Mathematical Biosciences and Engineering

MBE, 21(2): 2004–2023.
DOI: 10.3934/mbe.2024089
Received: 13 August 2023
Revised: 22 December 2023
Accepted: 29 December 2023
Published: 08 January 2024

http://www.aimspress.com/journal/MBE

Research article

Polyphonic sound event localization and detection based on Multiple Attention Fusion ResNet

Shouming Zhang[1], Yaling Zhang[1,2], Yixiao Liao[2,*] Kunkun Pang[2], Zhiyong Wan[2] and Songbin Zhou[2]

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2] Institute of Intelligent Manufacturing, Guangdong Academy of Science, Guangdong Key Laboratory of Modern Control Technology, Guangzhou 510030, China

* **Correspondence:** Email: yx.liao@giim.ac.cn.

Abstract: Sound event localization and detection have been applied in various fields. Due to the polyphony and noise interference, it becomes challenging to accurately predict the sound event and their occurrence locations. Aiming at this problem, we propose a Multiple Attention Fusion ResNet, which uses ResNet34 as the base network. Given the situation that the sound duration is not fixed, and there are multiple polyphonic and noise, we introduce the Gated Channel Transform to enhance the residual basic block. This enables the model to capture contextual information, evaluate channel weights, and reduce the interference caused by polyphony and noise. Furthermore, Split Attention is introduced to the model for capturing cross-channel information, which enhances the ability to distinguish the polyphony. Finally, Coordinate Attention is introduced to the model so that the model can focus on both the channel information and spatial location information of sound events. Experiments were conducted on two different datasets, TAU-NIGENS Spatial Sound Events 2020, and TAU-NIGENS Spatial Sound Events 2021. The results demonstrate that the proposed model significantly outperforms state-of-the-art methods under multiple polyphonic and noise-directional interference environments and it achieves competitive performance under a single polyphonic environment.

Keywords: sound event localization and detection; attention; Gated Channel Transformation; deep learning

1. Introduction

Sound event localization and detection (SELD) is the combination of sound event detection (SED) and sound source localization (SSL), which can simultaneously predict the category and location of sound event. SED refers to the task of categorizing sound event, whereas SSL estimates the direction of sound sources. Currently, SELD plays a crucial role in improving the quality of life while ensuring health and safety [1], which has been applied in various fields [2]. For example, noise pollution can be reduced by monitoring the noises in life [3]. By analyzing the sounds made by animals, it is possible to monitor their health status [4] and categorize animals [5]. Considering medical treatment, the relevant lung diseases can be diagnosed by judging whether the anomaly is present in the breathing sounds [6], using the 1D convolutional network to automate COVID-19 disease diagnosis [7]. In industrial productions, the operation of facilities [8] is closely monitored.

The SELD was first introduced in the 2019 Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [9]. The traditional SED methods include the Gaussian mixture model (GMM), the hidden Markov model (HMM) [10], and so on. The traditional SSL methods are mostly premised on the time difference of arrival. For example, Wang et al. [11] proposed the time difference of arrival (TDoA) indoor localization technology. Liu et al. [12] proposed the TDoA and frequency differences of arrival measurements of the given planar stationary radiation source. Due to the lack of expert knowledge, the performance of traditional SED and SSL methods heavily relies on feature engineering techniques, whereas deep learning can extract potential information from raw signals without expert knowledge. Hayashi et al. [13] combined deep learning and signal analysis for SED. By modeling the temporal structure associated with sound events, sequence-to-sequence detection was performed without solving domain values. Zhu et al. [14] addressed the transmission loss problem caused by the traditional feature extraction method, which improve the performance of the model.

However, their improvement reaches a significant extent while the accuracy of single-SELD tasks is enhanced in the absence of interference. However, the challenges posed by noises and polyphony have not been fully addressed yet. In recent years, some deep-learning-based approaches have been proposed for SELD research conducted in multiple polyphonic environments. Adavanne et al. [15] proposed an End-to-End convolutional recurrent neural network (CRNN), which consists of two branches: a classification branch (SED) and a regression localization branch (SSL). The model can be applied to various array structures, which became the Baseline for SELD tasks in the DCASE Challenge in 2020, and 2021. Komatsu et al. [16] proposed to address the incomplete effectiveness of features extracted by convolutional neural networks (CNNs). They introduced a CRNN combined with a gated linear unit (GLU). GLU is used to weigh the importance of CNNs input, which enhances the extraction of effective features. To reduce model computation, Spoorthy et al. [17] replaced ordinary convolutions with depth-separable convolutions. However, this change presents a risk of information loss.

There are up to two polyphonic SELD tasks that have been well performed. However, in more complex polyphonic environments with directional interference from noises, the predictive performance of the model is significantly compromised. Kim et al. proposed an AD-YOLO [18] model based on the YOLO [19] framework, which was initially used in the image detection of multiple targets, for SELD. This adaptation enhances the ability of the model, which evidences the inadequacies in detecting small objects within the framework. In addition, this can result in the oversight of transient sound events in SELD.

Despite significant progress has been made in SELD under a single polyphonic environment, there often exist multiple polyphony (with two or more polyphony) in the real world. In addition, real-world SELD task is susceptible to the interference of noises. These problems cause a significant challenge for SELD.

To solve the above difficulties, polyphonic sound event localization and detection are proposed in this paper based on Multiple Attention Fusion ResNet. The major contributions are as follows:

1) By combining multiple attentions, the model can effectively model contextual, channel, and spatial information. Thus, the model is capable of discriminating polyphony while suppressing noise interference, and accurately detecting and localizing sound events.

2) Multiple Attention and residual networks are combined to enhance the ability to distinguish the polyphony.

2. Related work

To address the limitations of traditional methods relying on prior knowledge, Zhang et al. [20] proposed a novel approach that is reliant on CNNs to extract spectrograms of arbitrary lengths from audio recordings. Through CNNs, the relevant spatial features were extracted to enhance SSL performance. Additionally, the recurrent neural networks (RNNs) were leveraged to determine the temporal dependencies, which enables the integration of sequential information over time. This combined CNNs and RNNs approach, known as convolutional recurrent neural networks (CRNNs), was successfully applied to both SED and SSL tasks. Phan et al. [21] proposed that both SED and SSL are expressed as regression problems and Adavanne et al. [22] investigated the joint localization, detection, and tracking of sound events using CRNNs, which outperformed traditional methods [23]. Therefore, it is of profound significance to explore the combination of SED and SSL for SELD.

Nguyen et al. [24] proposed Salsa, which is a method to detect sound events and estimate their arrival directions separately before a deep neural network is trained to match SED and SSL for joint optimization. Politis et al. [25] proposed SELDnet to jointly detect sound events and estimate the location of the sound source. Cao et al. [26] experimentally proved that the trained SED model can improve SSL performance by weight sharing. The network layers increase resulted in gradient explosion or vanishing, which would deteriorate the overall performance. To address this issue, the ResNet [27] is widely applied in SELD [28]. Ranjan et al. [29] combined the ResNet with RNN to estimate SED and SSL labels jointly for sound events with one or two active sound sources.

