INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Frame-wise dynamic threshold based polyphonic acoustic event detection

Xianjun Xia[1], *Roberto Togneri*[1], *Ferdous Sohel*[2], *David Huang*[1]

[1]School of Electrical, Electronic and Computer Engineering, The University of Western Australia
[2]School of Engineering and Information Technology, Murdoch University

Xianjun.Xia@research.uwa.edu.au,{Roberto.Togneri,David.Huang}@uwa.edu.au,
F.Sohel@murdoch.edu.au

Abstract

Acoustic event detection, the determination of the acoustic event type and the localisation of the event, has been widely applied in many real-world applications. Many works adopt multi-label classification techniques to perform the polyphonic acoustic event detection with a global threshold to detect the active acoustic events. However, the global threshold has to be set manually and is highly dependent on the database being tested. To deal with this, we replaced the fixed threshold method with a frame-wise dynamic threshold approach in this paper. Two novel approaches, namely contour and regressor based dynamic threshold approaches are proposed in this work. Experimental results on the popular TUT Acoustic Scenes 2016 database of polyphonic events demonstrated the superior performance of the proposed approaches.

Index Terms: acoustic event detection, multi-label classification, dynamic threshold.

1. Introduction

Acoustic event detection (AED) deals with the event type and the localization (determination of the start and end positions) of the acoustic events. Acoustic event detection has been widely applied in many real world applications, such as in surveillance systems [1], siren detection systems [2], chew event detection systems [3] and human-computer interaction [4][5][6]. Intra-class variations and the spectral-temporal properties across classes pose great challenges to acoustic event detection. Due to the significant real world applications of AED and the challenges being faced, some campaigns, such as CLEAR [7] and D-CASE [8][9] have attempted to capture the wide range of variations in the design of the acoustic event detection databases [10][11][12].

Many approaches are proposed based on the classification framework. Local acoustic features, such as zero-crossing rates, energy coefficients and Mel-frequency cepstral coefficients (MFCC) are extracted. Then, these local features are modelled by some representative models, such as Gaussian Mixture Models (GMM) [13] or Hidden Markov Models (HMM) [14]. In [15][16][17], random forest techniques were utilized to perform the acoustic event detection task. While testing, a segmented event is recognized under the criteria of maximum posterior probability. Recently, motivated by the successful application of neural networks in speech signal processing [18] and image processing [19], deep neural networks (DNN) and recurrent neural networks (RNN) based approaches have been proposed to deal with the challenging real world polyphonic acoustic event detection. In [20][21][22], the DNN was employed to tackle the problem of polyphonic acoustic event detection. Recurrent neural networks have been adopted in [23][24] to deal with the polyphonic acoustic event detection problem in DCASE 2016 [12].

When dealing with the polyphonic acoustic events using the neural network method, a threshold (applied to the output probabilities across different acoustic event types) is used to determine the presence of acoustic events. In this paper, the acoustic events are defined as active acoustic events when they show their presence within the frames under consideration. According to [21], the accuracy is high for high threshold values in the low polyphony levels, where the polyphony level reflects the number of active sources. On the other hand, the accuracy is high for low threshold value when the acoustic signal stream is highly polyphonic. The recall rate would decrease if the threshold is set too high and the precision would decrease if the threshold is set too low. However, the level of polyphony for the test audio stream is unknown and varies with each frame. In [21][22][23], the thresholds were manually set with values of 0.5, 0.95 and 0.5 respectively, which cannot capture the polyphonic level of the test acoustic stream at each frame.

To deal with the complex and polyphonic level changes across time with each frame during test, this paper proposes two frame-wise dynamic threshold approaches to automatically determine the threshold: i) A straightforward contour based dynamic threshold approach; and ii) A novel regressor based dynamic threshold approach. The contour based dynamic threshold approach utilizes the output probability information and the regressor based dynamic threshold approach adopts a regressor to estimate the frame-wise threshold for each frame index. There are two advantages by replacing the fixed threshold with a frame-wise dynamic threshold. To begin with, the frame-wise dynamic threshold can avoid setting the global threshold manually, which requires expert knowledge to set the threshold correctly for the database under consideration. Moreover, the frame-wise dynamic threshold approach can automatically deal with varying polyphonic levels and estimate the frame-wise threshold accordingly.

