Liu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:19
https://doi.org/10.1186/s13636-021-00206-7

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Frequency-dependent auto-pooling function for weakly supervised sound event detection

Sichen Liu[1,2], Feiran Yang[2,3*], Yin Cao[4] and Jun Yang[1,2]

Abstract

Sound event detection (SED), which is typically treated as a supervised problem, aims at detecting types of sound events and corresponding temporal information. It requires to estimate onset and offset annotations for sound events at each frame. Many available sound event datasets only contain audio tags without precise temporal information. This type of dataset is therefore classified as weakly labeled dataset. In this paper, we propose a novel source separation-based method trained on weakly labeled data to solve SED problems. We build a dilated depthwise separable convolution block (DDC-block) to estimate time-frequency (T-F) masks of each sound event from a T-F representation of an audio clip. DDC-block is experimentally proven to be more effective and computationally lighter than "VGG-like" block. To fully utilize frequency characteristics of sound events, we then propose a frequency-dependent auto-pooling (FAP) function to obtain the clip-level present probability of each sound event class. A combination of two schemes, named DDC-FAP method, is evaluated on DCASE 2018 Task 2, DCASE 2020 Task4, and DCASE 2017 Task 4 datasets. The results show that DDC-FAP has a better performance than the state-of-the-art source separation-based method in SED task.

Keywords: Sound event detection, Weakly supervised, Auto-pooling function, Depthwise separable convolution

1 Introduction

Sound event detection (SED) becomes an important research topic in auditory perception. It has many potential applications such as healthcare in smart home [1, 2], surveillance monitor in public area [3], and large-scale information retrieval [4]. The goal of SED is to predict event classes and corresponding time stamps, i.e., onset and offset times of sound events, while audio tagging (AT) aims at detecting what occurred in an audio clip. Therefore, it is desired that strongly labeled data that contains information of precise presence and absence time for sound classes can be used to train SED systems [5–8]. However, it is costly to acquire such strongly annotated data in realities. On the other hand, there is a large amount of weakly labeled data that is tagged with only types of sound events at clip-level and does not provide the corresponding temporal information. For instance, AudioSet released by Google consists of a collection of 2,084,320 human-labeled 10-s sound clips [9, 10].

Multiple instance learning [11–13] (MIL) is a common framework to train using weakly labeled data. In MIL methods for SED, the audio clip (bag) is divided into overlapped frames (instances), where only ground truth labels of clips are available. An audio clip is labeled positive if the clip contains at least one positive frame. MIL methods usually consist of two parts, a dynamic predictor for generating the present probability of the specific event in each frame and a pooling function for aggregating frame-level probabilities to a clip-level prediction. For the dynamic predictor, conventional support vector machine (SVM) [14], Gaussian mixture model (GMM) [15], and

*Correspondence: feiran@mail.ioa.ac.cn
[2]University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Beijing, China
[3]State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Beijing, China
Full list of author information is available at the end of the article

neural network approaches [16–19] are employed to perform prediction for each event class. The pooling function is used to reduce the dimension of the dynamic feature space, which has a great impact on the overall performance of the weakly supervised SED system. Several pooling functions are exploited in the literature. The global max pooling (GMP) focuses on instances with the highest probability, which is difficult to estimate onset and offset annotations for long-time events. The global average pooling (GAP) assumes that all the instances contribute equally, and hence, short-time events are likely to be underestimated. Attention pooling [20, 21] is a flexible weighting method, which adds a dense neural network to learn weights for each frame in parallel. However, a limitation of the attention pooling is that larger attention weights concentrate upon frames with smaller probabilities when the label is negative [19]. An auto-pooling (AP) function is developed in [22] by introducing a learnable parameter for each class to deal with the weakly labeled SED problem. The AP function reduces to min-, mean-, or max-operators with the increase of the learnable parameter, which can be interpreted as an automatic interpolation between different standard pooling behaviors.

Another aspect of research is based on source separation framework for non-overlapping case [23, 24] or overlapping case [25, 26]. As a starting point, [24] focuses on the non-overlapping sound events. Time-frequency (T-F) segmentation masks are learned from the clip-level tags and then aggregated over both the time and frequency indices to obtain present probabilities of sound events. The T-F segmentation mask is equivalent to ideal ratio mask (IRM) [27] in the context of speech enhancement and source separation. As a byproduct, each sound event can be separated from the mixed audio. In this method, a global weighted rank pooling (GWRP) [28] function is employed to aggregate the masks to clip-level predictions. The T-F bin with a larger value is assigned a larger weight. GWRP is a generalization of GMP and GAP in essence. But, the decay coefficient of GWRP function is manually chosen and may not be optimal in practice.

In this paper, we propose an improved source separation-based approach to solve the problem of weakly supervised SED. The proposed method has a similar framework as in [24], which consists of a segmentation mapping stage and a classification mapping stage. In the segmentation mapping stage, we employ a CNN to capture local patterns of the input spectrogram, i.e., to learn a T-F mask of each specific sound event from weakly labeled data. Concretely, we build a dilated depthwise separable convolution block, named as DDC-block. DDC-block first applies a single-layer dilated filter to each input channel and then applies a 1×1 convolution to combine the output of the previous layer. The presented DDC-block outperforms the "VGG-like" CNN originally used in [24] in terms of the detection performance and the complexity. In the classification mapping stage, we present a frequency-dependent auto-pooling function (FAP) to aggregate T-F masks to clip-level predictions of sound events. The FAP function inherently considers the fact that each sound event exhibits different frequency characteristics by introducing a learnable frequency-varying vector for each class. Furthermore, we show that there are close links between the proposed FAP and the commonly used GMP, the soft-max pooling, the GAP, and the AP functions. In this paper, the proposed method is not specifically designed for handling overlapping sounds. We first focus on the weakly labeled problem without considering the impact of overlapping. Next, we evaluate the proposed method on DCASE 2017 task 4 dataset and DCASE 2020 task 4 dataset, which are recorded in a realistic environment and contain overlapping sound events.

The remainder of the paper is organized as follows: In Section 2, we present the proposed method in detail, including DDC-block for producing T-F masks of sound events and FAP used to aggregate T-F masks to clip-level predictions. In Section 3, we carry out extensive experiments to evaluate the performance of the new method. Section 4 concludes the paper.

