SOUND EVENT LOCALIZATION AND DETECTION FOR REAL SPATIAL SOUND SCENES: EVENT-INDEPENDENT NETWORK AND DATA AUGMENTATION CHAINS

Technical Report

Jinbo Hu[1,2]*, Yin Cao*[3]*, Ming Wu*[1]*, Qiuqiang Kong*[4]*, Feiran Yang*[1]*, Mark D. Plumbley*[5]*, 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]Xi'an Jiaotong Liverpool University, Suzhou, China, yin.cao@xjtlu.edu.cn
[4]ByteDance Shanghai, China, kongqiuqiang@bytedance.com
[5]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
m.plumbley@surrey.ac.uk

ABSTRACT

Polyphonic sound event localization and detection (SELD) aims at detecting types of sound events with corresponding temporal activities and spatial locations. In DCASE 2022 Task 3, data types transition from computationally generated spatial recordings to recordings of real-sound scenes. Our system submitted to the DCASE 2022 Task 3 is based on our previous proposed Event-Independent eNetwork V2 (EINV2) and novel data augmentation method. To detect different sound events of the same type with different locations, our method employs EINV2, combining a track-wise output format, permutation-invariant training, and soft-parameter sharing. EINV2 is also extended using conformer structures to learn local and global patterns. To improve the generalization ability of the model, we use a data augmentation approach containing several data augmentation chains, which are composed of random combinations of several different data augmentation operations. To mitigate the lack of the real-scene recordings in the development dataset and the presence of sound events being unbalanced, we exploit FSD50K, AudioSet, and TAU Spatial Room Impulse Response Database (TAU-SRIR DB) to generate simulated datasets for training. The results show that our system is improved over the baseline system on the *dev-set-test* of Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22).

***Index Terms*—** Sound event localization and detection, real spatial sound scenes, Event-Independent Network, data augmentation chains, simulated datatasets

1. INTRODUCTION

Sound event localization and detection (SELD) contains two subtasks, sound event detection (SED) and direction-of-arrival (DoA) estimation. SED aims at detecting types of sound and their corresponding temporal activities. Whereas DoA estimation predicts spatial trajectories of different sound sources. SELD characterizes sound sources in a spatial-temporal manner that can be used in a wide range of applications, such as robot auditory, surveillance, and smart home.

SELD has received broad attention recently. Adavanne et al. [1] proposed a polyphonic SELD work using an end-to-end network, SELDnet, which was utilized for a joint task of SED and regression-based DoA estimation. SELD was then introduced in the Task 3 of the Detection and Classification of Acoustics Scenes and Events (DCASE) 2019 Challenge for the first time, which uses the TAU Spatial Sound Events 2019 dataset [2]. Most datasets of spatial sound events are computationally simulated, and these recordings are generated by convolving randomly chosen sound event examples with a corresponding random real-life spatial room impulse response (SRIR) to spatially place them at a given position [2–4]. Moreover, stronger reverberation, diversity of environment, dynamic scenes with both moving and static sound sources, ambient noise, sound events with the same type, and unknown directional interfering events out of the target classes were added into datasets to complicate the SELD task and brought each iteration of Task 3 of DCASE Challenge closer to real conditions. In 2022, the challenge task transitions from computationally simulated spatial recordings to real spatial sound scenes recordings. Sony-TAu Realistic Spatial Soundscapes 2022 (STARSS22) dataset is released to serve as the development and evaluation dataset of DCASE2022 Task 3 this year, which are manually annotated [5].

SELDnet has the limitation that it is unable to detect sound events of the same type but with different locations [1]. Event independent network (EIN) with track-wise output format was proposed to tackle this problem [6–8]. In EIN, there are several event-independent tracks, which means the prediction on each track can be of any event type. The number of tracks needs to be pre-determined according to the maximum number of overlapping events. EINV2 utilizes multi-head self-attention (MHSA) and soft parameter-sharing to achieve better performance compared with SELDnet [7].

In practical applications, training set cannot cover all actual instances from different spatial and sound environments, and mismatches between the training set and test set are common. To improve the generalization of the model, a novel data-augmentation method is used [8, 9]. The data-augmentation method is characterized by utilizing several augmentation operations. These data augmentation operations are sampled, layered, and combined randomly to produce a high diversity of augmented features.

In this study, our model exploits EINV2, combining a track-wise-output format, permutation-invariant training (PIT), and soft parameter-sharing (PS). The Conformer structure is utilized to ex-

tend EINV2 to learn local and global patterns. The data augmentation method is composed of several augmentation operations. These data augmentation operations are sampled and layered randomly to be combined to several data augmentation chains [8]. External data is allowed in this challenge. We generate simulated data by convolving stochastically chosen samples of sound event from AudioSet [10] and FSD50K [11] with measured SRIRs from TAU Spatial Room Impulse Responses Database[1] (TAU-SRIR DB). The experimental results show the proposed model with the novel data augmentation methods, which was trained in our simulated data, outperformed the DCASE2022 challenge Task 3 baseline model which was trained in official synthetic SELD mixtures[2].