The structure of this paper is as follows. In Section 2, an overview of the fixed threshold based AED systems is shown. Our proposed approach and algorithms are described in Section 3. In Section 4, we provide the experimental results followed by conclusion and future work in Section 5.

2. Fixed threshold based AED system

2.1. The task of the polyphonic AED system

Fig. 1 shows the task of polyphonic acoustic event detection. As shown in Fig. 1, each frame may correspond to more than one acoustic label ('people speaking' and 'car passing by' overlap with each other). In a polyphonic acoustic event detection system, the determination of the event type and position can be regarded as a multi-label classification problem.

Fig. 2 shows the flowchart of the multi-label classification

http://dx.doi.org/10.21437/Interspeech.2017-746

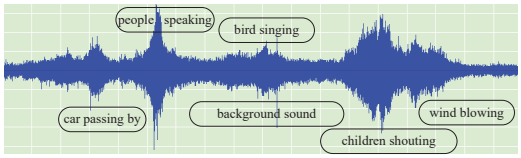

Figure 1: *Polyphonic acoustic event detection task.*

based acoustic event detection system [20]. As shown in Fig. 2, the multi-label classification based acoustic event detection system is made up of four components, namely feature extraction, frame-wise model training, event probability estimation and event type detection to determine the active acoustic event type. During the frame-wise feature extraction, each frame corresponds to one output training label and an input feature vector. The training labels, which can be obtained from the given labeled onset and offset time of the database, are in binary format. For each training frame, the corresponding output training label is a binary representation for each acoustic event type. The training label at frame k is expressed as $L_k = \{l_{k,1}, l_{k,2}, ..., l_{k,N}\}$, where $l_{k,n}$ ($n \in \{1, 2, ..., N\}$) is set to 1 when the nth event is active at frame index k and N is the number of event types of interest.

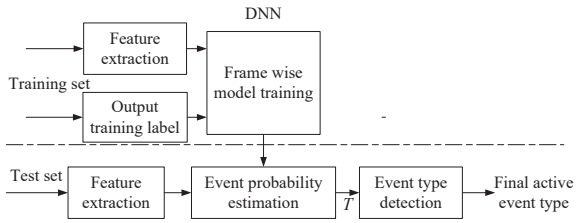

Figure 2: *The flowchart of a multi-label classification based AED system [20].*

2.2. Frame-wise model training in the AED system

In this work, a fully connected deep neural network is adopted as the classifier. The DNN classifier outputs the continuous probabilities representing the probability that each frame belongs to the event classes of interest.

For a deep neural network based acoustic event detection, we adopted [20] as our benchmark system. In this approach, the AED task is regarded as a multi-label classification problem and the class labels are converted into binary format units. Then the binary cross-entropy function [25] is adopted as the training criteria. The binary cross-entropy is the loss function of choice for multi-label classification problems and sigmoidal output units, which can be expressed as:

$$L = -t \times log(p) - (1 - t) \times log(1 - p) \quad (1)$$

where t is the target probability from the training database and p is the estimated probability that the current frame belongs to a certain event type. In this work, t is set to 1 if the training vector corresponds to the ground truth label and p is the sigmoidal output of the deep neural network.