To improve the performance of the model in a polyphonic environment, it is proposed in some research to perform the SELD task through various attention mechanisms. Among them is the squeeze-and-excitation (SE) network [30], in which the feature vectors of each channel are compressed through squeeze modules and then the feature representations of each channel are weighed through excitation modules. Huang et al. [31] combined the SE module with the ResNet in the SELD task to predict categories and locations in a polyphonic environment. However, the overall performance of the model remained unsatisfactory. Subsequently, Woo et al. [32] proposed the Convolutional Block Attention Module (CBAM) after the SE module, which added spatial attention based on channel-wise attention. Kim et al. [33] applied CBAM to SED tasks, which had higher computational complexity. However, there was only a slight improvement in the accuracy of recognition. Xu et al. [34] proposed the CECA model, which integrated an Efficient Channel Attention (ECA) [35] based on SE upgrade, incorporating it into residual blocks for use in SELD to capture channel-wise information in feature

maps. Although the application of the above attention mechanisms improves the performance of the model to a certain extent, ECA and SE are restricted to considering the importance of each channel. Also, CBAM is capable only of capturing local information, and long-range dependence information cannot be obtained. The predictive performance is slightly improved for the three types of attention modules. These constraints underscore the need for further research and the development of more robust attention mechanisms for SELD models.

In summary, despite significant progress made in SELD research, there remain challenges in daily life due to the presence of polyphony and noise interference. By achieving the simultaneous detection and localization of multiple sound events, it is achievable to distinguish between noise and non-noise events, which is worthy of further research.

3. Multiple Attention Fusion ResNet

The overall flow chart is shown in Figure 1. First, the features of the sound signal were extracted, which were sent to the proposed Multiple Attention Fusion ResNet (MAFR) for training. Then, the network outputs the sound category and the sound location. To enhance the performance of SELD in a polyphonic environment, this paper proposes the MAFR, the network structure of MAFR is shown in Figure 2.

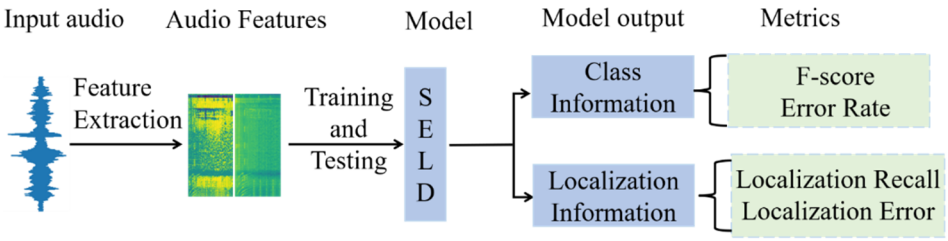

Figure 1. Overall flow chart of polyphonic sound event location and detection.

The network structure of MAFR is shown in Figure 2. A robust feature extraction capability serves as the cornerstone for SELD. While convolutions in the feature extraction process mainly attend to local information, SELD tasks demand consideration of both local and global features. In this paper, ResNet34 and Bidirectional Gated Recurrent Units (BIGRU) are selected as the basic network model. Gated Channel Transformation (GCT), Split Attention (SA), and Coordinate Attention (CA) are added to the basic residual block, where GCT is added to realize the effective modeling of sound context information and inter-channel information, SA is added to capture cross-channel information, and CA is added to enable the model focus both on channel and location information of the sound event. Through the effective combination of GCT, SA, and CA, the SELD performance in a polyphonic environment has been improved.

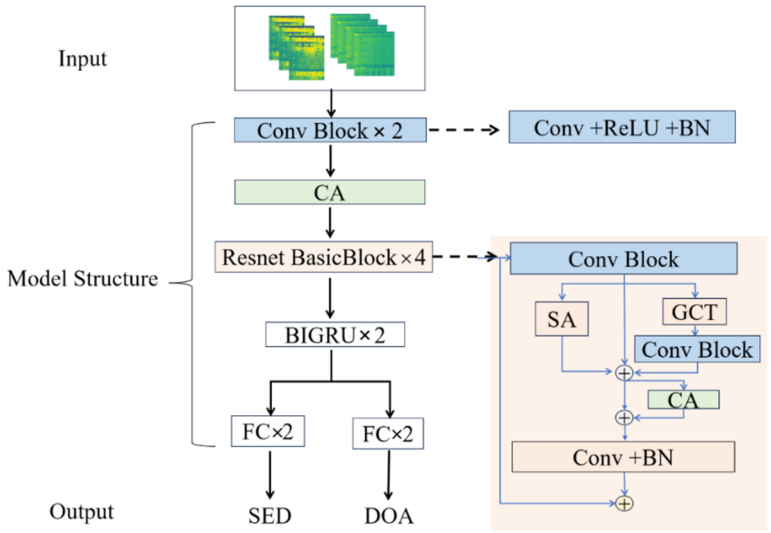

Figure 2. The overall structure of MAFR. Blue: Convolution block with kernel size 3, Green: Attention, grey: Basic residual block; CA: Coordinate Attention, SA: Split Attention, GCT: Gated Channel Transformation, Conv Block: Conv 3 × 3 + BN + ReLU, dashed arrows indicate the specific structure of a block.

3.1. Feature extraction

The Spatial cue-augmented log-spectrogram (Salsa) is used as the input feature for MAFR with a shape of [7, 200, 4800]. Where "7" represents the number of sound channels, "200" indicates the Mel-frequency range, and "4800" signifies that the audio is divided into 4800 segments. Salsa consists of two components: The log linear-frequency spectrogram of the first four-channel sound signals, and the normalized intensity vector of the spatial covariance matrix for the remaining three-channel sound signals.

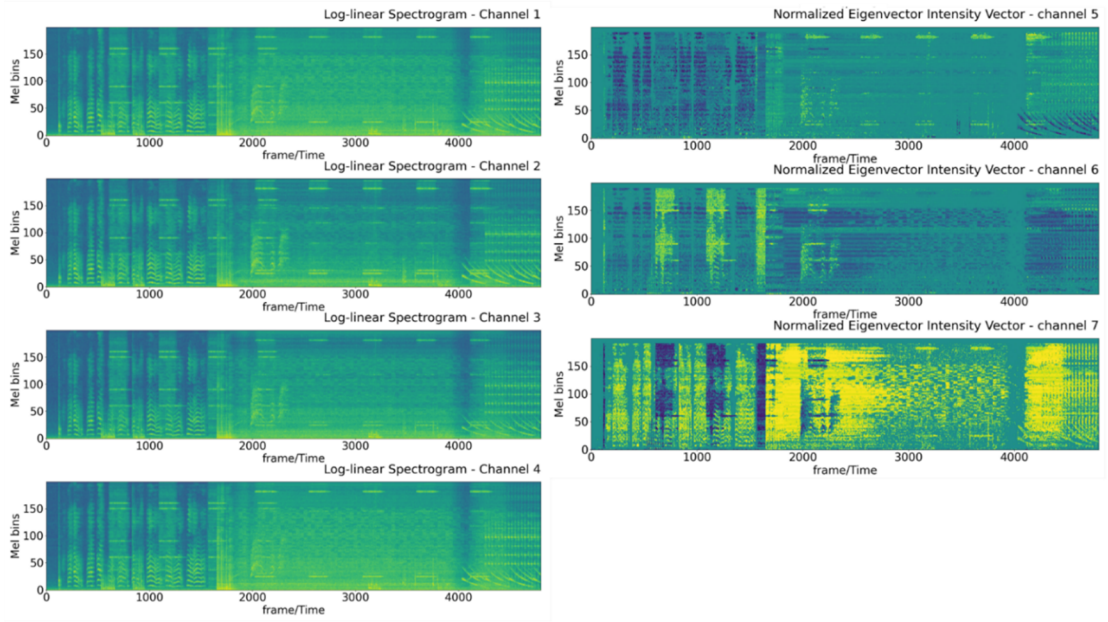

Figure 3. Salsa feature.

