TWO-STAGE SOUND EVENT LOCALIZATION AND DETECTION USING INTENSITY VECTOR AND GENERALIZED CROSS-CORRELATION

Technical Report

Yin Cao, Turab Iqbal*, Qiuqiang Kong, Miguel B. Galindo, Wenwu Wang, Mark D. Plumbley*

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{yin.cao, t.iqbal, q.kong, m.blancogalindo, w.wang, m.plumbley}@surrey.ac.uk

ABSTRACT

Sound event localization and detection (SELD) refers to the spatial and temporal localization of sound events in addition to classification. The Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Task 3 introduces a strongly labelled dataset to address this problem. In this report, a two-stage polyphonic sound event detection and localization method. The method utilizes log mel features for event detection, and uses intensity vector and GCC features for localization. Intensity vector and GCC features use the supplied Ambisonic and microphone array signals, respectively. This method trains SED first, after which the learned feature layers are transferred for direction of arrival (DOA) estimation. It then uses the SED ground truth as a mask to train DOA estimation. Experimental results show that the proposed method is able to localize and detect overlapping sound events in different environments. It is also able to improve the performance of both SED and DOA estimation, and performs significantly better than the baseline method.

Index Terms— Sound event localization and detection, direction of arrival, intensity vector, generalized cross correlation, convolutional recurrent neural networks.

1. INTRODUCTION

Sound event detection is a rapidly developing research area that aims to analyze and recognize a variety of sounds in urban and natural environments. Compared to sound tagging, event detection also involves estimating the time of occurrence of sounds. Automatic recognition of sound events would have a major impact in a number of applications [1]. For instance, sound indexing and sharing, bio-acoustic scene analysis for animal ecology, smart home automatic audio event recognition (baby cry detection, window break alarm), and sound analysis in smart cities (security surveillance).

In real-world applications, a sound event is always transmitted in a certain direction. Given this fact, it is reasonable to combine sound event detection and localization by not only identifying the type and temporal location of the sound but also estimating its spatial location. Therefore, it is worthwhile to study them together and investigate the effects and potential connections between them.

Task 3 of the DCASE 2019 challenge focuses on sound event localization and detection (SELD) for overlapping sound sources [2]. A recently developed system known as SELDnet was used as the baseline system. SELDnet uses magnitude and phase spectrograms as input features and trains the SED and DOA estimation objectives jointly [3]. Besides spectrograms, generalized cross-correlation (GCC) based features have also been used as input features [4–8] to solve sound event localization, which can effectively supply time difference information.

In this report, log mel features have been used for SED, while a type of intensity vector in log mel space and generalized cross-correlation features have been used for DOA estimation. A novel two-stage method for polyphonic sound event detection and localization is used [9]. This method trains sound event detection and localization in two stages: the SED stage and the DOA estimation stage, corresponding to the SED branch and the DOA estimation branch in the model, respectively. During training, the SED branch is trained first only for SED, after which the learned feature layers are transferred to the DOA estimation branch. The DOA estimation branch fine-tunes the transferred feature layers and uses the SED ground truth as a mask to learn only DOA estimation. During inference, the SED branch estimates the SED predictions first, which are used as the mask for the DOA estimation branch to infer predictions. The experimental results show that by using the proposed method, DOA estimation can benefit from the SED predictions; both SED and DOA estimation can be improved at the same time. The proposed method performs significantly better than the baseline method.

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

2. THE METHOD

In this report, a two-stage polyphonic sound event detection and localization network using log mel space intensity vector and generalized cross-correlation features is utilized for Task 3. The source code is released on GitHub [1][2].

2.1. Features

Task 3 provides two types of input data format: First-Order of Ambisonics (FOA) and tetrahedral microphone array [2]. In this report, a log mel feature is first used for SED, while an intensity vector in log mel space and a GCC with phase transform (GCC-PHAT) features are used for DOA estimation. The log mel space intensity vector utilizes FOA input data, whereas the GCC-PHAT utilizes the

* Equal contribution.

[1] https://github.com/yinkalario/DCASE2019-TASK3

[2] https://github.com/yinkalario/Two-Stage-Polyphonic-Sound-Event-Detection-and-Localization