2.3. Event type detection

Upon testing, with the trained acoustic classifier and the given test audio stream, each frame index k will correspond to N probabilities $p_{k,1}, p_{k,2}, ..., p_{k,N}$, where $p_{k,n}$ represents the probability that the current frame k belongs to the nth event type. For the monophonic acoustic event detection, the event type with the highest probability would be detected as the final active event. However, for the polyphonic acoustic event detection, a threshold T is often used to determine the active acoustic events. Fig. 3 shows the principle of the threshold in an AED system. The horizontal axis denotes the number of frames multiplied by the number of event classes N, where the kth vertical line is the kth frame and the subsequent probabilities to the next vertical line are those for the N event classes for that frame. The vertical axis represents the probability for each frame. The T, N and frame number are set to 0.2, 8 and 10 for the example in Fig. 3. The nth event is detected as an active event if $p_{k,n}$ is higher than T at frame index k. If $p_{k,n}$ for $n \in \{1, 2, ..., N\}$ are all lower than the threshold T, the system detects no active event.

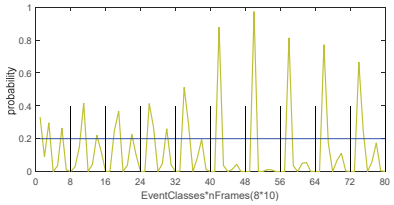

Figure 3: *The threshold principle in an AED system.*

3. The proposed dynamic threshold AED system

In this section, we propose both a contour based and a regressor based dynamic threshold approach for the polyphonic acoustic event detection task. Fig. 4 shows the real world acoustic event output probability from the classifier. As displayed in Fig. 4, if the threshold is set too high (red threshold), the recall rate will decrease (the active acoustic events which fall in the range of the red rectangular part will not be detected). If the threshold is set too low (black threshold), the precision will decrease because many acoustic events which are not active will be detected as active.

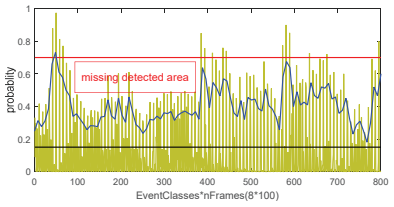

Figure 4: *Drawbacks of using fixed thresholds in an AED system.*

3.1. Contour-based dynamic threshold scheme

A straightforward approach to avoid the threshold being too high or too low is to adaptively use the contour information. The probability contour is derived by plotting for each frame

the highest output probability for that frame across the N event classes. If the contour peaks of the probabilities are directly used, the task would be for a monophonic problem. To deal with this, a global coefficient α is adopted. The threshold for the frame index k can be defined as:

$$T_k^{Con} = \alpha * Con(k) \tag{2}$$

$$Con(k) = max\{p_{k,1}, p_{k,2}, ..., p_{k,N}\} \tag{3}$$

where $n \in \{1, 2, ..., N\}$ and $p_{k,n}$ is the probability that the kth frame belongs to the event type n. Here, α is set globally according to the whole training database:

$$\alpha = \frac{C_{active}}{C_{total}} \tag{4}$$

where C_{active} and C_{total} denote the number of frames which correspond to more than one active event and the number of frames of the whole training set respectively. Then the dynamic threshold corresponding to each frame is used to detect the active acoustic events.

3.2. Regressor based dynamic threshold scheme

Another effective way to generate the frame- wise dynamic threshold is to use a regressor to estimate the threshold. In this work, we used the acoustic features as the input and the output probability information from the training set as the output to train the threshold estimator using Long-Short Term Memory (LSTM) based recurrent neural network. Fig. 5 shows the flowchart of the regressor based dynamic threshold scheme in our AED system. As shown in Fig. 5, we use a DNN classifier trained in the same way as the baseline system. Then the trained DNN classifier is used to evaluate all the training data. The training instances $j \in \{1, 2, ..., J\}$ are chosen to train the regressor where J is the total number of frames with their event types being correctly detected in the training set. The acoustic features of the chosen frames and the corresponding output probabilities from the DNN classifier are used as the input and target respectively for the regressor. However we found that the output probability distribution of the training data will usually differ from that of the test data. We address this by using the following unit normalised output probability as the target to the RNN regressor:

$$U_j = \frac{p_j - f_{min}}{(f_{max} - f_{min})} \tag{5}$$

where p_j is the highest probability among the N target event types at the frame index j. The p_j, f_{max} and f_{min} are defined as:

$$p_j = max(p_{j,1}, p_{j,2}, ..., p_{j,N}) \tag{6}$$

$$\begin{aligned} f_{max} &= max(p_1, p_2, ..., p_J) \\ f_{min} &= min(p_1, p_2, ..., p_J) \end{aligned} \tag{7}$$

