Applied Acoustics 148 (2019) 62–66

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Technical note

Acoustic event recognition using cochleagram image and convolutional neural networks

Roneel V. Sharan [a,*], Tom J. Moir [b]

[a] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia
[b] School of Engineering, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 4 September 2018
Received in revised form 31 October 2018
Accepted 4 December 2018

Keywords:
Acoustic event recognition
Cochleagram
Convolutional neural network
Mel-spectrogram
Spectrogram

ABSTRACT

Convolutional neural networks (CNN) have produced encouraging results in image classification tasks and have been increasingly adopted in audio classification applications. However, in using CNN for acoustic event recognition, the first hurdle is finding the best image representation of an audio signal. In this work, we evaluate the performance of four time-frequency representations for use with CNN. Firstly, we consider the conventional spectrogram image. Secondly, we apply moving average to the spectrogram along the frequency domain to obtain what we refer as the smoothed spectrogram. Thirdly, we use the mel-spectrogram which utilizes the mel-filter, as in mel-frequency cepstral coefficients. Finally, we propose the use of a cochleagram image the frequency components of which are based on the frequency selectivity property of the human cochlea. We test the proposed techniques on an acoustic event database containing 50 sound classes. The results show that the proposed cochleagram time-frequency image representation gives the best classification performance when used with CNN.

1. Introduction

Acoustic event recognition has many applications such as urban sound classification [1], fall detection [2], etc. As with other pattern recognition problems, a key challenge in acoustic event recognition is finding techniques for robust classification.

Convolutional neural networks (CNN) have produced encouraging results in image classification tasks such as on the ImageNet dataset [3], which has various image categories, and handwritten digit recognition [4]. This formed the inspiration to its extension to speech recognition tasks where it has shown to perform better than deep neural networks (DNN) [5]. Acoustic event recognition is a relatively new area of research and a number of techniques are inspired from other pattern recognition tasks, speech recognition, in particular. As such, some recent works have adopted CNN for sound classification applications [6,7].

CNN is an image classification technique and one of the major challenges in speech and acoustic event recognition has been how to best represent the audio signal using an image for this purpose. Two common approaches have been seen in addressing this problem. Firstly, the audio signal is converted to spectrogram images [6]. Secondly, a mel-filter, as used in computing mel-frequency cepstral coefficients (MFCC), is used to form an image-like representation [5]. We refer this as the mel-spectrogram. To ensure that all images are of an equal size, the audio signal is divided into a fixed number of frames [5] or image scaling options [7] are explored. In addition, techniques such as moving average have been applied to spectrograms [8].

The spectrogram offers equally spaced frequency components with equal bandwidth. This is not ideal for modeling the frequency characteristics of acoustic event recognition tasks where, depending on the application, most spectral energy lies in the lower frequency range. While the use of a mel-filter helps to some extent, in this work we propose the use of a cochleagram time-frequency representation for this purpose. The gammatone filter utilized in forming the cochleagram representation is modeled on the frequency selectivity property of the human cochlea. It offers narrow frequency components in the lower frequency range and wide frequency components in the upper frequency range. The finer frequency resolution in the lower frequency range helps reveal more spectral information without significantly losing spectral information in the upper frequency range [9]. The use of cochleagram has shown to be effective in audio classification and separation tasks [10,11].

The rest of the paper is organized as follows. Section 2 provides an overview of the sound signal image representation techniques

[*] Corresponding author.
E-mail addresses: r.sharan@uq.edu.au (R.V. Sharan), tom.moir@aut.ac.nz (T.J. Moir).

https://doi.org/10.1016/j.apacoust.2018.12.006
0003-682X/© 2018 Elsevier Ltd. All rights reserved.

and the CNN architecture. Baseline methods for this work are presented in Section 3. In Section 4, we present the experimental setup and results followed by conclusion in Section 5.

2. CNN Method

An overview of the proposed input layer and the CNN architecture employed in this work is given in Fig. 1. The generic architecture of CNN has been discussed in detail in a number of literature, such as [5], therefore, we only provide an overview of the CNN architecture and settings that we used. In this work, our primary focus is the formation of the input layer (image) of CNN for audio signal classification.

In this work, we consider spectrogram, smoothed spectrogram, mel-spectrogram, and cochleagram time-frequency images for the input layer. The dimension of the input layer image is chosen as 32 × 15, consistent with other similar work such as [5]. In addition, our sound database includes impulsive sounds and having more number of frames would make the frame size small, making it difficult to capture distinguishing frequency characteristics.

