on the *dev-test* set of the Sony-TAU Realistic Spatial Soundscapes 2023 (STARSSS23) dataset.

Index Terms— Sound event localization and detection, data generation, Event-Independent Network, data augmentation chains

1. INTRODUCTION

Sound event localization and detection (SELD) aims to detect categories, presence, and spatial locations of different sound sources. SELD characterizes sound sources in a spatial-temporal manner. It can be used in various fields, such as robot auditory systems, intelligent home surveillance, assistive technologies, environmental monitoring, and automotive systems.

In the first three iterations (2019-2021) of Task 3 of the Detection and Classification of Acoustics Scenes and Events (DCASE) Challenge, the datasets of spatial sound events were computationally simulated, and these recordings were generated by convolving sound event examples with collected real-scene spatial room impulse responses (SRIRs) [1–3]. In 2022, the datasets of the challenge were transformed into real spatial sound scene recordings. The Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) dataset was manually annotated and released to serve for DCASE2022 Task 3 [4]. This year the task remains similar to the last iteration, evaluated on manually annotated recordings of real sound scenes from the STARSS23 dataset[1] used in DCASE 2023. Compared to STARSS22, this version further includes simultaneous 360° video recordings for all the audio recordings and additional source distance information in labels.

In this report, we still pay attention to the audio-only track, continuing the SELD task setup of the previous year, where only multi-channel audio signals are used while both training and evaluating. Due to expensive manual annotations, the development set of STARSS23 is limited, compared with the synthetic datasets used in the first three iterations of DCASE Challenge Task 3. External datasets are allowed to improve model performance, so we concentrate on data generation methods to mitigate the distribution difference between the training set and test set. We synthesize multi-channel spatial recordings by convolving single-channel sound event examples with multi-channel SRIRs. The sound event examples are sampled from FSD50K [5] and AudioSet [6], and then cleaned to drop terrible data by PANNs [7]. On the other hand, the SRIRs are sourced from the TAU Spatial Room Impulse Response Database (TAU-SRIR DB)[2] dataset and computationally generated using the image source method (ISM) [8]. Additionally, we exploit our previously proposed Event-Independent Network V2 (EINV2) with data augmentation chains [9, 10]. Data augmentation chains are combined by some augmentation operations, which are randomly selected and linked in chain. EINV2 contains several event-independent tracks, which means the prediction on each track can be of any event type. EINV2 is adopted in our systems to make it suitable for handling simultaneous sound events of the same type from different positions.

2. THE METHOD

2.1. Event-Independent Network V2

Event-Independent Network V2 (EINV2) [11] consists of two branches, sound event detection and direction-of-arrival (DoA) es-

[1]https://zenodo.org/record/7880637
[2]https://zenodo.org/record/6408611

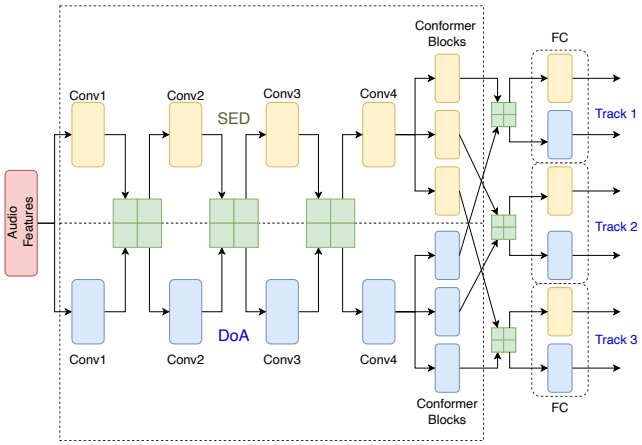

Figure 1: The architecture of the SELD network, which is a Conv-Conformer network. The upper half (yellow boxes) is the SED task. The lower half (blue boxes) is the DoA estimation task. The green boxes sandwiched between SED branch and DoA branch indicate soft connections between SED and DoA estimation.

timation. Both two branches have a Conv-Conformer architecture, and they are connected by a soft parameter-sharing strategy in multi-task learning. There are several event-independent tracks in each of the branches, which yields several track pairs. Each track pair can only predict a sound event with the corresponding DoA. The track-wise output format should utilize permutation-invariant training to tackle misaligned track problems between ground truth and predictions of sound events.

