Journal of Systems Engineering and Electronics
Vol. PP, No. 99, August 2023, pp.1 – 8

Sound event localization and detection based on deep learning

ZHAO Dada [1,2], DING Kai [2], QI Xiaogang [1,*], CHEN Yu [2], and FENG Hailin [1]

1. School of Mathematics and Statistics, Xidian University, Xi'an 710071, China;
2. Science and Technology on Near-Surface Detection Laboratory, Wuxi 214035, China

Abstract: Acoustic source localization (ASL) and sound event detection (SED) are two widely pursued independent research fields. In recent years, in order to achieve a more complete spatial and temporal representation of sound field, sound event localization and detection (SELD) has become a very active research topic. This paper presents a deep learning-based multi-overlapping sound event localization and detection algorithm in three-dimensional space. Log-Mel spectrum and generalized cross-correlation spectrum are joined together in channel dimension as input features. These features are classified and regressed in parallel after training by a neural network to obtain sound recognition and localization results respectively. The channel attention mechanism is also introduced in the network to selectively enhance the features containing essential information and suppress the useless features. Finally, a thourough comparison confirms the efficiency and effectiveness of the proposed SELD algorithm. Field experiments show that the proposed algorithm is robust to reverberation and environment and can achieve higher recognition and localization accuracy compared with the baseline method.

Keywords: sound event localization and detection (SELD), deep learning, convolutional recursive neural network (CRNN), channel attention mechanism.

DOI: 10.23919/JSEE.2023.000110

1. Introduction

Recognition of sound events category and occurrence time in audio records is a relatively active research topic which has a wide range of applications on many occasions [1]. Although sound event detection (SED) can reveal a lot about the recording environment, the spatial information of events is just as important. Discovering sound spatial information is what sound source localization is aimed for. Sound source localization is a classical signal processing task based on the propagation characteristics of sound and the signal relations between channels, without considering the types of sound representations. Sound event localization and detection (SELD) is a comprehensive task, which is capable of achieving a more complete spatial and temporal representation of the acoustic scene by combining SED with acoustic source localization. The spatial information makes SELD suitable for a wide range of machine listening tasks, such as inference of environmental types [2], robot positioning, navigation, tracking and mapping [3,4], and audio surveillance [5].

Traditionally, acoustic source localization and SED are two independent types of researches. Early research dealt with these two problems without attempting to correlate sound source location with the event type. SED tasks often use different supervised classification methods to predict sound class. Some methods contain hidden Markov model (HMM), Gaussian mixture model (GMM) [6], deep neural network (DNN) [7], recursive neural network (RNN) [8−11], and convolutional neural network (CNN) [12,13]. The latest research results are obtained through continuous stacking of CNN and RNN layers, which are jointly called convolutional recursive neural network (CRNN) [14−18]. Sound source localization uses classical array processing methods such as time difference of arrival (TDOA) [5], steering response power [19], and multiple signal classification (MUSIC) [20]. The classical array processing method performs poorly in scenarios with strong reverberation and noise [21].

However, there are certain defects in independent sound recognition and localization [22]. For instance, the system correctly detects two sound events, yet the spatial locations of the sound events are reversed. The use of independent recognition metrics can evaluate whether the system correctly predicts the presence of sound events regardless of their location. Similarly, independent localization metrics would evaluate the spatial error between the closest sound pairs regardless of the sound type.

An early attempt to combine these two tasks was presented in [23], where the beamforming output of a distributed array was used in combination with an HMM-

Manuscript received June 06, 2021.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (61877067), the Foundation of Science and Technology on Near-Surface Detection Laboratory (TCGZ2019A002;TCGZ2021C003; 6142414200511), and the Natural Science Basic Research Program of Shaanxi (2021JZ-19).

