EVENT-INDEPENDENT NETWORK FOR POLYPHONIC SOUND EVENT LOCALIZATION AND DETECTION

Technical Report

Yin Cao,[1] *Turab Iqbal,*[1] *Qiuqiang Kong,*[2] *Yue Zhong,*[1] *Wenwu Wang,*[1] *Mark D. Plumbley*[1]

[1]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{yin.cao, t.iqbal, y.zhong, w.wang, m.plumbley}@surrey.ac.uk
[2]ByteDance Menlo Park, US
kongqiuqiang@bytedance.com

ABSTRACT

Polyphonic sound event localization and detection is to not only detect what sound events are happening but to localize corresponding sound sources. This series of tasks was firstly introduced in DCASE 2019 Task 3. This year, the sound event localization and detection task brings additional challenges in moving sources and up to two overlapping sound events, which include cases of two same type of events with two different direction-of-arrival (DoA) angles. In this report, a novel event-independent network for polyphonic sound event localization and detection is proposed. Unlike the two-stage method that was proposed by us last year [1], this new network is fully end-to-end. Inputs to the network are first-order Ambisonics (FOA) time-domain signals, which are then fed into a 1-D convolutional layer to extract logmel spectrograms and intensity vectors. The network is then split into two parallel branches. The first branch is for the sound event detection (SED), and the second branch is for the DoA estimation. There are three types of predictions from the network, which are SED predictions, event activity detection (EAD) predictions that are used to combine the SED and DOA features for the on-set and off-set estimation, and DoA predictions. All of these predictions have the format of two tracks indicating that there are at most two overlapping events. Within each track, there could be at most one event happening. This architecture brings a problem of track permutation. To address this problem, a frame-level permutation invariant training method is used. Experimental results show that the proposed method can detect polyphonic sound events and their corresponding DoAs. The performance of Task 3 dataset is greatly increased compared with the baseline method.

***Index Terms*—** Sound event localization and detection, direction of arrival, intensity vector, permutation invariant training, event-independent

1. INTRODUCTION

Sound event localization (SED) and detection become a more and more popular research topic since DCASE 2019. It detects types of sound events and localizes corresponding sound sources frame-wisely. This year, DCASE 2020 Task 3 [2–4] brings additional challenges in moving sources and polyphonic cases that includes the same class of event but with different direction-of-arrivals (DoAs).

For DCASE 2019 Task3, we introduced a two-stage method for polyphonic sound event localization and detection [1]. Although it obtained a relatively high ranking, it was not designed as an actual polyphonic localization method for the reason that it lacks the ability to detect the case of the same type of event but with different DoAs. Besides, it is not a graceful end-to-end system.

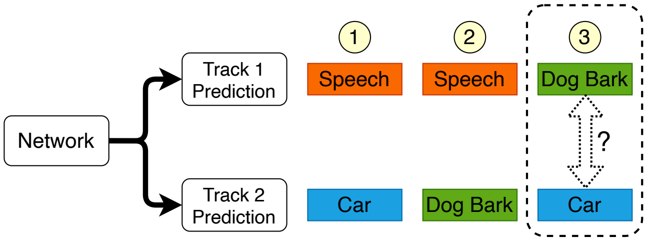

Figure 1: Illustration of track permutation problem. Numbers mean different group of labels.

In this report, we introduced a re-designed event-independent end-to-end system for polyphonic sound event localization and detection. It is designed for overlapping-event cases, especially for the case of the same type of event with different DoAs. It is also convenient to expand the system to the case of more than two overlapping events. The source code is released on GitHub[1]. The proposed system predicts overlapping events using track-wise outputs, that is, it predicts event and corresponding DoAs per each track. For DCASE 2020 Task3 with up to two overlapping events, the track number is two. Similar ideas were adopted by Nguyen [5]. However, to make the system more complete, it is reasonable to assume these tracks are event-independent. Thinking of a polyphonic prediction case illustrated in Fig. 1. The network has one prediction per each tack, within which there could only be maximally one event and corresponding DoA. There are also three groups of labels which are all potentially two events overlapping cases. Presumably, for the first group, the ”speech” label and the ”car” label are tied to track 1 and 2. For the second group, it is reasonable to still assign the ”speech” label to track 1, and the new ”dog bark” label to track 2. Nevertheless, for the third group of labels, it is hard to decide which tracks to assign ”dog bark” and ”car” labels to. In other words, track permutation problems emerge if track-wise predictions are used.