The log linear-frequency spectrogram contains information about the energy distribution of sound in frequency and time, which can help distinguish different sound events, and the normalized intensity vector of the spatial covariance matrix contains information such as inter-channel amplitude and phase differences, which facilitate source localization. In conclusion, Salsa contributes to the extraction of multi-channel features and the differentiation of overlapping sounds. The visualization of Salsa is presented in Figure 3.

3.2. Basic network architecture

This paper used ResNet as the basic network [36]. The optimized ResNet has 34 layers and mainly consists of convolutional blocks and basic residual blocks with skip connections. In the convolutional block, we replace the 7×7 kernel convolutional layer with two 3×3 kernel convolutional layers, aiming to improve the model's generalization ability. Specific modifications are depicted in Figure 4.

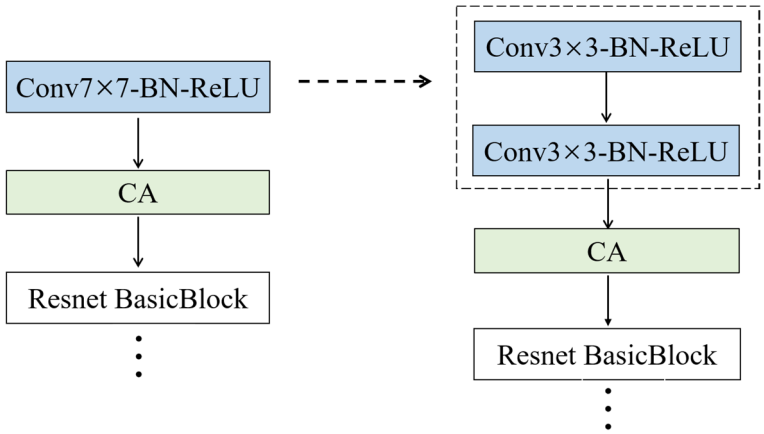

Figure 4. Structure diagram of partial changes in the residual network.

3.3. Gated Channel Transformation

In real-life scenarios, sound signals are complex, uncertain, and contain multiple polyphony. The presence of polyphony significantly impacts the performance of SELD. Inspired by Yang et al. [37], GCT enables more effective modeling of inter-channel and contextual information. In this paper, the GCT is positioned as shown in Figure 2.

The structure of the GCT module is shown in Figure 5. First, the global context information for each channel is aggregated and combined with the trainable parameter α_c, and the importance of different channels is controlled. When α_c approaches 0, the features of the *c-th* channel will not be propagated to the subsequent convolutional layers. Second, to reduce computational complexity, $l_2 - norm$ is used to establish competition among neurons. Finally, a gate mechanism is adopted to adapt the original features.

$$s_c = \alpha_c \left|| x_c \right||_2 \tag{1}$$

$$\hat{s}_c = f(C, s_c, \varepsilon) \tag{2}$$

$$\hat{x}_G = x_c[1 + \tanh(\gamma_c \hat{s}_c + \beta_c)] \tag{3}$$

where x_c is the corresponding input feature of the *c-th* channel, α_c is a trainable parameter, s_c is the output value after the global control of the context, ε is a constant close to 0 and is used to avoid the inverse being 0, C is the corresponding channel, $f(\cdot)$ denotes channel normalization operation, β_c and γ_c are trainable parameters, and $\hat{x}_c$ is the output of the *c-th* channel after GCT processing, $\hat{x}_G$ is the output transformed by the entire GCT module.

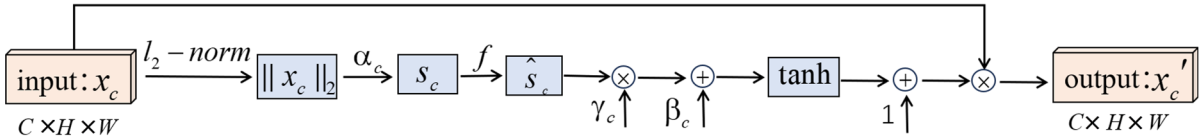

Figure 5. Gated Channel Transformation module.

By combining normalization and the gating mechanism, we model the competitive or cooperative relationships between different channels. When the specific channel's gating weight is activated (non-noise) and competes with features from other channels, it emphasizes that channel's features, leading to the attenuation of other channels and the reduction of noise interference. When polyphony occurs, the channels cooperate, combining multi-channel information for prediction, enabling polyphonic differentiation.

A richer representation of sound features can be achieved with GCT, which determines whether or not to pass the information of that channel to the convolutional layer. In addition, noisy channels and segments are suppressed, the ability to distinguish between different polyphony is enhanced, and directional interference is reduced, thus improving the predictive performance of the model.

3.4. Split Attention Module

Sound in complex environments is subject to noise and other interference. To accurately realize SELD, it is crucial to distinguish polyphony. SA combines the channel-wise attention strategy with a multi-path network layout, which can capture cross-channel feature correlations, in sound signals to enhance the ability of distinguishing polyphony. This paper incorporates the SA [38] module into the residual basic blocks. First, the input features are divided into *K* groups, and then each group is divided into *R* subgroups, a total of $G = KR$ subgroups of features are obtained, the weights of *K* groups weights were calculated and the corresponding features are fused. The *R* groups' intermediate features are obtained through this splitting transformation. The structure of the Split Attention Module is shown in Figure 6.

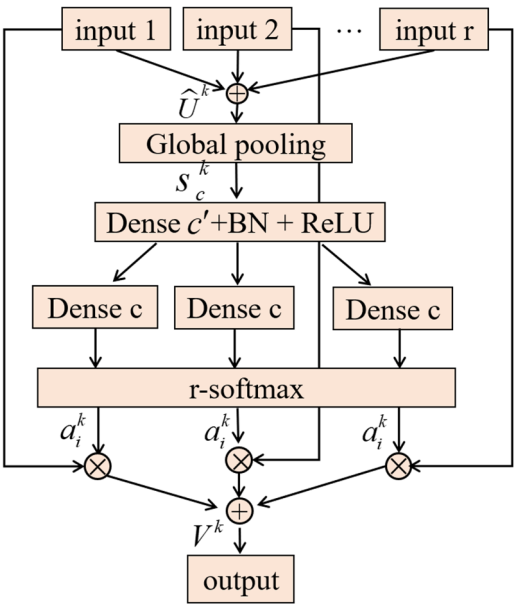

Figure 6. *k-th* subgroups Split Attention Module.

