Circuits, Systems, and Signal Processing (2019) 38:3433–3453
https://doi.org/10.1007/s00034-019-01094-1

A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection

Xianjun Xia[1] · Roberto Togneri[1] · Ferdous Sohel[2] · Yuanjun Zhao[1] · Defeng Huang[1]

Received: 3 September 2018 / Revised: 12 March 2019 / Accepted: 15 March 2019 /
Published online: 21 March 2019

Abstract
Recently, neural network-based deep learning methods have been popularly applied to computer vision, speech signal processing and other pattern recognition areas. Remarkable success has been demonstrated by using the deep learning approaches. The purpose of this article is to provide a comprehensive survey for the neural network-based deep learning approaches on acoustic event detection. Different deep learning-based acoustic event detection approaches are investigated with an emphasis on both strongly labeled and weakly labeled acoustic event detection systems. This paper also discusses how deep learning methods benefit the acoustic event detection task and the potential issues that need to be addressed for prospective real-world scenarios.

Keywords Deep learning · Acoustic event detection · Strongly labeled · Weakly labeled

✉ Xianjun Xia
xianjun.xia@research.uwa.edu.au

Roberto Togneri
roberto.togneri@uwa.edu.au

Ferdous Sohel
F.Sohel@murdoch.edu.au

Yuanjun Zhao
yuanjun.zhao@research.uwa.edu.au

Defeng Huang
david.huang@uwa.edu.au

[1] School of Electrical, Electronic and Computer Engineering, University of Western Australia, 35 Stirling Hwy, Perth, WA 6009, Australia

[2] School of Engineering and Information Technology, Murdoch University, 90 South St, Murdoch, WA 6150, Australia

1 Introduction

Acoustic event detection (AED), which determines both the types and the happening times (beginning and end times) of different acoustic events, enables automatic systems to obtain a better understanding of what is happening in an acoustic scene. It has been applied to many applications including security [12,73,80], life assistance [36,61] and human–computer interaction [78,94]. Due to the important and indispensable role of AED, there has been an increasing research activity in this area [29,59,62,85]. Various evaluation campaigns including CLEAR [75], DCASE 2013 [26], DCASE 2016 [83] and DCASE 2017 [4] were organized to address the challenges in and to promote research on AED.

The AED task is typically treated as a frame-based classification or regression problem with each frame corresponding to an acoustic event type and its continuous-valued localizations. Frame-wise classification of acoustic events is applied over sliding time windows with statistical models to represent the acoustic features. A straightforward idea is to use conventional machine learning algorithms, such as GMMs [93] to model the means and variances of the acoustic features for each event type. During testing, the likelihoods from GMMs are summed across time and the type with the highest probability is chosen as the final detected result. Other statistical machine learning algorithms such as hidden Markov models (HMMs) [72,90], support vector machines (SVMs) [57,79] and nonnegative matrix factorization (NMF) [19,40] are also applied to perform the classification task. The non-speech sounds were also detected using the fuzzy integral (FI) [77], which have shown comparable results to the high performing SVM feature-level fusion in [77]. In [62], the authors proposed a technique for the joint detection and localization of non-overlapping acoustic events using random forest (RF). Multivariable random forest regressors are learned for each event category to map each frame to continuously estimate the onset and offset time of the events. Heittola et al. [30] proposed two iterative approaches based on the expectation maximum (EM) algorithm [91] to select the most likely stream to contain the target sound: one by always selecting the most likely stream and the other by gradually eliminating the most unlikely streams from the training data.

Although some improvements have been made using the aforementioned learning algorithms, these conventional machine learning techniques show a limited power when the AED task becomes more challenging if the acoustic events are polyphonic or strongly labeled annotations (both the presence and the localizations of the acoustic events are given in the training process) are not available. Inspired by the successful applications of deep learning techniques in computer vision, speech signal and natural language processing, the AED task is also popularly performed using different neural network-based deep learning approaches.

Neural network-based deep learning algorithms are originated in the 1940s leading to the first wave of artificial intelligence (AI) algorithms with the creation of the single-layer perceptron (SLP) and the multilayer perceptron (MLP) [66,67]. In [32], a new layer-wise greedy learning-based training method was proposed for the deep neural network (DNN). Hidden layers in a network are pretrained one layer at a time using the unsupervised learning approach, and this considerably helps to accelerate subsequent supervised learning through the back-propagation algorithm [23,46]. A convolutional neural network (CNN)-based approach achieved a new error record of 0.39% on the handwriting digits database MNIST in [64], which marks a significant progress in performance since the classical prototype LeNet-5 [47]. In [33], the authors proposed the auto-encoders (AEs) to pretrain the feed-forward neural network (FNN). Afterward, variations of AEs including the de-noising auto-encoder (DAE) [81,82], the sparse auto-encoder (SAE) [50] and the variational auto-encoder (VAE) [38] which was proposed to enhance the ability of feature learning and representation. However, the VAE introduces potentially restrictive assumptions about the approximate posterior distribution making the generated data blurry [37]. Generative adversarial networks (GANs) were proposed to offer a distinct approach by focusing on the game-theoretic formulation while training the generative model [27] and to produce high-quality images [18,65]. GANs [31,58,65,71] have been popularly adopted to generate training data in [6,55,92].

The aforementioned neural network-based deep learning approaches have shown their superior performances in AED tasks. In [9], the DNN was used to perform the polyphonic acoustic event detection task. The deep neural network-based system outperformed the conventional learning method using nonnegative matrix factorization at the preprocessing stage and HMM as a classifier. The recurrent neural network (RNN) was applied to the AED system in [49,60] to capture the context information deep in time. In [60], the authors presented a technique based on the bidirectional long short-term memory (BLSTM). The multi-label BLSTM was trained to map acoustic features of multiple classes to binary activity indicators of each event class. In [34], the authors presented a polyphonic AED system with a multi-model system. In that work, one DNN was used to detect acoustic event of "car" and five bidirectional gated recurrent units–recurrent neural networks (BGRU-RNN) were used to detect other acoustic events. The CNN [48] was used to extract the high-level features that are invariant to local spectral and temporal variations in [35]. Authors in [4,59] combined the RNN and CNN by adopting the convolutional recurrent neural network (CRNN) to model the audio features and achieved the state-of-the-art performance.

Tables 1 and 2 comprehensively list the recent works with conventional machine learning- and neural network-based deep learning approaches applied to the AED.