2. THE METHOD

2.1. Input features

In this work, log-mel spectrograms are first used for SED, while IV in log-mel space is used for DoA estimation [6, 8, 12, 13]. FOA includes four channels of signals, i.e., omni-directional channel $\mathbf{w}$, and three directional channels $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. Log-mel spectrograms are computed from the short-time Fourier transform spectrograms of four-channel signals, and intensity vectors are computed from cross-correlation of $\mathbf{w}$ with $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ in log-mel space. These features are directly calculated online using a 1-D convolutional layer, which supports data augmentation on raw waveform.

2.2. Network Architecture

The trackwise output format was introduced in our previous works [6–8]. It can be defined as

$$\boldsymbol{Y}_{\text{Trackwise}} = \left\{ (y_{\text{SED}}, y_{\text{DoA}}) \mid y_{\text{SED}} \in \mathbb{O}_{\mathbf{S}}^{M \times K}, y_{\text{DoA}} \in \mathbb{R}^{M \times 3} \right\} \quad (1)$$

where M is the number of tracks, K is the number of sound-event types, $\mathbb{O}_{\mathbf{S}}^{M \times K}$ is one hot encoding of K classes, $\mathbf{S}$ is the set of sound events, and the number of dimensions of Cartesian coordinates is 3.

The number of tracks needs to be pre-determined according to the maximum number of overlapping events. Each track can only detect a sound event and a corresponding location. While a model with track-wise output format is trained, sound events are not always predicted in a fixed track. It may result in a problem that sound events predicted in a track may not be aligned to its ground truth. This may be due to the track permutation problem. Permutation-invariant training (PIT) can be utilized for the problem. The PIT loss is defined as

$$\mathscr{L}_{PIT}(t) = \min_{\alpha \in \mathbf{P}(t)} \sum_{M} \{ \lambda \cdot \ell_{\text{SED}}(t, \alpha) + (1 - \lambda) \cdot \ell_{\text{DoA}}(t, \alpha) \} \quad (2)$$

where $\alpha \in \mathbf{P}(\mathbf{t})$ indicates one of the possible permutations and λ is a weight between SED loss and DoA loss. ℓ_{SED} is binary cross entropy loss for the SED task, and ℓ_{DoA} is mean square error for the DoA task. The lowest loss will be chosen by finding a possible permutation, and the back-propagation is then performed.

[1] https://doi.org/10.5281/zenodo.6408611
[2] https://doi.org/10.5281/zenodo.6406873

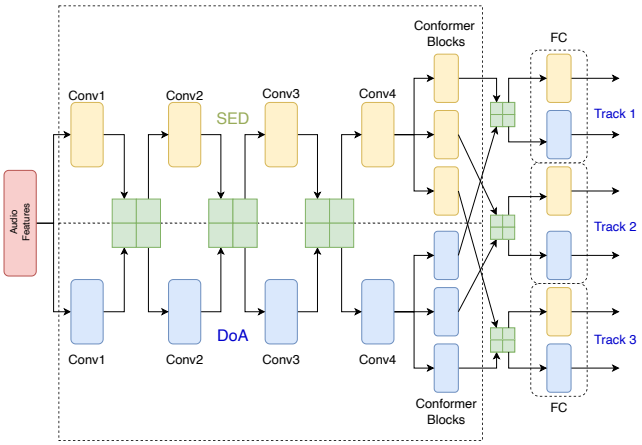

Figure 1: The architecture of the SELD network, which is a Conv-Conformer network. Dashed-yellow is the SED task. Dashed-blue is the DoA estimation task. The green boxes indicate soft connections between SED and DoA estimation.

From an multi-task learning (MTL) perspective, joint SELD learning can be mutually beneficial. Hard PS and soft PS are two typical methods to implement MTL. Hard PS means subtasks use the same feature layers, while soft PS means subtasks use their own feature layers with connections existing among those feature layers. In [7], experimental results show that soft PS using cross-stitch is more effective.

EINV2, which combines the track-wise output format, PIT, and soft PS, is utilized for our system. We extend EINV2 to three tracks to address up to three overlapped sound events. We then utilize Conformer blocks to replace the multi-head self-attention (MHSA) blocks in EINV2. Conformer consists of two feed-forward layers with residual connections sandwiching the MHSA and convolution modules, where MHSA and convolution modules can capture global and local patterns, respectively [8, 14]. Our proposed network is shown in Fig. 1.

