A TRACK-WISE ENSEMBLE EVENT INDEPENDENT NETWORK FOR POLYPHONIC SOUND EVENT LOCALIZATION AND DETECTION

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

ABSTRACT

Polyphonic sound event localization and detection (SELD) aims at detecting types of sound events with corresponding temporal activities and spatial locations. In this paper, a track-wise ensemble event independent network with a novel data augmentation method is proposed. The proposed model is based on our previous proposed Event-Independent Network V2 and is extended by conformer blocks and dense blocks. The track-wise ensemble model with track-wise output format is proposed to solve an ensemble model problem for track-wise output format that track permutation may occur among different models. The data augmentation approach contains several data augmentation chains, which are composed of random combinations of several data augmentation operations. The method also utilizes log-mel spectrograms, intensity vectors, and Spatial Cues-Augmented Log-Spectrogram (SALSA) for different models. We evaluate our proposed method in the Task of the L3DAS22 challenge and obtain the top ranking solution with a location-dependent F-score to be 0.699. Source code is released[1].

Index Terms— Sound event localization and detection, event-independent network, track-wise ensemble model, data augmentation chains

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. 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 [1]. 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 [1, 2]. The Learning 3D Audio Sources (L3DAS) project held the L3DAS21 and L3DAS22 challenges in 2021. The main novelty is to use two Ambisonics microphone arrays.

SELDnet employed multi-channel magnitude and phase spectrograms as input features [1]. Subsequently, multi-channel log-mel spectrograms, intensity vectors (IV) in log-mel space for first-order Ambisonics (FOA) format signals, and generalized cross-correlation with phase transform (GCC-PHAT) for microphone array (MIC) signals were demonstrated to be more effective features for SELD [3–7]. Nguyen et al. proposed a novel feature called Spatial Cue-Augmented Log-Spectrogram (SALSA), which was composed of multi-channel log spectrograms stacked with normalized principal eigenvectors of a spatial covariance matrix. [8, 9].

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 [4, 10]. 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) to achieve better performance compared with SELDnet [10]. Other network structures, such as DenseNet [11] and Conformer [12], can also be employed. Our proposed model uses DenseNet and Conformer to extend EINV2.

To further improve the performance of trained models, post-processing methods can be utilized during inference. A spatial augmentation technique is utilized for test-time augmen-

[1]https://github.com/Jinbo-Hu/L3DAS22-TASK2

tation (TTA) [8, 9]. Test samples are augmented by 16-pattern rotations [13]. Output is then computed by a mean of all 16 outputs. Average ensemble and weighted ensemble compute a mean or weighted mean output of several different trained models [8, 14]. Stacking is also an ensemble method. It is to train inhomogeneous models using the original dataset at first, and then train an ensemble model using predictions of these inhomogeneous models [3, 15].

In this paper, we propose an ensemble Event Independent Network based on previously proposed EINV2 and a novel data augmentation approach. The method utilizes log-mel, IV, and SALSA featrues for different models. Our proposed model exploits EINV2, combining a track-wise-output format, permutation-invariant trainint (PIT), and soft parameter-sharing (PS). The Conformer structure is employed to learn local and global patterns. The DenseNet structure is utilized to increase the diversity of different models for ensemble. The proposed 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. We propose a track-wise ensemble method for the track-wise output format to improve the system performances. The proposed system obtained a Top ranking in the Task 2 of the L3DAS22 challenge [16].

2. THE METHOD

2.1. Input Features

The dataset provided by the L3DAS22 Task 2 uses two FOA microphone arrays that is placed in the center of a room. In this paper, two types of features are used for ensemble models. Log-mel spectrograms and IV in log-mel space from first-order Ambisonics (FOA) are used as the first set of features. SALSA is used as the second set of features. We extract features for both FOA and concatenate them as input features.

Log-mel spectrograms are first used for SED, while IV in log-mel space is used for DoA estimation [3–5]. 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.

SALSA is composed of two major components: multi-channel log-linear spectrograms and normalized principal eigenvectors. The detailed information can be found in [8, 9].

2.2. Network Architecture

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.

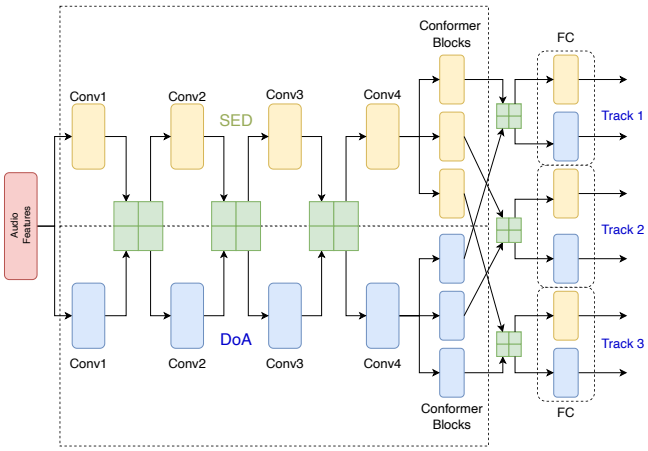

Fig. 1: The architecture of the SELD network, which is a Conv-Conformer network. Dashed-yellow is the SED task. Dashed-blue is the DoA estimation task. The green boxes indicate soft connections between SED and DoA estimation. The convolution blocks can be extended to dense blocks. The network can be adapted to either log-mel spectrograms and IV input features, or SALSA input features.