The number of tracks must be pre-defined according to the maximum overlap. While higher numbers of overlapping events (up to 6) can occur but are rare, occurrences of up to 3 simultaneous events are fairly common. As a result, three tracks are adopted to address up to three overlapped sound events. Our proposed network is shown in Fig. 1

2.2. Data Augmentation Chains

We randomly sample $k = 3$ augmentation chains, each of which is a random combination of augmentation operations. Augmentation operations contain Mixup [12], Random Crop [13], SpecAugment [14], and frequency shifting [15].

Mixup trains a neural network on convex combinations of pairs of spectrogram and their labels. We use Mixup on both raw waveforms and spectrograms to improve the performance of detecting overlapped sound events. Random Crop produces several rectangular masks on the spectrograms, while SpecAugment produces time and frequency stripes to mask the spectrograms. Frequency shifting in the frequency domain is similar to pitch shifting in the time domain, and it randomly shifts input spectrograms of all the channels up or down along the frequency axis by several bands. Rotation of First Order Ambisonics (FOA) signals is an additional augmentation method [16], excluded by augmentation chains. It rotates FOA format signals and enriches DoA labels without losing physical relationships between steering vectors and observers.

2.3. Data Generation

Development set STARSS23, which contains roughly 7.5 hours of recordings, has less data compared with the development set in DCASE 2021, which contains roughly 13 hours of synthetic recordings. [3]. Considering the complexity of the real-scene environment, we use additional datasets to improve model performance. As there is very little publicly accessible FOA data recorded in line with the official microphone setup, the data generation method plays a crucial role in model training. We generated simulated data using the generator code provided by DCASE[3].

Samples of sound events are selected from AudioSet [6] and FSD50K [5], based on the affinity of the labels in those datasets to target classes in STARSS23. PANNs [7] are then exploited to clean sound event examples. We use pre-trained PANNs to infer these examples and select high-quality examples based on output probability.

We use both extracted SRIRs from TAU-SRIR DB and computationally generated SRIRs. TAU-SRIR DB contains SRIRs captured in various spaces at Tampere University. It was used for synthetic datasets in DCASE 2019-2021. The computational generation method consists of two steps, RIRs simulation and Ambisonics format converter.

The RIRs simulation is based on the image source method [8]. This method replaces reflection on walls with virtual sources playing the same sound as the original source and builds an RIR from the corresponding delays and attenuations. As the microphones are mounted on an acoustically-hard spherical baffle in the official setup, the frequency response of the h-th microphone with a wave number of k on a rigid baffle of radius R for l-th image source is obtained as:

$$H_{hl}\left(k, \psi_{hl}\right) = \sum_{n=0}^{\infty} \mathrm{i}^{n}(2n+1) b_n(kR) P_n\left(\cos \psi_{hl}\right) \quad (1)$$

where ψ_{hl} denotes the angle between the DoA of the l-th sound source and the orientation of the h-th microphone, P_n denotes the Legendre polynomial [17]. The b_n is the mode strength term for a rigid baffle array written as

$$b_n(kR) = \frac{\mathrm{i}}{(kR)^2 h_n^{(1)'}(kR)} \quad (2)$$

where $h_n^{(1)'}$ denotes the derivate of the n-th-order spherical Hankel function of the first kind [17].

Ambisonics format convertion transforms the above mentioned MIC format signals to FOA format signals. The spherical harmonic representation of the RIRs can be computed by using the following encoding process [17, 18]:

$$\mathbf{a}(k) = \mathbf{B}(k)^{-1} \mathbf{Y}^{\dagger} \mathbf{x}(k) \quad (3)$$

with

$$\mathbf{B}(k) = \begin{pmatrix} b_0 & 0 & 0 & 0 \\ 0 & b_1 & 0 & 0 \\ 0 & 0 & b_1 & 0 \\ 0 & 0 & 0 & b_1 \end{pmatrix} \quad (4)$$

where $\mathbf{x}(k)$ denotes MIC format signal, $\mathbf{Y} \in \mathbb{C}^{M \times (N+1)^2}$ denotes spherical harmony matrics, N is the order of spherical harmony,

[3] https://github.com/danielkrause/DCASE2022-data-generator

e.g., $N = 1$ corresponds to FOA, M is the number of microphones, $(\cdot)^\dagger$ represents the Moore-Penrose pseudo inverse, and $\mathbf{a}(k)$ denotes FOA format signal.