Table 1 Some conventional machine learning approaches applied to the acoustic event detection

References	Feature	Modeling	Metric performance
Phan et al. [62]	Super-frame vector	Random forest	Error rate 0.3979
Xia et al. [85]	Super-frame vector	Random forest	Error rate 0.2490
Zhuang et al. [93]	Learned features using UBM	GMM	*F*-score 6.0%
Vuegen et al. [84]	MFCC	GMM	*F*-score 48.0%
Schröder et al. [72]	Gabor filter bank	HMM	*F*-score 73.0%
Diment et al. [21]	MFCC	HMM	*F*-score 61.6%
Nogueira et al. [57]	MFCC	SVM	*F*-score 32.5%
Komatsu et al. [40]	Event-wise aggregated activation	NMF	*F*-score 60.0%
Gemmeke et al. [25]	Mel-magnitude spectrograms	NMF	*F*-score 31.9%

In this table, different acoustic features, modeling approaches, metrics and performances are shown

Table 2 Neural network-based deep learning approaches in acoustic event detection

References	Feature	Modeling	Metric performance
Cakir et al. [9]	Mel-band energy	DNN	Accuracy 63.8%
Dai Wei et al. [14]	MFCC	DNN	*F*-score 3.6%
Laffitte et al. [44]	MFCC	DNN	Error rate 6.2%
Kong et al. [41]	Mel-band energy	DNN	*F*-score 36.3%
Xia et al. [86]	Mel-band energy	DNN	*F*-score 57.3%
Xia et al. [87]	Mel-band energy	DNN	*F*-score 32.7%
Takahashi et al. [68]	Mel-band energy	CNN	Model accuracy 92.8%
Chou et al. [11]	Mel-spectrogram	CNN	*F*-score 32.8%
Su et al. [76]	Mel-spectrogram	CNN	Clip accuracy 51.73%
Jeong et al. [35]	Mel-spectrogram	CNN	*F*-score 67%
Meyer et al. [51]	TF representation	CNN	Accuracy 85.1%
Xia et al. [88]	Mel-band energy	CNN	*F*-score 45.3%
Lu et al. [49]	MFCC + Pitch + LMS	Bi-GRU	Error rate 0.79
Parascandolo et al. [60]	Raw audio	BLSTM	*F*-score 65.5%
Kim et al. [39]	Multi-channel features	GRU	*F*-score 50.3%
Cakir et al. [8]	Mel-band energy	CRNN	ROC curve 88.5%
Adavanne et al. [4]	Binaural features	CRNN	*F*-score 42.9%
Adavanne et al. [5]	Mel-band energy	CRNN	*F*-score 43.3%
Adavanne et al. [2]	Spatial features	CRNN	Error rate 0.43
Cakir et al. [10]	TF representation	CRNN	*F*-score 61.0%

In this table, different acoustic features, scoring strategies, modeling approaches, metrics and performances are shown

In Tables 1 and 2, different evaluation datasets and metrics were adopted. In order to highlight the advantages of the deep learning approaches over the conventional machine learning techniques, Table 3 lists selected top systems based on the unified evaluation databases in DCASE Challenges from the years 2013 to 2017. The first block in Table 3 shows the system performances using the conventional machine learning approaches. It is worth noting that the deep learning approaches were not applied to the AED system in 2013. The second block shows that the neural network-based deep learning approaches outperformed the conventional machine learning techniques. The detection performance was pushed further with improved error rates using the deep learning approaches in 2017. The number of deep learning-based systems also increased dramatically from 0 in 2013 to 33 in 2017, which dominated the 36 submitted systems. To show the computational load of different training approaches, Table 4 shows the detailed information of some typical neural network structures and the corresponding detection error rates when the Mel-band energy is adopted as the acoustic feature. The aforementioned trend motivated us to write this survey of neural network-based deep learning approaches applied to the acoustic event detection task.

The purpose of this paper is to provide a comprehensive survey for the neural network-based deep learning approaches on the acoustic event detection task. Two

Table 3 Selected top systems using conventional machine learning and deep learning approaches on the DCASE Challenge evaluation databases from the years 2013 to 2017

References	Feature	Modeling	*F*-score (%)	Error rate
Vuegen et al. [84]	MFCC	GMM	24.6	1.76
Niessen et al. [56]	MFCC+ZCR+LPC	RF	33.5	1.58
Gemmeke et al. [25]	Mel-spectrogram	NMF	13.2	1.56
Schröder et al. [72]	Gabor filter	HMM	41.5	1.51
Gorin et al. [28]	Mel-band energy	CNN	41.1	0.97
Elizalde et al. [22]	MFCC	RF	33.6	0.96
Kong et al. [41]	MFCC	DNN	36.3	0.95
Lai et al. [43]	MFCC	Fusion	34.5	0.92
Zoehrer et al. [43]	spectrogram	GRNN	39.6	0.90
Adavanne et al. [1]	Mel-band energy	RNN	47.8	0.80
Heittola et al. [53]	Mel-band energy	DNN	42.8	0.94
Lu et al. [49]	MFCC	RNN	39.6	0.83
Meyer et al. [35]	Mel-band energy	CNN	40.8	0.81
Adavanne et al. [4]	Mel-band energy	CRNN	41.7	0.79

A higher *F*-score and a lower error rate correspond to a better system. The systems are sorted based on the error rate

Table 4 Some typical neural network structures and the corresponding detection error rates in acoustic event detection

Reference	Neural network structure	Size	Error rate
Gorin et al. [28]	Convolution layer	80 filters (6 × 60)	0.97
	Convolutional layer	80 filters (1 × 3)	
	Feed-forward layer	1024 hidden units	
	Feed-forward layer	1024 hidden units	
Heittola et al. [53]	Feed-forward layer	50 hidden units	0.94
	Feed-forward layer	50 hidden units	
Adavanne et al. [1]	Recurrent layer (LSTM)	32 hidden units	0.80
	Recurrent layer (LSTM)	32 hidden units	
Meyer et al. [35]	6 convolution layers	64 filters (1 × 3)	0.81
Adavanne et al. [4]	3 convolution layers	128 filters (3 × 3)	0.79
	2 recurrent layers	32 hidden units)	

types of acoustic event detection, namely the strongly and the weakly labeled acoustic event detection, are surveyed in this paper. Our survey is different to that of [15] by including works on the important weakly labeled acoustic event detection problem (where only the presence of acoustic events is given in the training process) which is one of the new tasks for the DCASE Challenge 2017.

This survey is organized as follows. Some common acoustic event databases and evaluation metrics are introduced in Sect. 2. In Sect. 3, the strongly labeled acoustic event detection is first introduced. Afterward, applications of some state-of-art deep learning approaches on the strongly labeled acoustic event detection are elaborated. In Sect. 4, we introduce the weakly labeled acoustic event detection task and the recent advances in that area. The reasons why neural network-based deep learning approaches benefit the AED task and the issues to be studied further are given in Sect. 5. We conclude the paper in Sect. 6.

2 Metrics and Databases in Acoustic Event Detection

2.1 Evaluation Metrics

There are two ways of evaluation in the system performance, namely the segment-based and event-based statics [54] when the system output and the ground truth label are compared in fixed length intervals or at event instance level.