GMM classifier. Over the past decade, DNNs have become the most mature approach to SED, providing ample modeling flexibility and surpassing traditional machine learning approaches when training with sufficient data [24]. Recently, DNNs have also been used for source localization [25−27], with promising results. Therefore, DNNs seem to be a good choice for joint modeling of localization and detection in SELD tasks. The first works we know of using this method are references [28,29]. Hirvonen proposed to set joint modeling as a multi-label-multi-class classification problem, which maps two event classes into eight discrete angles on azimuth [28] and uses CNN to infer the probability of each sound class in each position. Then Hirvonen used a predefined threshold to determine the existence and location of the final sound event class [28]. This method is called HIRnet. The angle resolution is subjected to the predefined direction because it cannot detect angle values that are not seen in the training data. For larger data sets, with more sound events and higher angular resolution, this approach can result in a lot of output nodes. Training DNN with so many output nodes, where the number of positive labels is much lower than the number of negative class labels, can lead to the problem of dataset imbalance. In addition, handling so many acoustic types needs a large scale data set, where sufficient data is needed for each class. Adavanne et al. proposed a SELDnet network architecture [29] to solve the problem of joint sound event localization and detection. There are two output branches in the result, one performing SED and the other performing localization. This approach solves the problem of data association due to sound overlap in the SELD task [24]. However, features used in this method cannot sufficiently represent the sound and channel information, thus the recognition and localization accuracy still needs to be improved.

This paper proposes a sound recognition and localization combination algorithm based on CRNN, and the network structure is based on SELDnet. The proposed algorithm is aimed at a circular array of omnidirectional microphones. All the datasets used in this paper are generated by a circular array of omnidirectional microphones. Different from [29], this paper uses the more effective Log-Mel spectrum and generalized cross-correlation (GCC) features in the field of SED and sound source localization, which better represents the information required by the SELD task. In addition, since the output of the convolutional layer does not consider the relationship between each channel, this paper introduces the channel attention mechanism to selectively enhance the features with essential information and suppress useless features, and finally achieves a result better than [29]. The rest of this paper is organized as follows. Section 2 describes the proposed method based on deep learning. Experimental results and analysis are presented in Section 3. Finally, conclusions are drawn in Section 4.

2. SELD algorithm

The input is a multi-channel audio signal. Then Log-Mel and GCC spectra as features are extracted from the input. The method takes a series of features in the continuous Log-Mel and GCC spectra frames as input of the network, predicts all the active sound event classes in each input frame and their respective spatial location, and generates the time activity and arrival angle of each sound event class. CRNN maps feature sequences to two outputs in parallel. The first output performs SED as a multi-label classification task, allowing the network to estimate the simultaneous presence of multiple sound event classes in each frame. The second output takes direction of arrival (DOA) estimation in a continuous three-dimensional (3D) space as a multi-output regression task, where each sound event class is associated with three regression variables. These variables make up the DOA's 3D coordinates on the unit ball. For each acoustic type in the data set, the SED output range belong to 0 to 1, and the threshold value k is set. If the SED output is greater than k, this sound event class is considered to exist, and vice versa. The corresponding acoustic source localization (ASL) outputs for those present acoustic types provide their space position. Feature extraction and the method proposed in this paper are described as follows.

2.1 Feature extraction

The features used in the SELDnet cannot enough represent the sound and channel information. Therefore, more effective Log-Mel spectrum and GCC phase transfrom (GCC-PHAT) features in the field of SED and sound source localization are used, which better represents the information required by the SELD task.

2.2 Network architecture

The Log-Mel and GCC-PHAT spectra are stacked in the channel dimension as input features. Note that the number of signal channels is M, the frame number is N, and the number of Mel filters is L, then the input feature dimension is

$$N \times L \times \sum_{i=1}^{M} i.$$

Input the features into the neural network as shown in Fig. 1. Each CNN layer contains P filters of 3×3 dimensional receptive fields, which are activated by rectified linear unit (ReLU). After the completion of each layer of CNN, batch normalization [30] is used to normalize the activation output, and then channel attention

module is added to improve the network performance [31]. Finally, maximum pooling is used to reduce the dimension along the frequency dimension while keeping the time dimension N unchanged. After passing through three layers of CNN, the feature size becomes $N\times 2\times P$.

