INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust Sound Event Detection in Continuous Audio Environments

Haomin Zhang[1], *Ian McLoughlin*[2,1], *Yan Song*[1]

[1]National Engineering Laboratory of Speech and Language Information Processing
The University of Science and Technology of China, Hefei, PRC
[2] School of Computing@Medway, The University of Kent, UK

zhm1991@mail.ustc.edu.cn, ivm@kent.ac.uk, songy@ustc.edu.cn

Abstract

Sound event detection in real world environments has attracted significant research interest recently because of it's applications in popular fields such as machine hearing and automated surveillance, as well as in sound scene understanding. This paper considers continuous robust sound event detection, which means multiple overlapped sound events in different types of interfering noise. First, a standard evaluation task is outlined based upon existing testing data sets for the sound event classification of isolated sounds. This paper then proposes and evaluates the use of spectrogram image features employing an energy detector to segment sound events, before developing a novel segmentation method making use of a Bayesian inference criteria. At the back end, a convolutional neural network is used to classify detected regions, and this combination is compared to several alternative approaches. The proposed method is shown capable of achieving very good performance compared with current state-of-the-art techniques.

Index Terms: sound event detection, convolutional neural network, Bayesian inference, segmentation.

1. Introduction

Continuous sound event detection is an extension of classification methods which are trained to recognise isolated and well separated sounds, allied with a segmentation technique to extract same-sound regions from continuous audio. Within the sound event detection field, traditional features and methods introduced from the speech recognition domain such as MFCCs and HMMs have been shown to not perform as well as spectrogram image features and image classifiers [1] for classification of noise-corrupted sounds. Deep learning has also been applied in this field, achieving excellent results [2], [3], again particularly for classification of noisy sounds. Previous research has evaluated both deep neural networks (DNN) [2] and convolutional neural networks (CNN) [3], with the latter achieving slightly better performance on isolated sound event evaluation tasks.

It is necessary to consider several aspects of real world environmental conditions for robust sound event detection, for example overlapping sound events and background noises. In addition, when operating a system in real environments, usually it is not possible to know a priori which sound events might occur together or overlap. We also don't know in advance what kinds of noise will occur, nor know the signal-to-noise (SNR) ratio of the sounds in the given noise environment.

For training purposes, the above classifiers are trained using data which in most cases (including the standard evaluation task discussed in Section 4) is presented in individual files, each of which contains an isolated sound event without added noise. Several SIFs (spectrogram image features) are obtained from each labelled sound, downsampled, and used to train a CNN which is described in Section 2. Meanwhile testing material should be much more realistic in being noisy, continuous and overlapped. Unlike the evaluation of well-separated sounds, continuous sound events need to be detected and isolated or segmented first, and only then can be classified.

1.1. Contribution

The specific contributions of this paper are firstly to formulate a standard evaluation task for robust, continuous and overlapping sound event detection, constructed from the same underlying data as the current standard evaluation task for isolated sound event classification as used in [1, 4, 2, 3] (see Section 4.1). Secondly, the current state-of-the-art SIF-based CNN classifier, described in Section 2, is evaluated with an energy-based segmentation front-end (Section 3.1). While this will be shown able to perform reasonably well, early testing revealed that some portions of the background noise were inherently similar to one or more sound classes. To reduce the influence of these sounds on the final classification results, an adaptive background penalty is developed in Section 3.2. Finally, a novel BIC segmentation method is proposed (Section 3.3) to isolate individual sound regions prior to classification and evaluated in Section 4.2, demonstrating excellent classification performance overall.

2. SIF-CNN based classification system

2.1. Spectrogram image feature

Unlike speech, sound events contain more random time-frequency structures which make the spectrogram look more like an unstructured image to the human eye. Fortunately, both DNN and CNN classifiers have been shown very capable of extracting discriminative information from spectrogram features. The best performing SIF extraction process from [2, 3] is;

- Take FFT of highly overlapped window sequence to obtain a spectrogram.
- Downsample the spectrogram by averaging frequency regions.
- Smooth the downsampled spectrogram.
- Denoise by subtracting the minimum amplitude in each smoothed frequency channel.
- Detect high energy frames and their immediate context to form a rectangular time-frequency image for classification.

http://dx.doi.org/10.21437/Interspeech.2016-392

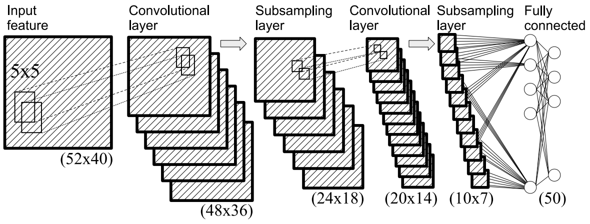

Figure 1: *Diagram of the CNN classifier using SIF features.*

The final image for CNN classification has a dimension of 52×40, where 52 is the number of downsampled frequency channels and 40 defines the context size, in overlapped frames.