- Segment-based metric For segment-based metric, the predicted active acoustic events are determined in a fixed short time interval with the true positive (tp), false positive (fp), false negative (fn) and true negative (tn) defined, respectively. The true positive means that the acoustic event exists in both the system output and the ground truth label simultaneously. The false positive denotes that the system determines the acoustic event as active, while the true label for the acoustic event is inactive. The false negative means the system fails to detect the acoustic event when the reference indicates the acoustic event to be active. The true negative means the system and the ground truth both determine the acoustic event as inactive.
- Event-based metrics For event-based evaluation metric, the system output and the ground true label are compared event by event. Similarly, the true positive, false positive, false negative and true negative are defined. All these mentioned statics are calculated based on the fact whether the system output which has a temporal position overlaps with the temporal position in the ground true label. A tolerance with respect to the ground true label is usually allowed.

The F-score and the error rate (ER) are commonly adopted as the final evaluation metrics when the segment-based and event-based statics are available.

- F-score
  Based on the segment-based and event-based statics, the segment-based and event-based F-score can be calculated as:

$$F = \frac{2 \times P \times R}{P + R} \tag{1}$$

where the P and R denote the precision and recall accuracy, respectively. The P and R are expressed as:

$$\begin{aligned} P &= \frac{tp}{tp + fp} \\ R &= \frac{tp}{tp + fn} \end{aligned} \tag{2}$$

Table 5 Some commonly adopted acoustic event detection databases

Reference	Name	Event classes	Event count
Adavanne et al. [4]	TUT Sound Events 2017	6	729
Beltrán et al. [7]	Sound Events	20	1367
Cotton et al. [13]	FBK-IRST isolated database	16	576
Gemmeke et al. [24]	AudioSet	632	2, 084, 320
Piczak et al. [63]	ESC-50	50	2000
Salamon et al. [69]	Urban-sound	10	3075
Salamon et al. [70]	Urban-sed	10	50, 000
Xia et al. [85]	UPC-TALP isolated database	14	1026

– Error rate
The error rate measures the number of prediction errors regarding the insertions (I), the deletions (D) and the substitutions (S). The errors are calculated segment by segment. In a segment seg, the number of insertions $I(seg)$ is the number of incorrect system outputs, the number of deletions $D(seg)$ is the number of ground truth events that are not correctly identified, and the number of substitutions $S(seg)$ means the number of acoustic events for which some other acoustic events are the outputs rather than the correct acoustic events. The error rate can be calculated as:

$$ER = \frac{\sum_{seg=1}^{seg=S} I(seg) + \sum_{seg=1}^{seg=S} D(seg) + \sum_{seg=1}^{seg=S} S(seg)}{N(seg)} \tag{3}$$

where seg denotes the seg^{th} segment, $N(seg)$ is the number acoustic events annotated as active in the segment seg, and the $I(seg)$, $D(seg)$ and $S(seg)$ can be expressed as:

$$\begin{aligned} I(seg) &= \max(0, fp(seg) - fn(seg)) \\ D(seg) &= \max(0, fn(seg) - fp(seg)) \\ S(seg) &= \min(fn(seg), fp(seg)) \end{aligned} \tag{4}$$

2.2 Datasets

In this part, several commonly used acoustic event detection databases are listed in Table 5. In Table 5, the number of acoustic event classes, the acoustic event segments and the corresponding references are shown. A detailed description of the acoustic event detection databases can be referred in.[1]

[1] http://www.cs.tut.fi/~heittolt/datasets.

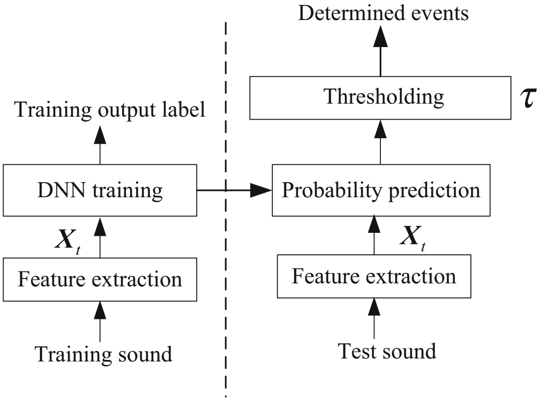


Fig. 1 The flowchart of the DNN-based strongly labeled acoustic event detection system

3 Strongly Labeled Acoustic Event Detection

For strongly labeled acoustic event detection, the acoustic event types and the event localizations are annotated in the training set. The task is to detect the acoustic event types and the happening times given an audio stream during testing.

The inputs of the AED system are the acoustic features $\boldsymbol{X}_t$ and the acoustic features of each frame are associated with one output label vector, which can be written in binary format as:

$$\boldsymbol{y}_t = \{y_{t,1}, y_{t,2}, \ldots, t_{t,e}, \ldots, t_{t,E}\} \tag{5}$$

where $y_{t,e}$ is equal to 1 when the eth event type is active at time index t. Otherwise, $y_{t,e}$ is set to 0. The E is the total number of acoustic event types of interest.

The training space Ω_{strong} for the strongly labeled AED system training can be expressed as:

$$\Omega_{strong} = \{\boldsymbol{X}_t, \boldsymbol{y}_t\} \tag{6}$$

Figure 1 shows the general flowchart of the DNN-based strongly labeled AED system. As shown in Fig. 1, each frame corresponds to one input feature vector $\boldsymbol{X}_t$ and one output training label $\boldsymbol{y}_t$. The neural network classifier is trained in a supervised way and outputs the continuous probabilities representing the probability that each frame belongs to the event classes of interest. The binary cross-entropy function [17] is adopted as the training criteria, which can be expressed as:

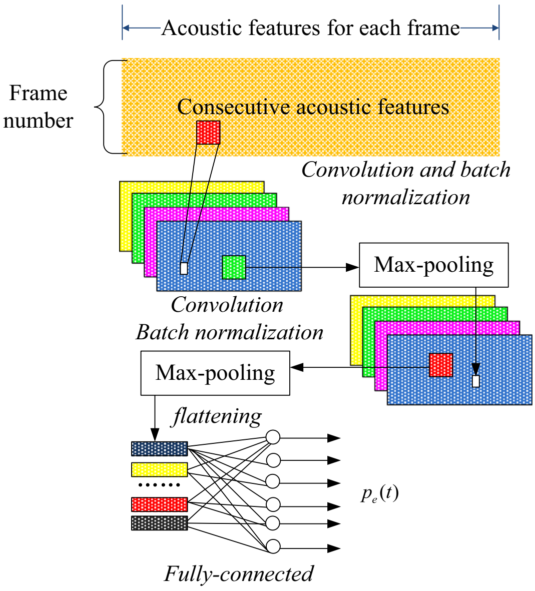
$$L = -q \times log(p) - (1 - q) \times log(1 - p) \tag{7}$$

where q is the target probability from the training database and p is the estimated probability that the current frame belongs to a certain event type. The q is equal to 1 if the training vector corresponds to the ground truth label and p is the sigmoidal output of the deep neural network.