The R groups' intermediate features s_c^k are summed up and grasp the global context information by globally pooling. The assignment of weights $\alpha_i^k(c)$ between feature channels are performed according to the activation function, and the corresponding intermediate features are interacted to generate the corresponding weights V_c^k for each group.

$$V_c^k = \sum_{i=1}^{R} \alpha_i^k(c) U_{R(k-1)+i} \tag{4}$$

$$V = concat\{V^1, V^2, \cdots V^K\} \tag{5}$$

where $U_{R(k-1)+i}$ is the k-th input, where $\alpha_i^k(c)$ denotes inter-channel weights in the k group $concat\{\cdot\}$ denotes concatenated along the channel dimension.

$$\hat{x}_c = V + x \tag{6}$$

In a polyphonic environment, by splitting and combining the features along the channel dimension, corresponding channel weights V_c^k are obtained to measure the importance of each channel. This process enables the separation of various sound events occurring simultaneously. The SA module is placed alongside the GCT convolutional layer, and the specific integration of the modules is depicted in Figure 6. The outputs of the SA, the first convolutional layer, and the outputs through GCT and the second convolutional layer are summed together, as described in Eq (7).

$$\hat{x} = x + \hat{x}_s + \hat{x}_G \tag{7}$$

where x denotes the input raw features, $\hat{x}_s$ denotes the features after passing through the SA module, $\hat{x}_G$ denotes the features after passing through the GCT module.

3.5. Coordinate Attention Module

The sound features are further processed by both the SA module and the GCT module, which suppresses noises and distinguishes overlapping sounds. However, there exists an issue of incomplete separation of overlapping sounds.

To enhance the model's representation capability and balance the importance of each channel feature, we adopt the CA [39]. CA not only balances the importance of various channels but also captures favorable spatial feature information. The structure of the CA module is illustrated in Figure 7.

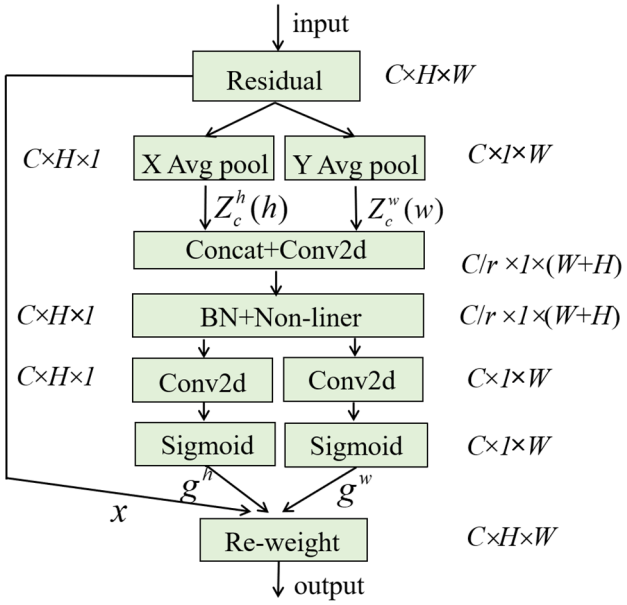

Figure 7. Structure diagram of Coordinate Attention module.

The CA decomposes the channel attention into two one-dimensional feature encoding processes, aggregating features along two spatial directions. In the time domain, it captures long-range dependencies ($Z_c^h(h)$), while in the frequency domain, it preserves precise positional information ($Z_c^w(w)$). The feature map is encoded into a pair of attention maps that are direction-aware and position-sensitive, thereby forming feature maps with specific directional information and accurately highlighting regions of interest. To fully utilize the extracted positional information and handle the inter-channel relationships, the CA module concatenates the outputs from two directions after two pooling layers. The dimension is consistent with the input dimension through the convolution layer with convolution kernel 1, and the output weights in both directions are as follows.

$$g^h = \sigma\left(F_h(f^h)\right) \tag{8}$$

$$g^w = \sigma\left(F_w(f^w)\right) \tag{9}$$

where f^h and f^w denote the tensors in two spatial directions, F_w and F_h are both convolution transformations with convolution kernel 1, and σ is the sigmoid function.

By weighting the horizontal and vertical directions together, the final feature contains inter-channel information, horizontal spatial information, and vertical spatial information.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i,j) \times g_c^w(i,j) \tag{10}$$

where $g_c^h(i,j)$ and $g_c^w(i,j)$ denote the feature weights through the horizontal and vertical directions, respectively, $x_c(i,j)$ denotes the input to the CA, and $y_c(i,j)$ denotes the output through the CA module, where $i \in (0,W)$, $j \in (0,H)$, W and H are the width and height of the feature map, respectively.

The spatial information is complemented with the channel information and applied to the input feature map to enhance the representation of the object of interest. In a polyphonic environment, multiple sound events may occur simultaneously. By combining spatial information in both horizontal and vertical directions, the focus is on local positions to determine if they contain interference. This helps reduce the probability of erroneous predictions in a polyphonic environment. In this proposed method, the CA is integrated even in shallow layers of the network, as shown in Figure 2. Applying CA in the shallow layers allows for the initial learning of both channel information and positional information. This leads to the preliminary reduction in noise interference and separates overlapping sound events. Moreover, the CA module is embedded after the GCT and SA modules within the basic residual blocks.

4. Dataset and experimental parameters

4.1. Dataset

To verify the model's generalization ability in a polyphonic environment, we conducted experiments using the TAU-NIGENS Spatial Sound Events 2020 [40] and TAU-NIGENS Spatial Sound Events 2021 datasets [41]. The sound events included different categories of sounds such as alarms, dog barks, etc. There were 14 sound categories in the 2020 dataset and 12 sound categories in the 2021 dataset. The official data provided two different acquisition methods: 4-channel microphone acoustics and first-order ambisonics acoustics. We used the first-order ambisonics audio format for model training and evaluation. Any sound events in the data that are different from the specified categories will be considered interfering noises, including sounds of the running engine and the burning fire.

Both datasets have a common characteristic of being one-minute long and non-continuous; the differences between them include the number of polyphony, the presence of directional interference, and event durations. Relevant dataset details are provided in Table 1, where "long" and "short" represent whether the duration of a single sound event is long or short.

1) Polyphony: In the former dataset, there is typically one sound event per second, with a maximum of two overlapping sound events at the same moment. In contrast, the latter dataset may have two to three different sound events occurring simultaneously at a given moment.

2) Interfering noises: The former has no directional noise interference.

3) Duration: The sound events in the former dataset have relatively longer durations compared to the sound events in the latter dataset, which tend to have shorter durations.