2.3. Data Augmentation Chains

Our proposed data-augmentation is characterized by utilizing several augmentation operations [8, 9, 15]. We randomly sample k augmentation chains, where $k = 3$ is used by default. Each augmentation chain is constructed by composing from some randomly selected augmentation operations. Augmentation operations that we used include Mixup [16], SpecAugment [17], Cutout, frequency shifting [18] and rotation of FOA signals [19]. The diagram of data augmentation chains is shown in Fig. 2

Mixup trains a neural network on convex combinations of pairs of feature vectors and their labels. We use Mixup on both raw waveforms and features to improve the generalization for detecting overlapping sound events. While random Cutout produces several rectangular masks on features, SpecAugment produces time and frequency stripes to mask on features. Frequency shifting in the frequency domain is similar to pitch shift in the time domain, and it randomly shifts input features of all the channels up or down along the frequency dimension by several bands. We also use a spatial augmentation method, which is rotation of FOA signals. It rotates FOA format signals and enriches DoA labels without losing physical relationships between steering vectors and observer. We use z axis as the rotation axis, which leads to 16 types of channel rotation.

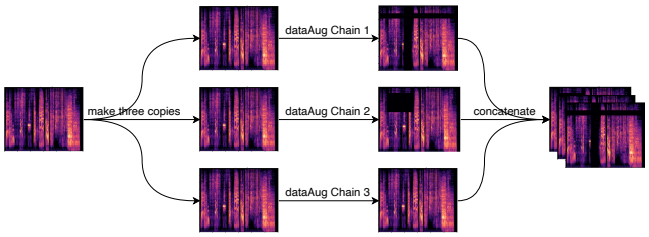

Figure 2: Diagram of data augmentation chains

2.4. Simulated Data

Since manual annotations are expensive and the duration of STARSS22 is limited compared to the synthetic datasets used in previous years, external dataset is allowed to improve model performance. We generated simulated data using generator code provided by DCASE 2022[3].

Samples of sound event chosen are mainly sources from FSD50K dataset, based on affinity of the labels in that datasets to the target classes. The target class *background music*, and interference classes *shuffling cards* are not in FSD50K dataset, hence we use AudioSet as a supplement. Spatial events spatialized in 9 unique rooms, using collected SRIRs from TAU-SRIR DB. The ambient noise from the same room was additional mixed at varying signal-to-noise rations (SNR) from 30 dB to 6 dB. The maximum polyphony of target classes is 3, excluding additional polyphony of interference classes.

We select all sound event samples whose labels are only corresponded to the target classes. Since samples of each class are much unbalanced, sound event samples of each class are randomly stratified sampled. Each sound event sample also has a different energy gain for mixing. By setting different ranges of gain and choosing different samples, we generate three dataset, A, B and C. All of these synthetic datasets have 2700 1-minute clips.

3. EXPERIMENTS

3.1. Datasets

The STARSS22 contains recordings of real scenes, and the density of sound event samples and the presence of each class varies greatly. The maximum number of overlap is 5, but those samples are very rare [5]. Occurences of up to 3 simultaneous events are fairly common, so we ignore overlapping events with polyphony degree of more than 3 that occur. During development stage, we train our proposed model on mixed dataset of synthetic recordings and *dev-set-train* of STARSS22, and evaluate those systems using *dev-set-test* of STARSS22. During evaluation stage, synthetic recordings, *dev-set-train* and *dev-set-test* of STARSS22 are all used for training.

3.2. Hyper-parameters

The sampling frequency of the dataset is 24 kHz. We used a 1024-point Hanning window with a hop size of 400 and 128 mel bins for log-mel spectrograms and IV features. Audio clips are segmented to have a fix length of 5 seconds with no overlap for training and inferring. AdamW optimizer is used. The learning rate is set to 0.0003 for the first 70 epochs and is adjusted to 0.00003 for the following 20 epochs. The threshold for SED is set to 0.5 to binarize predictions. The loss weight between SED and DoA is 0.5.

3.3. Evaluation Metrics

We use official evaluation metrics to evaluate the SELD performance [20, 21]. The evaluation metrics uses a joint metric of localization and detection: location-sensitive detection metrics $\mathrm{F}_{\leq T^\circ}$ and $\mathrm{ER}_{\leq T^\circ}$, and class-sensitive localization metrics $\mathrm{LR}_{\mathrm{CD}}$ and $\mathrm{LE}_{\mathrm{CD}}$. $\mathrm{F}_{\leq T^\circ}$ and $\mathrm{ER}_{\leq T^\circ}$ consider true positives predicted under a spatial threshold T° from the ground truth. As for $\mathrm{LE}_{\mathrm{CD}}$ and $\mathrm{LR}_{\mathrm{CD}}$, the detected sound class has to be correct in order to count the corresponding localization predictions.

Contrary to the previous challenges, the evaluation metrics are micro-averaged, which gives equal weight to each individual decision and affected by the performance on the larger classes. In this challenge, macro-averaging of evaluation metrics are used. Macro-averaging gives equal weight to each class, and emphasize the system behavior on the smaller classes [22].