During testing, with the trained acoustic model and the given test audio stream, each time index t will correspond to E output probability predictions, which are expressed as:

$$\hat{\boldsymbol{y}}_t = \{\hat{y}_{t,1}, \hat{y}_{t,2}, \ldots, \hat{y}_{t,e}, \ldots, \hat{y}_{t,E}\} \tag{8}$$

Fig. 2 The training process of the CNN-based strongly labeled acoustic event detection system (Figure extracted from [88])



where $\hat{y}_{t,e}$ represents the probability that the current frame t belongs to the eth event type. Afterward, a global threshold τ, which is empirically set, is applied to $\hat{\boldsymbol{y}}_t$. Event classes with a higher probability than the global threshold are detected as the final active acoustic events.

3.1 CNN in Strongly Labeled AED

The flowchart in Fig. 1 can be applied to the CNN-based strongly labeled AED when the DNN classifier is replaced by the CNN classifier in the training process. Figure 2 shows the training process of the CNN-based strongly labeled AED system. The convolutional neural network model structure includes convolutional layers, max-pooling layers, a flattening layer and a sigmoid output layer. The convolution operation performs the high-level feature extraction. The sub-sampling operation is performed, and max-pooling operations are carried out over the entire sequence length. Typically, the Relu or the sigmoid activation function is used for the kernels. As there may be more than one acoustic event happening at the same time index, a sigmoid layer composed of fully connected neurons is used. The binary cross-entropy is adopted as the loss function in training.

3.2 RNN in Strongly Labeled AED

The same flowchart in Fig. 3 can be applied to the RNN-based strongly labeled AED when the DNN-based classifier is replaced with the RNN classifier. The RNN classifier is adopted in order to utilize the long context information.

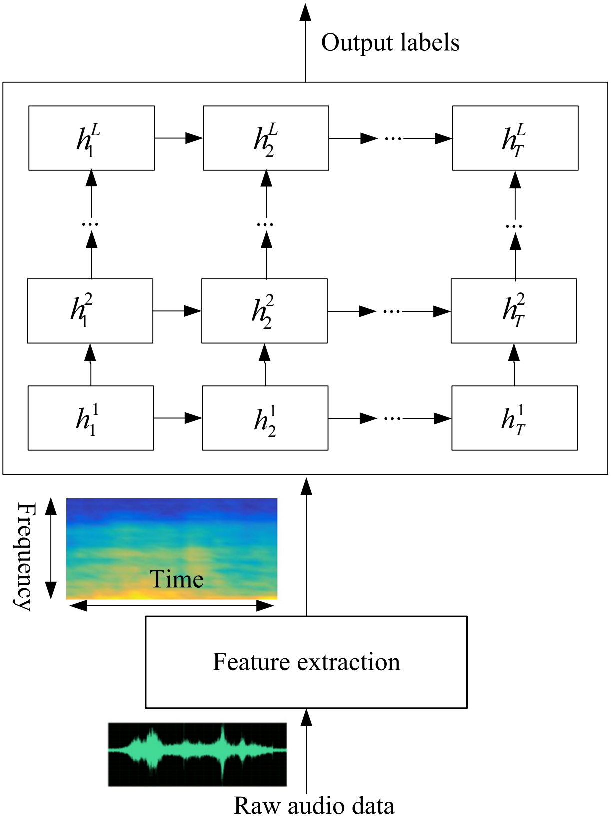


Fig. 3 Training process for the RNN-based strongly labeled acoustic event detection. The current hidden layer output depends on both the input and the previous hidden neurons, which effectively utilizes the long context information

Figure 3 shows the basic concept of the RNN training process. As shown in Fig. 3, the current hidden layer depends on both the input and the previous hidden neurons. Multiple RNN hidden units are stacked on top of each other. The hidden state sequence of the lower layer can be computed as:

$$h_t^l = H(h_t^{l-1}, h_{t-1}^l)\ (1 \leq l \leq L) \tag{9}$$

Here, h^l becomes the first input layers when l equals to 0. The output of the RNN is expressed as:

$$\mathbf{o} = W_{h,y} h_T^L + b_y \tag{10}$$

where $W_{h,y}$, h_T^L and b_y are the weight parameters between the output layer and the last hidden layer, the last hidden layer output and the bias, respectively. Afterward, the $\mathbf{o}$ is presented to a sigmoid layer to get the predicted probability $\hat{\boldsymbol{y}}_t$.

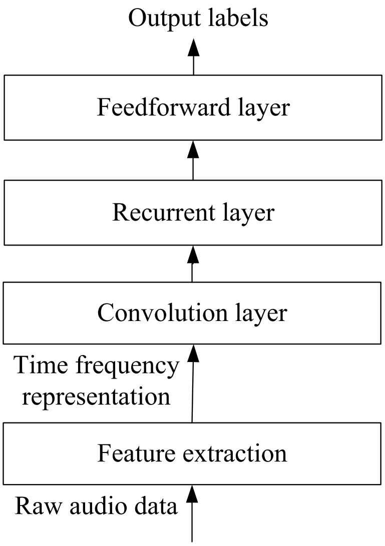


Fig. 4 Training process of the CRNN-based strongly labeled AED. The extracted acoustic features of consecutive frames are fed to the convolutional layers. Stacked outputs of the convolutional layers are fed to the recurrent network, activations of which act as the inputs to the feed-forward layers

3.3 CRNN in Strongly Labeled AED

Figure 4 shows the training process of the CRNN-based strongly labeled AED system. The testing process is the same as testing process in Fig. 1. From Fig. 4, there are three function blocks used for the training, namely the convolution layer block, the recurrent layer block and the feed-forward layer block. The convolution layer block extracts the high-level features with the acoustic features of consecutive frames as the input. The stacked features from the convolutional and max-pooling layers are then fed to the recurrent layer block. The feed-forward layer block with sigmoid activation function is used for the classification as the output layer, and the cross-entropy is adopted as the loss function, which is expressed as:

$$J_{CE}(\mathbf{W}, \boldsymbol{b}) = -\sum_{t}\sum_{e} log\ v_{e,t}^{L} \tag{11}$$

Here $v_{e,t}^{L}$ is the probability estimated from the neural network $P_{NN}(e|\boldsymbol{X}_t)$, which is the RNN output in the training process.

During testing, with the trained acoustic neural network model and the given test audio stream, each time index t will correspond to E output probabilities which is expressed as:

$$\hat{y}_t = \{\hat{y}_{t,1}, \hat{y}_{t,2}, \ldots, \hat{y}_{t,e}, \ldots, \hat{y}_{t,E}\} \tag{12}$$

where $\hat{y}_{t,e}$ represents the probability that the current frame t belongs to the eth event type.