2.2. Convolutional neural network

CNNs are generally known to be good at learning structures in images, and this has been demonstrated in image processing [5, 6], speech recognition [7, 8] and similar domains. In this machine hearing application, a typical CNN structure is used, namely 5-layers (2 convolutional layers, 2 subsampling layers and 1 full connected layer), with $52 \times 40 = 2080$ input dimensionality and 50 output classes from the final fully connected (FC) layers. Apart from where noted, the structure, shown as a block diagram in Fig. 1 is as described in [3].

3. Energy detector and BIC separation

3.1. Energy detector

We only select and detect high energy frames plus context from the continuous sound files. This is reasonable in that we also only use the high energy frames for training (i.e. the classification features are the high energy frames and their immediate context, as described in Section 2). In operation, any frame is detected whose energy exceeds a threshold and is also the maximum over the 80 frames before and 80 frames after. The threshold is set relative to the long term average energy, to confer a degree of noise resistance, and the hold-off period is designed to ensure that loud sounds spanning multiple frames do not dominate over quieter sounds occurring elsewhere. The feature images, which are centred on the detected frames, are classified by CNN for recognition. This is a low-complexity detector, but is also effective in practice, surprisingly even for noisy conditions.

The CNN classifier yields a posterior probability vector P_i, for each detected high energy region, where $i = 1...50$. Index $j = \arg\max(P_i), i = 1...50$ identifies the highest probability class, but this is only accepted if $P_j > P_{th}$, otherwise this sound event is classed as noise. The influence of P_{th} is discussed in Section 4.2.

3.2. Penalty on background probability

In order to reduce the influence of background noise, one technique is to classify all frames using the CNN to obtain a background probability distribution. We define P_{ij} where $i = 1...50$ being the class number and $j = 1...N$ is the frame number where N is the total number of frames. Then we define $Pc_k = \Sigma_j P_{ij}/N$ where $k = 1...50$ to denote the average probability of each class k. Now, when we need to detect frame j, instead of using $P_{kj} = \max(P_{ij}), i = 1...50$, we compute

$$P_{kj} = \max(P_{ij} - \lambda(Pc_i - \Sigma_l Pc_l/50)), i = 1...50, \lambda = 0.2$$

and compare this against P_{th} as before. If $P_{kj} > P_{th}$, we detect frame j as sound event class k, otherwise we assume it is background noise. Thus, if the background probability Pc_k is very high, which means this class k is very likely to be found in the background noise of this sound file, the output probability P_{kj} will be reduced and this frame may not actually be detected as class k in the end, but instead as another class. Again, this technique will be evaluated in Section 4.2 and shown to perform better than the energy detector alone, when tested with the same classifier and same test material.

3.3. BIC separation

The final contribution of this paper revolves around the need to segment continuous audio prior to classification, and proposes a novel segmentation method based on Bayesian inference. During testing, each frame of input sound contains background noise (apart from the 'clean' condition) and 0, 1 or more sound classes. High energy frames are very important decision points for classification as we have seen, but the classification of these frames applies the classification result to all frames within a detected audio region. The high energy frames themselves do not necessarily reveal the start and end points of a particular sound - especially when noise corrupted. Thus we develop a segmentation heuristic which is inspired by speaker region separation techniques for diarization [9]. This Bayesian inference criteria (BIC) decides in a probabilistic sense between two alternative hypotheses that are namely whether an input array z follows a single Gaussian distribution or can be separated into two parts x and y that follow two different Gaussian distributions. The criterion determines which hypothesis is a better match to the underlying data, on the assumption that the sounds can be represented by Gaussian distributions, and that these differ for different sounds. If $\mathcal{H}_0$ and $\mathcal{H}_1$ denotes these two hypotheses, we can define,

