CONFIDENCE BASED 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

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 the multi-label classification technique to perform the polyphonic acoustic event detection with a global threshold to detect the active acoustic events. However, the manually labeled boundaries are error-prone and cannot always be accurate, especially when the frame length is too short to be accurately labeled by human annotators. To deal with this, a confidence is assigned to each frame and acoustic event detection is performed using a multi-variable regression approach in this paper. Experimental results on the latest TUT sound event 2017 database of polyphonic events demonstrate the superior performance of the proposed approach compared to the multi-label classification based AED method.

Index Terms— acoustic event detection, multi-label classification, confidence, multi-variable regression

1. INTRODUCTION

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

Many approaches are proposed based on the multi-label 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) [8] or Hidden Markov Models (HMM) [9]. In [10], random forest techniques were utilized to perform the acoustic event detection task. During testing, a segmented event is recognized under the criteria of maximum posterior probability. Recently, motivated by the successful application of neural networks in speech and image processing, deep neural networks (DNN) [11][12] and recurrent neural networks (RNN) [13][14] based approaches have been proposed to deal with the challenging problem of real world polyphonic acoustic event detection.

When the acoustic event detection is performed using the multi-label classification approach, the manually labeled boundaries are converted into frame based training samples corresponding to different acoustic event labels. Usually the frame length varies from 5ms [12] to 100ms [15], which requires the manually labeled boundaries to be accurate when the frame length is short. However, the frame wise labeling accuracy around the event boundaries cannot be always guaranteed due to labelling errors from human annotation, especially when the acoustic events are overlapped, which makes the multi-label classification based acoustic event detection more challenging.

In this work we propose a novel confidence measure which is assigned to each frame. If the current frame is closer to the middle of the manually labeled event boundaries, a higher confidence is assigned to the current frame. By doing this we can achieve: 1) soft boundaries rather than hard boundaries, which makes the acoustic event detection system more tolerant to the label inaccuracy at the boundaries; 2) continuous confidence assigned according to the event position containing rich event position information, which ideally match the event detection task for the determination of the event happening time. After different confidences are assigned to different frames, the training outputs are real-valued variables rather discrete-valued labels. In this paper, we adopt simple parabolic functions to obtain the confidence with a preset probability at the onset and offset frame. Afterwards, the multi-variable regression rather than multi-label classification method is used to perform the acoustic event detection task.

The structure of this paper is as follows. In Section 2, an overview of the multi-label classification based AED system is shown. Our proposed approach and algorithm are described in Section 3. In Section 4, we provide the experimental results followed by the conclusion and future work in Section 5.

2. MULTI-LABEL CLASSIFICATION BASED ACOUSTIC EVENT DETECTION

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 problem.

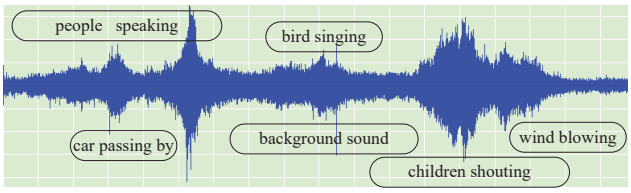

Fig. 1. Polyphonic acoustic event detection task.

2.2. Multi-label classification based AED

In this section, multi-layer perceptron (MLP) and convolutional neural network (CNN) based acoustic event detection systems from our previous work [16] are adopted as the baseline systems.

The audio signal is represented by the log-mel energies corresponding to one output training label. 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 output training label at frame k is expressed as:

$$L_k = \{l_{k,1}, l_{k,2}, ..., l_{k,N}\} \quad (1)$$

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.

For the MLP based acoustic event detection, the multi-layer perceptron with two layers, each with 50 units, is used to construct the multi-class classification based acoustic event detection system. Acoustic features of 5 consecutive frames are concatenated to form the input space. The L_k is the output space for the classifier. The cross entropy is adopted as the loss function and the dropout strategy with a value of 0.2 is used while training the multi-label classifier.