2 Proposed method

We now describe the proposed source separation-based framework and how this method can be used to solve the weakly supervised SED as well as AT problems.

The training process of the new framework consists of two steps, i.e., segmentation mapping and classification mapping. In the segmentation mapping stage, a log-mel spectrogram of the audio clip $x(n)$ is extracted to obtain a feature matrix $\mathbf{X} = [\,|X(t,f)|\,] \in \mathbb{R}_{+}^{T\times F}$, where $t = 1, 2, ..., T$ and $f = 1, 2, ..., F$ represent frame and frequency indices, respectively; T denotes the number of audio clip frames; and F is the number of frequency bands. Then, a segmentation mapping of $\mathbf{X} \rightarrow \hat{\mathbf{M}}$ is modeled via a deep neural network, where $\hat{\mathbf{M}} = [\,\hat{M}_k(t,f)] \in \mathbb{R}_{+}^{K\times T\times F}$ is the estimation of the IRM for each sound event class, k is the index of class, and K represents the number of predefined classes. In [24], a "VGG-like" CNN is employed to complete the transformation from the input feature to the specific mask of each sound event. In this paper, we utilize DDC-blocks for this task to obtain a better performance. The details of implementation is presented in Section 2.1. Because only the clip-level tag $\mathbf{y} = [\,y_1, y_2, \cdots, y_K]^{\mathrm{T}} \in \mathbb{R}_{+}^{K\times 1}$ is available for weakly supervised problems, a global pooling function should be designed to transform the estimated T-F mask into the presence probability of the kth sound event. In the classification mapping stage, we map the estimated T-F mask into the clip-level prediction, i.e., $\hat{\mathbf{M}} \rightarrow \hat{\mathbf{y}}$, where $\hat{\mathbf{y}} = [\,\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_K]^{\mathrm{T}} \in \mathbb{R}_{+}^{K\times 1}$ denotes the clip-level

probability. The objective is to minimize the binary cross-entropy between $\hat{y}_k$ and clip-level tag y_k, and the loss function is given by:

$$L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{k=1}^{K} y_k \log(\hat{y}_k) + (1 - y_k)\log(1 - \hat{y}_k). \quad (1)$$

Several global pooling functions such as GMP, GAP, and GWRP have been adopted for the classification mapping. We present a frequency-dependent pooling function to fully exploit the potentials of the source separation-based approach in Section 2.2.

Both AT and SED tasks share the same training stage as described above. Once the training is completed, we can obtain the prediction of the trained model as the AT result. However, for SED task, extra operations need to be carried out to get frame-level probabilities at inference process. Since the estimated mask $\hat{M}_k(t,f)$ contains the information of sound event activities, the frame-level probability can be obtained by aggregating the mask across frequency axis as:

$$\hat{y}_k(t) = \sum_{f=1}^{F} w_k(f)\hat{M}_k(t,f), \quad (2)$$

where $w_k(f)$ denotes the weight of the fth frequency band for the kth class, and $\hat{y}_k(t)$ is the estimated frame-level probability. In [24], Kong et al. average over frequency axis of the mask with $w_k(f) = 1/F$, whereas we utilize the learned vector $\mathbf{w}_k = [w_k(1), w_k(2), \cdots, w_k(F)]^{\mathrm{T}} \in \mathbb{R}_{+}^{F\times 1}$ of FAP to calculate the weighted average along the frequency dimension. To produce smooth frame-level predictions [24], we first select a frame t as a seed where $\hat{y}_k(t) \geq 0.2$. Then, we merge the neighboring frames on both sides in a region-growing style until the frame t' where $\hat{y}_k(t') \leq 0.1$. The diagram of the proposed source separation-based method is shown in Fig. 1.

2.1 The segmentation mapping stage

The segmentation mapping performs feature transformation via the deep neural network. The model consisting of VGG-blocks has been proven quite promising since it can capture local patterns of input features. However, the utilization of VGG-block leads to a high computational cost. To solve this problem, we build a convolutional block, i.e., DDC-block, which employs depthwise separable convolution [29, 30] with dilated filters instead of the typical CNN as in [24]. The architecture of DDC-block is shown in Fig. 2. Each of the three convolution operations is followed by a non-linearity activation and a batch normalization process. The stack of depthwise and pointwise convolution is called depthwise separable convolution, which is considered to be a single convolution layer as the typical CNN. Thus, the number of convolution layers in DDC-block is the same as that in VGG-block.

For the depthwise convolution, the number of filters is required to be equal to that of input channels, which means the spatial convolution is performed independently in each input channel. For the pointwise convolution, 1×1 filters are used to project the outputs of the depthwise layer onto a new feature space. It has been demonstrated that the stack of depthwise and pointwise layer can reduce the number of parameters by a factor of $1/N + 1/(w_{\text{width}} \cdot w_{\text{height}})$, where w_{width} and w_{height} represent width and height of the filter, respectively, and N is the number of output channels [31]. Such a design can significantly alleviate the over-fitting problem. Additionally, we propose to use dilated filters in depthwise layer. Dilated filters can increase the size of receptive field without introducing extra parameters. We carried out experiments (not shown here) and found that the utilization of dilation has a positive impact on both AT and SED tasks. The process of depthwise separable convolution with dilation rate is shown in Fig. 3.