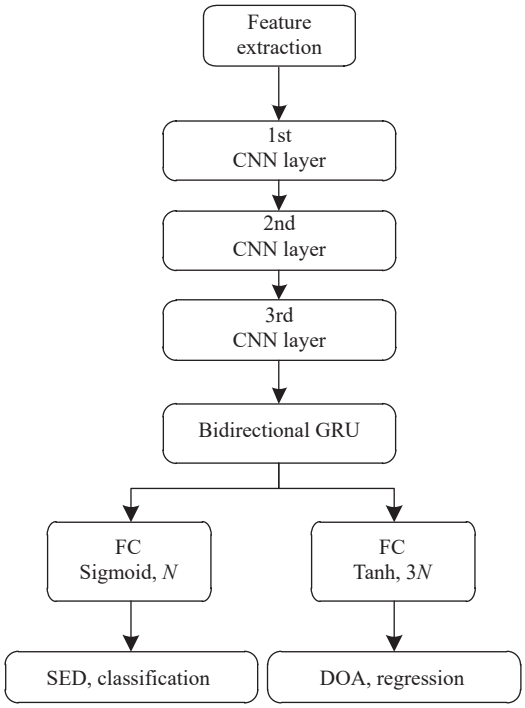


Fig. 1 Algorithm flow chart

The output is reshaped as $N\times 2P$ features of the CNN layer is fed to the bi-directional RNN layer for learning time information. The RNN layer uses gated recurrent unit (GRU) with Q nodes, and Tanh is used in this layer. Next, two parallel full connected (FC) layers are connected for SED and DOA estimation, respectively. The FC layer used for SED has H nodes, and the activation function is Sigmoid. The FC layer used for DOA estimation has $3H$ nodes, which are activated by Tanh, corresponding to x, y, and z values of H sound event classes, and the value range of $[-1,1]$.

2.3 Channel attention mechanism

The attention mechanism was first proposed and used in natural language processing and machine translation to align text, and achieved good results. In the field of computer vision, scholars have explored how to use attention mechanism to improve network performance in CNNs.

The output of the convolutional layer does not consider the relationship between each channel, while the channel attention mechanism is to model the correlation of each channel. Therefore, we can selectively enhance the features with essential information and suppress useless features by using the channel attention mechanism [31]. The principle of the channel attention mechanism is shown in Fig. 2.

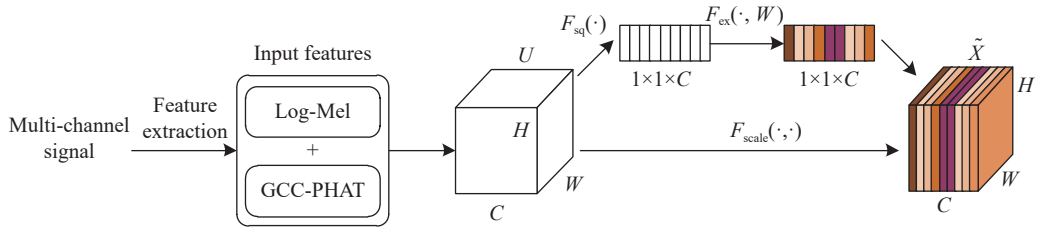


Fig. 2 Schematic diagram of the channel attention mechanism

First, we investigate the features of each channel, squeeze the global spatial information into the channel descriptor, and use global average pooling to generate the statistics of each channel.

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{1}$$

where H, W, and C denote the 3D of feature U, z_c denotes the global average pooling result in channel c, and u_c denotes the plane of feature U on channel c.

Second, investigate the degree of dependence of each channel. This is achieved through a threshold mechanism with a sigmoid activation function. In order to limit the complexity of the model and enhance the generalization ability, two FC layers in the form of bottleneck layer are used in the threshold mechanism. The first FC layer reduces its dimension to $1/R$, where R is the hyperparameter, and the second FC layer increases it to R. The final sigmoid function is the weight of each channel. Adjusting the weight of each channel feature according to the input data helps to enhance the distinguishability of features.

$$s = F_{ex}(z,W) = \sigma(g(z,W)) = \sigma(W_2\delta(W_1 z)) \tag{2}$$

where s is the degree of dependence, W_1 and W_2 are two FC layers, and σ is the Sigmoid activation function.

Multiply the weight of each channel with the original features to get the features needed.

$$\tilde{x}_c = F_{\text{scale}}(u_c, s_c) = s_c \cdot u_c \tag{3}$$

where $\tilde{x}_c$ is the feature after processing.