Data visualization is shown in Figures 8 and 9. Top to bottom, they are Waveform, Spectrogram, and Mel Spectrogram, and the fourth subgraph represents sound events occurring within the 60 s, the length

of the short horizontal line represents the duration, and the color represents the sound category. "2020" in Table 1 represents the TAU-NIGENS Spatial Sound Events 2020 dataset, and "2021" in Table 1 represents the TAU-NIGENS Spatial Sound Events 2021 dataset.

Table 1. Detailed information about the two datasets.

Dataset name	Audio duration	sample number	Total number of events	Number of sound events occurring at the same time			Duration	Directional interferers
				1	2	3		
2020	60 s	600	267,855	181,412	86,443	0	Long	None
2021	60 s	600	302,119	125,303	**123,808**	**53,008**	**Short**	**Exist**

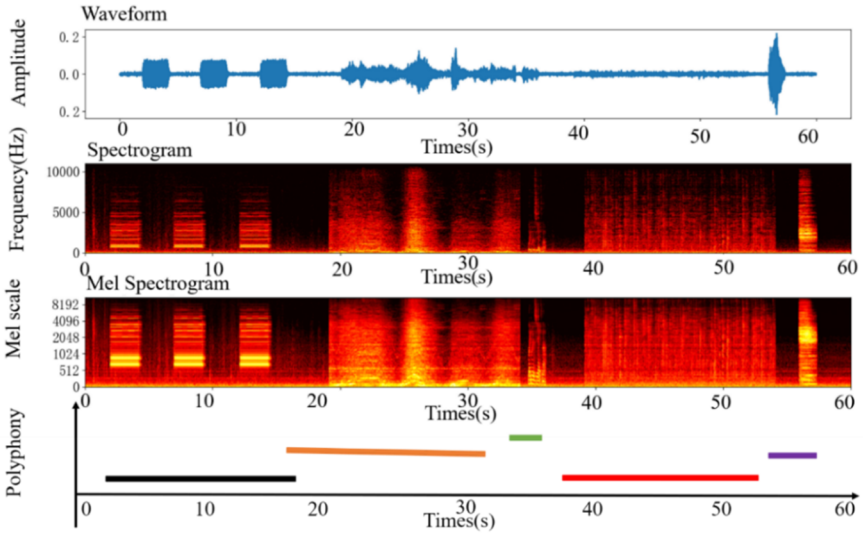

Figure 8. Visualization of TAU-NIGENS Spatial Sound Events 2020.

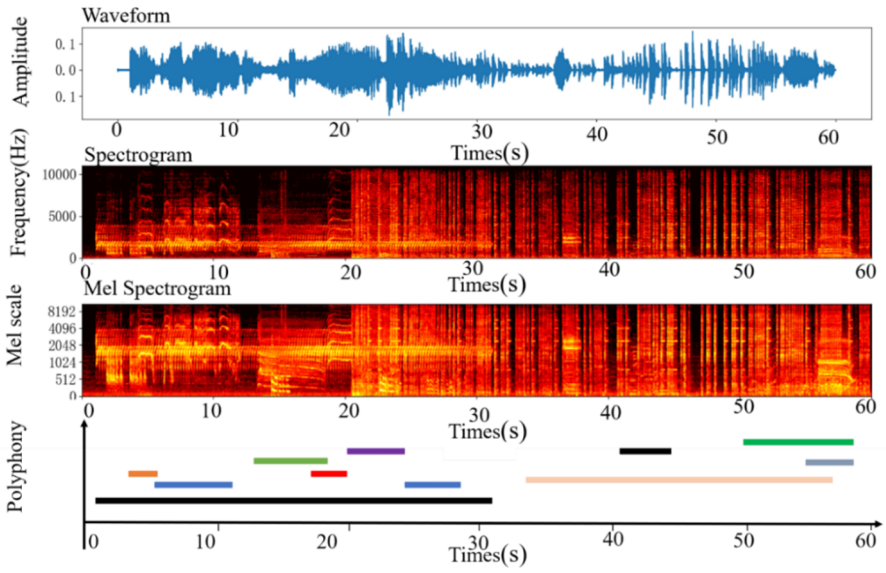

Figure 9. Visualization of TAU-NIGENS Spatial Sound Events 2021.

4.2. Experimental process and parameter design

In this experiment, the Adam optimizer is used for model training, with a learning rate of 0.0003 for the first 70% of epochs and 0.0001 for the remaining 30% of epochs. The batch size is 16, and the epochs are 90.

The comparative methods, Baseline, AD-YOLO, Salsa, and CECA, are all intended for DCASE. The parameters in the training process of the comparative methods are consistent with the training parameters in the corresponding papers. All the results of comparative methods are obtained from the corresponding paper. It should be noted that the experiments of the CECA model are conducted only on the TAU-NIGENS Spatial Sound Events 2021 dataset in the original paper. In this paper, the CECA model is reproduced and conducted on the TAU-NIGENS Spatial Sound Events 2020 dataset to get corresponding results, with the same parameters provided in the corresponding paper.

4.3. Evaluation metrics

Multiple evaluation metrics can better evaluate the model performance [42]. In this paper, F_1 Score (*F1-Score*), Error Rate (*ER*) Localization Frame Recall (*LR*), and Localization Error (*LE*) are used to evaluate the SELD performance of models [43], F_1 and *ER* for the detection task and *LR* and *LE* for the localization task. The calculation formulas are as follows: Eqs (11) to (14).

$$F_1 = \frac{2\sum_{k=1}^{k} TP(k)}{2\sum_{k=1}^{k} TP(k)+\sum_{k=1}^{k} FP(k)+\sum_{k=1}^{k} FN(k)} \tag{11}$$

$$ER = \frac{\sum_{k=1}^{k} FP(k)+\sum_{k=1}^{k} FN(k)}{\sum_{k=1}^{k} TP(k)+\sum_{k=1}^{k} FP(k)+\sum_{k=1}^{k} TN(k)+\sum_{k=1}^{k} FN(k)} \tag{12}$$

where True Positive (*TP*) refers to the number of correctly identified positive samples, False Positive (*FP*) refers to the number of incorrectly identified positive samples, and False Negative (*FN*) refers to the number of incorrectly identified negative samples.

$$LE = \frac{1}{\sum_{k=1}^{K} D_p^k}\sum_{k=1}^{K} H\left(DOA_R^k, DOA_P^k\right) \tag{13}$$

where K denotes the frame length, DOA_R^k denotes the location true angle of the sound event, DOA_P^k denotes the location prediction angle of the sound event, and D_P^K denotes the total number of angles DOA_R^k at the *k-th* moment, $H(\cdot)$ is the Hungarian algorithm.

$$LR = \frac{\sum_{k}^{k} 1\left(D_R^K = D_P^K\right)}{K} \tag{14}$$

where D_R^K denotes the total number of angles DOA_R^k at the *k-th* frame, D_P^K denotes the total number of angles DOA_P^k at the *k-th frame*. If $D_R^K = D_P^K$ then $1(D_R^K = D_P^K) = 1$, K denotes all the moments. The subscript R denotes the true value and the subscript P denotes the predicted value.