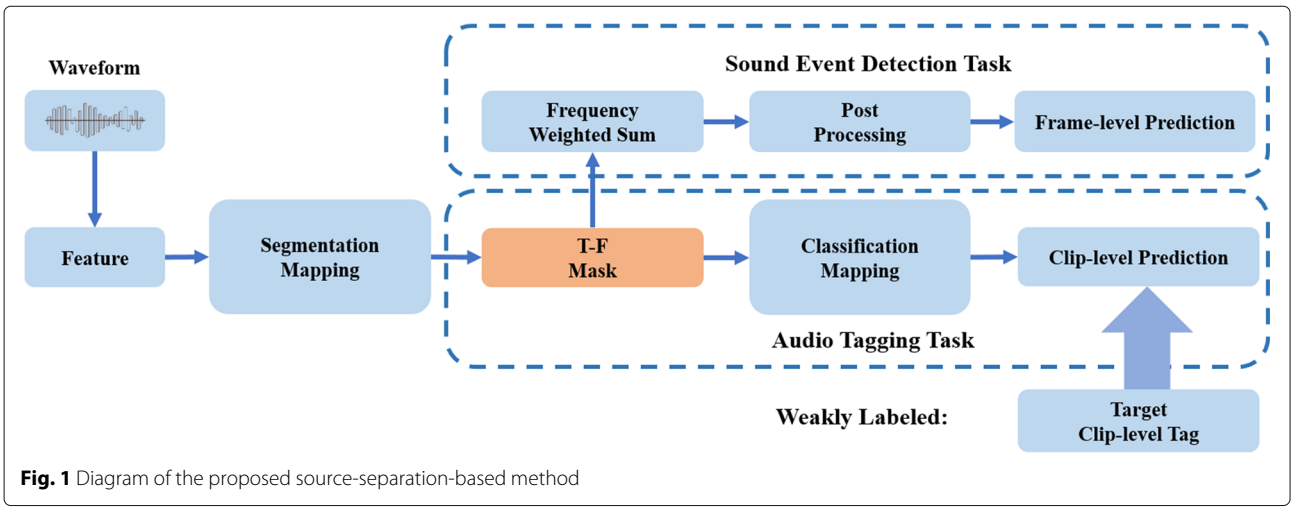

Fig. 1 Diagram of the proposed source-separation-based method

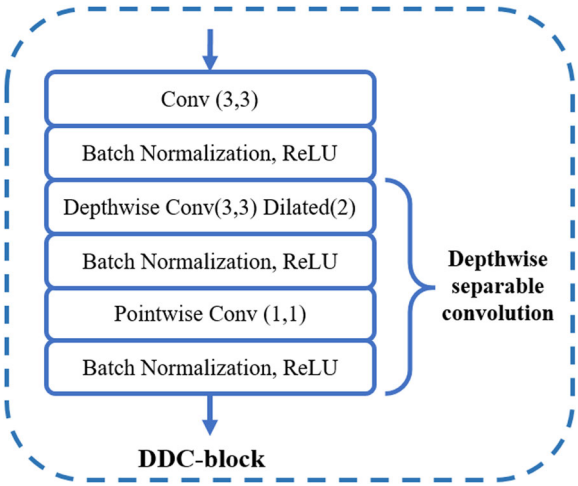

Fig. 2 The internal structure of DDC-block module

At this point, we present a detailed configuration of the model which is used in the segmentation mapping stage. We first apply four DDC-blocks on the input log-mel spectrogram, and the channel numbers of the four blocks are 32, 64, 128, and 128, respectively. Then, a K-channel convolution layer with 1×1 filter is used to convert the output of the last DDC-block to T-F segmentation masks through sigmoid activation functions. Finally, a global pooling function is used to aggregate the estimated mask to the clip-level prediction. The proposed network has a similar depth compared to the "VGG-like" network in [24], which leads to a fair comparison. The detail of the proposed model architecture is summarized in Table 1.

2.2 The classification mapping stage

In this subsection, we model the classification mapping of $\hat{\mathbf{M}} \rightarrow \hat{\mathbf{y}}$ via the pooling function. To this end, pooling functions such as GMP, GAP, and GWRP [26] are commonly used in SED field. GMP only concerns on the T-F bin with the maximum probability, leading to the

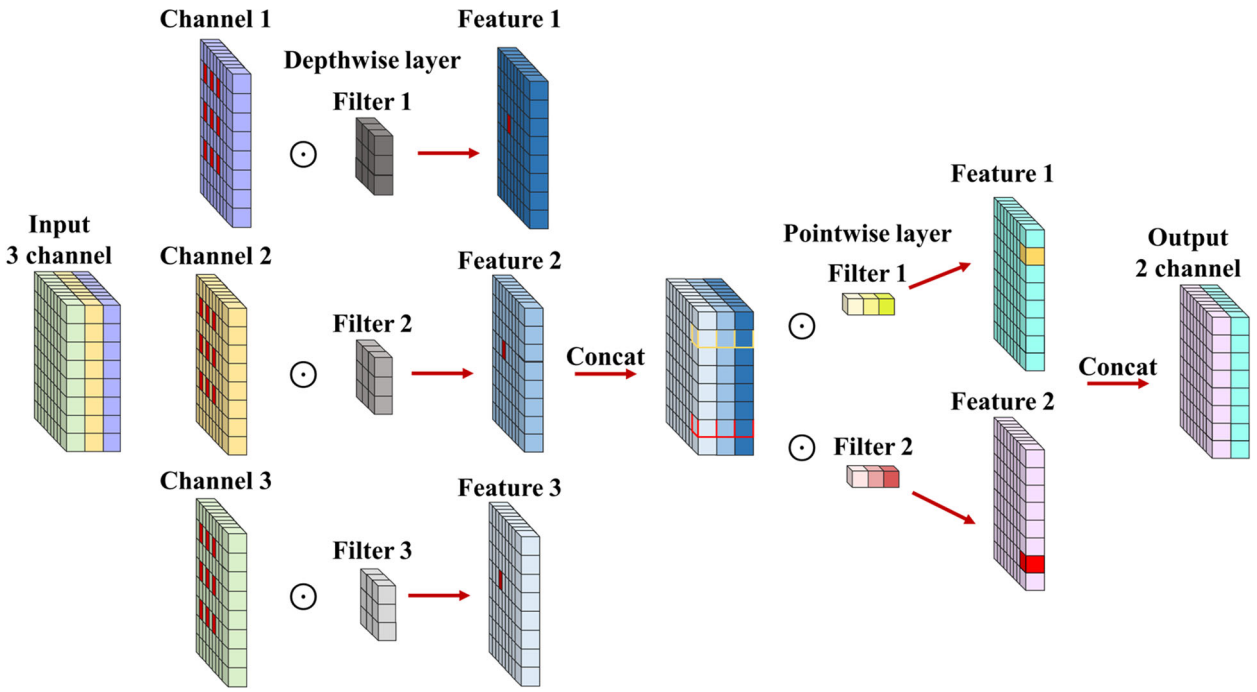

Fig. 3 Block diagram of the depthwise separable convolution with dilated filter in DDC-block

Table 1 Model architecture

Block (filter, channels, dilated)	Output shape (channels × frequency × frame)
Input log-mel spectrogram	1 × 64 × 311
DDC-block(3 × 3, 32, 2)	32 × 64 × 311
DDC-block(3 × 3, 64, 2)	64 × 64 × 311
DDC-block(3 × 3, 128, 2)	128 × 64 × 311
DDC-block(3 × 3, 128, 2)	128 × 64 × 311
CNN (1 × 1, *K*, 1)	*K* × 64 × 311
Global pooling function	*K* × 1

constrained gradient path and inefficient computation. GAP assumes that all the T-F bins contribute equally to the clip-level prediction, which means GAP is unable to focus on the specific T-F bins.