We then utilize Conformer 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. To increase the diversity of model ensembles, we replace convolution blocks with dense blocks. In dense blocks, each layer is connected with all preceding layers to obtain additional inputs, and delivers its own feature maps to all subsequent layers. This feed-forward structure can strengthen forward propagation of features and back propagation of gradients. Our proposed network is shown in Fig. 1. The detail of the network architecture can be found in the released code.

2.3. Data Augmentation

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, we propose a novel data-augmentation method.

Our proposed data-augmentation is characterized by utilizing several augmentation operations [17, 18]. 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 [19], SpecAugment [20], Cutout, and rotation of FOA signals [13].

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. 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 x, y, and z axis as the rotation axis, respectively, which leads to 48 types of channel rotation.

3. POST PROCESSING

3.1. Track-wise Output Format

The trackwise output format was introduced in previous works [4, 10]. 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.

3.2. Track-wise Ensemble Model

For the track-wise output format, the prediction of a sound event may be allocated to a track randomly. Methods like TTA, average or weighted ensemble [8, 14] cannot align predictions from different tracks, which makes these methods inapplicable to the track-wise output format. We propose a novel post-processing method named track-wise ensemble model. The track-wise ensemble model is a trainable model shown in Fig. 2. The inputs to the ensemble model are the outputs from different single models. The ensemble model then predicts results in a manner of the track-wise output format. The

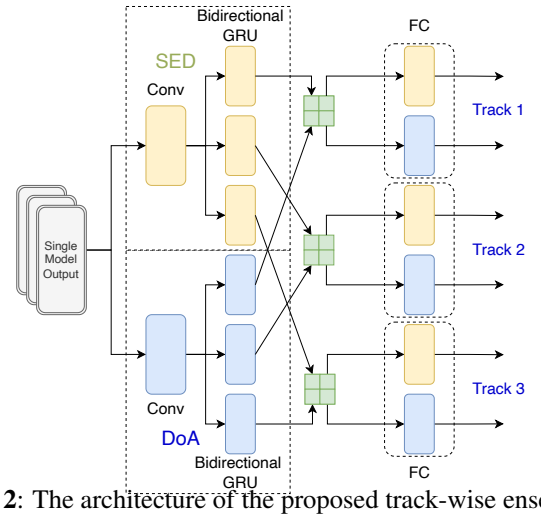

Fig. 2: The architecture of the proposed track-wise ensemble model

model structure is a simplified version of EINV2 with SED and DoA estimation branches. Each branch is consisted of a convolutional-recurrent neural network (CRNN) with soft PS connecting two branches. CRNN consists of 2D convolution layers with 128 output channels and kernel size of 3×3, and bidirectional GRU with hidden size of 64.

After training N inhomogeneous models, we get N predictions $[\mathbf{y}_{\text{SED}}^1, \mathbf{y}_{\text{SED}}^2, \ldots, \mathbf{y}_{\text{SED}}^N]$ and $[\mathbf{y}_{\text{DoA}}^1, \mathbf{y}_{\text{DoA}}^2, \ldots, \mathbf{y}_{\text{DoA}}^N]$. We concatenate and flatten predictions of SED and DoA, which are then sent as the inputs to the ensemble model. The model also outputs track-wise format predictions.

4. EXPERIMENTS

4.1. Dataset

We verify our proposed method using the dataset provided by the L3DAS22 Task 2 [16]. The dataset is split into 3 subsets, which consist of 600, 150, and 150 30-second-long audio recordings for the train, validation and test splits, respectively. There are 14 types of sound events that are selected from the FSD50K dataset. The maximum overlapping sound events are three. The room impulse response (RIR) is sampled in an office room with the dimension to be around 6 m (length) by 5 m (width) by 3 m (height). FOA microphone arrays are placed in the center of the room. The position of the FOA microphone arrays is set to be the origin of the coordinates.

4.2. Hyper-parameters and Evaluation Metrics

The sampling frequency of the dataset is 32 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, and a 512-point Hanning window with a hop size of 400 for SALSA features. Audio clips are segmented to have a fix length of 5 seconds with no overlap for training. AdamW optimizer is used. The learning rate is set to 0.0003 for the first 90 epochs

Table 1: The performance of our proposed model on the validation set

System	Models	Features	F score$_{\leq 1m}$		F score$_{\leq 2m}$	
			Single FOA	Double FOAs	Single FOA	Double FOAs
#1	ConvBlock-Conformer	log-mel + IV	0.667	0.677	0.695	0.700
#2	DenseBlock-Conformer	log-mel + IV	0.653	0.668	0.674	0.689
#3	ConvBlock-Conformer	SALSA	0.651	0.661	0.681	0.685

and is adjusted to 0.00003 for the following 10 epochs. The threshold for SED is set to 0.5 to binarize predictions. For two FOA microphone arrays, we extract 4 channels of log-mel spectrograms and 3 channels of IV features, and 4 channels of log-linear spectrograms and 3 channels of normalized principal eigenvectors for SALSA, respectively. Then we concatenate these features to make the input channels to be 14.