The SED and SSL evaluation metrics are combined to assess the overall model performance using the value indicated in Eq (15).

$$SELD = \frac{ER+(1-F_1)+(1-LR)+\frac{LE}{180^\circ}}{4} \tag{15}$$

When F_1 approaches 1, and *ER* approaches 0, it indicates more accurate sound event category predictions. When *LR* approaches 1 and *LE* approaches 0, it signifies more accurate event localization predictions. it indicates the better overall performance of SELD.

5. Analysis of experimental results

5.1. Comparative experiment

To validate the effectiveness of the proposed MAFR in a polyphonic environment, we conducted comparative experiments with Baseline, AD-YOLO, Salsa, and CECA methods. The Baseline is the official DCASE baseline for the SELD task in 2020 and 2021. The AD- YOLO model utilizes the YOLO network architecture for SELD tasks. Salsa combines pannresnet [44] and BIGRU as the network architecture. CECA is an improvement over Salsa, adding CA and ECA modules, and adopting the L1 loss function as the loss function. The results of the comparative experiments can be found in Tables 2 and 3.

In the following tables (from Tables 2 to 7), the up arrows denote that larger values indicate better model performance in the corresponding columns; conversely, the down arrows denote that smaller values indicate better model performance in the corresponding columns.

Table 2. Comparative experimental results of TAU-NIGENS Spatial Sound Events 2020.

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
Baseline [15]	0.720	37.40%	22.80°	60.70%	0.466
AD-YOLO [18]	0.482	61.27%	8.60°	69.75%	0.305
Salsa [24]	0.338	**74.80%**	**7.90°**	78.40%	0.226
CECA [34]	0.372	73.40%	8.57°	78.20%	0.225
MAFR	**0.336**	**74.80%**	8.05°	**78.65%**	**0.212**

Table 3. Comparative experimental results of TAU-NIGENS Spatial Sound Events 2021.

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
Baseline [15]	0.690	33.90%	24.10°	43.90%	0.690
AD-YOLO [18]	0.519	54.35%	13.54°	64.70%	0.351
Salsa [24]	0.404	72.40%	12.51°	72.70%	0.255
CECA [34]	0.393	72.00%	**11.71°**	72.80%	0.253
MAFR	**0.369**	**73.53%**	13.85°	**74.91%**	**0.240**

Table 2 shows the experimental results of the TAU-NIGENS Spatial Sound Events 2020 dataset. MAFR has shown the best overall performance among these methods, where *ER* is 0.336, *LE* is 8.05°, F_1 is 74.8%, and *LR* is 78.65%. Specifically, MAFR shows significant improvement in F_1 and *LR* (13.53% and 8.9%) compared with AD-YOLO. MAFR increases by 0.25% in *LR* compared with Salsa and achieves an increase in *ER* by 0.036 compared with CECA. These demonstrated that the MAFR is superior in SELD tasks when one or two sound events occur at the same time.

Table 3 shows the experimental results of the TAU-NIGENS Spatial Sound Events 2021 dataset. MAFR had gain best overall results, where *ER* is 0.369, *LE* is 13.85°, F_1 is 73.53%, and *LR* is 74.91%. MAFR achieved the best performance in *ER*, F_1, and *LR*. In particular, MARF shows tremendous

performance improvement compared with AD-YOLO, in which *ER* decreased by 0.15, *SELD* decreased by 0.111, F_1 increased by 19.18%, and *LR* increased by 10.21%. Compared with CECA, GCT, and SA modules added to MAFR, F_1 increased by 1.53% and *LR* increased by 2.21%. It shows that the GCT and SA in MAFR are effective for SELD in multiple polyphonic. Compared to Salsa, the proposed MAFR has the same F_1 as Salsa's model in the TAU-NIGENS Spatial Sound Events 2020 dataset; in the TAU-NIGENS Spatial Sound Events 2021 dataset, the proposed MAFR has an increase in F_1 by 1.13%. From the experimental results, it can be seen that the model shows more significant improvement in the TAU-NIGENS Spatial Sound Events 2021 dataset than in the TAU-NIGENS Spatial Sound Events 2020 dataset when compared with the state-of-the-art methods. In conclusion, the proposed MARF outperforms the state-of-the-art methods in SELD tasks under multiple polyphony environments with directional interference.

5.2. Analysis of the CA module

To verify the effectiveness of the CA module, we use the optimized ResNet34 (Section 3.1) as the basis network. Four attention mechanisms (SE, CBAM, ECA, and CA) are combined with it for SELD. Each attention module is placed at the position of the green box in Figure 2, and the experimental results are presented in Tables 4 and 5. In the table header, "2020" represents the TAU-NIGENS Spatial Sound Events 2020 dataset, and "2021" represents the TAU-NIGENS Spatial Sound Events 2021 dataset.

Table 4. Comparison of the effects of four different attention models (2020).

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
None	0.375	71.66%	8.85°	76.60%	0.235
CBAM	0.362	72.84%	8.98°	77.70%	0.227
SE	0.364	72.95%	**8.32°**	77.29%	0.227
ECA	0.375	72.05%	9.29°	77.04%	0238
CA	**0.350**	**74.39%**	8.88°	**78.58%**	**0.217**

Table 5. Comparison of the effects of four different attention models (2021).

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
None	0.422	69.46%	**14.23°**	**72.66%**	0.270
CBAM	0.407	69.93%	14.90°	71.87%	0.268
SE	0.403	70.52%	14.86°	73.23%	**0.261**
ECA	0.410	69.45%	15.42°	72.73%	0.268
CA	**0.394**	**70.92%**	14.84°	72.40%	**0.261**

The results show that compared with the model without adding the attention module, the model performance is improved after adding CA, CBAM, SE, and ECA, respectively. CA can both focus on channel information and capture the direction perception of each sound source under multiple polyphonic environments. In the TAU-NIGENS Spatial Sound Events 2020 dataset, compared with not adding attention, F_1 increased by 2.73% and *LR* increased by 2% when adding CA. In the TAU-NIGENS Spatial Sound Events 2021 dataset, compared with not adding attention, F_1 increased by 1.46% after adding CA, and the performance of adding the CA module is optimal. The results show that only a

small performance improvement is achieved in the SELD task when SE or ECA is added. Since SE and ECA modules only focus on the channel information and cannot completely extract all the useful information. CBAM is unable to obtain the global spatial information, and experiments show that the CABM module is not suitable for this task. In summary, in the SELD task, the use of the CA module enables the whole network to better distinguish polyphony and reduce noise interference by comprehensively learning the channel information and location information.

5.3. Ablation study

To verify the impact of each module in the proposed MAFR, the following ablation experiments are conducted. The basic network structure is called ResNet34-Bigru, abbreviated as RB (7 × 7). To replace the convolutions in RB (7 × 7) with two convolutional layers using a kernel size of 3 × 3, the resulting network is referred to RB. SA, CA, and GCT modules are introduced to RB to form RBS, RBC, and RBG, respectively. Then, the CA module is incorporated into RBS to get RBSC. Finally, the proposed MAFR is constructed by integrating GCT into RBSC. To assess the effectiveness of all proposed modules, ablation experiments are conducted. The results are presented in Tables 6 and 7.