(1) Global weighted rank pooling, GWRP can be understood as a generalization of GMP and GAP. The main idea of GWRP is to put high weights on the T-F bins with high values [26]. The T-F bins of the kth sound event $\hat{\mathbf{M}}_{\mathbf{k}}$ are sorted in a descending order, and the corresponding jth element of the sorted sequence is denoted by $\hat{M}_{k,j}$. The clip-level prediction can be represented as:

$$\hat{y}_k = \frac{1}{\sum_{j=1}^{T\cdot F} r_k^{j-1}} \sum_{j=1}^{T\cdot F} r_k^{j-1} \cdot \hat{M}_{k,j}, \tag{3}$$

where $r_k \in [0,1]$ is a hyper parameter that controls the behavior of GWRP function. Notice that the GWRP reduces to GMP for $r_k = 0$ and GAP for $r_k = 1$. Since the value of weight increases as $\hat{M}_{k,j}$ becomes large, the aggregation performance is improved compared with GMP and GAP. However, the performance of GWRP highly depends on the interpolation coefficient r_k which is difficult to be chosen in practice. We thus propose a FAP function which can determine the interpolation coefficient automatically as required.

(2) Frequency-dependent auto-pooling, FAP function is actually an improved version of soft-max pooling, which introduces a learnable parameter vector $\boldsymbol{\alpha}_k = [\alpha_k(1), \alpha_k(2), \cdots \alpha_k(f)]^{\mathrm{T}} \in \mathbb{R}^{F\times 1}$ as weighting coefficients for the kth class. FAP treats $\boldsymbol{\alpha}_k$ as a free vector that can be learned during training. The expression of FAP is:

$$\hat{y}_k = \sum\nolimits_{t,f} \hat{M}_k(t,f) \cdot \frac{\exp\left(\alpha_k(f)\cdot \hat{M}_k(t,f)\right)}{\sum_{t,f}\exp\left(\alpha_k(f)\cdot \hat{M}_k(t,f)\right)}, \tag{4}$$

where $\alpha_k(f)$ is the weight of the fth frequency band. Note that the frequency-varying $\alpha_k(f)$ is shared among all frames in each frequency band.

We now show the relationship between the proposed FAP and several well-known pooling functions. The proposed FAP function can be treated as an extension of AP proposed in [22]. FAP is specifically used for 2D data like spectrograms, so the prior information of frequency distribution for event classes can be considered during aggregation. That is, FAP can focus on the crucial frequency bands adaptively by learning the vectors. Additionally, FAP reduces to GMP, soft-max pooling, and GAP when $\alpha_k(f) \to \infty$, $\alpha_k(f) = 1$, and $\alpha_k(f) = 0$, respectively.

A word on the bound of the parameter $\alpha_k(f)$ is appropriate here. For $\alpha_k(f) < 0$, FAP is similar to min-pooling, and hence, T-F bins with smaller values attract much more attention, which is not desired. On the other hand, for $\alpha_k(f) \to \infty$, FAP simplifies to GMP which may result in the gradient explosion problem. We thus propose to set the parameter $\alpha_k(f)$ to $0 < \alpha_k(f) < \alpha_{\max}$, where $\alpha_{\max}$ is a predefined constant. In this paper, $\alpha_{\max} = 10$ is empirically chosen and achieves a satisfactory performance.

3 Experimental results

3.1 Data preparation

We utilize the audio clips of DCASE 2018 Task 1 dataset as background noises which are recorded in 10 scenes such as metro station and shopping mall. For the sound events, 3710 manually verified clips that include 41 categories are obtained from DCASE 2018 Task 2. These events involve various human activities, household events, instrument events, etc. All of the audio clips are sampled at 32 kHz. More details on the preparation of data are shown in Table 2. We fix the duration of every sound events to 2 s which is same as [24] to make sure that the generated clips are non-overlapping. To be specific, the events shorter than 2 s are padded with zeros to 2 s. For the events longer than 2 s, we extract the first 2 s of them as training data and remove the other parts of the recording. Three randomly selected events are mixed with the background noise without any overlapping at 0-dB SNR. The onsets of events are 0.5 s, 3 s, and 5.5 s, respectively. Thus, we can ensure that the impact of overlapping can be avoided, which will help us to focus on the weakly labeled task first. We synthesize 8000 audio clips and divide them into 4 cross-validation folds.

Table 2 Setting of experiment data

	Acoustic scene	Sound event
Dataset	DCASE2018 Task 1	DCASE2018 Task 2
Instances	8640	3710 (manually verified)
Length	10 s	0.3–30 s
Class	10	41

3.2 Setup

We choose 64 bands log-mel spectrum as the input data representation to our model. A 64-ms-long Hanning window is employed for STFT with 50% overlap. Then, each frame is converted into a 64-dimensional vector by a log-mel filter bank. This process converts a 10-s audio clip into a 64×311 dimensional log-mel spectrogram representation. Learning rate is initially set to 1e−3 and automatically reduced to 0.9 times of the previous value per 1000 iterations. Xavier initialization is used to initialize the model. The experiment setting is shown in Table 3, and each result shown in this paper is obtained by averaging over 10 independent experiments.

We observed that the prediction result of AT task is more convincing than that of SED task. In order to reduce false positives for SED task, we first evaluate the clip-level predictions of each clip. Only the classes that are predicted as active at clip-level can be selected to evaluate the frame-level predictions [24]. Since the lengths of most events are longer than 10 frames, we treat the predictions which are shorter than 10 frames as false-positive cases and remove them to reduce inserts. Moreover, some classes such as "Knock", occur discontinuously in audio clips, but their ground-truth frame-level labels are always continuous. Thus, the events or the silence gap of events shorter than 10 frames are removed or merged. We set the onset collar of 200 ms and an offset collar of 200 ms/50% to count the true positives of the prediction, which is similar to the configuration of [32].