$$\begin{aligned} \Delta B &= BIC(\mathcal{H}_1) - BIC(\mathcal{H}_0) \\ &= Nlog|\Sigma_z| - \frac{1}{2}\lambda(d + d(d+1)/2)logN \\ &- N_y log|\Sigma_y| - N_x log|\Sigma_x| \end{aligned} \tag{1}$$

where N, N_x and N_y are the lengths of arrays z, x, and y ($N = N_x + N_y$) while d is the feature dimension and the penalty term for the model complexity, λ, is set to 1.0 for all evaluations. Next, we compute ΔB for every possible x and y within limits. If $\max(\Delta B) > 0$, then hypothesis $\mathcal{H}_1$(x and y follow two Gaussian distributions separately) is true and $t = \arg\max(\Delta B)$ marks a separation point whereas if $\max(\Delta B) <= 0$, then hypothesis $\mathcal{H}_0$ is true and there is no partition in array z.

In this task, z is actually an array of features, while d is the dimension of the features and N is the number of frames in the array. Here we use 39-dimension MFCC features (13 MFCCs, Δs and $\Delta\Delta$s) for segmentation – however we do not use MFCC data for classification. This process will result in at most one possible separation point. If we advance the window forward at the desired resolution and repeat the process, we will get all possible separations from the entire array of continuous audio.

Finally, the energy detector described in Section 3.1 is applied to the detected segments using the same criteria: the highest energy frames in each segment are detected. Within one

Table 1: *Precision, recall and F_1 of the CNN classifier using only energy detector for feature selection, for a range of different probability thresholds.*

P_{th}	0.9	0.8	0.7	0.5	0.3	0.1
Precision	92.2	84.6	78.8	68.7	65.0	64.8
Recall	62.7	71.3	75.8	80.9	82.3	82.3
F_1	74.7	77.3	77.3	74.3	72.6	72.5

segment, the features have similar statistical distributions (in an MFCC sense) and are thus more likely belong to the same sound event, and this is naturally represented by the highest energy region (bearing in mind also that the energy trace has been smoothed as part of the spectrogram processing). The following section will now evaluate each of these segmentation and detection processes in turn.

4. Experiments and Results

4.1. The evaluation task

The sound material contains 50 different sound event classes with 80 files per class, randomly selected from the Real World Computing Partnership (RWCP) Sound Scene Database in Real Acoustic Environments [10] according to selection criteria in [1]. Of the 80 files, 50 are randomly selected to be the training set ($50 \times 50 = 2500$) and the remainder ($30 \times 50 = 1500$) used for evaluation.

The process for forming the test is that we first create 100 separate 1-min long empty files. Then we add 15 random sound events into each file at random time points. Finally, we choose one type of background noise to add to each file randomly from four different NOISEX-92 noises (specifically "Destroyer Control Room", "Speech Babble", "Factory Floor 1" and "JetCockpit 1"). Given 4 different noise conditions (clean, 20dB, 10dB and 0dB SNR), there are now 400 multi-sound testing files in total. The classifier used in this paper is implemented using the CNN toolbox with 5 layers in total as specified in Section 2.2 [11].

To evaluate our system, we define precision as $P = M/N$, where M is the number of sound events we detect correctly, and N is the number of all the detection we make, as well as recall $R = K/1500$, where K is the total number of sound events we detect among the 1500 events per noise condition. Besides this, we combine both scores to derive a single overall metric $F_1 = 2/(P^{-1} + R^{-1})$.

4.2. Results and discussion

The baseline score will be that obtained using only an energy detector allied with the CNN classifier. For simplicity, we will reproduce only the average performance of the 4 different noise SNRs (clean, 20dB, 10dB and 0dB SNR), with F_1 results shown in Table 1 for different values of P_{th}.

Looking at the table, we can see that the best F_1 is achieved at a P_{th} of 0.8 or 0.7, however the best Precision has a higher P_{th} and the best recall is at a lower P_{th}.

Next, we apply the probability penalty to the CNN output with results as shown in Table 2. Again the best F_1 score has a P_{th} around 0.7 to 0.8, whereas the best precision and recall are also at the extremes of the table. Clearly, the P_{th} setting is operating as a tradeoff between the two conflicting demands of better recall or better precision.