2.1. Spectrogram

In forming the spectrogram image, discrete Fourier transform (DFT) is applied to the windowed signal as

$$X(k,r) = \sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-2\pi ikn}{N}}, \quad k = 0, \ldots, N-1 \tag{1}$$

where N is the length of the window, $x(n)$ is the time-domain signal, $X(k,r)$ is the k^{th} harmonic corresponding to the frequency $f(k) = kF_s/N$ for the r^{th} frame, F_s is the sampling frequency, and $w(n)$ is the window function.

The spectrogram values are obtained from log of the magnitude of the DFT values as

$$S(k,r) = \log|X(k,r)|. \tag{2}$$

To get the same time-frequency image resolution for all signals, each sound event signal is divided into 15 frames with 50% overlap between frames. A Hamming window is then applied to the frames and Fourier transform performed using 64 points so that the final image is 32 × 15 dimensional.

2.2. Smoothed spectrogram

Non-overlapping moving average computation is performed on the frequency components of the spectrogram in each frame to obtain a smoothed spectrogram. That is,

$$X_S(a,r) = \sum_{v=(a-1)W+1}^{aW} |X(k_v, r)|, \quad a = 1, \ldots, N/2W \tag{3}$$

where W is the length of the moving average window.

For the smoothed spectrogram image, each sound event signal is divided into 15 frames with 50% overlap between frames. A Hamming window is then applied and Fourier transform performed using 1024 points. For the smoothed spectrogram, the moving window length, W, is set to 16 to get 32 non-overlapping moving windows and a final image size of 32 × 15. The log values are then calculated as in Eq. (2).

2.3. Mel-spectrogram

The mel-spectrogram image intensity values are computed similar to MFCC [12] but without applying the discrete cosine transform (DCT), that is, using the filter bank energies. The mel-filter bank output of the m^{th} filter can be determined as

$$E(m,r) = \sum_{k=0}^{\frac{N}{2}-1} V(m,k)|X(k,r)|, \quad m = 1, 2, \ldots, M \tag{4}$$

where $E(m,r)$ is the filter bank energy of the m^{th} filter in the r^{th} frame, $V(m,k)$ is the normalized filter response of the triangular filter banks which are equally spaced on the mel-scale [13], and M is the total number of mel-filters.

For the mel-spectrogram, the spectrogram image used in the smoothed spectrogram is computed and the number of mel-filters, M, is set to 32 to obtain a 32 × 15 image. The log values are then calculated as in Eq. (2).

2.4. Cochleagram

In the cochleagram representation, the frequency components in the time-frequency image are based on the frequency selectivity property of the human cochlea and modeled by a gammatone filter as [14]

$$h(t) = At^{j-1}e^{-2\pi Bt}\cos(2\pi f_c t + \phi) \tag{5}$$

where A is the amplitude, j is the order of the filter, B is the bandwidth of the filter, f_c is the center frequency of the filter, ϕ is the phase, and t is the time.

The equivalent rectangular bandwidth (ERB), a psychoacoustic measure of the auditory filter width at each point along the cochlea, is used to describe the bandwidth of each cochlea filter in [14]. In this work, we use the ERB filter model as described in [15] which was shown to produce the best results in [16].

After filtering the signal using the gammatone filter, the implementation of which can be found in [16,17], a representation similar to the spectrogram is obtained by adding the energy in the windowed signal for each frequency channel as

$$C(g,r) = \sum_{n=0}^{N-1} |\hat{x}(g,n)|w(n), \quad g = 1, 2, \ldots, G \tag{6}$$

where $\hat{x}(g,n)$ is the gammatone filtered signal, $C(g,r)$ is the g^{th} harmonic corresponding to the center frequency f_{cg} for the r^{th} frame, and G is the number of gammatone filters.

With the cochleagram representation, we set the number of gammatone filters to 32. The filtered signal is divided into 15 frames with 50% overlap between frames. The energy in each

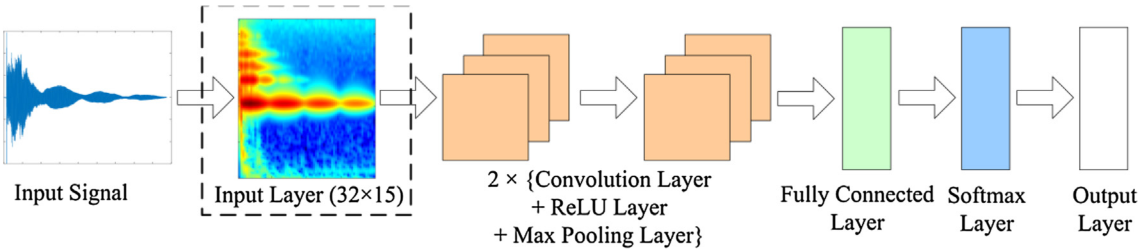

Fig. 1. An overview of the proposed input layer and CNN architecture.

frame is added to obtain a cochleagram image of size 32 × 15. The log values are then calculated as in Eq. (2).