4 Weakly Labeled Acoustic Event Detection

The weakly labeled acoustic event detection research is a recent hot topic area [4,74], and only the presence of the acoustic events is annotated in each audio segment which makes the acoustic event detection more challenging. Let $\boldsymbol{R} = \{R_s : s = 1 \rightarrow N\}$ be the audio recording collections and $\boldsymbol{EV}_s = \{EV_{s1}, EV_{s2}, \ldots, EV_{se}, \ldots, EV_{sE}\}$ be the corresponding acoustic events presented in the sth audio recording. Here the N and E denote the number of audio recordings and event types of interest. For each R_s, the presence of acoustic events $\boldsymbol{EV}_s$ is annotated but without annotating the localization of each event in R_s (weak label for each acoustic event).

The AED system inputs are the acoustic features of one certain audio stream $\boldsymbol{X}_s$, and the training outputs are the recording-wise-based label $\boldsymbol{EV}_s$ rather than the frame-wise-based labels $\boldsymbol{y}_t$. The training space Ω_{weak} for the strongly labeled AED system can be expressed as:

$$\Omega_{\text{weak}} = \{\boldsymbol{X}_s, \boldsymbol{EV}_s\} \tag{13}$$

Here s is the audio recording index and $\boldsymbol{X}_s$ is the acoustic feature vector for the sth audio recording. The $\boldsymbol{EV}_s$ are the training output labels represented as binary vectors.

Although the localization of the active acoustic events is not known in the training set for the weakly labeled AED, the task of the weakly labeled AED is exactly the same as the strongly labeled acoustic events which is to predict both the types and the localizations of the active acoustic events in the test audio stream.

Since the happening times of each present acoustic events are not known in the training set, it is impossible to use the audio segments that contain only the events of interest to train the acoustic models in a supervised way. In this section, how to train the acoustic models using the weakly labeled data to perform the acoustic event detection task will be elaborated.

4.1 Multiple Instance Learning for Weakly Labeled AED

One common technique to perform the weakly labeled AED is to treat the task as a multiple instance learning problem [20,42], in which labels are known for a collection of instances.

4.1.1 Multiple Instance Learning

Multiple instance learning is based on bag–label pairs rather than the instance–label pairs. Here, the "bag" is a collection of instances. Two types of bags, namely the positive bag and the negative bag, are used in the training process. The positive bags contain at least one positive instance which belongs to the target class to be classified. The negative bags are only collections of negative instances.

Let the bag–label pairs be (B_s, Y_s), where B_s and Y_s denote the sth bag and the assigned label for this bag. The B_s contain multiple instances a_{sj} where j is from 1 to n_s and n_s is the number of instances in the B_s. The bag B_s can be expressed as:

$$B_s = \{a_{sj} : j = 1 \rightarrow n_s\} \tag{14}$$

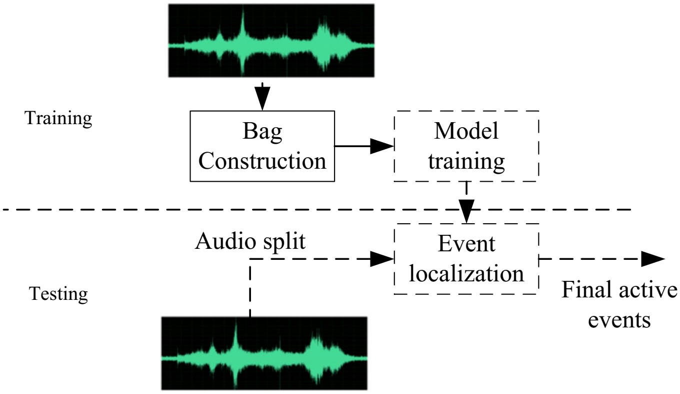


Fig. 5 The weakly labeled AED system based on the multiple instance learning

If all the instances in B_s are negative, the label for B_s is -1. The label for B_s is 1 if there is at least one positive instances in B_s. The label Y_s for the B_s can be expressed as:

$$Y_s = \max_{1 \le j \le n_s} \{y_{sj}\}$$

where y_{sj} denotes the actual label for the jth instance in the sth bag.

4.1.2 Multiple Instance Learning for Weakly Labeled AED

Figure 5 shows the general flowchart of the multiple instance learning-based weakly labeled AED. As shown in Fig. 5, the bag construction, model training and the event localization constitute the AED system.

- Bag construction
  To perform the weakly labeled AED, each audio recording R_s and the label EV_s can be treated as one bag and the corresponding bag label. The audio recording R_s is segmented into a number of short sub-segments $\{R_{s,1}, R_{s,2}, \ldots, R_{s,k}, \ldots, R_{s,K}\}$ where k and K denote the kth audio segment and the total number of instances in the bag.
- Model training
  In the training process, the conventional instance-level-based loss function is replaced by the multiple instance learning-based loss function due to the fact that only the bag-level labels are provided in the training set. The multiple instance learning-based loss function can be expressed as:

$$J_{\text{loss}} = \sum_{s=1}^{s=S} J_{s,\text{loss}} \tag{15}$$

The $J_{s,\text{loss}}$ is the loss of the bag B_s. The $J_{s,\text{loss}}$ is computed as:

$$J_{s,\text{loss}} = \frac{1}{2}\left(\max_{1 \le j \le n_s} o_{s,j} - d_s\right)^2$$

where $o_{s,j}$ is the neural network output for the jth instance in the sth bag and the d_s is the manually annotated bag label for the sth bag.
It is worthy noting that the weight parameters are updated after all instances in bag have been fed forward to the network. The process is then continued until the overall divergence falls to a desired tolerance or the maximum iteration has been reached. By training the acoustic models based on the bag–label pairs, the trained model outputs both the instance-wise probabilities o_{sj} for each instance in the bag and the bag labels if maximal-scoring strategy is applied to the instance-wise probabilities.

– Event localization
Once the training is complete, the trained models can classify the individual instances by outputting the instance probability $o_{s,j}$, which means the constructed system can not only detect the presence of an event in a test recording but also in individual segments. In order to perform the localization task, the testing recording RT is first split into K short sub-segments $\{RT_1, RT_2, \ldots, RT_K\}$. The localization of the sub-segment RT_k is between $(k-1) \times l^{'}$ and $(k-1) \times l^{'} + l$ where l is the length of the segment in seconds and $l^{'}$ is the overlapping length with previous segment. If one acoustic event is detected in RT_k, then the localization of the detected acoustic event is $(k-1) \times l^{'}$ and $(k-1) \times l^{'} + l$.

4.2 Variational Deep Learning Approaches in Weakly Labeled AED

Deep learning approaches from the strongly labeled AED can also be applied to the weakly labeled AED with different scoring or training strategies.