Given a test audio stream, the DNN classifier will output the event type probability $p_{k,n}$ for each acoustic event n and the RNN based regressor will output the normalised probability estimation $\hat{U}_k$ at each frame index k. From the DNN classifier output for the test audio stream we also derive the $\hat{f}_{max}$ and $\hat{f}_{min}$ values. The dynamic threshold for the frame index k is expressed as:

$$T_k^{reg} = \hat{f}_{min} + \hat{U}_k \times (\hat{f}_{max} - \hat{f}_{min}) \tag{8}$$

which is then used to determine the presence of the acoustic events.

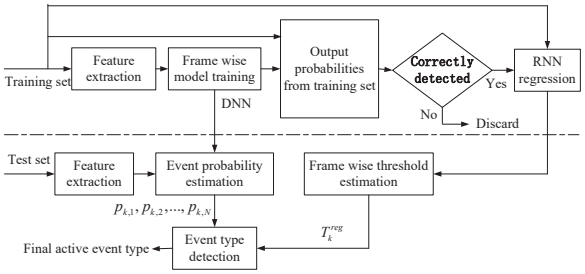

Figure 5: *The flowchart of the proposed regressor based dynamic threshold in the AED system.*

4. Experimental results and analysis

4.1. Evaluation database

Our two proposed threshold schemes are evaluated on the acoustic events from the TUT Acoustic Scenes 2016 database for polyphonic acoustic events [12], which is a popular overlapped acoustic database. The complex acoustic events consist of two common everyday environments, namely the residential area and home environment. According to [12], these two environments are present in outdoor surveillance and indoor human activity monitoring. With the residential area as the environment, 8 acoustic event types, (object) banging, bird singing, car passing by, children shouting, people speaking, people walking, wind blowing and background audio are recorded and annotated. For home environment acoustic events, 12 acoustic event types including the background are recorded and annotated. The recorded audio events with home environment as the background are (object) rustling, (object) snapping, cupboard, cutlery, dishes, drawer, glassing jingling, object impact, people walking, washing dishes and water tap running. Details can be found in [12].

In the provided development subset, the acoustic events across different environments are partitioned into four folds of training and test data. Each recording is used only once in the test data and the classes in the test data are a subset of the training set.

4.2. Evaluation metric

In this work, segment-based F-score in a fixed time grid [26] is adopted as the evaluation metric. A segment-based metric is performed in short segments. In a short segment, the number of true positive (TP), false positive (FP) and false negative (FN) are calculated to get the segment-based F-score, which can be expressed as:

$$\text{F-score} = \frac{2 \times P \times R}{P + R} \tag{9}$$

where

$$P = \frac{TP}{TP + FP} \quad \text{and} \quad R = \frac{TP}{TP + FN} \tag{10}$$

True positive denotes that the ground truth (reference) and system prediction both indicate an event is active in a short segment. False positive means the system judges an acoustic event which is inactive as being active in one segment. False negative denotes that the system fails to detect the active acoustic events in one segment. The length of the short segment is set to 100ms as in [27].

Table 1: *Different F-scores with different set of thresholds on the residential area data.*

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Contour based	Regressor based
fold 1	43.9%	46.5%	52.2%	57.5%	62.7%	63.4%	64.2%	67.3%	**69.7%**	63.2%	68.2%
fold 2	30.2%	31.8%	33.3%	34.5%	35.4%	37.2%	**38.9%**	38.6%	35.1%	35.4%	38.6%
fold 3	39.4%	42.8%	43.8%	**45.0%**	44.2%	39.3%	38.0%	37.6%	37.1%	43.5%	44.7%
fold 4	47.0%	48.2%	51.2%	50.7%	49.5%	49.0%	47.9%	43.0%	42.4%	49.6%	**53.1%**
Average	40.1%	42.3%	45.1%	46.9%	48.0%	47.2%	47.3%	46.6%	46.1%	47.9%	**51.2%**