2.4 Model training

In the part of the SED, the target value is 1 if there is a sound event class in the frame, and 0 otherwise. Since there may be multiple overlapping sound events, we use binary cross-entropy loss to predict losses. In the DOA estimation, if the corresponding event is non-existent, the values are set as: $x = 0, y = 0, z = 0$. We use mean square error (MSE) to estimate the losses. Adavanne et al. [29] proposed that it is theoretically more beneficial to use Cartesian coordinates instead of azimuth and elevation regressions when predicting omnidirectional or full-elevation DOA. This is because the angle is discontinuous at the border. The flaw makes network learning even worse. Therefore, Cartesian coordinates are used to replace azimuth angle and elevation angle for regression. We train the model with a dropout rate of 0.3 and a dynamic learning rate using the Adam optimizer for 250 epochs with MSE and binary cross-entropy loss. The learning rate drops to 80% of its original value every 30 epochs with an initial value of 0.003. In order to extract input features, the sampling rate of short time Fourier transform (STFT) is set at 44.1 kHz. A 1 024-point Hanning window is used, with a sliding size of 512 points. In the dataset used, the maximum microphone distance is 10 cm. The number of Mel filters and the GCC-PHAT delay are set to 96. For eight channels of sound signals, 36 channels of input features are transmitted to the neural network. This network is implemented by TensorFlow [32].

2.5 Dataset

The dataset consists of static point sources, each of which is associated with a spatial coordinate. And it is generated from a circular array of omnidirectional microphones. All datasets contain three sub-datasets, at most one (OV1), two (OV2), and three (OV3) overlapping sound events. Each sub-dataset has three cross-validation splits (Split1, Split2, and Split3). Refer to [29] for details.

(i) Circular Array, Anechoic and Synthetic Impulse Response (CANSYN) dataset: a circular array with a radius of 5 cm in which eight microphones are evenly located.

(ii) Circular Array, Reverberation and Synthetic impulse Response (CRESYN) dataset: it is similar to CANSYN dataset, and the only difference is the addition of reverberation.

2.6 Evaluation metrics

For SED, we use the standard polyphonic SED metrics, error rate (ER) and F-score calculated without overlap within the one-second segment [33]. Mathematically, the F-score is calculated as follows:

$$F = \frac{2\sum_{h=1}^{H}\text{TP}(h)}{2\sum_{h=1}^{H}\text{TP}(h) + \sum_{h=1}^{H}\text{FP}(h) + \sum_{h=1}^{H}\text{FN}(h)} \tag{4}$$

where $\text{TP}(h)$ denotes the total number of sound event classes that are existing in both the label and the prediction in the hth 1-second segment. $\text{FP}(h)$ denotes the total number of sound event classes that are existing in the prediction but not in the label. $\text{FN}(h)$ denotes the number of acoustic types which are not existing in the prediction but in the label. ER score is calculated as follows:

$$\text{ER} = \frac{\sum_{h=1}^{H}S(h) + \sum_{h=1}^{H}D(h) + \sum_{h=1}^{H}I(h)}{\sum_{h=1}^{H}N(h)}, \tag{5}$$

where for the hth 1-second fragment, $N(h)$ denotes the number of existing acoustic events in the label,

$$\begin{cases} S(h) = \min(\text{FN}(h), \text{FP}(h)) \\ D(h) = \max(0, \text{FN}(h) - \text{FP}(h)) \\ I(h) = \max(0, \text{FP}(h) - \text{FN}(h)) \end{cases} \tag{6}$$

where $S(h)$ is the number of times an event detected but given the wrong level. The remaining false positives and false negatives, if any, are counted as insertions $I(h)$ and deletions $D(h)$ respectively.

The ideal value of F-score is 1 and the ER score is 0.

For the DOA estimation (x_E, y_E, z_E) of the label (x_G, y_G, z_G), the center angle and frame recall rate are used to evaluate as follows:

$$\text{frame recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

$$\sigma = 2\arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right) \times \frac{180}{\pi}, \tag{8}$$

where true positive TP denotes the total frame number of predicted DOA equal to the reference DOA, false negative FN denotes the total frame number of predicted DOA not equal to the reference DOA, and $\Delta x = x_G - x_E$, $\Delta y = y_G - y_E$, $\Delta z = z_G - z_E$. DOA error is calculated as follows:

$$\text{DOA error} = \frac{1}{K} \cdot \sum_{k=1}^{K}\sigma_k \tag{9}$$

where K is the number estimated by DOA.

3. Experiments

3.1 Baseline

(i) SELDnet

Spectrogram and phase spectrogram are introduced

into the CRNN network as input features and mapped to two branches. One branch produces the time activity of all the acoustic types on each time frame. The other branch is positioned by DOAE.