– **Scoring Strategy**
In [45], the authors used a global-input CNN model and the separated-input model to perform the weakly labeled AED. The global-input CNN model takes the spectrogram as the input and the bag labels as the output. The separated-input models are trained using n-second segmented waveform as the input. All the short sub-segments that make up the audio recording R_s are assigned with the same label (bag label). As the global-input CNN model is expected to have better performance than the separated-input models by using the correct label, predictions of the global-input model are spread evenly and then subsequently averaged with the predictions of other separated-input models. The work of [16] also proposed to use the sub-segments of each audio recording to train the separated acoustic models.

– **Training Strategy**
In [89], the authors proposed to a gated convolutional neural network (GCRNN) in the weakly labeled AED. In that work, in order to get the localization information, an additional feed-forward neural network with softmax as the activation function is introduced. The pooling operation was only applied to the frequency

domain rather than the time domain to keep the time resolution of the whole audio spectrogram. Authors in [3] also adopted the same strategy by performing the max-pooling along the frequency domain to preserve the input time resolution. During training, the weak labels help with controlling the learning of strong labels by weighting the loss at the weak and the strong outputs differently.

5 Discussions

The recent advances in neural network-based deep learning for AED are reviewed in this paper. The introduction of neural network-based deep learning approaches undoubtedly has boosted the acoustic event detection performance. In this section, we briefly discuss the key reasons behind the success of the neural network-based deep learning methods and several potential issues for further consideration in the area of acoustic event detection.

5.1 Benefits of Deep Learning

Here, we list three main advantages of the neural network-based deep learning in AED as follows:

- **Effective Representations** Neural network-based deep learning approaches in acoustic event detection can learn more comprehensive and informative information from the raw data, which greatly benefits the acoustic event detection when the various event classes and the noise-like characteristic of the acoustic events challenge the acoustic event detection performance. Intraclass variations and spectral-temporal properties across classes pose challenges to acoustic event detection. The deep learning approaches effectively deal with the aforementioned challenges. The state-of-the-art acoustic event detection system [59] adopted the CNN to extract the high-level information from the spectrogram information with a subsequent recurrent neural network to utilize the long context information. Due to the effective learning and representation, the neural network-based deep learning approaches greatly improve and extend the frontier of the acoustic event detection.
- **Powerful Relationship Modeling** With the various types of activation functions, neural networks have the ability to model nonlinear and complicated relationships between inputs and outputs. This is a great advantage for dealing with natural signals sampled from real-world scenarios and for predicting unseen data. In the real-world acoustic event detection, some acoustic events sound similar but are actually different, such as the "car" and "bus". The deep learning approaches can successfully learn the inherent relationship between different event classes rather than training the acoustic models based on one specific event class as adopted in the conventional machine learning methods. That is one key reason why the deep learning approaches are achieving higher detection performance than the conventional machine learning methods, such as GMM and SVM in the area of acoustic event detection in recent campaigns [52]
- **Flexible Setting of Networks** Neural network-based deep learning approaches can be applied to the AED task using a flexible architecture with diverse combinations

of different neural networks. The deep learning approaches in the strongly labeled acoustic event detection system where the training process is instance–label pair-based can be flexibly transferred to the weakly labeled acoustic event detection with some variations. The works [3,89] applied the CRNN network from the strongly labeled acoustic event detection to the weakly labeled acoustic event detection by only pooling the frequency domain and keeping the time resolution fixed thus performing the detection task when the annotations are weakly labeled. Authors in [16,45] similarly adopted the CNN structures from the strongly labeled acoustic event detection to perform the task of weakly labeled acoustic event detection by splitting the audio recordings into multiple segments with the global-input and separated-input models trained, respectively.

5.2 Future Issues

Although the neural network-based deep learning approaches have been successfully applied to the acoustic event detection task, there are still some issues that need to be resolved in order to further extend the frontier of the acoustic event detection. Here, we list two main challenges facing the deep learning-based acoustic event detection as following:

- **Weakly Labeled and Imbalanced Training Data** A powerful neural network is always associated with a large amount of training data. However, in the area of acoustic event detection, the audio recordings can be obtained easily while the annotation process is always expensive especially when the precise localizations of the polyphonic acoustic events need to be labeled, which leads to the weakly labeled data problem. The other problem is the imbalanced training data. For some acoustic events, such as "break squeaking", the audio collection process is not as easy as the collection process of common events such as "people speaking" and "car". How to effectively deal with the limited and imbalanced training data using the neural network-based deep learning approaches are facing the acoustic event detection.
- **Hyper-parameters and Architectures** High-performance acoustic models are associated with good neural network structures and fine-tuning strategies. The deep learning models are influenced by various aspects, such as network topology, training method and hyper-parameters. In the area of acoustic event detection, the large amount of event classes and the uncertainties when different acoustic events overlap with each other require the deep learning approaches to deal with the hyper-parameters and the neural network structures in order to avoid the general traps, such as the over-fitting or local optimum problem.

6 Conclusion

In this paper, the recent neural network-based deep learning approaches on the acoustic event detection task are reviewed. Two types of acoustic event detection, namely the strongly labeled acoustic event detection and the weakly labeled acoustic event detection, are first introduced with subsequent elaboration on different deep learning approaches applied to these two acoustic event detection tasks.

Neural network-based deep learning approaches have demonstrated remarkable success in acoustic event detection task and outperformed other conventional machine learning techniques. Meanwhile, the advances in the hardware equipments, such as high-performance GPUs, also accelerate the development of the deep learning approaches in acoustic event detection task. However, there are still many challenges, such as the limited training data and the hyper-parameter fine-tuning, facing the deep learning-based acoustic event detection.

Acknowledgements This work was supported by the International Postgraduate Research Scholarship (IPRS) from the University of Western Australia.

