
Downloaded from https://kar.kent.ac.uk/55020/ The University of Kent's Academic Repository KAR

The version of record is available from https://doi.org/10.1109/ICASSP.2015.7178031

This document version
Author’s Accepted Manuscript

DOI for this version

Licence for this version
CC BY (Attribution)

Additional information

Versions of research works

Versions of Record
If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts
If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in Title of Journal, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries
If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).
Traditional sound event recognition methods based on informative front end features such as MFCC, with back end sequencing methods such as HMM, tend to perform poorly in the presence of interfering acoustic noise. Since noise corruption may be unavoidable in practical situations, it is important to develop more robust features and classifiers. Recent advances in this field use powerful machine learning techniques with high dimensional input features such as spectrograms or auditory image. These improve robustness largely thanks to the discriminative capabilities of the back end classifiers. We extend this further by proposing novel features derived from spectrogram energy triggering, allied with the powerful classification capabilities of a convolutional neural network (CNN). The proposed method demonstrates excellent performance under noise-corrupted conditions when compared against state-of-the-art approaches on standard evaluation tasks. To the author’s knowledge this is the first application of CNN in this field.

Index Terms—Machine hearing, auditory event detection, convolutional neural networks

1. INTRODUCTION

Sound event classification is a developing research field which has traditionally benefitted from advances in more mature research in related areas, such as automatic speech recognition (ASR). Detecting sound events in noise is potentially very useful in daily life, such as in allowing a computer to hear and eventually understanding environmental sounds like a human, and from this to infer what is happening in the environment. This technology has implications for improving ASR in many noisy real world scenarios, in security and healthcare monitoring, in intelligent building or city management, and in environmental analysis [1].

Unlike in spoken language, sound events are more random, both periodic and aperiodic, with less well defined occurrence patterns. Sound events also exhibit much wider frequency and amplitude ranges, since they are not constrained by production from the human vocal apparatus [2]. These factors make the task of sound event detection and recognition inherently more difficult than ASR. In fact, ASR-inspired techniques such as MFCC, PLP, ZCR, LSPs [3] have featured prominently in the field [4, 5, 6]. However state-of-the-art robust performance has been achieved only when using higher dimensionality representations such as auditory images [7], spectrogram image features [8] and spectrogram-derived subband power distribution [9]. Feature vectors derived from these representations are used in conjunction with machine learning techniques including SVM [10], kNN [9], PAMIR [7] and so on. The objective of these systems is for powerful machine learning capabilities to infer discriminative relationships from less refined but higher dimensionality input features. A baseline comparison of many techniques on standard evaluation tasks, has been performed recently by Dennis [9].

It is notable that, for the robust task (i.e. recognition of sounds in noise), the best performing input features are in fact images [11]. This provides support for adopting machine learning algorithms from the image processing domain. This was the stated reason for adoption of PAMIR with stabilised auditory images (SAI) in [12]. Similarly, the current paper proposes the use of convolutional neural networks (CNN) with a novel spectrogram image feature (SIF), based upon the observation that CNN-based techniques have recently performed well in related image processing tasks [13, 14]. In particular, the fact that general sounds are not precisely localised in the time-frequency spectrogram, but may preserve strong local relationships, means that the global convolution and subsampling approach inherent to the CNN has advantages. Therefore, this paper develops and evaluates a novel CNN back-end classifier and SIF feature extraction front-end.

2. IMAGE FEATURE BASED ON SPECTROGRAM

This section will detail the formation of SIF vectors from a spectrogram of a sampled sound. Firstly, a spectrogram is generated by stacking fast Fourier transform (FFT) magnitudes from the original sound’s highly overlapped analysis windows. Given length N analysis frame s(n) and Hamming window w(n), the short time spectral representation of the lth frame f(l, k) is obtained, for k = 0...⌊N/2⌋ as follows:
The feature extraction process flow is illustrated in Fig. 1, from where the three maximum energy indices are identified for each frame. This process will therefore yield 18 separate features, SIF, each of which is an \( L \times B \) dimension down-sampled, de-noised image, irrespective of the length of the original sound array:

\[
\text{SIF} = f_{dn}\{\kappa - \lfloor L/2 \rfloor : \kappa = 1 + \lfloor L/2 \rfloor, 1 : B \}
\]

where \( \kappa = J_j - 2 : J_j + 3 \) for \( j = 1 \ldots 3 \). The entire feature extraction process flow is illustrated in Fig. 1, from top to bottom, showing an input sound waveform, forming the overlapped spectrogram, smoothing, down-sampling and de-noising followed by computation of frame-by-frame energy and subsequent pruning. The spectrogram is shown here in colour purely for purposes of illustration.

Note that the authors have investigated a number of alternative pruning methods which are not detailed here for reasons of lack of space. The use of the entire un-pruned stack of down-sampled images for classification was found to be not viable since it takes much longer to train the CNN, which is then much more difficult to achieve convergence. It is not the intention of the authors to claim that the pruning method is optimal, but simply to demonstrate that it is effective. It constitutes the first published application of CNN classification to sound event recognition.