As for evaluation metrics, we use F-score, area under the curve (AUC), mean average precision (mAP) [32], and error rate (ER) to evaluate the performance of AT and SED tasks. The F-score is computed as a harmonic mean between precision and recall. The AUC is the area under ROC curve which plots true-positive rate (TPR) versus false-positive rate (FPR). The mAP is the average of precision at different recall values regardless of thresholds. The mAP can evaluate the model comprehensively and is widely used in the weakly supervised SED field. Error rate (ER) measures the number of errors including deletions (D) and insertions (I). ER is a score rather than a percentage which can become larger than 1 in the case when the system makes more errors than correct predictions.

Table 3 Experiment setting

Fourier setting		Training parameters setting	
Sampling rate	32 kHz	Batch size	18
Window size	2048	Iterations	20,000
Overlap	1024	Optimizer	ADAM
Mel bands	64	Learning rate	1e−3, 1000 iter*0.9

3.3 Performance evaluation of the pooling functions

In this section, we evaluate the performance of five pooling functions including GMP, GAP, GWRP, AP, and FAP. To fairly validate the effectiveness of FAP function, we employ the same model in the segmentation mapping stage. Specifically, the DDC-blocks in Table 1 are replaced with four VGG-blocks. For GWRP function, the interpolation coefficient r_k is set to 0.9998 as in [24].

As seen from Table 4, GMP and GAP are inferior to the other approaches due to the impractical assumptions. As expected, FAP achieves the highest scores in all the involved methods in terms of the mAP and AUC for both AT and SED tasks. Especially, FAP achieves a significant performance improvement over GWRP and AP. This is mainly because FAP function can automatically interpolate between different pooling behaviors through the learnable frequency-wise vectors.

Figure 4 illustrates the parameters learned by the proposed FAP function. Clearly, the learned weights vary with frequency bands for certain acoustic events, and the weight vectors $\boldsymbol{\alpha}_k$ exhibit different distribution characteristics for different sound events. For instance, the energy of *keys_jangling* events and *bark* events are mainly distributed in the high- and low-frequency bands, respectively. The corresponding vectors $\boldsymbol{\alpha}_k$ show a consistent tendency with the energy distribution. It is observed in the experiments that $\alpha_k(f) \leq 3$ for most cases, and we present the results for $\alpha_{\max} = 3, 5, 10$, respectively. Table 5 investigates the effect of the upper bound of $\alpha_k(f)$ on the overall performance. In the case of $\alpha_{\max} = 10$, FAP function achieves the best performance, and hence, it is used for the proposed method in the other experiments.

3.4 Performance evaluation of the proposed method

We compare the performance of DDC-blocks and VGG-blocks based on GWRP, AP, and FAP functions. Two well-known examples of MIL methods, i.e., Attention [20] and TALNet [19], are also involved to make a comprehensive comparison. The results are shown in Table 6.

Table 4 Performance of different pooling functions

Metrics		GMP	GAP	GWRP	AP	FAP
AT	*F*-score	0.433	0.386	0.572	0.538	**0.590**
	AUC	0.797	0.883	0.923	0.909	**0.923**
	mAP	0.450	0.452	0.635	0.639	**0.672**
SED	*F*-score	0	0.252	**0.429**	0.352	0.407
	AUC	0.675	0.669	0.803	0.823	**0.848**
	mAP	0.078	0.214	0.372	0.362	**0.385**
Error	ER	1	2.610	1.991	1.886	**1.776**
rate	D	1	0.896	**0.780**	0.844	0.823
	I	0	1.718	1.210	1.042	**0.952**

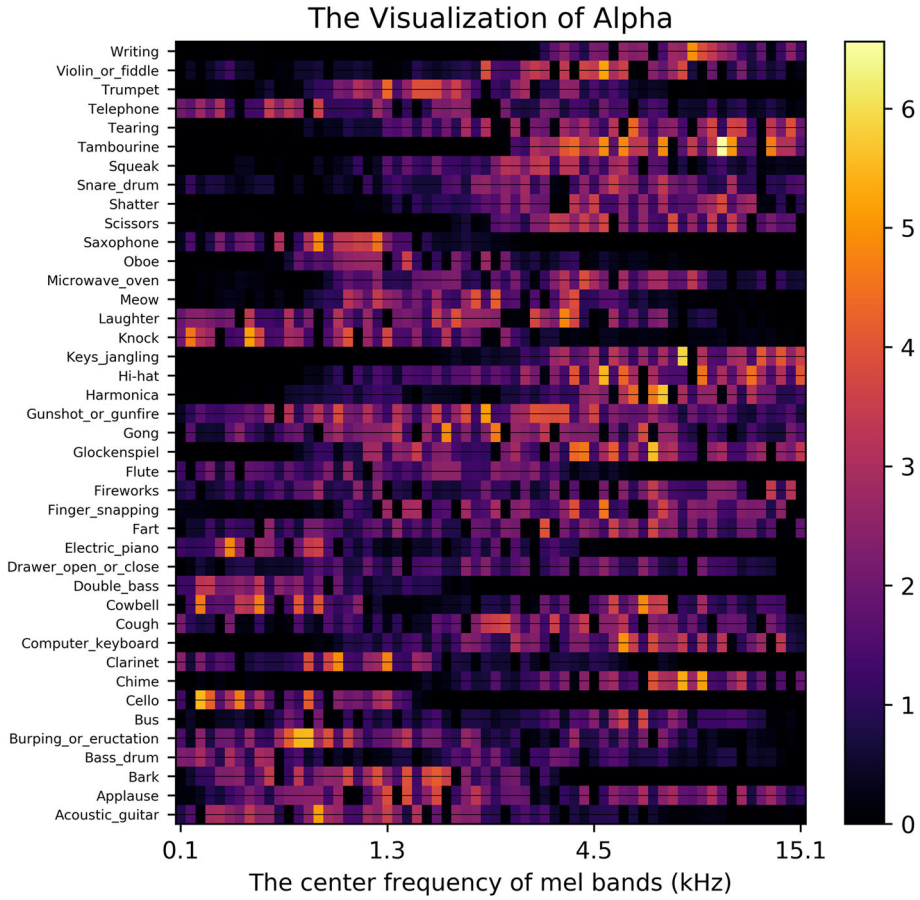

Fig. 4 Learned $\alpha_k(f)$ parameters for each event class

We first compare the performance of DDC-blocks and VGG-blocks. Using the same pooling function, the methods with DDC-blocks outperform that with VGG-blocks in terms of the F1-score, AUC, and mAP. As for the model size, the required parameters for DDC-block-based approaches decrease by 49.5% compared with the VGG-block-based methods. Thus, the utilization of DDC-blocks significantly reduces the number of parameters while it achieves a better performance in segmentation. Moreover, DDC-FAP achieves the highest mAP and AUC, the lowest ER, and the least insertion among all the source separation-based methods. It turns out that the combination of DDC-block and FAP achieves a significant performance improvement compared with the method in [24].