Illustration of spectrogram, mel-spectrogram, and cochleagram images of a sample sound signal, mapped to the jet colorspace for better visualization, are given in Fig. 2(a), (b), and (c), respectively. All three representations have the same frequency range from 0 Hz to the Nyquist frequency of 22,050 Hz. However, in the cochleagram representation, the finer resolution of the frequency components in the lower frequency range helps reveal more spectral information in the lower frequency range for the sound signals considered in this work without losing the spectral information in the upper frequency range. Refer to [9] for more information on cochleagram image formation and a detailed comparison between spectrogram and cochleagram representations.

2.5. CNN

The layout of the CNN was determined after a number of considerations and experimentations. The model is trained using stochastic gradient descent with momentum [18]. The network includes two convolution layers, each of which includes a rectified linear unit (ReLU) [19] and followed by a max pooling layer [20]. The filter size for both convolution layers is 3 × 3, stride 1 × 1, and padding 1 × 1. The number of filters in each layer is set to 16 after experimenting with a number of filters. Similarly, the max pooling layer size is 2 × 2, stride 1 × 1, and padding 1 × 1. This is followed by a fully connected layer and a softmax layer [18] of size 50 and an output layer of size 50.

The settings for other parameters are as follows: image normalization = zero-centering, momentum = 0.4, initial learn rate = 0.01, learn rate schedule = piecewise, learn rate drop factor = 0.6, learn rate drop period = 6, L2 regularization = 0.05, mini batch size = 50, data shuffle = once, and max epochs = 20. The parameters were optimized based on the training/validation performance. The training stops after the maximum number of epochs is reached.

3. Baseline Method

We use three baseline feature sets (MFCC, time-frequency image feature extraction, and raw time-frequency image) and two baseline classifiers (K-nearest neighbor (KNN) and support vector machines (SVM)) to compare the performance of the proposed techniques.

3.1. MFCC

Firstly, we use MFCC as features which are computed as the DCT of the log compressed filter bank energies given as

$$c(l,r) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \log(E(m,r)) \cos\left(\frac{\pi l}{M}(m-0.5)\right) \tag{7}$$

which is evaluated from $l = 1, 2, \cdots, L$, where L is the order of the cepstrum.

The number of mel-filters, M, was set to 32 to be consistent with the approach taken in forming the mel-spectrogram image for CNN classification. Two approaches were then taken for feature representation. In the first method, each signal is divided into frames of 1024 points with 50% overlap between frames. A Hamming window is then applied, Fourier transform performed using 1024 points, and MFCC values computed in each frame. The final feature vector is represented using the mean and standard deviation across each of the 32 dimensions resulting in a 64 dimensional feature vector. In the second method, the difference is that each signal is divided into 15 frames with 50% overlap between frames and the final feature vector represented using raw values of dimension 480 (32 × 15).

3.2. Time-frequency image derived features

Secondly, we use time-frequency image feature extraction. We use central moments to capture the spectral distribution in the spectrogram, a technique which produced encouraging results in [21]. The central moments are computed for the spectrogram and cochleagram representations and the corresponding features referred as spectrogram image feature (SIF) [21] and cochleagram image feature (CIF) [16], respectively.

In computing the SIF, we divide each signal into 1024 points with 50% overlap. Fourier transform is performed using 1024 points. The spectrogram image is divided into subbands and second and third central moments are extracted as features in each subband. Various number of subbands were experimented with but best results were achieved when using 64 subbands which results in a 128 dimensional feature vector.

For the cochleagram representation, each sound signal is filtered using 512 gammatone filters to have same number of frequency components as the spectrogram image. The filtered signal is divided into frames of size 1024 with 50% overlap and the energy in each frame computed as per Eq. (6). The same feature extraction procedure is then followed as the spectrogram image.