Fig. 2 shows the flowchart of the multi-label classification based acoustic event detection system using CNN. The convolutional neural network model structure includes two convolutional layers, two max-pooling layers, two batch normalization layers, a flattening layer and a sigmoid output layer. The first layer performs a convolution over the input acoustic features with 16 kernels characterized by 3 by 3. The second convolutional layer is the same as the first one except that the number of kernels is set to 32 in order to obtain a higher level representation. The sub-sampling operation is performed and max-pooling operations is done over the entire sequence length. In both convolutional layers, the Relu [17] 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 N fully-connected neurons is used. The binary cross entropy is adopted as the loss function in training and the Adam [18] is used to optimize the network weights.

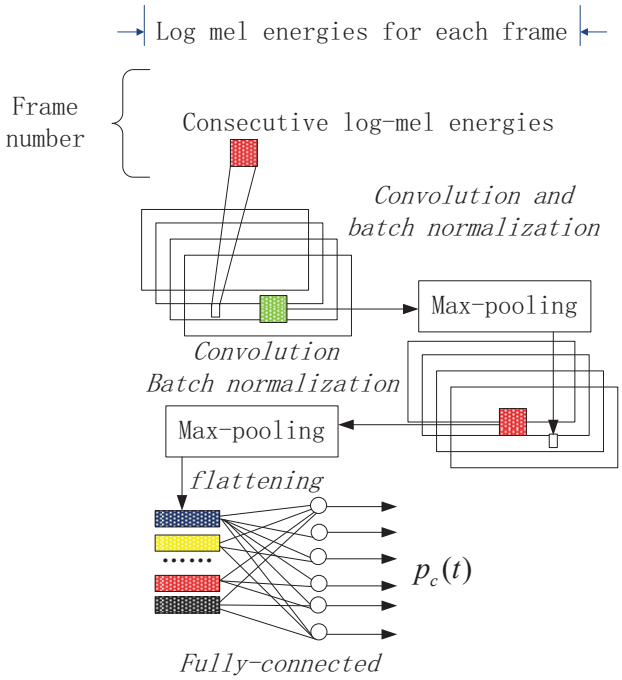

Fig. 2. The convolutional neural network structure for the baseline AED system.

In testing, the audio signal is represented by the log-mel energies. Then the representations are fed to the trained model and the model outputs a probability $p_c(t)$ for each event class c at frame index t. A global threshold is applied to determine the active acoustic events at each frame. If the output probability $p_c(t)$ is higher than the global threshold, then the event class c is detected as active. Otherwise, the event class c is regarded as inactive at the current frame t.

3. CONFIDENCE BASED ACOUSTIC EVENT DETECTION

3.1. Confidence

Two acoustic events, both when overlapped and when not overlapped, are taken as an example to show how the confidences are calculated. As shown in Fig. 3, the red and blue lines denote different acoustic events. The rectangular solid boxes denote the manually labeled boundaries of the acoustic events and the dotted parabolic lines denote the corresponding confidence for each frame. The LB and LE are the manually labeled onset and offset time for the acoustic event respectively. However, the actual happening time for the acoustic event may be different from the LB and LE due to the labeling inaccuracy caused by the limitations of human annotation at the frame level. To deal with this, we assume a soft boundary with different confidences. The closer the current frame is to the centre of the manually labeled acoustic event, the higher the confidence will be. The VB and VE denote the virtual onset and offset time respectively.

The confidence curve represented by the dotted line for a specific acoustic event in Fig. 3, with LB and LE as the manually labeled boundaries, can be expressed as:

$$p(t) = (1 - (1 - v)(\frac{2t - LB - LE}{(LE - LB)})^2)f(t) \quad (2)$$

where $f(t)$ is defined as:

$$f(t) = \begin{cases} 1 \ (VB <= \ t \ <= VE) \\ 0 \ (otherwise) \end{cases} \quad (3)$$

To get different parabolic functions for acoustic events with different durations, the confidence at the manually labeled boundaries are fixed to v, which can be expressed as:

$$v = p(t = LB) = p(t = LE) \quad (4)$$

In this paper, the v is experimentally set to 0.3.