Table 1: The SELD metrics of our proposed methods on *dev-test* set of STARSS23. The *dev-train* set of STARSS23 is mixed into synthetic training set by default. Only FOA format signals are used.

	Datasets	ER_{20°	F_{20°	$\mathrm{LE}_{\mathrm{CD}}$	$\mathrm{LR}_{\mathrm{CD}}$
Baseline	Official	0.57	29.9 %	22°	47.7 %
System #1	A	0.481	47.3 %	16.1°	62.6 %
System #2	B	0.474	48.0 %	16.2°	63.7 %
System #3	B+C	0.475	48.7 %	14.9°	63.0 %

3. EXPERIMENTS AND RESULTS

We evaluate our proposed method on the *dev-test* set of STARSS23. We generate a large amount of data using the abovementioned data generation method, including 2700 1-minute clips from TAU-SRIR DB (dataset A and B) and 50000 5-second clips (dataset C) from computationally generated SRIRs, where the sound event examples of B and C are cleaned by PANNs. Table 1 shows the experimental results on different simulated datasets A, B, and C. The official dataset means the synthetic mixtures for baseline training. As shown in the table, according to System #1 and System #2, the data-cleaning strategy has a positive effect on model training. By comparing System #2 with System #3, computationally generated SRIRs exactly enrich DoA labels and achieve further performance.

4. CONCLUSION

In this report, based on our previous EINV2 and data augmentation chains, we propose a data generation method. The spatial sound events are simulated by convolving cleaned sound events samples from FSD50K and AudioSet using PANNs with SRIRs from TAU-SRIR DB and computational generation by ISM. Our proposed method is evaluated in the *dev-test* set of STARSS23 and outperforms the baseline systems significantly. The experiments also show the effectiveness of the data generation method.

5. REFERENCES

[1] S. Adavanne, A. Politis, and T. Virtanen, "A multi-room reverberant dataset for sound event localization and detection," in *Proc. DCASE 2019 Workshop*, 2019, pp. 10–14.

[2] A. Politis, S. Adavanne, and T. Virtanen, "A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection," in *Proc. DCASE 2020 Workshop*, 2020, pp. 165–169.

[3] A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, and T. Virtanen, "A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection," in *Proc. DCASE 2021 Workshop*, 2021, pp. 125–129.

[4] A. Politis, K. Shimada, P. Sudarsanam, S. Adavanne, D. Krause, Y. Koyama, N. Takahashi, S. Takahashi, Y. Mitsufuji, and T. Virtanen, "STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events," in *Proc. DCASE 2022 Workshop*, 2022, pp. 161–165.

[5] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 30, pp. 829–852, 2021.

[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 *Proc. IEEE ICASSP 2017*, 2017, pp. 776–780.

[7] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[8] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[9] J. Hu, Y. Cao, M. Wu, Q. Kong, F. Yang, M. D. Plumbley, and J. Yang, "A track-wise ensemble event independent network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP 2022*, 2022, pp. 9196–9200.

[10] ——, "Sound event localization and detection for real spatial sound scenes: Event-independent network and data augmentation chains," in *Proc. DCASE 2022 Workshop*, 2022, pp. 46–50.

[11] Y. Cao, T. Iqbal, Q. Kong, F. An, W. Wang, and M. D. Plumbley, "An improved event-independent network for polyphonic sound event localization and detection," in *Proc. IEEE ICASSP 2021*, 2021, pp. 885–889.

[12] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. ICLR 2018*, 2018.

[13] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. of AAAI 2020*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[14] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613 – 2617.

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

[16] L. Mazzon, Y. Koizumi, M. Yasuda, and N. Harada, "First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation," in *Proc. DCASE 2019 Workshop*, 2019, pp. 154–158.

[17] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015.

[18] Y. Koyama, K. Shigemi, M. Takahashi, K. Shimada, N. Takahashi, E. Tsunoo, S. Takahashi, and Y. Mitsufuji, "Spatial data augmentation with simulated room impulse responses for sound event localization and detection," in *Proc. IEEE ICASSP 2022*, 2022, pp. 8872–8876.