### 3. CNN FOR SOUND EVENT RECOGNITION

CNNs are a class of multi-layer neural networks which contain convolution layers, subsampling layers and fully connected layers. While the network complexity is high due to the large amount of connectivity, the use of shared weights within layers assists in reducing the number of parameters that need to be trained. However, CNNs share the need, with deep neural networks (DNN), for large amounts of training data. In general, for a convolutional layer \( l-1 \), we form layer output maps from

\[
x_j^l = f(\sum_{i \in M_j} x_{i-1}^l \ast k_{ij} + b_j^l),
\]

where \( x_j^l \) is the \( j \)th output map, \( x_{i-1}^l \) is the \( i \)th input map, \( k_{ij} \) denotes the kernel that is applied, and \( M_j \) represents a selection of input maps [15]. The subsampling layer is simpler, \( x_j^l = f(\beta_j \downarrow(x_{i-1}^l) + b_j^l) \) with \( \downarrow(.) \) representing sub-sampling and \( \beta \) and \( b \) are biases [15].

The fully connected output layer is effectively a dual layer multi-layer perceptron (MLP) network, with input layer size depending upon the total number of nodes in the final CNN subsampling layer, but otherwise formed as a typical MLP. Like an MLP, the CNN can be learned by gradient descent using the back-propagation algorithm. As mentioned above,

![Fig. 1. Block diagram of the image feature extraction process.](image1)

![Fig. 2. CNN structure used for SIF classification.](image2)
4. EXPERIMENTS AND RESULTS

4.1. The evaluation task

The sound and noise corpora used in this paper are chosen to match those used to evaluate current state-of-the-art SIF-based methods, as defined by Dennis et al. [18, 9]. 50 sound classes and 80 sound files are selected randomly from the Real Word Computing Partnership (RWCP) Sound Scene Database in Real Acoustic Environments [19]. Four different environments of noise are chosen from the NOISEX-92 database, namely “Destroyer Control Room”, “Speech Babble”, “Factory Floor 1” and “Jet Cockpit 1”.

50 of the 80 files in each class are designated to be a training set (total 2500 files), with the rest forming the testing set (total 1500 files). During testing, randomly-chosen noise is added from random starting points to the sounds at levels of 20, 10 and 0 dB SNR (plus one test with no noise added). However training uses only clean sounds. The mismatched noise conditions make the task more challenging, but are arguably more similar to the situation in reality.

At a 16kHz sample rate, we choose an FFT analysis window length of 1024, which means one frame lasts for 1024/16kHz= 64ms. While speech may be considered pseudo-stationary for around 20ms [2], general environmental sounds are more agile, so we use highly overlapped analysis windows spaced 64 samples apart. This time difference between two frames in the spectrogram array is therefore only 64/16kHz= 4ms, allowing important instantaneous information to be captured.

A typical CNN structure form is chosen to match those used by other authors in the ASR domain. This comprises two convolutional layers with outputmaps of size 6 and 12, a convolution kernel size of 5×5 and a subsampling kernel size of 2×2. The CNN toolbox [20] is used for all experiments.

4.2. Results and discussion

While the CNN classifier and the input feature representation both involve many parameters which could be individually tuned to improve performance, the following subsections investigate only the effect of different frequency and time resolutions in the input SIF, the effect of smoothing, and use of Mel-filterbanks to form the CNN input feature. Each test required the creation of a custom-sized CNN which were, apart from the feature under test, identical in other aspects.

4.2.1. The effect of SIF time-span on performance

The number of frames in a feature defines how much time one SIF spans. Since the test data set includes a range of sounds from very short to very long duration, it is not immediately clear what is the optimal time span. We therefore investigate full performance (i.e. in both clean and noisy conditions) with the number of frames in the SIF (L) set from 16 to 48. Results are shown in Table 1, where the frequency resolution is maintained at 24. We can see that performance first rises with L, then drops as it becomes too big, and within the central region of the table, the performance is relatively flat. We will therefore set the baseline L = 40 for future experiments. This value appears to be long enough to contain the necessary timespan, but short enough to maintain sufficient time resolution. It yields highest accuracy in clean conditions and yet still maintains good accuracy in noisy conditions.

4.2.2. The effect of frequency resolution on performance

The frequency solution defines how many frequency bands there are in an image feature. We begin by setting the number of frames in the SIF to 40. Then we compute performance as the frequency resolution, B, is swept from 48 to 68 in steps of 4. The results are shown in Table 2. It is clear that best overall performance – in both clean and noisy conditions – is achieved when B = 52.

Therefore, 52 × 40 seems to be a suitable SIF dimensionality for the given experimental conditions, dataset and clas-
The final results, particularly for SIF-IS-CNN, confirm the benefits of a SIF representation, including smoothing and de-noising, on creating a noise-robust sound event detection method. An excellent 85% accuracy is achieved in 0dB SNR, and mean accuracy exceeding 94%.

## 5. CONCLUSION

The paper has proposed the use of a convolutional neural network (CNN) for robust sound event detection, motivated by the inherent image-like nature of the spectrogram representation – and encouraged by recently reported good CNN performance for similar ASR tasks. A dimension reduction process has been developed to convert the arbitrary length spectrogram obtained from a sound recording into smoothed and de-noised spectrogram image feature (SIF) blocks of size $52 \times 40$. Both the frequency domain resolution and the time span of these blocks have been investigated in terms of classification performance using appropriately sized CNNs. Use of a standard evaluation task adopted by other authors has allowed direct comparison with other sound event recognition systems, and has revealed that the proposed CNN formulation, using smoothed and de-noised SIF features, is capable of yielding excellent classification accuracy, especially for the challenging 0dB SNR noise condition. To the author’s knowledge, this paper describes the first published application of CNN to this domain, and yields the best accuracy reported to date from spectrogram features.

## 6. ACKNOWLEDGEMENTS

The authors would like to acknowledge, the following for supporting this work: National Nature Science Foundation of China (grant 61172158), Chinese Universities Scientific Fund (grants no WK2100060008 and WK2100000002).
7. REFERENCES