References

1. S. Adavanne, G. Parascandolo, P. Pertila, T. Heittola, T. Virtanen, Sound event detection in multichannel audio using spatial and harmonic features. arXiv preprint arXiv:1706.02293 (2017)
2. S. Adavanne, P. Pertila, T. Virtanen, Sound event detection using spatial features and convolutional recurrent neural network. arXiv preprint arXiv:1706.02291 (2017)
3. S. Adavanne, T. Virtanen, Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. arXiv preprint arXiv:1710.02998 (2017)
4. S. Adavanne, T. Virtanen, A report on sound event detection with different binaural features, in *Workshop on DCASE Challenge, Tech. Rep.* (2017)
5. S. Adavanne, T. Virtanen, Sound event detection using weakly labeled dataset with stacked convolutional and recurrent neural network. arXiv preprint arXiv:1710.02998 (2017)
6. A. Antoniou, A. Storkey, H. Edwards, Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017)
7. J. Beltran, E. Chavez, J. Favela, Scalable identification of mixed environmental sounds, recorded from heterogeneous sources. Pattern Recognit. Lett. **68**, 153–160 (2015)
8. E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, T. Virtanen, Convolutional recurrent neural networks for bird audio detection, in *IEEE Signal Processing Conference (EUSIPCO)* (2017), pp. 1744–1748
9. E. Cakir, T. Heittola, H. Huttunen, T. Virtanen, Polyphonic sound event detection using multi label deep neural networks, in *International Joint Conference on Neural Networks (IJCNN)* (2015), pp. 1–7
10. E. Cakir, T. Virtanen, End-to-end polyphonic sound event detection using convolutional recurrent neural networks with learned time-frequency representation input. arXiv preprint arXiv:1805.03647 (2018)
11. S.Y. Chou, S.R. Jang, Y.H. Yang, FrameCNN: a weakly-supervised learning framework for frame-wise acoustic event detection and classification. Recall **14**, 55–64 (2017)
12. C. Clavel, T. Ehrette, G. Richard, Events detection for an audio-based surveillance system, in *IEEE International Conference on Multimedia and Expo (ICME)* (2005), pp. 1306–1309
13. C.V. Cotton, D.P. Ellis, Spectral vs. spectro-temporal features for acoustic event detection, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2011), pp. 69–72
14. J.L. Dai Wei, P. Pham, S. Das, S. Qu, F. Metze, Sound event detection for real life audio DCASE challenge, in *Proceedings of the Workshop Detection and Classification of Acoustic Scenes and Events* (2016)
15. A. Dang, T.H. Vu, J.C. Wang, A survey of deep learning for polyphonic sound event detection, in *IEEE International Conference on Orange Technologies (ICOT)* (2017), pp. 75–78
16. A. Dang, T.H. Vu, J.C. Wang, Deep learning for DCASE2017 challenge. Workshop on DCASE2017 Challenge, Tech. Rep. (2017)
17. P.T. De Boer, D.P. Kroese, S. Mannor, R.Y. Rubinstein, A tutorial on the cross-entropy method. Ann. Oper. Res. **134**(1), 19–67 (2005)

18. E.L. Denton, S. Chintala, R. Fergus, Deep generative image models using a laplacian pyramid of adversarial networks, in *Advances in neural information processing systems (NIPS)* (2015), pp. 1486–1494
19. A. Dessein, A. Cont, G. Lemaitre, Real-time detection of overlapping sound events with non-negative matrix factorization, in *Matrix Information Geometry* (2017), pp. 341–371
20. T.G. Dietterich, R.H. Lathrop, T. Lozano Perez, Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1–2), 31–71 (1997)
21. A. Diment, T. Heittola, T. Virtanen, Sound event detection for office live and office synthetic AASP challenge, in *Proceedings of the IEEE AASP Challenge on Detection Classif. Acoust. Scenes Events (WASPAA)* (2013)
22. B. Elizalde, K. Anurag, S. Ankit, B. Rohan, V. Emmanuel, R. Bhiksha, L. Ian, Experimentation on the DCASE challenge 2016: Task 1 Acoustic scene classification and task 3 Sound event detection in real life audio. DCASE Challenge, Tech. Rep.(2016)
23. D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
24. J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human labeled dataset for audio events, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), pp. 776–780
25. J.F. Gemmeke, L. Vuegen, P. Karsmakers, B. Vanrumste, An exemplar-based NMF approach to audio event detection, in *IEEE Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2013), pp. 1–4
26. D. Giannoulis, D. Stowell, E. Benetos, M. Rossignol, M. Lagrange, M.D. Plumbley, A database and challenge for acoustic scene classification and event detection, in *IEEE Signal Processing Conference (EUSIPCO)* (2013), pp. 1–5
27. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems* (2014), pp. 2672–2680
28. A. Gorin, N. Makhazhanov, N. Shmyrev, DCASE sound event detection system based on convolutional neural network. Workshop on DCASE Challenge, Tech. Rep. (2016)
29. R. Grzeszick, A. Plinge, G.A. Fink, Bag-of-features methods for acoustic event detection and classification. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1242–1252 (2017)
30. T. Heittola, A. Mesaros, T. Virtanen, M. Gabbouj, Supervised model training for overlapping sound events based on unsupervised source separation, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 8677–8681
31. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, S. Hochreiter, GANs trained by a two time-scale update rule converge to a nash equilibrium. arXiv preprint arXiv:1706.08500 (2017)
32. G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
33. G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
34. Y. Hou, S. Li, Sound event detection in real life audio using multimodel system. DCASE Challenge, Tech. Rep. (2017)
35. I.Y. Jeong, S. Lee, Y. Han, K. Lee, Audio event detection using multiple-input convolutional neural network, in *Workshop on DCASE Challenge, Tech. Rep.* (2017)
36. F. Jin, F. Sattar, S. Krishnan, Log-frequency spectrogram for respiratory sound monitoring, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2012), pp. 597–600
37. D.P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, M. Welling, Improving variational inference with inverse autoregressive flow, in arXiv preprint arXiv:1606.04934 (2016)
38. D.P. Kingma, M. Welling, Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
39. H.G. Kim, J.Y. Kim, Acoustic event detection in multichannel audio using gated recurrent neural networks with high resolution spectral features. ETRI J. **39**(6), 832–840 (2017)
40. T. Komatsu, Y. Senda, R. Kondo, Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation, in *IEEE Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 2259–2263
41. Q. Kong, I. Sobieraj, W. Wang, M. Plumbley, Deep neural network baseline for DCASE challenge (2016)