Table 5 The different upper bound setting of FAP

Metrics		FAP-3	FAP-5	FAP-10
AT	*F*-score	0.560	0.589	**0.590**
	AUC	**0.924**	0.922	0.923
	mAP	0.654	0.667	**0.672**
SED	*F*-score	**0.429**	0.412	0.407
	AUC	0.840	0.847	**0.848**
	mAP	0.385	0.384	**0.385**
Error	ER	2.229	1.845	**1.776**
Rate	D	**0.772**	0.821	0.823
	I	1.457	1.024	**0.952**

In AT task, DDC-FAP is comparable with Attention [20] and TALNet [19] . The proposed method achieves a somewhat higher AUC score in the AT task. It indicates that DDC-FAP makes fewer false-negative predictions, which is consistent with the observation that DDC-FAP gets fewer deletions compared to Attention [20]. As for the SED task, it is apparent that DDC-FAP achieves the highest mAP (0.427) and AUC (0.868) and hence outperforms the other methods. To provide a better illustration, we summarize the results of 10 independent experiments to draw the box plot of the main metrics in Fig. 5. It shows that the results of DDC-FAP are relatively stable in these metrics and superior to the other methods in SED task.

Table 6 The results of MIL and source-separation-based methods

Method		Parameters	Audio tagging			Sound event detection			Error rate		
		10 k	*F*-score	AUC	mAP	*F*-score	AUC	mAP	ER	D	I
MIL	Attention [20]	54.15	0.671	0.923	0.723	0.341	0.861	0.348	1.574	0.885	0.689
	TALNet [19]	94.06	0.646	0.911	0.687	0.397	0.849	0.390	1.339	0.865	0.474
Source	VGG-GWRP [24]	58.76	0.572	0.923	0.635	0.429	0.803	0.372	1.991	**0.780**	1.210
separation	VGG-AP	58.76	0.538	0.909	0.639	0.352	0.823	0.362	1.886	0.844	1.061
-based	VGG-FAP	58.76	0.590	0.923	0.672	0.407	0.848	0.385	1.776	0.823	0.952
	DDC-GWRP	**28.84**	0.626	0.931	0.689	**0.468**	0.808	0.404	1.850	0.813	1.037
	DDC-AP	28.84	0.573	0.919	0.684	0.382	0.845	0.398	1.831	0.853	0.978
	DDC-FAP	29.10	**0.633**	**0.931**	**0.719**	0.446	**0.868**	**0.427**	**1.689**	0.845	**0.844**

In order to show the performance of the aforementioned methods more intuitively, we let the model parameters as the abscissa and the mAP in SED task as the ordinate as shown in Fig. 6. An ideal SED system should require fewer parameters and achieves a higher mAP. It can be seen from Fig. 6 that the method using the DDC-block outperforms all the other approaches.

3.5 Performance evaluation on DCASE 2020 Task 4

Compared with the aforementioned synthetic dataset, some of the events in DCASE 2020 Task 4 are overlapped. DCASE 2020 Task 4 dataset mainly consists of a FUSS dataset and a DESED dataset. The FUSS dataset used for sound separation task does not provide labels for event classes, so that DDC-FAP method cannot utilize it for training. The DESED dataset is used for the SED task, which consists of strong labeled, weakly labeled, and unlabeled audio clips. We evaluate the proposed method on the weakly labeled training set of DESED. The results are shown in Table 7. The performance of the mentioned methods on DCASE 2020 Task 4 dataset is similar with that in Section 3.4. For SED task, DDC-FAP still achieves the best results for all metrics. For AT task, DDC-FAP ranks 2nd, which is slightly worse than Attention. Experimental results show that although the proposed method is not specifically designed for overlap, it still has good performance for the overlapping case.

3.6 Performance evaluation on DCASE 2017 Task 4

Data imbalance is also a challenging problem for SED task. We utilize DCASE 2017 Task 4 dataset to verify the effectiveness of DDC-FAP on unbalanced situation. We add