Based on the Eq. (2), the onset VB and offset VE can be obtained when the confidence becomes 0 ($p(t) = 0$). Given the maximum frame number L of one audio file, the VB and VE can be expressed as:

$$VB = max\{0, \frac{LE + LB - (LE - LB)\frac{1}{\sqrt{1-v}}}{2}\} \quad (5)$$

$$VE = min\{L, \frac{LE + LB + (LE - LB)\frac{1}{\sqrt{1-v}}}{2}\} \quad (6)$$

3.2. Confidence Regression

After the confidences are assigned to different frames, the training outputs are real-valued variables rather discrete-valued labels. For each training frame, the corresponding

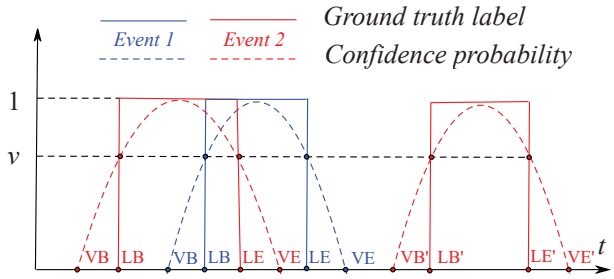

Fig. 3. Confidence for different frame indexes.

output is a continuous representation for each acoustic event type. The training output at frame k is expressed as:

$$P_k = \{p_{k,1}, p_{k,2}, ..., p_{k,N}\} \quad (7)$$

where $p_{k,n}$ ($n \in \{1, 2, ..., N\}$) is the confidence at the frame k for the nth event class.

The same CNN framework and configurations have been adopted as in the benchmark system introduced in Section 2 except that the binary cross entropy loss function be replaced by the mean squared error E expressed as:

$$E = \frac{1}{K} \sum_{k=1}^{k=K} \sum_{n=1}^{n=N} (p_{k,n} - \hat{p}_{k,n})^2 \quad (8)$$

where K is the total number of the training samples, $p_{k,n}$ is the confidence and $\hat{p}_{k,n}$ is the predicted probability for the nth class at the frame index k respectively.

During testing, a global threshold is applied to the predicted probability $\hat{p}_{k,n}$ to determine the final active acoustic events. In this paper, the global threshold is set to 0.5 as in [19].

4. EXPERIMENTAL RESULTS

4.1. Database

The TUT sound event 2017 database [20] is used to evaluate the performance of different systems. The recordings are carried out in a street environment and recorded at 44.1kHz sampling rate and 24 bit resolution. The annotators were instructed to annotate the audio signals with start time and end time. The detailed description of the data recording and annotation procedure can be found in [20].

For the acoustic event detection task in the database, the selected 6 target acoustic event classes are: “brake squeaking”, “car”, “children”, “large vehicle”, “people speaking” and “people walking”.

4.2. Evaluation metrics

The segment-based error rate and F-score are used to evaluate the different AED systems. The segment-based error rate and

F-score are calculated with respect to a segment. In this paper, the duration for the evaluation segment is set to 100ms. A lower error rate or a higher F-score indicates a better AED system. Detailed definitions about the error rate and F-score are described in [21].

4.3. Experimental results and analysis

To demonstrate the effectiveness of the proposed regression based acoustic event detection method using the confidence, different systems including the DCASE 2017 task3 baseline system [19] are constructed for comparison.

To begin with, a multi-layer perceptron with two layers, each with 50 units, is used to construct the multi-class classification and regression based acoustic event detection systems respectively. These two systems are named as *MLP-C* and *MLP-R* in this paper. An average of 0.69 error rate and 56.15% F-score are achieved for the *MLP-C*. For the *MLP-R*, the average error rate decreased by 5% to 0.64 and the F-score increased to 60.14%, which proves the effectiveness of the confidence based AED system.