The evaluation metric uses a joint metric of localization and detection: F-score based on the location-sensitive detection. It counts true positives predicted when the label of a sound event is correct and its location is within a Cartesian distance threshold from its reference location [21].

4.3. Experimental Results

We trained the proposed model with different configurations using the training set of the L3DAS22 Task 2 dataset and evaluated the performance on the validation set. The configurations and results are shown in Table 1. Our system is developed based on the spatial error threshold to be 1 m and is evaluated at the threshold to be both 1 m and 2 m. We also show the performance of our system with one FOA array and two FOA arrays. The performance of two FOA arrays is slightly better.

Table 2 shows the performance of proposed data augmentation and ensemble model. We evaluate the performance of proposed data augmentation using the configuration of System #1 in Table 1. Several data-augmentation operations, which are directly linked in series, has a significant performance improvement compared to the method without any augmentations, but the method of augmentation chains performs much better. The average ensemble model is not stable and performs much worse than any of the single models that are shown in Table 1. This is due to the track permutation problem. On the other hand, the track-wise ensemble model, which uses PIT to solve the track permutation problem among different models, can improve the performance compared with any single model.

The final submitted system is trained on both L3DAS21 and L3DAS22 datasets. The evaluation results on the blind test set are shown in Table 3. It can be seen that the performance of the proposed method outperforms the baseline method significantly on the blind test set.

Table 2: The performance of proposed data augmentation and ensemble model on the validation set.

Methods	F score$_{\leq 1m}$	F score$_{\leq 2m}$
System #1 without dataAug	0.515	0.566
System #1 with dataAug in series	0.616	0.659
System #1 with dataAug in chains	**0.677**	**0.700**
Average Ensemble	0.499	0.607
Track-wise Ensemble	**0.685**	**0.715**

Table 3: Evaluation results for submitted systems on the blind test set. The official spatial threshold is fixed to 2 m.

System	Precision	Recall	F score
Baseline	0.423	0.289	0.343
Track-wise ensemble model	**0.706**	**0.691**	**0.699**

5. CONCLUSION

We have presented a track-wise ensemble event independent network with a novel data augmentation approach for 3D polyphonic sound event localization and detection. The proposed data augmentation method contains several augmentation chains. Each augmentation chain contains several randomly sampled augmentation operations. In addition, the proposed ensemble model is based on single models that is extended from Event-Independent Network V2. We use log-mel spectrograms, intensity vectors, as well as Spatial Cues-Augmented Log-Spectrogram as input features to these single models. The performance of these single models is also improved by using conformer blocks and dense blocks. We adopt a trainable ensemble model with the track-wise output format to tackle with the track permutation problem for different models. Experimental results show that the proposed method can achieve location-dependent F-score of 0.699, which is the top ranking in the Task 2 of the L3DAS22 challenge. The proposed track-wise ensemble method can solve the track permutation problem well and outperforms the average ensemble method by a large margin.

6. ACKNOWLEDGEMENT

This work was supported in part by Frontier Exploration project independently deployed by Institute of Acoustics, Chinese Academy of Sciences (No. QYTS202009), National Natural Science Foundation of China (Grant No. 11804365), and EPSRC Grant EP/T019751/1 “AI for Sound”.

7. 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] 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.

[4] 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.

[5] 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.

[6] L. Perotin, R. Serizel, E. Vincent, and A. Guerin, "CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings," *IEEE J Sel Top Signal Process*, pp. 22–33, 2019.

[7] M. Ricciardi Celsi, S. Scardapane, and D. Comminiello, "Quaternion neural networks for 3D sound source localization in reverberant environments," in *Proc. IEEE MLSP 2020*, 2020, pp. 1–6.

[8] T. Nguyen, K. Watcharasupat, N. K. Nguyen, D. L. Jones, and W. S. Gan, "DCASE 2021 Task 3: Spectrotemporally-aligned features for polyphonic sound event localization and detection," *arXiv preprint arXiv:2106.15190*, 2021.

[9] T. Nguyen, K. 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," *arXiv preprint arXiv:2110.00275*, 2021.

[10] 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.

[11] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE CVPR 2017*, 2017, pp. 4700–4708.

[12] 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.

[13] 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.

[14] K. Shimada, N. Takahashi, Y. Koyama, S. Takahashi, E. Tsunoo, M. Takahashi, and Y. Mitsufuji, "Ensemble of ACCDOA- and EINV2-based systems with D3Nets and impulse response simulation for sound event localization and detection," in *arXiv preprint arXiv:2106.10806*, 2021.

[15] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

[16] E. Guizzo, C. Marinoni, M. Pennese, X. Ren, X. Zheng, C. Zhang, B. Masiero, A. Uncini, and D. Comminiello, "L3DAS22 Challenge: Learning 3D audio sources in a real office environment," in *Proc. IEEE ICASSP 2022*, 2022.

[17] 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.

[18] 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.

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

[20] 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.

[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.