Table 6. TAU-NIGENS Spatial Sound Events 2020 ablation experiment.

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
RB (7 × 7)	0.459	67.00%	14.58°	72.10%	0.287
RB	0.375	71.66%	8.85°	76.60%	0.235
RBS	0.367	72.32%	9.06°	77.05%	0.231
RBG	0.345	74.56%	8.60°	78.89%	0.215
RBC	0.349	74.80%	8.38°	78.23%	0.216
RBSG	0.347	74.80%	8.31°	78.31%	0.214
MAFR	**0.336**	**74.80%**	**8.05°**	**78.65%**	**0.212**

Table 7. TAU-NIGENS Spatial Sound Events 2021 ablation experiment.

Method	*ER* ↓	F_1 ↑	*LE* ↓	*LR* ↑	*SELD* ↓
RB (7 × 7)	0.457	66.80%	15.72°	71.00%	0.291
RB	0.422	69.46%	14.23°	72.66%	0.270
RBS	0.412	69.80%	15.43°	72.30%	0.269
RBG	0.391	71.84%	14.28°	73.93%	0.253
RBC	0.396	70.93%	15.06°	73.31%	0.259
RBSG	0.394	71.50%	14.90°	73.50%	0.255
MAFR	**0.369**	**73.53%**	**13.85°**	**74.91%**	**0.240**

Table 6 shows that RB outperformed RB (7 × 7), where *ER* and *LE* decreased by 0.084 and 5.73°, respectively; and F_1 and *LR* increased by 4.66% and 4.5%, respectively. The comparison indicates that using multiple layers with smaller convolutional kernels to replace one layer with a large convolutional kernel can enhance the ability to extract key sound features. Compared to RB, RBC reduced the *ER* and *LE* by 0.026 and 0.47°, respectively, while increasing the F_1 and *LR* by approximately 3.14% and 0.65%, respectively. It demonstrated that CA is helpful in SELD tasks. Compared to RB, the F_1 and *LR* of RBS improved by 0.66% and 0.45%, respectively; and the F_1 and *LR* of RBG improved

by 2.9% and 2.29%, respectively. RBS and RBG both gain performance improvement compared with RB. Additionally, when CA, SA, and GCT were simultaneously integrated into RB, the *ER*, and *LE* were reduced by 0.011 and 0.26°, respectively, while the F_1 remained unchanged, and the *LR* increased by 0.34%.

Table 7 shows that all modules are effective in SELD under a complex polyphonic environment. Compared to RB (7 × 7), the RB model reduced the *ER* by 0.035 and increased the F_1 and *LR* by 2.66% and 1.66%, respectively. Compared with RB, the RBC model reduced the *ER* by 0.026 and increased the F_1 and *LR* by 1.47% and 1.63%, respectively; while the performance of RBS and RBG also showed improvement. Adding both CA and SA to RB has better performance than adding only one of the modules. It is worth noting that when SA is added to RBC, the model performance improves more significantly in the environment with multiple polyphony and noise directional interference than in the environment with one or two polyphony and no noise directional interference. When CA, SA, and GCT were added to RB, the *ER* and *LE* were 0.369, 13.85°, F_1 and *LR* were 73.53% and 74.91%, respectively. MAFR showed the best performance.

In summary, when the duration of sound events is short and there are multiple overlapping sound and noise directional interferences, SA can increase the receptive field of the whole model; GCT can connect the context, which reduces the number of mispredictions of brief sound events and weakens the interference of polyphonic and noises on the model. Through CA, the channel information and direction and position information of sound features can be captured. Furthermore, the effective combination of SA, GCT, and CA with ResNet34 can further improve the performance of the SELD task.

6. Conclusions

Due to the interference of polyphony and noise, accurately predicting the sound event category and the occurrence locations in SELD becomes challenging. In this paper, we propose the MAFR to achieve satisfactory performance in SELD task under multiple polyphonic environments with noise-induced interference, which combines GCT, SA, and CA with ResNet34. Experimental results demonstrate that GCT selectively captures spatial features, enabling a better extraction of global information. SA increases the model's receptive field, facilitating joint learning of cross-channel features. CA enhances feature complementarity between channel and positional features. On dataset TAU-NIGENS Spatial Sound Events 2020, MAFR achieved *ER* and *LE* of 0.336 and 8.05°, respectively; F_1 and *LR* of 74.80% and 78.65%, respectively. On dataset TAU-NIGENS Spatial Sound Events 2021, MAFR achieved *ER* and *LE* of 0.369 and 13.85°, respectively; F_1 and *LR* of 73.53% and 74.91%, respectively.

In summary, in multiple polyphonic environments with noise-induced interference, MAFR significantly outperforms the state-of-the-art methods in terms of comprehensive performance in SELD task. Moreover, the MAFR also shows competitive performance compared to the state-of-the-art methods in a single polyphonic environment. For future work, we would like to focus on the SELD task with moving sound sources.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgments

The authors are grateful to the anonymous referees for their careful reading, valuable comments, and helpful suggestions, which have contributed to improving the presentation of this work significantly. This work was supported by the Key-Area and Development Program of Guangdong Province (Grant number: 2019B010154002), Guangdong Basic and Applied Basic Research Foundation (Grant number: 2022A1515011559), Key-Area and Development Program of Dongguan (Grant number: 20201200300062), and GDAS' Project of Science and Technology Development (Grant number: 2022 GDASZH 2022010108).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. T. K. Chan, C. S. Chin, A comprehensive review of polyphonic sound event detection, *IEEE Access*, **8** (2020), 103339–103373. https://doi.org/10.1109/ACCESS.2020.2999388
2. A. Mesaros, T. Heittola, T. Virtanen, M. D. Plumbley, Sound event detection: A tutorial, *IEEE Signal Process Mag.*, **38** (2021), 67–83. https://doi.org/10.1109/MSP.2021.3090678
3. J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, et al., Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution, *Commun. ACM*, **62** (2019), 68–77. https://doi.org/10.1145/3224204
4. T. Hu, C. Zhang, B. Cheng, X. P. Wu, Research on abnormal audio event detection based on convolutional neural network (in Chinese), *J. Signal Process.*, **34** (2018), 357–367. https://doi.org/10.16798/j.issn.1003-0530.2018.03.013
5. D. Stowell, M. Wood, Y. Stylianou, H. Glotin, Bird detection in audio: A survey and a challenge, in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, (2016), 1–6. https://doi.org/10.1109/MLSP.2016.7738875
6. K. K. Lell, A. Pja, Automatic COVID-19 disease diagnosis using 1D convolutional neural network and augmentation with human respiratory sound based on parameters: Cough, breath, and voice, *AIMS Public Health*, **8** (2021), 240. https://doi.org/10.3934/publichealth.2021019
7. K. K. Lella, A. Pja, Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: Cough, voice, and breath, *Alexandria Eng. J.*, **61** (2022), 1319–1334. https://doi.org/10.1016/j.aej.2021.06.024
8. G. Chen, M. Liu, J. Chen, Frequency-temporal-logic-based bearing fault diagnosis and fault interpretation using Bayesian optimization with Bayesian neural network, *Mech. Syst. Signal Process.*, **145** (2020), 1–21. https://doi.org/10.1016/j.ymssp.2020.106951
9. S. Adavanne, A. Politis, T. Virtanen, A multi-room reverberant dataset for sound event localization and detection, preprint, arXiv:1905.08546.
10. S. R. Eddy, What is a hidden Markov model, *Nat. Biotechnol.*, **22** (2004), 1315–1316. https://doi.org/10.1038/nbt1004-1315