We use an aggregated SELD metric which was computed as

$$\epsilon_{\mathrm{SELD}} = \frac{1}{4}\left[\mathrm{ER}_{\leq T^\circ} + (1 - \mathrm{F}_{\leq T^\circ}) + \frac{\mathrm{LE}_{\mathrm{CD}}}{180^\circ} + (1 - \mathrm{LR}_{\mathrm{CD}})\right] \tag{3}$$

A good SELD system should have low $\mathrm{ER}_{\leq T^\circ}$, high $\mathrm{F}_{\leq T^\circ}$, low $\mathrm{LE}_{\mathrm{CD}}$, high $\mathrm{LR}_{\mathrm{CD}}$, and low aggregated SELD metric ϵ_{SELD}.

3.4. Experimental Results

The official spatial threshold is set to $T = 20^\circ$. Table 1 shows the performance on *dev-set-test* of STARSS22. Official dataset means official synthetic SELD mixtures for baseline training[4]. System baseline, EINV2 without dataAug chains, and EINV2 with dataAug chains all use the same dataset for training. EINV2 without data augmentation chains outperforms the baseline model, whereas EINV2 with data augmentation chains performs better.

All configurations of system #1 - #4 are the same as system EINV2 with dataAug chains, except for training set. The results also demonstrate the effectiveness of our simulated data over the official dataset, but there are not significant improvement of metric scores among different datasets.

4. CONCLUSION

We have presented Event-Independent Network V2 (EINV2) with a novel data augmentation approach for real-life sound event localization and detection. EINV2 is extended by conformer blocks, and the novel data augmentation approach contains several augmentation chains. Each augmentation chain contains several randomly sampled augmentation operations. In addition, external data is permitted in this challenge, hence samples of sound event from FSD50K and AudioSet are convolved with measured spatial room impulse responses from TAU Spatial Room Impulse Responses Database (TAU-SRIR DB) to generate simulated data. Our model with data augmentation chains performs better than the baseline model. Furthermore, experimental results show further improvement with our synthetic dataset.

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

[4] https://zenodo.org/record/6406873

Table 1: The SELD performance on the *dev-set-test* of STARSS22. The *dev-set-train* of STARSS22 is mixed into training set by default.

System	Datasets	Macro-average					Micro-average				
		ER_{20°	F_{20°	LE_{CD}	LR_{CD}	ϵ_{SELD}	ER_{20°	F_{20°	LE_{CD}	LR_{CD}	ϵ_{SELD}
Baseline FOA [5]	Official	0.71	0.21	29.3°	0.46	0.55	0.71	0.36	-	-	-
EINV2 w/o dataAug chains	Official	0.75	0.32	24.0°	0.56	0.50	0.75	0.36	25.6°	0.62	0.48
EINV2 w/ dataAug chains	Official	0.56	0.42	19.3°	0.61	0.41	0.56	0.53	19.1°	0.71	0.36
System #1	A+B+C	0.50	0.48	19.5°	0.66	0.37	0.50	0.57	18.7°	0.72	0.33
System #2	A+B	0.50	0.51	16.4°	0.66	0.36	0.50	0.59	17.2°	0.72	0.32
System #3	A	0.53	0.48	17.8°	0.63	0.38	0.53	0.55	18.2°	0.69	0.35
System #4	B	0.53	0.45	17.4°	0.63	0.39	0.53	0.57	17.5°	0.69	0.34

5. 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 J Sel Top Signal Process*, vol. 13, pp. 34–48, 2018.

[2] 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.

[3] 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.

[4] 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.

[5] 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 *arXiv:2206.01948*, 2022.

[6] Y. Cao, T. Iqbal, Q. Kong, Y. Zhong, W. Wang, and M. D. Plumbley, "Event-independent network for polyphonic sound event localization and detection," in *Proc. DCASE 2020 Workshop*, 2020, pp. 11–15.

[7] 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.

[8] 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.

[9] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in *Proc. ICLR 2020*, 2020.

[10] 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.

[11] 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.

[12] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. D. Plumbley, "Polyphonic sound event detection and localization using a two-stage strategy," in *Proc. DCASE 2019 Workshop*, 2019, pp. 30–34.

[13] Q. Wang, J. Du, H. Wu, J. Pan, F. Ma, and C. Lee, "A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection," *arXiv preprint arXiv:2101.02919*, 2021.

[14] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036 – 5040.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML 2020*, 2020, pp. 1597–1607.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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 Trans. on Audio, Speech, and Lang. Process.*, vol. 29, pp. 684–698, 2020.

[21] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, "Joint measurement of localization and detection of sound events," in *Proc. IEEE WASPAA 2019*, 2019, pp. 333–337.

[22] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.