(ii) HIRnet

The logarithmic spectral powers are fed into CNN and maps it to each of the two classes (voice and music) as an all-around eight angles for a multi-label classification task. In order to compare with the algorithm proposed in this paper, the azimuth resolution of the proposed algorithm is improved to 10°.

(iii) MUSIC

MUSIC is a high-resolution DOA estimator. Specifically, it is very generic in terms of array geometry, directional characteristics. It is based on matrix eigenspace decomposition. From the geometric point of view, the observation space of signal processing can be decomposed into signal subspace and noise subspace, which are obviously orthogonal.

(iv) Two-Stage method

The two-stage method [34] deals with sound event detection and localization in two stages: the SED stage and the DOAE stage, corresponding to the SED branch and the DOAE 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 DOAE branch. The DOAE branch fine-tunes the transferred feature layers and uses the SED ground truth as a mask to learn only DOAE.

3.2 Experimental results

Fig. 3 shows the confusion matrix of the proposed algorithm for the estimation of the number of sound event classes per frame. As shown in Fig. 3(e), frames with three sound sources in the label are estimated by the network as three sound sources in 44% (true positives). The frame recall rate is a value representing these confusion matrixes. In Fig. 3, the true positives percentage decreases with the increase in the number of sources, and the decrease is more significant in the case of reverberation. However, compared to the baseline frame recall rate metrics in Table 1, the algorithm presented in this paper performs better for a higher number of overlapping sound events, especially under reverberation conditions.

Fig. 3 Confusion matrixes of the proposed algorithm on two datasets

Table 1 Comparation of the proposed algorithm with baseline

Algorithm	Evaluation metrics	CANSYN			CRESYN		
		OV1	OV2	OV3	OV1	OV2	OV3
SELDnet	ER	0.11	0.18	0.19	0.13	0.22	0.30
	F score	93.0	86.6	85.3	90.4	82.2	78.0
	DOA error	29.5	31.3	34.3	28.4	33.7	41.0
	Frame recall	**97.9**	78.8	67.0	96.4	75.7	60.7
HIRnet	ER	0.41	0.45	0.62	0.43	0.46	0.50
	F score	60.0	54.9	58.8	59.3	60.2	58.6
	DOA error	**5.2**	**16.3**	33.0	**7.4**	**18.6**	43.3
	Frame recall	60.2	35.9	18.4	56.9	20.5	10.7
MUSIC	DOA error	26.4	28.9	**31.1**	38.6	49.5	61.9
Two-stage	ER	**0.07**	0.17	0.20	**0.12**	0.21	0.28
	F score	**95.9**	91.0	84.7	92.4	84.2	81.0
	DOA error	**27.6**	31.3	36.2	27.0	33.5	39.8
	Frame recall	**98.0**	82.1	65.0	96.3	78.1	62.7
The proposed algorithm	ER	**0.08**	**0.16**	**0.18**	**0.12**	**0.18**	**0.23**
	F score	**95.8**	**91.0**	**89.5**	**93.7**	**89.5**	**86.2**
	DOA error	28.5	31.0	32.6	27.2	32.4	**38.4**
	Frame recall	97.8	**85.5**	**72.1**	**96.6**	**82.4**	**68.5**

A comparison of the algorithms on the two datasets is listed in Table 1. For the SED metrics, the proposed algorithm almost outperforms all the baselines on the datasets used, and is just slightly lower than the two-stage algorithm on CANSYN OV1, which shows an acceptable performance of our algorithm to polyphonic events and reverberation. In terms of DOA metrics, the frame recall rate of the proposed algorithm is the highest on OV2 and OV3, and slightly lower than that of SELDnet and two-stage algorithm on CANSYN OV1. On the dataset of CANSYN, the DOA errors of the MUSIC algorithm are the lowest except HIRnet because of the knowing number of acoustic sources and the environment without reverberation. The DOA error of the proposed algorithm is the best on OV3. The DOA error of others on OV1 and OV2 is higher than HIRnet because it only evaluates the azimuth angle, while the rest of the algorithms evaluate the azimuth and elevation angle. Moreover, the regression method used for both the proposed algorithm and SELDnet localization may not fully learn the complete mapping between input features and continuous DOA space. However, the proposed algorithm shows better performance than SELDnet in localization and recognition. In general, the proposed algorithm achieves good results in recognition and localization, and it is robust to reverberation and polyphonic events.