3.3. Raw time-frequency image feature

Finally, the raw time-frequency image (32 × 15) values were concatenated into a 480 dimensional feature vector for classification. The spectrogram, smoothed spectrogram, mel-spectrogram, and cochleagram representations, as described in Sections 2.1–2.4, were used for this purpose.

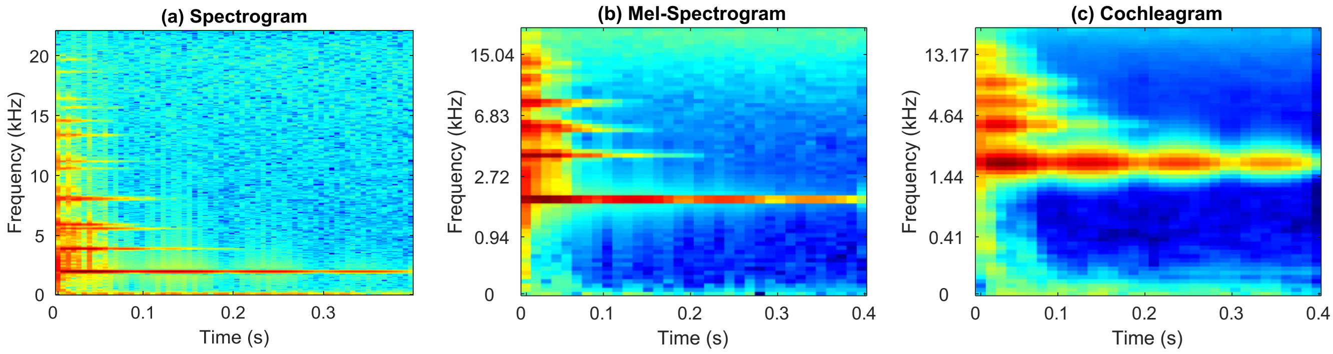

Fig. 2. (a) Spectrogram, (b) mel-spectrogram, and (c) cochleagram image of a sample sound signal. The frequency range in each case is 0–22,050 Hz.

3.4. Baseline classifiers

The Euclidean distance measure was used for the KNN classifier and the value of K was set to 1 after experimenting with a number of values. For the SVM classifier, nonlinear SVM with a Gaussian RBF kernel in a one-against-all strategy is used as it was found to give the best results. SVM parameters, the penalty parameter and the width of the Gaussian function [22], were tuned using Bayesian optimization [23] using 5-fold cross-validation on the training data.

4. Experimental evaluation

4.1. Dataset

The sound event files used in this work are taken from the Real World Computing Partnership (RWCP) Sound Scene Database (SSD) in Real Acoustical Environments [24]. The sound database has a total of 4000 manually segmented sound event files belonging to 50 classes, 80 files per class. All signals in the database have 16-bit resolution and a sampling frequency of 44,100 Hz.

Each file has one sound event the duration of which depends on the class of the sound. The sound event duration varies from a minimum of 12.6 ms to a maximum of 3.84 s.

4.2. Experimental setup

For all experimentations, the classifier is trained and validated with 50 samples per class with the remaining 30 samples used for testing. As such, a total of 2500 samples are used for training and validating the classifier and the remaining 1500 samples used for testing the trained model. The classification accuracy is reported in percentage as *number of correctly classified test samples* divided by the *total number of test samples*. The classification accuracy in the case of CNN is averaged over 10 runs.

4.3. Baseline results

The classification accuracy value using the baseline methods is given in Table 1. The best baseline classification accuracy of 95.07% is achieved using the cochleagram image feature vector and SVM classifier. The cochleagram image feature vector or cochleagram image derived feature vector were seen to be more robust than features using other time-frequency representations.

4.4. Results using CNN

In this subsection, spectrogram, smoothed spectrogram, mel-spectrogram, and cochleagram time-frequency representations are considered for classification using CNN. Results using the four time-frequency representations and CNN for classification are given in Table 2. For spectrogram, smoothed spectrogram, mel-spectrogram, and cochleagram, the best classification accuracy of 93.46%, 96.34%, 95.35%, and 98.03%, respectively, is achieved.

The classification accuracy using the smoothed spectrogram, mel-spectrogram, and cochleagram exceed the best baseline accuracy of 95.07%. The classification accuracy value using the mel-spectrogram is better than the spectrogram. Similarly, the classification accuracy value using smoothed spectrogram is observed to be better than the mel-spectrogram. The best classification accuracy of 98.03% is achieved using the cochleagram image. This shows the suitability of the proposed cochleagram image in acoustic event recognition.