11. J. Wang, S. Sun, Y. Ning, M. Zhang, W. Pang, Ultrasonic TDoA indoor localization based on Piezoelectric Micromachined Ultrasonic Transducers, in *2021 IEEE International Ultrasonics Symposium (IUS)*, (2021), 1–3. https://doi.org/10.1109/IUS52206.2021.9593813
12. C. Liu, J. Yun, J. Su, Direct solution for fixed source location using well-posed TDOA and FDOA measurements, *J. Syst. Eng. Electron.*, **31** (2020), 666–673. https://doi.org/10.23919/JSEE.2020.000042
13. T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, K. Takeda, BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection, in *2017 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, (2017), 766–770. https://doi.org/10.1109/ICASSP.2017.7952259
14. H. Zhu, H. Wan, Single sound source localization using convolutional neural networks trained with spiral source, in *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*, (2020), 720–724. https://doi.org/10.1109/CACRE50138.2020.9230056
15. S. Adavanne, A. Politis, J. Nikunen, T. Virtanen, Sound event localization and detection of overlapping sources using convolutional recurrent neural networks, *IEEE J. Sel. Top. Signal Process.*, **13** (2019), 34–48. https://doi.org/10.1109/JSTSP.2018.2885636
16. T. Komatsu, M. Togami, T. Takahashi, Sound event localization and detection using convolutional recurrent neural networks and gated linear units, in *2020 28th European Signal Processing Conference (EUSIPCO)*, (2021), 41–45. https://doi.org/10.23919/Eusipco47968.2020.9287372
17. V. Spoorthy, S. G. Koolagudi, A transpose-SELDNet for polyphonic sound event localization and detection, in *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, (2023), 1–6. https://doi.org/10.1109/I2CT57861.2023.10126251
18. J. S. Kim, H. J. Park, W. Shin, S. W. Han, AD-YOLO: You look only once in training multiple sound event localization and detection, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, (2023), 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096460
19. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91
20. H. Zhang, I. McLoughlin, Y. Song, Robust sound event recognition using convolutional neural networks, in *2015 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, (2015), 559–563. https://doi.org/10.1109/ICASSP.2015.7178031
21. H. Phan, L. Pham, P. Koch, N. Q. K. Duong, I. McLoughlin, A. Mertins, On multitask loss function for audio event detection and localization, preprint, arXiv:2009.05527.
22. S. Adavanne, A. Politi, T. Virtanen, Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network, preprint, arXiv:1904.12769.
23. Z. X. Han, Research on robot sound source localization method based on beamforming (in Chinese), *Nanjing Univ. Inf. Sci. Technol.*, **2022** (2022). https://doi.org/10.27248/d.cnki.gnjqc.2021.000637
24. T. N. T. Nguyen, K. N. Watcharasupat, N. K. Nguyen, D. L. Jones, W. S. Gan, SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **30** (2022), 1749–1762. https://doi.org/10.1109/TASLP.2022.3173054

25. A. Politis, A. Mesaros, S. Adavanne, T. Heittola, T. Virtanen, Overview and evaluation of sound event localization and detection in DCASE 2019, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 684–698. https://doi.org/10.1109/TASLP.2020.3047233
26. Y. Cao, Q. Kong, T. Iqbal, F. An, W. Wang, M. D. Plumbley, Polyphonic sound event detection and localization using a two-stage strategy, preprint, arXiv:1905.00268.
27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
28. J. Naranjo-Alcazar, S. Perez-Castanos, J. Ferrandis, P. Zuccarello, M. Cobos, Sound event localization and detection using squeeze-excitation residual CNNs, preprint, arXiv:2006.14436.
29. R. Ranjan, S. Jayabalan, T. Nguyen, W. Gan, Sound event detection and direction of arrival estimation using Residual Net and recurrent neural networks, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, (2019), 214–218. https://doi.org/10.33682/93dp-f064
30. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372
31. D. L. Huang, R. F. Perez, Sseldnet: A fully end-to-end sample-level framework for sound event localization and detection, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, (2021), 1–5.
32. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: Convolutional Block Attention Module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
33. J. W. Kim, G. W. Lee, C. S. Park, H. K. Kim, Sound event detection using EfficientNet-B2 with an attentional pyramid network, in *2023 IEEE International Conference on Consumer Electronics (ICCE)*, (2023), 1–2. https://doi.org/10.1109/ICCE56470.2023.10043590
34. C. Xu, H. Liu, Y. Min, Y. Zhen, Sound event localization and detection based on dual attention (in Chinese), *Comput. Eng. Appl.*, **2022** (2022), 1–11.
35. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 11534–11542. https://doi.org/10.1109/CVPR42600.2020.01155
36. J. Jia, M. Sun, G. Wu, W. Qiu, W. G. Qiu, DeepDN_iGlu: Prediction of lysine glutarylation sites based on attention residual learning method and DenseNet, *Math. Biosci. Eng.*, **20** (2023), 2815–2830. https://doi.org/10.3934/mbe.2023132
37. Z. Yang, L. Zhu, Y. Wu, Y. Yang, Gated Channel Transformation for visual recognition, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11791–11800. https://doi.org/10.1109/CVPR42600.2020.01181
38. H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, et al., ResNeSt: Split-attention networks, in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2022), 2735–2745. https://doi.org/10.1109/CVPRW56347.2022.00309
39. Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 13708–13717. https://doi.org/10.1109/CVPR46437.2021.01350
40. A. Politis, S. Adavanne, T. Virtanen, A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection, preprint, arXiv:2006.01919.

41. A. Politis, S. Adavanne, D. Krause, A. Deleforge, P. Srivastava, T. Virtanen, A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection, preprint, arXiv:2106.06999.
42. A. Politis, A. Mesaros, S. Adavanne, T. Heittola, T. Virtanen, Overview and evaluation of sound event localization and detection in DCASE 2019, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **29** (2021), 684–698. https://doi.org/10.1109/TASLP.2020.3047233
43. K. Liu, X. Zhao, Y. Hu, Y. Fu, Modeling the effects of individual and group heterogeneity on multi-aspect rating behavior, *Front. Data Domputing*, **2** (2020), 59–77. https://doi.org/10.11871/jfdc.issn.2096-742X.2020.02.005
44. Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, PANNs: Large-scale pretrained audio neural networks for audio pattern recognition, *IEEE/ACM Trans. Audio Speech Lang. Process.*, **28** (2020), 2880–2894. https://doi.org/10.1109/TASLP.2020.3030497

AIMS Press