Finally, we explore the BIC separation method as a front-end segmentation technique before the energy detector, and including the application of a probability penalty. The results are shown in Table 3 and reveal that the optimum P_{th} for overall F_1 score is now lower at about 0.5. Interestingly, while the precision score has improved substantially over other methods, the recall score is slightly worse. The final combined F_1 score achieves over 80% accuracy.

Table 2: *Precision, recall and F_1 of the CNN classifier, with a probability penalty applied to combat noise-like sound classes, for a range of different probability thresholds.*

P_{th}	0.9	0.8	0.7	0.5	0.3	0.1
Precision	95.0	90.0	83.9	72.1	65.4	64.7
Recall	60.4	70.0	75.5	80.7	82.3	82.3
F_1	73.8	78.7	79.5	76.2	72.9	72.4

Table 3: *Precision, recall and F_1 of the CNN classifier with a BIC segmentation front end, using energy detection and with a probability penalty, shown for a range of different probability thresholds.*

P_{th}	0.9	0.8	0.7	0.5	0.3	0.1
Precision	96.5	94.0	91.2	86.7	78.4	77.7
Recall	57.4	66.5	71.5	75.0	78.1	78.1
F_1	72.0	77.9	80.2	80.4	78.2	77.9

Comparing these results with several alternative systems, Table 4 shows the SIF-CNN baseline (SIF-CNN/Baseline) and final CNN classifier system with BIC segmentation (SIF-CNN/Final), HMM based on MFCC features (MFCC-HMM), SIF with an SVM classifier (SIF-SVM), and SIF features with a DNN classifier from [2] (SIF-DNN), in the 4 different levels of noise. In this case, we set $P_{th} = 0.7$ for the proposed CNN classifiers. While mean recall of the final system is slightly worse than the SIF-DNN and SIF-CNN systems, the average precision is very much higher, particularly in high levels of noise. Consequently, the overall F_1 score of the proposed system is improved over the baseline as well as over existing methods.

To better understand the trade-offs inherent in the three proposed SIF-CNN systems, two graphs are presented to explore the P_{th} tradeoff. Fig. 2 shows ROC (receiver operating curve) plots of the three techniques introduced in this paper in which the Y-axis shows precision while the X-axis shows recall. The BIC-based SIF-CNN/Final system is clearly superior to the system with a probability penalty, which in turn outperforms the basic energy detector (SIF-CNN/Baseline). Meanwhile Fig. 3 reveals how the overall F_1 score changes with P_{th}. On the whole, the probability penalty and the BIC separation improve system performance. However when P_{th} becomes too high, for example 0.8 or 0.9, the results are confusing. It is worth bearing in mind that it is probably not sensible to set P_{th} so high, because it means that many true sound events are ignored, although the fewer that are detected are more often classified correctly, thus a good balance would be achieved with P_{th} around 0.5 or 0.6.

5. Conclusion

This paper has proposed a method of robust continuous sound event detection. Firstly, spectrogram based image features are used rather than traditional auditory features such as MFCC, in order to obtain a better two-dimensional description of sound

Table 4: *Comparison between proposed system and other systems for different levels of background noise.*

System	Precision					Recall					F_1
SNR	clean	20dB	10dB	0dB	mean	clean	20dB	10dB	0dB	mean	
MFCC-HMM	28.12	08.69	06.60	04.57	12.00	94.87	79.20	60.47	38.53	68.27	20.41
SIF-SVM	90.84	85.87	57.32	27.51	65.39	86.93	86.80	85.60	71.20	82.63	73.01
SIF-DNN	87.70	82.53	53.69	24.63	62.14	84.87	84.33	81.33	64.13	78.67	69.43
SIF-CNN/Baseline	93.66	92.03	77.99	51.67	78.84	81.80	81.67	79.33	60.47	75.82	77.30
SIF-CNN/Final	95.79	94.95	89.67	84.40	91.20	76.67	77.73	75.53	56.20	71.53	80.18