4. Conclusions

This paper proposes a CRNN based multi-overlapping SELD algorithm in 3D space. Log-Mel spectrum and GCC spectrum are used as features to feed to the neural network, and a channel attention module is added to enhance the network performance. Compared with the other three SELD algorithms and the MUSIC algorithm on both datasets, the proposed algorithm achieves better performance in localization and recognition than those algorithms in the vast majority of aspects. Experiments show that the proposed algorithm is robust to reverberation and polyphonic events.

References

[1] MESAROS A. Sound event detection in the DCASE 2017 challenge. IEEE/ACM Trans. on Audio, Speech, Language Processing, 2019, 27(6): 992–1006.

[2] SALAMON J, BELLO J P. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 2017, 24(3): 279–283.

[3] EVERS C, NAYLOR P A. Acoustic SLAM. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2018, 26(9): 1484–1498.

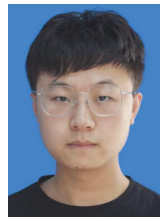
[4] MACK W, BHAEADWAJ U, CHAKRABARTY S, et al. Signal-aware broadband DOA estimation using attention mechanisms. Proc. of the IEEE International Conference Acoustics, Speech and Signal Processing, 2020: 4930–4934.

[5] VALENZISE G, GEROSA L, TAGLIASACCHI M, et al. Scream and gunshot detection and localization for audio-surveillance systems. Proc. of the IEEE Conference on Advanced Video & Signal Based Surveillance, 2007: 21–26.

[6] MESAROS A, HEITTOLA T, ERONEN A, et al. Acoustic event detection in real-life recordings. Proc. of the European

Signal Processing Conference, 2010: 1267–1271.
[7] AKIR E C, HEITTOLA T, HUTTUNEN T, et al. Polyphonic sound event detection using multi-label deep neural networks. Proc. of the IEEE International Joint Conference on Neural Networks, 2015. DOI: 10.1109/IJCNN.2015.7280624.
[8] PARASCANDOLO G, HUTTUNEN H, VIRTANEN T. Recurrent neural networks for polyphonic sound event detection in real life recordings. Proc. of the IEEE International Conference Acoustics, Speech and Signal Processing, 2016: 6440–6444.
[9] ADAVANNE S, PARASCANDOLO G, PERTILA P, et al. Sound event detection in multichannel audio using spatial and harmonic features. Proc. of the Workshop on Detection Classification Acoustic Scenes and Events, 2016. DOI: 10.48550/arXiv.1706.02293.
[10] HAYASHI T, WATANABE S, TODA T, et al. Duration-controlled LSTM for polyphonic sound event detection. IEEE/ACM Trans. on Audio, Speech, Language Processing, 2017, 25(11): 2059–2070.
[11] ZOHRER M, PERNKOPF F. Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks. Proc. of the Interspeech, 2017: 493–497.
[12] ZHANG H M, MCLOUGHLIN I, SONG Y. Robust sound event recognition using convolutional neural networks. Proc. of the IEEE International Conference Acoustics, Speech and Signal Processing, 2015: 559–563.
[13] PHAN H, HERTEL L, MAASS M, et al. Robust audio event recognition with 1-max pooling convolutional neural networks. Proc. of the Interspeech, 2016. DOI: 10.48550/arXiv.1604.06338.
[14] ADAVANNE S, POLITIS A, VIRTANEN T. Multichannel sound event detection using 3D convolutional neural networks for learning inter-channel features. Proc. of the IEEE International Joint Conference on Neural Networks, 2018. DOI: 10.1109/ISCNN.2018.8489542.
[15] LIM H, PARK J, LEE K, et al. Rare sound event detection using 1D convolutional recurrent neural networks. Proc. of the Detection and Classification of Acoustic Scenes and Events, 2017: 80–84.
[16] CAKIR E, PARASCANDOLO G, HEOTTOLA T, et al. Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. on Audio, Speech, Language Processing, 2017, 25(6): 1291–1303.
[17] ADAVANNE S, VIRTANEN T. A report on sound event detection with different binaural features. Proc. of the Detection and Classification Acoustic Scenes and Events, 2017. DOI: 10.1109/ICASSP.2017.7952260.
[18] ADAVANNE S, PERTILA P, VIRTANEN T. Sound event detection using spatial features and convolutional recurrent neural network. Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2017: 771–775.
[19] BUTKO T, PLA F G, SEGURA C, et al. Two-source acoustic event detection and localization: online implementation in a smart-room. Proc. of the 19th Eurpean Signal Processing Conference, 2011: 1317–1321.
[20] SCHMIDT R O. Multiple emitter location and signal parameter estimation. IEEE Trans. on Antennas and Propagation, 1986, 34(3): 276–280.
[21] DIBIASE J H, SILVERMAN H F, BRANDSTEIN M S. Robust localization in reverberant rooms in microphone arrays. Microphone Arrays Signal Processing Techniques & Applications, 2001, 2: 157–180.
[22] POLITIS A, MESAROS A, ADAVANNE S, et al. Overview and evaluation of sound event localization and detection in DCASE 2019. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 684–698.
[23] CHAKRABORTY R, NADEU C. Sound-model-based acoustic source localization using distributed microphone arrays. Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2014: 619–623.
[24] MESAROS A. Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE/ACM Trans. on Audio, Speech and Language Processing, 2018, 26(2): 379–393.
[25] NGUYEN T N T, GAN W S, RANJAN R, et al. Robust source counting and DOA estimation using spatial pseudo-spectrum and convolutional neural network. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2020, 28: 2626–2637.
[26] ZHAO X Y, CHEN S W, ZHOU L, et al. Sound source localization based on SRP-PHAT spatial spectrum and deep neural network. Computers, Materials and Continua, 2020, 64(1): 253–271.
[27] CHAKRABARTY S, HABETS E A. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. IEEE Journal Selected Topics in Signal Processing, 2019, 13(1): 8–21.
[28] HIRVONEN T. Classification of spatial audio location and content using convolutional neural networks. Proc. of the Audio Engineering Society Convention, 2015: 9294.
[29] ADAVANNE S, POLITIS A, NIKUNEN J, et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(1): 34–48.
[30] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proc. of the International Conference on Machine Learning, 2015: 448–456.
[31] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023.
[32] ABADI M, AGARWAL A, BARHAM P, et al. TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/.
[33] MESAROS A, HEITTOLA T, VIRTANEN T. Metrics for polyphonic sound event detection. Applied Science, 2016, 6(6): 162.
[34] CAO Y, KONG Q X, IQBAL T, et al. Polyphonic sound event detection and localization using a two-stage strategy. Proc. of the Detection Classification Acoustic Scenes and Events, 2019: 30–34.