For track-wise predictions, tracks are event-independent. Frame-level permutation invariant training (denoted as tPIT), which was first proposed for speaker-independent source separation [6–8], can address this problem by examining all possible label permutations

[1] `https://bit.ly/3fAGANB`

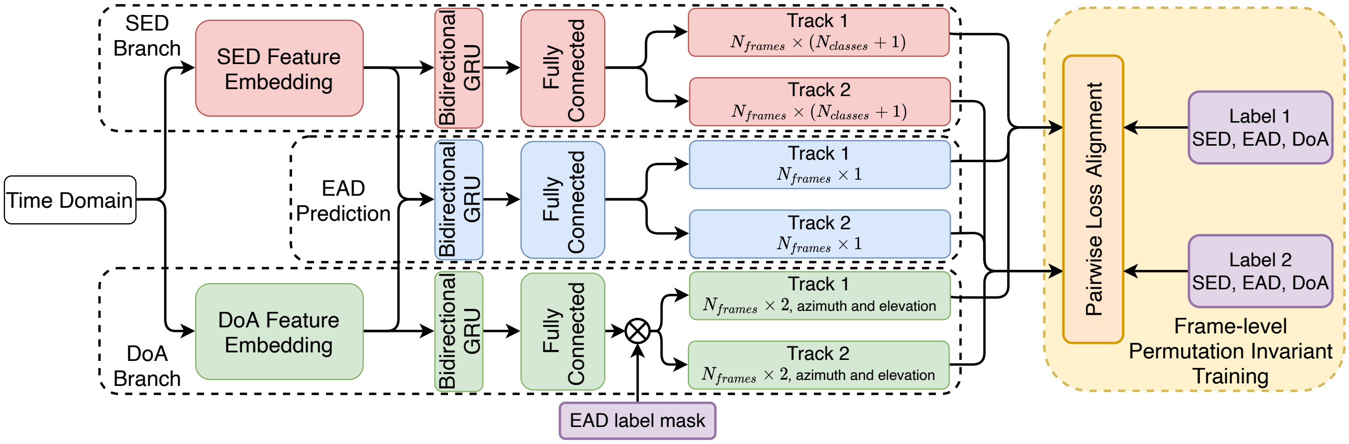

Figure 2: Network Architecture. N_{frames} is the number of frames. $N_{classes}$ if the number of classes of events. In SED branch, there is one additional class of event that is silence.

in each frame during training. It then selects the lowest frame-level loss among these label permutations for the backward propagation to train the model. In this way, the optimum local assignment of track-event pairs can be reached, thus leading to the excellent SED and DoA prediction performance frame-wisely.

In order to utilize the combined feature information from both SED and DoA estimation branches, besides the SED prediction and DoA prediction, an additional event activity detection (EAD) prediction is also adopted. The aim of this EAD prediction is to constrain the detection of the existence of events (or DoAs) not only from SED features but also from DoA features. That means SED and DoA predictions do not solely dependent in one way, but both information is used to detect the existence of events (or DoAs). With the proposed system, experimental results show that the performance is greatly increased compared with the baseline system.

The rest of the report is arranged as follows. In Section 2, the proposed learning method is described in detail, including features, network architecture, permutation invariant training, and hyper-parameters. Development results are shown in Section 3. Finally, conclusions are summarized in Section 4.

2. THE METHOD

2.1. Features

Task 3 provides two types of the input data format: First-Order of Ambisonics (FOA) and tetrahedral microphone array. In this report, a logmel spectrogram feature is used for SED, while an intensity vector from FOA data in logmel space is used for DoA estimation.