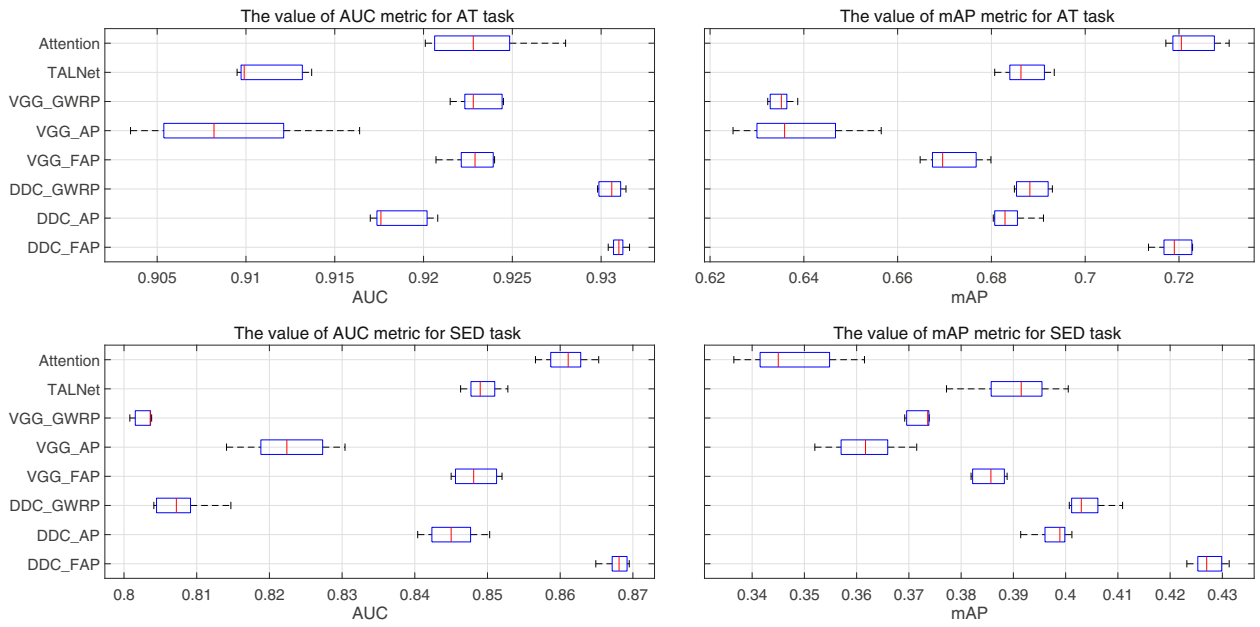

Fig. 5 The metrics box plot of the mentioned methods

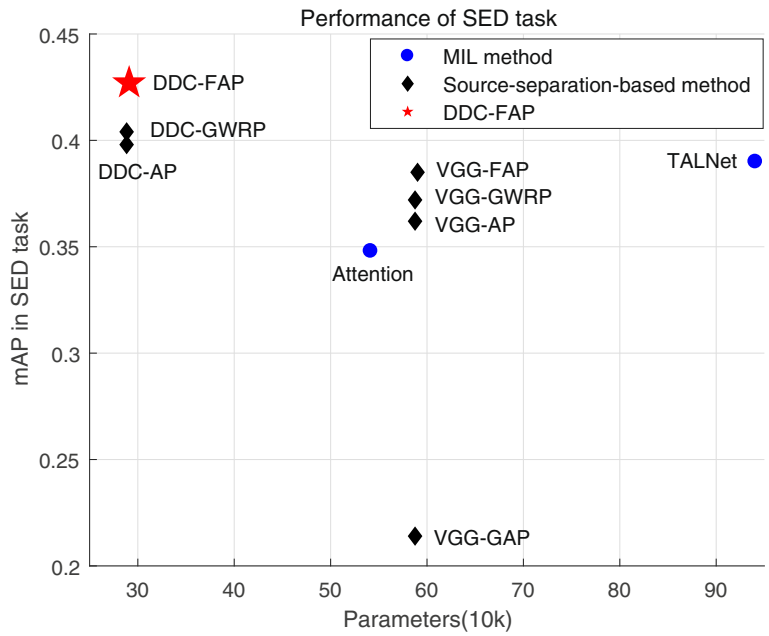

Fig. 6 The number of parameter and performance in the SED task

a mini-batch data balancing operation to ensure that the number of the most frequent events is at most five times than the least frequent samples in a mini-batch. To be consistent with [19], the class-specific thresholds which achieve the highest *F*-score of AT task are utilized to make the clip-level predictions. The performance of AT and SED task is evaluated on the clip level and 1-s segment level, respectively. In addition to the mentioned methods, the adaptive distance-based pooling function has been proposed recently. It compensates non-relevant information of audio events by applying an adaptive transformation in temporal axis. For a comprehensive comparison, we show the results of the mentioned methods in Table 8.

It can be seen that the proposed method achieves the best *F*-score in both AT and SED task. Besides the unbalanced property, some of sound events are overlapping in the dataset. Moreover, the events of DCASE 2017 Task 4 have not been padded or trimmed, which contain a more natural and diverse distribution of duration. The results show that DDC-FAP performs well in these situations, which indicates its robustness to complex scenes.

4 Conclusion

In this paper, we proposed a novel source separation-based method for weakly supervised SED. In segmentation mapping stage, we designed a model consisting of four DDC-blocks to convert the input feature to the T-F mask of each sound event. To utilize the prior frequency information, we proposed the FAP function which introduces learnable vectors to find the key bands when aggregating the T-F masks. Both of the temporal location of the predefined events and the separated waveform can be obtained from the trained T-F mask. Extensive experiments demonstrated that the DDC-block is more effective and computationally lighter than the VGG-block in segmentation mapping stage, and the FAP function outperforms the widely used pooling operators. The proposed DDC-FAP method achieves a better performance than the

Table 7 The results of DCASE 2020 Task 4 dataset

Method	Parameters	Audio tagging			Sound event detection			Error rate		
	10 k	*F*-score	AUC	mAP	*F*-score	AUC	mAP	ER	D	I
Attention [20]	54.15	0.719	0.946	0.810	0.557	0.867	0.572	1.715	0.755	0.960
TALNet [19]	94.06	0.672	0.917	0.741	0.536	0.851	0.516	1.523	0.773	0.750
VGG-GWRP [24]	58.76	0.675	0.938	0.772	0.499	0.843	0.578	1.866	0.769	1.096
DDC-FAP	29.10	0.694	0.941	0.795	0.595	0.881	0.610	1.755	0.790	0.965

Table 8 The results of DCASE 2017 Task 4 dataset

Method	AT			SED	
	F-score	Precision	Recall	*F*-score	ER
Attention [20]	0.561	0.559	0.585	0.476	0.866
TALNet [19]	0.525	0.556	0.537	0.463	0.893
VGG-GWRP [24]	0.561	0.518	0.641	0.470	0.964
DDC-FAP	0.572	0.595	0.583	0.482	0.962
Adaptive distance-based pooling [33]	0.487	0.677	0.465	–	–

state-of-the-art source separation-based methods in various situations such as the non-overlapped, overlapped, and unbalanced cases.

Abbreviations

SED: Sound event detection; DDC-block: Dilated depthwise separable convolution block; T-F: Time-frequency; FAP: Frequency-dependent auto-pooling; AT: Audio tagging; MIL: Multiple instance learning; SVM: Support vector machine; GMM: Gaussian mixture model; GMP: Global max pooling; GAP: Global average pooling; AP: Auto-pooling; IRM: Ideal ratio mask; GWRP: Global weighted rank pooling; AUC: Area under the curve; mAP: Mean average precision; ER: Error rate; TPR: True-positive rate; FPR: False-positive rate; D: Deletion; I: Insertion

Acknowledgements

Not applicable.

Authors' contributions

SCL conducted the research and performed the experiments. FRY and YC supervised the experimental work and polished the structure as well as the text of the manuscript. The guidance of the whole work was performed by JY. Moreover, all authors read and approved the final manuscript.