Biographies



ZHAO Dada was born in 1997. He received his B.S. degree in statistics from the School of Mathematical Sciences, Shanxi University, Taiyuan, China, in 2015. He is pursuing his M.S. degree in statistics with Xidian University. His research interests are signal processing and acoustic source localization.
E-mail: ddzhao@stu.xidian.edu.cn

DING Kai was born in 1983. He received his Ph.D. degree in weapon science and technology from Army Engineering University of the PLA in 2013. He is an engineer in the Science and Technology on Near-Surface Detection Laboratory. His current research interests include passive target recognition and intelligent network.
E-mail: winfast113@sina.com

QI Xiaogang was born in 1973. He is a professor and Ph.D. supervisor in the School of Mathematics and Statistics, Xidian University. He received his Ph.D. degree in applied mathematics from Xidian University in 2005 where he joined as a faculty member in 2002. He became an associate professor in 2006. From September 2012 to August 2013, he is a visiting scholar in the School of Electrical, Computer and Energy Engineering of Arizona State University. His research interests include system modeling and simulation, resource management and scheduling, performance evaluation and optimization algorithm design, and fault diagnosis in various networks.
E-mail: xgqi@xidian.edu.cn

CHEN Yu was born in 1980. He received his M.S. degree in physical electronics from National University of Defense Technology in 2005. He is an assistant researcher in the Science and Technology on Near-Surface Detection Laboratory. His current research interests include passive target recognition and intelligent network.
E-mail: cy0520tool@sohu.com

FENG Hailin was born in 1966. She received her B.S. degree in mathematics from Yan'an University, Yan'an, China, in 1988, and M.S. and Ph.D. degrees in applied mathematics from Xidian University, Xi'an, China, in 1991 and 2004, respectively. She is a professor in the School of Mathematics and Statistics, Xidian University. Her current research interests include system reliability modeling and survival data analysis.
E-mail: hlfeng@xidian.edu.cn