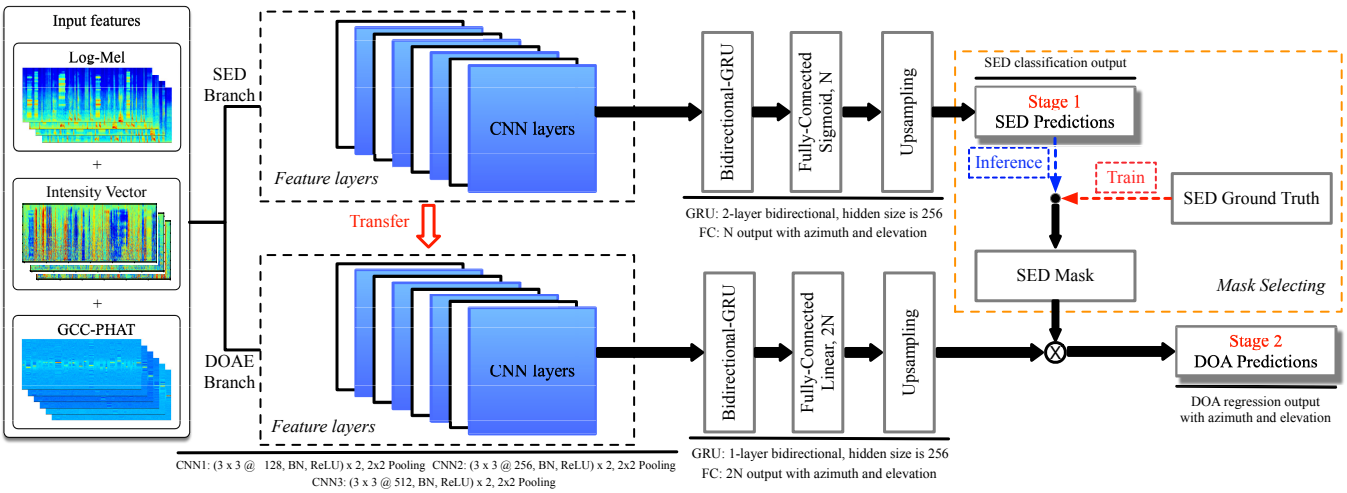

Figure 1: The diagram of the proposed two-stage sound event detection and localization network. SED ground truth is used as the mask to train DOA estimation branch. SED predictions are used as the mask to infer DOA.

microphone array input data. The detailed features are described in the following sections.

2.1.1. *Intensity vector*

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 [10].

In order to concatenate the log mel and the intensity vector features to input to the 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 neural network.

2.1.2. *Generalized Cross-Correlation*

GCC is based on microphone array signals, and is widely used in time difference of arrival (TDOA) estimation by means of maximizing the cross-correlation function to obtain the lag time between two microphones. The cross-correlation function is calculated through the inverse-FFT of the cross power spectrum. GCC-PHAT is the phase-transformed version of GCC, which whitens the cross power spectrum to eliminate the influence of the amplitude, leaving only the phase. GCC-PHAT can be expressed as

$$GCC_{ij}(t,\tau) = \mathcal{F}^{-1}_{f\to\tau}\frac{X_i(f,t)\cdot X_j^*(f,t)}{|X_i(f,t)\|X_j(f,t)|}, \tag{3}$$

where $\mathcal{F}^{-1}_{f\to\tau}$ is the inverse-FFT from f to τ, $X_i(f,t)$ is the Short-Time Fourier Transform (STFT) of the i-th microphone signal. $GCC_{ij}(t,\tau)$ can also be deemed as a GCC spectrogram, with τ corresponding to the number of mel-band filters. That is, GCC-PHAT can be stacked with a log mel spectrogram as the input features. In order to determine the size of GCC-PHAT, the largest distance between two microphones, $d_{\max}$, needs to be used. The maximum delayed samples corresponding to $\Delta\tau_{\max}$ can be estimated by $d_{\max}/c\cdot f_s$, where c is the sound speed and f_s is the sample rate. In this paper, log mel and GCC-PHAT will be stacked as the input features, considering the possibility of the advance and the delay of GCC. The number of mel-bands, therefore, should be no smaller than double the number of delayed samples plus one [11].

2.2. Network architecture

The network is shown in Fig. 1, and has two branches, the SED branch and the DOA estimation branch. During training, the extracted features, which have shape $C\times T\times F$, are first sent to the SED branch. C indicates the number of feature maps, T is the size of time bins, and F is the number of mel-band filters or delayed samples of GCC-PHAT. The CNN layers, which are also called feature layers in this report, are constructed with 4 groups of 2D CNN layers (Convs) with 2×2 average-pooling after each of them. Each Convs' group consists of two 2D Convs, with a receptive field of 3×3, a stride of 1×1, and a padding size of 1×1 [12]. Each single CNN layer is followed by a batch normalization layer [13] and a ReLU activation. After the CNN layers, the data is then sent to a global average-pooling layer to reduce the dimension of F followed by a bidirectional GRU. The output size is maintained and is sent to a fully-connected layer with an output size of K, which is the number of event classes. The sigmoid activation function is used afterwards

with an upsampling in the temporal dimension to ensure the output size is consistent with T. The SED predictions can now be obtained through an activation threshold. Binary cross-entropy is used for this multi-label classification task.

The DOA estimation branch is then trained. The CNN layers are transferred from the SED branch and are fine-tuned. The output of the fully-connected layer for the DOA estimation branch is a vector of $K \times 2$ linear values, which are azimuth and elevation angles for K events. They are then masked by the SED ground truth during training to determine if the corresponding angles are currently active. Finally, the mean absolute error is chosen as the DOA estimation regression loss.

During inference, the SED branch will first compute the SED predictions, which are then used as the SED mask to obtain the DOA estimation. For more detailed descriptions, readers can refer to [9].