FOA, which is also known as B-format, includes four channels of signals, w, x, y and z. These four channel signals indicates omni-directional, x-directional, y-directional and z-directional components, respectively. The instantaneous sound intensity vector can be expressed as $\boldsymbol{I} = p\mathbf{v}$, where p is the sound pressure and can be obtained with w, $\mathbf{v} = (\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z)^{\mathrm{T}}$ is the particle velocity vector and can be estimated using x, y and z. Intensity vector carries the information of the acoustical energy direction of a sound wave, its inverse direction can be interpreted as the DOA, hence the FOA based intensity vector can be directly utilized for DOA estimation [1].

In order to concatenate the logmel and the intensity vector features to input to the proposed neural network, the intensity vector is also calculated in the STFT domain and the mel space as

$$\boldsymbol{I}(f,t) = \frac{1}{\rho_0 c}\Re\left\{ \mathrm{W}^*(f,t) \cdot \begin{pmatrix} \mathrm{X}(f,t) \\ \mathrm{Y}(f,t) \\ \mathrm{Z}(f,t) \end{pmatrix} \right\}, \tag{1}$$

$$\boldsymbol{I}_{norm,mel}(k,t) = -\boldsymbol{H}_{mel}(k,f)\frac{\boldsymbol{I}(f,t)}{\|\boldsymbol{I}(f,t)\|}, \tag{2}$$

where, ρ_0 and c are the density and velocity of the sound, W, X, Y, Z are the STFT of w, x, y, z, respectively, $\Re\{\cdot\}$ indicates the real part, * denotes the conjugate, $\|\cdot\|$ is a vector's ℓ_2 norm, k is the index of the mel bins, $\boldsymbol{H}_{mel}$ is the mel-band filter banks. In this report, the three components of the intensity vector are taken as three additional input channels for the network.

2.2. Network architecture

The proposed event-independent network has two branches which are the SED branch and the DoA branch. The network architecture is shown in Fig. 2. FOA time-domain signals are used as the input and are firstly fed into two branches. Each branch has a feature embedding layer. For both of the feature embedding layers. A 1-D convolutional layer is used to extract logmel spectrograms and intensity vectors features. Then they are normalized using a batch-normalization layer. For SED feature embedding, 4 groups of convolutional blocks are used to extract SED embedding. Each convolutional block contains 2 2-D convolutional layers with a kernel size of 3x3, a batch norm layer, and an average-pooling layer. For DoA feature embedding, ResNet 18 is used. The two branches are then used to generate three predictions, SED predictions, EAD predictions, and DoA predictions. For SED and DoA predictions, the SED and the DoA feature embeddings are fed into a bidirectional GRU and a fully-connected layer, respectively, to generate two tracks predictions. Each track has at most one event and one DoA. For SED, the predictions have $N_{classes}$ + 1 types of events. Here, $N_{classes}$ is the number of event classes. The additional class of event indicates the silence class (no event is happening). The softmax activation function is used after each track for SED. For DoA, the two predictions tracks contain azimuth and elevation angles. Linear activation is used (no activation). For EAD prediction, the SED feature embedding and DoA feature embedding are concatenated to input to a bidirectional GRU and a fully-connected layer to generate event activity predictions for two tracks. Each track prediction indicates if the event is happening on that track. Sigmoid is used after each track. The existence of EAD predictions is very important in two folds. First, it combines the SED and DoA feature embeddings to predict on-set and off-set information; second, it constrains the SED and DoA feature embedding layers to unify track-sequence. During training, the DoA predictions are masked by EAD labels to filter those frames with actual events happening. But for inference, the DoA predictions are masked by EAD predictions.

After track predictions are obtained, the frame-level permutation invariant training (tPIT) is used to simultaneously assign labels to different tracks and select whichever the lowest loss as the actual loss to calculate the backpropagation. Assume there are P possible permutations pairs of predictions and labels, $\mathbf{Y}$ indicates one of the predictions of SED, EAD or DoA, and $\tilde{\mathbf{Y}}_{\alpha(t)}$ indicates the possible labels, where $\alpha(t) \in P$ is one of the possible permutations at time t. The tPIT loss can be written as Eq. 3.

$$L_t^{tPIT} = \min_{\alpha(t)\in P} \sum_{SED,EAD,DoA} \left\| \mathbf{Y} - \tilde{\mathbf{Y}}_{\alpha(t)} \right\| \tag{3}$$

Therefore, the process of tPIT is to not only perform the classification or regression training but also to pair the most possible targets and predictions at each frame. In this case, the event-independent track predictions can be excellently matched with the corresponding targets, hence the track permutation problem can be solved.