42. A. Kumar, B. Raj, Audio event detection using weakly labeled data, in *ACM Proceedings on Multimedia Conference* (2016), pp. 1038–1047
43. Y.H. Lai, C.H. Wang, S.Y. Hou, B.Y. Chen, Y. Tsao, Y.W. Liu, DCASE report for task 3 Sound event detection in real life audio, in *Workshop on DCASE Challenge, Tech. Rep.* (2016)
44. P. Laffitte, D. Sodoyer, C. Tatkeu, L. Girin, Deep neural networks for automatic detection of screams and shouted speech in subway trains, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 6460–6464
45. D. Lee, S. Lee, Y. Han, K. Lee, Ensemble of convolutional neural networks for weakly-supervised sound event detection using multiple scale input. Workshop on DCASE2017 Challenge, Tech. Rep. (2017)
46. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature **521**(7553), 436 (2015)
47. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. IEEE Proceedings **86**(11), 2278–2324 (1998)
48. X. Lin, J. Liu, X. Kang, Audio recapture detection with convolutional neural networks. IEEE Trans. Multimed. **18**(8), 1480–1487 (2016)
49. R. Lu, Z. Duan, Bidirectional GRU for sound event detection. Workshop on DCASE2017 Challenge, Tech. Rep. (2017)
50. A. Makhzani, B. Frey, K-sparse autoencoders. arXiv preprint arXiv:1312.5663 (2013)
51. M. Meyer, L. Cavigelli, L. Thiele, Efficient convolutional neural network for audio event detection. arXiv preprint arXiv:1709.09888 (2017)
52. A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge. IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) **26**(2), 379–393 (2018)
53. A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, T. Virtanen, DCASE 2017 challenge setup: tasks, datasets and baseline system, in DCASE2017 Challenge, Tech. Rep. (2017)
54. A. Mesaros, T. Heittola, T. Virtanen, Metrics for polyphonic sound event detection. Appl. Sci. **6**(6), 162 (2016)
55. S. Mun, S. Park, D.K. Han, H. Ko, Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane, in *Proc. DCASE* (2017), pp. 93–97
56. M.E. Niessen, T.L. Van Kasteren, A. Merentitis, Hierarchical sound event detection, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2013)
57. W. Nogueira, G. Roma, P. Herrera, Automatic event classification using front end single channel noise reduction, MFCC features and a support vector machine classifier, in *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events* (2013), pp. 1–2
58. A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, in *Proceedings of the 34th International Conference on Machine Learning*. **70**, 2642–2651 (2017)
59. 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**(6), 1291–1303 (2017)
60. G. Parascandolo, H. Huttunen, T. Virtanen, Recurrent neural networks for polyphonic sound event detection in real life recordings, in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 6440–6444
61. S. Passler, W.J. Fischer, Food intake monitoring: Automated chew event detection in chewing sounds. IEEE J. Biomed. Health Informat. **18**(1), 278–289 (2014)
62. H. Phan, M. Maass, R. Mazur, A. Mertins, Random regression forests for acoustic event detection and classification. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 20–31 (2015)
63. K.J. Piczak, Environmental sound classification with convolutional neural networks, in IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2015), pp. 1–6
64. C. Poultney, S. Chopra, Y.L. Cun, Efficient learning of sparse representations with an energy-based model, in *Advances in neural information processing systems (NIPS)* (2007), pp. 1137–1144
65. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
66. F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain. Psychol. Rev. **65**(6), 386 (1958)
67. D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep. (1985)

68. J. Salamon, J.P. Bello, Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Process. Lett. **24**(3), 279–283 (2017)
69. J. Salamon, C. Jacoby, J.P. Bello, A dataset and taxonomy for urban sound research, in *Proceedings of the ACM international conference on Multimedia* (2014), pp 1041–1044
70. J. Salamon, D. MacConnell, M. Cartwright, P. Li, J.P. Bello, Scaper: A library for soundscape synthesis and augmentation, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2017), pp. 344–348
71. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in *Advances in Neural Information Processing Systems* (2016), pp. 2234–2242
72. J. Schroder, B. Cauchi, R., M. Schadler, N. Moritz, K. Adiloglu, J. Anemuller, S. Doclo, B. Kollmeier, S. Goetze, Acoustic event detection using signal enhancement and spectro-temporal feature extraction. in *IEEE Workshop on Applicat. Signal Process. Audio Acoust. (WASPAA)* (2013)
73. J. Schroder, S. Goetze, V. Grutzmacher, J. Anemuller, Automatic acoustic siren detection in traffic noise by part-based models, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 493–497
74. R. Serizel, N. Turpault, H. Eghbal-Zadeh, A.P. Shah, Large-scale weakly labeled semi-supervised sound event detection in domestic environments. arXiv preprint arXiv:1807.10501 (2018)
75. R. Stiefelhagen, K. Bernardin, R. Bowers, R.T. Rose, M. Michel, J. Garofolo, The CLEAR 2007 evaluation, in *Multimodal Technologies for Perception of Humans* (2017), pp. 3–34
76. T.W. Su, J.Y. Liu, Y.H. Yang, Weakly-supervised audio event detection using event-specific gaussian filters and fully convolutional networks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2017), pp. 791–795
77. A. Temko, D. Macho, C. Nadeu, Fuzzy integral based information fusion for classification of highly confusable non-speech sounds. Pattern Recognit. **41**(5), 1814–1823 (2008)
78. A. Temko, C. Nadeu, Acoustic event detection in meeting-room environments. Pattern Recognit. Lett. **30**(14), 1281–1288 (2009)
79. A. Temko, C. Nadeu, Classification of acoustic events using SVM-based clustering schemes. Pattern Recognit. **39**(4), 682–694 (2006)
80. G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, A. Sarti, Scream and gunshot detection and localization for audio-surveillance systems, in *IEEE Advanced Video and Signal Based Surveillance (AVSS)* (2007), pp. 21–26
81. P. Vincent, H. Larochelle, Y. Bengio, P.A. Manzagol, Extracting and composing robust features with denoising autoencoders, in *ACM Proceedings of the 25th international conference on Machine learning* (2008), pp. 1096–1103
82. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**(Dec), 3371–3408 (2010)
83. T. Virtanen, A. Mesaros, T. Heittola, M. Plumbley, P. Foster, E. Benetos, M. Lagrange, in *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)* (2016)
84. L. Vuegen, B.V.D. Broeck, P. Karsmakers, J.F. Gemmeke, B. Vanrumste, H.V. Hamme, An MFCC-GMM approach for event detection and classification, in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2013), pp. 1–3
85. X. Xia, R. Togneri, F. Sohel, D. Huang, Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features. Pattern Recognit. **81**, 1–13 (2018)
86. X. Xia, R. Togneri, F. Sohel, D. Huang, Frame wise dynamic threshold based polyphonic acoustic event detection, in *Proc. Interspeech* (2017), pp. 474–478
87. X. Xia, R. Togneri, F. Sohel, D. Huang, Class wise distance based acoustic event detection. Tech. Rep., DCASE Challenge (2017)
88. X. Xia, R. Togneri, F. Sohel, D. Huang, Confidence based acoustic event detection, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 306–310
89. Y. Xu, Q. Kong, W. Wang, M.D. Plumbley, Large-scale weakly supervised audio classification using gated convolutional neural network, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 121–125
90. Y. Yang, J. Jiang, Bi-weighted ensemble via HMM-based approaches for temporal data clustering. Pattern Recognit. **76**, 391–403 (2018)
91. J. Yu, C. Chaomurilige, M.S. Yang, On convergence and parameter selection of the EM and DA-EM algorithms for Gaussian mixtures. Pattern Recognit. **77**, 188–203 (2018)

92. X. Zhu, Y. Liu, Z. Qin, J. Li, Data augmentation in emotion classification using generative adversarial networks. arXiv preprint arXiv:1711.00648 (2017)
93. X. Zhuang, J. Huang, G. Potamianos, M. Hasegawa-Johnson, Acoustic fall detection using gaussian mixture models and GMM supervectors (2019), pp. 69–72
94. X. Zhuang, X. Zhou, M.A. Hasegawa-Johnson, T.S. Huang, Real-world acoustic event detection. Pattern Recognit. Lett. **31**(12), 1543–1551 (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.