Table 1
Results using baseline methods.

Method		Feature Dimension	Classification Accuracy
MFCC	MFCC–SVM (Method 1)	64	94.27
	MFCC–SVM (Method 2)	480	93.67
Time-frequency image features	SIF–KNN	128	70.27
	SIF–SVM	128	84.07
	CIF–KNN	128	82.40
	CIF–SVM	128	88.00
Raw time-frequency image feature vector (32 × 15)	Spectrogram–KNN	480	63.73
	Spectrogram–SVM		86.87
	Smoothed Spectrogram–KNN		85.00
	Smoothed Spectrogram–SVM		93.47
	Mel-Spectrogram–KNN		89.87
	Mel-Spectrogram–SVM		93.33
	Cochleagram–KNN		95.00
	Cochleagram–SVM		95.07

Table 2
Results using different time-frequency representations and CNN.

Method	Classification Accuracy
Spectrogram–CNN	93.46
Smoothed Spectrogram–CNN	96.34
Mel-Spectrogram–CNN	95.35
Cochleagram–CNN	98.03

5. Conclusions

Four different time-frequency representations, spectrogram, smoothed spectrogram, mel-spectrogram, and cochleagram, are considered for classification using CNN in an acoustic event recognition task. The performance with cochleagram image was determined to be better than spectrogram, smoothed spectrogram, and mel-spectrogram. A classification accuracy of 98.03% could be achieved which is an improvement over the best baseline classification accuracy of 95.07% using cochleagram-SVM. In summary, cochleagram time frequency representation is determined to be more suited for classification using CNN then the other time-frequency image representations for the acoustic event recognition task considered in this work.

References

[1] Ye J, Kobayashi T, Murakawa M. Urban sound event classification based on local and global features aggregation. Appl Acoust 2017;117:246–56.

[2] Adnan SM, Irtaza A, Aziz S, Ullah MO, Javed A, Mahmood MT. Fall detection through acoustic Local Ternary Patterns. Appl Acoust 2018;140:296–300.

[3] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012:1097–105.

[4] Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. p. 3642–9.

[5] Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, Yu D. Convolutional neural networks for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2014;22(10):1533–45.

[6] Zhang H, McLoughlin I, Song Y. Robust sound event recognition using convolutional neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 559–63.

[7] Ozer I, Ozer Z, Findik O. Noise robust sound event classification with convolutional neural network. Neurocomputing 2018;272:505–12.

[8] Kovács G, Tóth L, Van Compernolle D, Ganapathy S. Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout. Pattern Recogn Lett 2017;100:44–50.

[9] Sharan RV, Moir TJ. Subband time-frequency image texture features for robust audio surveillance. IEEE Trans Inf Forensics Secur 2015;10(12):2605–15.

[10] Gao B, Woo WL, Khor LC. Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation. J Acoust Soc Am 2014;135(3):1171–85.

[11] Sharan RV, Moir TJ. Pseudo-color cochleagram image feature and sequential feature selection for robust acoustic event recognition. Appl Acoust 2018;140:198–204.

[12] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 1980;28(4):357–66.

[13] O'Shaughnessy D. Speech Communication: Human and Machine. Addison-Wesley Pub. Co.; 1987.

[14] Patterson RD, Robinson K, Holdsworth J, McKeown D, Zhang C, Allerhand M. Complex sounds and auditory images. In: Cazals Y, Demany L, Horner K, editors. Auditory physiology and perception. Oxford: Pergamon; 1992. p. 429–46.

[15] Greenwood DD. A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Am 1990;87(6):2592–605.

[16] Sharan RV, Moir TJ. Cochleagram image feature for improved robustness in sound recognition. In: Proceedings of the IEEE International Conference on Digital Signal Processing (DSP). p. 441–4.

[17] Slaney M. Auditory Toolbox for Matlab, Interval Research Corproation, 1998. Technical Report 1998–010.

[18] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2006.

[19] Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on International Conference on Machine Learning, Haifa, Israel. p. 807–14.

[20] Jarrett K, Kavukcuoglu K, Ranzato M, LeCun Y. What is the best multi-stage architecture for object recognition? In: 2009 IEEE 12th International Conference on Computer Vision. p. 2146–53.

[21] Dennis J, Tran HD, Li H. Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process Lett 2011;18 (2):130–3.

[22] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[23] Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems. p. 2951–9.

[24] Nakamura S, Hiyane K, Asano F, Nishiura T, Yamada T. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In: Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, Greece. p. 965–8.