To further demonstrate the effectiveness of the proposed method, the CNN based framework, as introduced in Section 2, is used to perform the multi-label classification and confidence regression respectively. Similarly, we name the CNN based AED systems as *CNN-C* and *CNN-R* respectively. An average of 0.67 error rate and 56.17% F-score are achieved for *CNN-C*. For *CNN-R*, the average error rate and the F-score are 0.63 and 61.02% respectively. Detailed results for the multi-label classification and confidence regression based approaches on the development dataset are shown in Table 1.

Table 1. Error rates (ER) and F-scores for the different AED systems on the development dataset.

System	ER	F-score
DCASE Task3 Baseline[19]	0.69	56.70%
MLP-C[16]	0.69	56.15%
CNN-C[16]	0.67	56.17%
MLP-R	0.64	60.14%
CNN-R	**0.63**	**61.02%**

Table 2 shows the detection results on the DCASE 2017 Challenge Task 3 evaluation dataset. As shown in the Table 2, the detection accuracy slightly increased and the error decreased to 0.87 and 0.84 when the confidence measure is applied to DNN and CNN respectively.

As can be seen in Table 1 and Table 2, the confidence regression based method has a lower error rate and a higher F-score than the multi-label classification based approach. This can be explained by observing that the soft boundary based confidence can be more tolerant to the manually labelling inaccuracy at the event boundaries. Furthermore, the confidence is positively correlated with the centre of the manually labeled acoustic events, which makes the confidence contains not only capture the acoustic event type information but also the acoustic event localisation information.

Table 2. Error rates (ER) and F-scores for the different AED systems on the evaluation dataset.

System	ER	F-score
DCASE Task3 Baseline[19]	0.94	42.8%
MLP-C	0.95	43.5%
CNN-C	0.94	41.1%
MLP-R	0.87	43.8%
CNN-R	**0.84**	**45.3%**

5. CONCLUSION AND FUTURE WORK

This paper proposed a confidence concept and the regression based acoustic event detection approach. The hard boundary is replaced by a soft boundary and a confidence is assigned to each frame, which makes the regressor more tolerant to manually labeled data and be able to utilize not only the acoustic event type but also the acoustic event localisation information. Experimental results demonstrate the superior performance of the proposed approach. How to train a joint model regarding the binary label and continuous confidence will be our future research direction.

6. ACKNOWLEDGMENT

This work was supported by the International Postgraduate Research Scholarship (IPRS) from the University of Western Australia. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

7. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, pp. 162–178, 2016.

[7] 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.

[8] 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.

[9] 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 Applicat. Signal Process. Audio Acoust.(WASPAA)*. IEEE, 2013.

[10] X. J. Xia, R. Togneri, F. Sohel, and D. Huang, "Random forest classification based acoustic event detection," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 163–168.

[11] Q. Q. Kong, I. Sobieraj, W. W. Wang, and M. Plumbley, "Deep neural network baseline for DCASE challenge 2016," in *Workshop on Detection and Classification of Acoustic Scenes and Events*. DCASE, September 2016.

[12] X. J. Xia, R. Togneri, F. Sohel, and D. Huang, "Frame wise dynamic threshold based polyphonic acoustic event detection," *Proc. Interspeech 2017*, pp. 474–478, 2017.

[13] S. Adavanne, G. Parascandolo, P. Pertilä, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Workshop on the Detection and Classification of Acoustic Scenes and Events*. DCASE, November 2017.

[14] S. Adavanne and T. Virtanen, "A report on sound event detection with different binaural features," Tech. Rep., DCASE Challenge, September 2017.

[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, "Class wise distance based acoustic event detection," Tech. Rep., DCASE Challenge, November 2017.

[17] 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.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Workshop on the Detection and Classification of Acoustic Scenes and Events*. DCASE, November 2017.

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

[21] G. Poliner and D.P.W. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, , no. 1, pp. 154–154, 2007.