2.3. Ensemble method

After training the proposed models, ensemble averaging was used to combine the predictions of these models. Let $\mathbf{y}_1, ..., \mathbf{y}_M$ be the predictions of the M models for some instance $\mathbf{x}$, where $\mathbf{y}_i \in \mathcal{Y}$, and $\mathcal{Y}$ is some prediction space. An ensemble function, $f : \mathcal{Y}^M \rightarrow \mathcal{Y}$, computes a new prediction as a function of $\mathbf{y}_1, ..., \mathbf{y}_M$. A simple example is the mean ensemble function, defined as the mean of the predictions. Although the mean ensemble is effective, other functions can exploit the different characteristics of the models.

Stacking [14] is an ensemble method in which the function f is learned using a machine learning algorithm. This allows exploiting the characteristics of the models, such as class-wise performance, in a data-driven manner. In our system, a neural network was used to model f, with the concatenation of the predictions, $[\mathbf{y}_1, ..., \mathbf{y}_M]$, as the input. Two types of inputs were investigated. Since the models output a prediction for each time bin, one could let $\mathbf{y}_i$ correspond to such a prediction, i.e. $\mathcal{Y} \coloneqq \mathbb{R}^K$. However, this means that temporal relations are discarded. Instead, one could let $\mathbf{y}_i$ correspond to several of these predictions, i.e. $\mathcal{Y} \coloneqq \mathbb{R}^{B \times K}$, where B is the number of time bins to consider jointly. In the experiments, this was found to improve the performance. As such, it is the approach used in the proposed system, which means the input of the network is a $B \times KM$ matrix. To train the network, the predictions of the training set were used by adopting the cross-validation setup.

The neural network used is a convolutional recurrent neural network with three layers. It is a 1D convolutional layer followed by batch normalization, a bidirectional GRU layer, and a final fully-connected layer with a sigmoid activation. The convolutional layer takes the input to be a sequence of length B with KM channels, and outputs a sequence of length B with 128 channels. The kernel size of the convolution is 7. The bidirectional GRU layer outputs 64 channels in each direction so that the total number of channels is also 128. Finally, the fully-connected layer takes each 1×128 slice and outputs a vector of length K. Concatenating these K-vectors gives the desired prediction with shape $B \times K$.

The stacking and mean ensemble methods are implemented for SED and DOA estimation, respectively. Stacking is first used to improve the performance of SED, and its output is binarized with a threshold later to form the SED mask for DOA inference. Mean ensemble is then used for the DOA predictions.

2.4. Hyper-parameters

To extract the input features, the sample rate of the STFT is set to 32 kHz. A 1024-point Hanning window with a hop size of 320 points is utilized. The number of mel-band filters and the delays of GCC-PHAT is set to 128. For 4 channels of FOA and microphone array signals, there are 8 channels of log mel features from FOA and microphone array signals in total, 3 channels of intensity vector features, and 6 channels of GCC-PHAT features, hence up to 17 input channels of signals are sent to the network. The audio clips are segmented to have a fixed length of 5 seconds with a 50 % overlap for training. The learning rate is set to 0.001 for the first 40 epochs and is then decayed by 10 % after each epoch that follows. The final results are calculated after 80 epochs. A threshold of 0.5 is used to binarize the SED predictions.

3. DEVELOPMENT RESULTS

Polyphonic sound event detection and localization were evaluated with individual metrics for SED and DOA estimation. For SED, segment-based error rate (ER) and F-score [15] were calculated in one-second lengths. A lower ER or a higher F-score indicates better performance. For DOAE, DOA error and frame recall were used. A lower DOA error and a higher frame recall are better.

Using the cross-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: Cross-validation results for the development set.

	Error rate	F score	DOA error	Frame recall
baseline-Ambisonic	0.34	0.799	28.5°	0.854
baseline-Microphone array	0.35	0.800	30.8°	0.840
Two-Stage	**0.13**	**0.930**	**6.61°**	**0.894**

4. CONCLUSION

The goal of the DCASE 2019 Task 3 is to recognize and localize sound events by determining their onset and offset times as well as their direction of arrival. In this report, a two-stage polyphonic sound event localization and detection method was proposed. The method uses log mel features for SED, and uses intensity vector in mel space and GCC-PHAT features for DOA estimation. The experimental results for the development dataset show that the proposed method outperforms the baseline methods by a significant 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.

6. REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018, ch. 1, pp. 3–12.

[2] http://dcase.community/challenge2019/task-sound-event-localization-and-detection.

[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, 2019.

[4] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[5] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

[6] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[7] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[8] Y. Sun, J. Chen, C. Yuen, and S. Rahardja, "Indoor sound source localization with probabilistic neural network," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 8, pp. 6403–6413, 2018.

[9] 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," *arXiv preprint arXiv:1905.00268*, 2019.

[10] J. Ahonen, V. Pulkki, and T. Lokki, "Teleconference application and b-format microphone array for directional audio coding," in *Audio Engineering Society Conference: 30th International Conference: Intelligent Audio Environments*. Audio Engineering Society, 2007.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015.

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

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162–178, 2016.