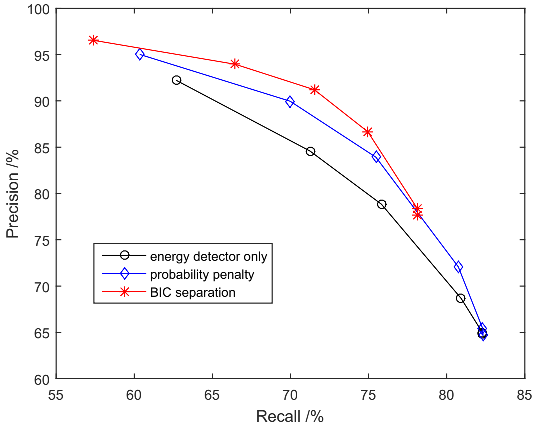

Figure 2: *ROC curves of the three systems.*

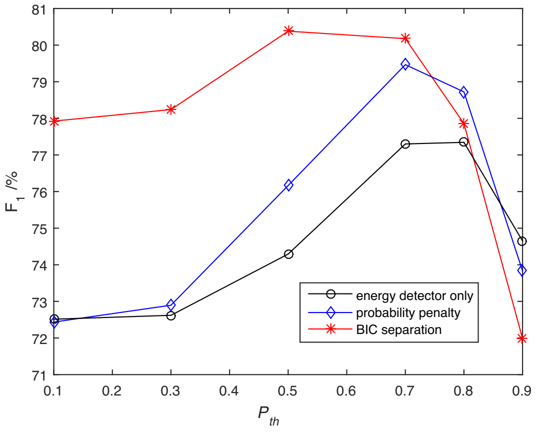

Figure 3: F_1 *values of the three systems.*

events. Because of the acknowledged ability of convolutional neural networks (CNN) to learn the discriminative structures in images, we apply this for classification on the image features. One main characteristic of continuous sound recognition that is unimportant for the classification of isolated sounds is the segmentation of the continuous audio into same-sound regions that are then classified. We present and discuss three methods of performing this task. The first, which will act as our baseline system, is a low-complexity energy detector with fixed context region. To improve the noise immunity of that method, particularly for noise-like sounds, a probability penalty is introduced to use background probability throughout the continuous sound file to reduce the influence of mis-classified background noise. Finally, a Bayesian approach is developed, inspired by the same/different speaker segmentation methods used in diarization research. This Bayesian inference criteria (BIC) is used for segmentation prior to the energy detector and application of probability penalty. The performance of each system was evaluated and shown to work well in noise.

5.1. Future work

It is notable that many of the RWCP sounds are percussive or scraping in nature, and are thus very similar to periods of background noise. It is therefore highly likely that such systems would be susceptible to high levels of noise, in particular when sound events are highly overlapped. Small degrees of overlap are handled well, but the BIC method is unable to separate sound regions when they are almost completely overlapping, and the CNN classifier is currently unable to assign a single region to two different classes. In future we aim to address these issues by exploring rules for multiple classifications per same-sound region, and better methods of background noise adaptation.

6. Acknowledgements

Yan Song is supported by the National Natural Science Foundation of China (grant 61172158) and Chinese Universities Scientific Fund (grant WK2100060008) for this work.

7. References

[1] J. Dennis, H. D. Tran, and H. Li, “Spectrogram image feature for sound event classification in mismatched conditions,” *Signal Processing Letters, IEEE*, vol. 18, no. 2, pp. 130–133, 2011.

[2] I. McLoughlin, H.-M. Zhang, Z.-P. Xie, Y. Song, and W. Xiao, “Robust sound event classification using deep neural networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 540–552, Mar. 2015.

[3] H. Zhang, I. McLoughlin, and Y. Song, “Robust sound event recognition using convolutional neural networks,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, no. 2635. IEEE, Apr. 2015, pp. 559–563.

[4] J. Dennis, H. D. Tran, and E. S. Chng, “Image feature representation of the subband power distribution for robust sound event classification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 367–377, 2013.

[5] Y. LeCun and Y. Bengio, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” in *Acoustics, Speech and Signal*

Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012, pp. 4277–4280.

[8] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep convolutional neural networks for lvcsr,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[9] Y. Xu, I. McLoughlin, Y. Song, and K. Wu, “Improved i-vector representation for speaker diarization,” *Circuits, Systems, and Signal Processing*, pp. 1–12, 2015.

[10] S. Nakamura, K. Hiyane, F. Asano, T. Yamada, and T. Endo, “Data collection in real acoustical environments for sound scene understanding and hands-free speech recognition,” in *EUROSPEECH*, 1999, pp. 2255–2258.

[11] R. B. Palm, “Prediction as a candidate for learning deep hierarchical models of data,” *Technical University of Denmark, Palm*, 2012.