Funding

This work was supported by the Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2018027, National Natural Science Foundation of China under Grants 11804368 and 11674348, IACAS Young Elite Researcher Project QNYC201812, National Key R&D Program of China under Grant 2017YFC0804900, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDC02020400.

Availability of data and materials

The datasets analyzed during the current study are available in the [DCASE2018] repository, [http://dcase.community/challenge2018/task-general-purpose-audio-tagging]

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Beijing, China. [2]University of Chinese Academy of Sciences, No.19(A) Yuquan Road, Beijing, China. [3]State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, No. 21 North 4th Ring Road, Beijing, China. [4]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK.

Received: 30 November 2020 Accepted: 14 April 2021
Published online: 17 May 2021

References

1. T. Virtanen, M. D. Plumbley, D. Ellis, *Computational analysis of sound scenes and events*. (Springer, Heidelberg, 2018)
2. Y. Lavner, R. Cohen, D. Ruinskiy, H. IJzerman, in *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*. Baby cry detection in domestic environment using deep learning, (2016), pp. 1–5
3. G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. Scream and gunshot detection and localization for audio-surveillance systems, (2007), pp. 21–26
4. A. Jati, D. Emmanouilidou, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Supervised deep hashing for efficient audio event retrieval, (2020), pp. 4497–4501
5. M. Ravanelli, B. Elizalde, K. Ni, G. Friedland, in *2014 22nd European Signal Processing Conference (EUSIPCO)*. Audio concept classification with hierarchical deep neural networks, (2014), pp. 606–610
6. H. Zhang, I. McLoughlin, Y. Song, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Robust sound event recognition using convolutional neural networks, (2015), pp. 559–563
7. E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen, Convolutional recurrent neural networks for polyphonic sound event detection. IEEE/ACM Trans. Audio. Speech. Lang. Process. **25**, 1291–1303 (2017)
8. D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, Detection and classification of acoustic scenes and events. IEEE Trans. Multimed. **17**, 1733–1746 (2015)
9. A. Mesaros, T. Heittola, T. Virtanen, in *2016 24th European Signal Processing Conference (EUSIPCO)*. TUT database for acoustic scene classification and sound event detection, (2016), pp. 1128–1132
10. J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Audio set: an ontology and human-labeled dataset for audio events, (2017), pp. 776–780
11. A. Kumar, B. Raj, in *Proceedings of the 24th ACM International Conference on Multimedia*. Audio event detection using weakly labeled data, (2016)
12. M. Ilse, J. M. Tomczak, M. Welling, in *International conference on machine learning*. Attention-based deep multiple instance learning, (2018), pp. 2127–2136
13. T.-W. Su, J.-Y. Liu, Y.-H. Yang, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks, (2017), pp. 791–795
14. S. E. Küçükbay, M. Sert, in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. Audio-based event detection in office live environments using optimized MFCC-SVM approach, (2015), pp. 475–480
15. A. Kumar, B. Raj, in *2017 International Joint Conference on Neural Networks (IJCNN)*. Audio event and scene recognition: a unified approach using strongly and weakly labeled data, (2017), pp. 3475–3482
16. S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, T. Virtanen, in *Scenes and Events 2016 Workshop (DCASE2016)*. Sound event detection in multichannel audio using spatial and harmonic features, (2016), p. 6
17. M. Espi, M. Fujimoto, K. Kinoshita, T. Nakatani, Exploiting spectro-temporal locality in deep learning based acoustic event detection. EURASIP J. Audio. Speech. Music. Process. **2015**, 26 (2015)
18. D. de Benito-Gorron, A. Lozano-Diez, D. T. Toledano, J. Gonzalez-Rodriguez, Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. EURASIP J. Audio. Speech. Music Process. **2019**, 9 (2019)
19. Y. Wang, J. Li, F. Metze, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling, (2019), pp. 31–35
20. Y. Xu, Q. Kong, W. Wang, M. D. Plumbley, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Large-scale weakly supervised audio classification using gated convolutional neural network, (2018), pp. 121–125
21. C. Yu, K. S. Barsim, Q. Kong, B. Yang, Multi-level attention model for weakly supervised audio classification. CoRR. **abs/1803.02353** (2018). http://arxiv.org/abs/1803.02353

22. B. McFee, J. Salamon, J. P. Bello, Adaptive pooling operators for weakly labeled sound event detection. IEEE/ACM Trans. Audio. Speech. Lang. Process. **26**, 2180–2193 (2018)
23. Q. Kong, Y. Xu, W. Wang, M. D. Plumbley, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A joint separation-classification model for sound event detection of weakly labelled data, (2018), pp. 321–325
24. Q. Kong, Y. Xu, I. Sobieraj, W. Wang, M. D. Plumbley, Sound event detection and time–frequency segmentation from weakly labelled data. IEEE/ACM Trans. Audio. Speech. Lang. Process. **27**, 777–787 (2019)
25. T. Heittola, A. Mesaros, T. Virtanen, A. Eronen, in *2011 Machine Listening in Multisource Environments*. Sound event detection in multisource environments using source separation, (2011), pp. 36–40
26. F. Pishdadian, G. Wichern, J. Le Roux, Finding strength in weakness: learning to separate sounds with weak supervision. IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 2386–2399 (2020)
27. A. Narayanan, D. Wang, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. Ideal ratio mask estimation using deep neural networks for robust speech recognition, (2013), pp. 7092–7096
28. A. Kolesnikov, C. H. Lampert, in *European Conference on Computer Vision*. Seed, expand and constrain: three principles for weakly-supervised image segmentation, (2016), pp. 695–711
29. F. Chollet, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Xception: deep learning with depthwise separable convolutions, (2017)
30. J. Guo, Y. Li, W. Lin, Y. Chen, J. Li, Network decoupling: From regular to depthwise separable convolutions. CoRR. **abs/1808.05517** (2018). http://arxiv.org/abs/1808.05517
31. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR. **abs/1704.04861** (2017). http://arxiv.org/abs/1704.04861
32. A. Mesaros, T. Heittola, T. Virtanen, Metrics for polyphonic sound event detection. Appl. Sci. **6**, 162 (2016)
33. I. Martín-Morató, M. Cobos, F. J. Ferri, Adaptive distance-based pooling in convolutional neural networks for audio event classification. IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 1925–1935 (2020)

Publisher's Note