Table 2: *Different F-scores with different set of thresholds on the home environment data.*

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	Contour based	Regressor based
fold 1	22.6%	24.4%	26.0%	26.8%	**32.0%**	29.6%	28.0%	26.1%	24.2%	28.5%	31.2%
fold 2	17.2%	18.6%	23.0%	24.6%	25.8%	27.2%	28.5%	29.0%	**30.8%**	30.7%	30.4%
fold 3	22.7%	23.8%	23.9%	24.0%	22.3%	21.0%	19.0%	17.3%	16.7%	23.8%	**26.3%**
fold 4	26.9%	28.4%	29.4%	32.8%	35.2%	36.5%	37.0%	38.4%	39.7%	38.6%	**42.8%**
Average	22.4%	23.8%	25.6%	27.0%	28.8%	28.6%	28.1%	27.7%	27.9%	30.0%	**32.7%**

4.3. Experimental configurations

The same configuration as that in [20] is used in this work to train the fully connected neural network with three hidden layers. For each hidden layer, the number of hidden units is set to 500. To utilize the time sequence information, the mel-filter bank coefficients of 10 frames are concatenated as the input to the neural network. To avoid the over-fitting during the training, a dropout strategy [28] with a value of 0.1 is adopted. The ReLu activation function [29] and the Rmsprop optimizer [30] are used to train the frame-wise classifier. While training the regressor, one layer with 50 LSTM cells is used and the learning rate is set to 0.001. Stochastic Gradient Descent (SGD) [31] is used to minimize the Root Mean Square Error (RMSE).

4.4. Experimental results

To show that the variable thresholds lead to different system performance, the fixed threshold is manually set from 0.1 to 0.9 while determining which acoustic event is active. Table 1 and Table 2 show the different F-score under the environments of residential area and home. As shown in Table 1 and Table 2, performance varies with different thresholds under different environments (residential area and home). In the residential area, thresholds with a value of 0.9, 0.7, 0.4 and 0.3 correspond to the highest F-score on the relevant evaluation fold. For the home environment acoustic event detection systems, thresholds with a value of 0.5, 0.9, 0.4 and 0.9 correspond to the best performance. The performance is random with fixed thresholds on different folds of the database.

When the contour based dynamic threshold method is adopted, the F-score for the four folds obtained on the residential area acoustic events are 63.2%, 35.4%, 43.5% and 49.6% respectively. Results under the home environment are 28.5%, 30.7%, 23.8% and 38.6%, which fall within the range of the relatively high ranked systems with the thresholds being manually set.

To verify the effectiveness of the regressor based dynamic threshold in an AED system, the same experiments are performed on the same four folds. The F-score evaluated on the residential area and home environment acoustic events are 68.2%, 38.6%, 44.7%, 53.1% and 31.2%, 30.4%, 26.3%, 42.8% respectively, which show that the dynamically estimated frame-wise threshold achieves superior performance compared to the fixed threshold approaches. This is because the dynamic thresholds are adaptive to the different portions of the database and this helps to improve the overall AED performance. Moreover, the AED system with the automatic and dynamic threshold can outperform the best manually configured system when evaluations are performed on the 4th fold in the residential area and 3rd, 4th fold under the home environment.

To compare our proposed approaches with other methods, Table 3 shows the average performance on the four folds for the baseline system from D-CASE 2016 [12] (Gaussian Mixture Model based approach), the DNN system from [20] and our proposed system using dynamic threshold approaches. As shown in Table 3, the AED system with the regressor based dynamic threshold scheme achieved the best performance, with average F-scores of 51.2% and 32.7% under the residential area and home environments respectively.