2.3. Hyper-parameters

To generate the weights of the 1-D convolutional layer for logmel and intensity feature extraction, the sample rate of the signal is set to 24 kHz. A 1024-point Hanning window with a hop size of 480 points is utilized. The audio clips are segmented to have a fixed length of 5 seconds with a 80 % overlap for training. The learning rate is set to 0.0005 for the first 60 epochs and is then adjusted to 0.0001 after each epoch that follows. The final results are calculated after 80 epochs. A threshold of 0.5 is used to binarize the EAD predictions.

3. DEVELOPMENT RESULTS

This year, polyphonic sound event detection and localization are evaluated with modified metrics that consider the joint nature of localization-and-detection [3]. There are two metrics for SED which are F-score ($F_{\leq T^\circ}$) and Error Rate ($ER_{\leq T^\circ}$). But they are also location-dependent, considering true positives predicted only under a distance threshold $T = 20^\circ$. For localization part, there are other two metrics who are also classification-dependent. These metrics are location error LE_{CD} and localization recall metric LR_{CD}.

Using the validation split provided for this task, Table 1 shows the development set performance for the proposed method. As shown in the table, the performance of the proposed method outperforms the two baseline methods for both sound event detection and DOA estimation by a large margin.

Table 1: Test results for the development set.

	$\mathbf{ER_{20^\circ}}$	$\mathbf{F_{20^\circ}}$	$\mathbf{LE_{CD}}$	$\mathbf{LR_{CD}}$
baseline-Ambisonic	0.72	37.4 %	22.8°	60.7 %
baseline-Microphone array	0.78	31.4 %	27.3°	59.0 %
Event-Independent	**0.47**	**61.5 %**	**16.7°**	**75.4 %**

4. CONCLUSION

We proposed a new end-to-end event-independent network for polyphonic sound event localization and detection. The network treats the polyphonic cases as multiple-track problems, with each track has at most one event and the corresponding direction-of-arrival. In order to solve the problem of track permutation, a frame-level permutation invariant training strategy is employed. The network outputs three predictions which are sound event detection, event activity detection, and direction-of-arrival. The event activity detection encompasses the feature embedding information from both SED and DoA, hence it is able to predict on-set and off-set time of events more accurately. Experimental results show that the proposed system outperforms the baseline methods by a large margin.

5. ACKNOWLEDGMENT

This research was supported by EPSRC grant EP/N014111/1 “Making Sense of Sounds” and was partially supported by the H2020 Project entitled AudioCommons funded by the European Commission with Grand Agreement number 688382. This work was supported by National Natural Science Foundation of China (Grant No. 11804365).

6. REFERENCES

[1] Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, and M. Plumbley, “Polyphonic sound event detection and localization using a two-stage strategy,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 30–34.

[2] A. Politis, S. Adavanne, and T. Virtanen, “A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection,” *arXiv e-prints: 2006.01919*, 2020. [Online]. Available: https://arxiv.org/abs/2006.01919

[3] 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, March 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8567942

[4] A. Mesaros, S. Adavanne, A. Politis, T. Heittola, and T. Virtanen, “Joint measurement of localization and detection of sound events,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, Oct 2019, accepted.

[5] T. N. T. Nguyen, D. L. Jones, and W.-S. Gan, “A sequence matching network for polyphonic sound event localization and detection,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 71–75.

[6] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, “Permutation invariant training of deep models for speaker-independent multi-talker speech separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 241–245.

[7] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[8] Y. Liu and D. Wang, “Divide and conquer: A deep casa approach to talker-independent monaural speaker separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2092–2102, 2019.