Table 3: *Performance of the different AED systems.*

	GMM [12]	DNN [20]	Contour based	Regressor based
Residential area	35.2%	47.0%	47.9%	**51.2%**
Home	18.1%	29.2%	30.0%	**32.7%**

5. Conclusions and future work

In this paper, we proposed two dynamic threshold based approaches to perform the AED. The contour based dynamic threshold strategy and the innovative employment of a regressor to estimate the output probability have demonstrated their superior performance compared with the baseline system. How to utilize the output probabilities of training frames to train the classifier will be our future research direction.

6. References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2007, pp. 21–26.

[2] J. Schröder, S. Goetze, V. Grutzmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models." in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 493–497.

[3] S. Päßler and W. J. Fischer, "Food intake monitoring: Automated chew event detection in chewing sounds," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 278–289, 2014.

[4] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognition Letters*, vol. 30, no. 14, pp. 1281–1288, 2009.

[5] X. D. Zhuang, Z. Xi, A. H. J. Mark, and S. H. Thomas, "Real-world acoustic event detection," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1543–1551, 2010.

[6] M. Bugalho, J. Portelo, I. Trancoso, T. Pellegrini, and A. Abad, "Detecting audio events for semantic video search," in *Interspeech*. 2009, pp. 1151–1154.

[7] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 311–322.

[8] D. Giannoulis, S. Dan, B. Emmanouil, R. Mathias, L. Mathieu, and D. P. Mark, "A database and challenge for acoustic scene classification and event detection," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2013, pp. 1–5.

[9] B. Elizalde, A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj, and I. Lane, "Experiments on the DCASE challenge 2016: Acoustic scene classification and sound event detection in real life recording," *arXiv preprint arXiv:1607.06706*, 2016.

[10] A. Temko, D. Macho, C. Nadeu, and C. Segura, "Upc-talp database of isolated acoustic events," *Internal UPC report*, vol. 85, 2005.

[11] Z. Christian and O. Maurizio, "Acoustic event detection-itc-irst aed database," *Internal ITC report*, 2005.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Signal Processing Conference (EUSIPCO), 24th European*. IEEE, 2016, pp. 1128–1132.

[13] Z. Xiaodan, J. Huang, G. Potamianos, and M. Hasegawa-Johnson, "Acoustic fall detection using gaussian mixture models and gmm supervectors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 69–72.

[14] J. Schröder, B. Cauchi, M. R. Schädler, N. Moritz, K. Adiloglu, J. Anemüller, S. Doclo, B. Kollmeier, and S. Goetze, "Acoustic event detection using signal enhancement and spectro-temporal feature extraction," in *Proc. Workshop Application of Signal Processing to Audio and Acousticc(WASPAA)*. IEEE, 2013.

[15] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[16] X.-J. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest regression based acoustic event detection with bottleneck features," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, Accepted on 27th February.

[17] X.-J. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest classification based acoustic event detection," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, Accepted on 27th February.

[18] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[19] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[20] Q. Kong, I. Sobieraj, W. Wang, and M. Plumbley, "Deep neural network baseline for DCASE challenge 2016." In Proc. Workshop Detection and Classification of Acoustic Scenes and Events 2016.

[21] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2015, pp. 1–7.

[22] J. L. Dai Wei, P. Pham, S. Das, S. Qu, and F. Metze, "Sound event detection for real life audio DCASE challenge," In Proc. Workshop Detection and Classification of Acoustic Scenes and Events 2016.

[23] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," In Proc. Workshop Detection and Classification of Acoustic Scenes and Events 2016.

[24] T. H. Vu and J.-C. Wang, "Acoustic scene and event recognition using recurrent neural networks," In Proc. Workshop Detection and Classification of Acoustic Scenes and Events 2016.

[25] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of Operations Research*, vol. 134, no. 1, pp. 19–67, 2005.

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

[27] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *Proc. Workshop Application of Signal Processing to Audio and Acousticc(WASPAA)*. IEEE, 2013, pp. 1–4.

[28] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, 2010, pp. 807–814.

[30] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, vol. 4, no. 2, 2012.

[31] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.