# Acoustic based Scene Event Identification Using Deep Learning CNN

**R. Abinaya[a], D.N.V.S.L.S. Indira[b], Dhanalakshmi Lanka[c]**

[a]Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India. E-mail: abinayamalar@gmail.com
[b]Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India. E-mail: indiragamini@gmail.com
[c]Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India. E-mail: dhanayarlagadda@gmail.com

**Article History**: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

**Abstract:** Deep learning is becoming popular nowadays on solving the classification problems when compared with conventional classifiers. Large number of researchers are exploiting deep learning regarding sound event detection for environmental scene analysis. In this research, deep learning convolutional neural network (CNN) classifier is modelled using the extracted MFCC features for classifying the environmental event sounds. The experiment results clearly show that proposed MFCC-CNN outperform other existing methods with a high classification accuracy of 90.65%.

**Keywords:** Scene Recognition, Convolution Neural Network Cepstral Coefficient, Deep Learning.

## 1. Introduction

Building computers with sense (touch, vision, hearing, taste and smell) like humans is a long- awaited goal among the researchers. Simply, computers are made mindful of their surroundings as human beings are. Thereby making sensible computers, we can produce robots which are conscious of their surroundings, design environment recognizing hearing aids, or even design self-driving car for hazardous situations, acoustic-based surveillance systems etc. But acoustic surroundings in real is hard to decode due to presence of simultaneous sounds or high background noise, or due to long distance to the sound source.

This work towards mining valuable information from environmental audio recordings. The proposed system is targeted towards recognizing different classes of environmental sounds which are nothing but noisy sounds that we hear in our day-to-day time (i.e. dog barks, car engine running, etc.).

## 2. Related Work

Feature extraction is the base of acoustic data classification, based on its temporal resolution they fall into three subdivisions. First one is frame-level features, which are derived from short analysis frames/windows of sample size between 10ms to 100ms for representing the local characteristics. Examples includes cepstral, spectral and temporal features like LPC, LPCC, MFCC, time energy.

ZCR, Centroid, Roll off, Flatness, etc. Second one is segment-level features, which apprehend the sound’s texture characteristics since their analysis windows are long enough when compared with frame-level (Tzanetakis et al., 2002), they are also named as texture windows. (Seyerlehner et al., 2010). Third one is clip-level features, which represent the global audio characteristics of the signal. These features are same for all frames extracted from a single audio clip. For example, fundamental frequency, pitch, (Costa et al., 2012) LBP (Ojala et al., 2002). (Wu et al., 2011) Gabor filter bank.

Artificial neural networks (Leon et al., 2002) are widely involved in solving complex problems. But recently, many researchers tried neural network for audio data and found promising results. (Meng et al., 2005) investigated linear neural network in music genre classification. (McKay et al., 2004) used BPNN in hierarchical classification system. This works utilized MFCC feature representation (Chu et al.,2009) and deep learning CNN for performing acoustic event detection.

## 3. Outline of The Work

### Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is an compact, frame level feature which tries to mimic the human auditory system which is derived from the audio through discrete fourier transform (DFT). Fourier spectrum is modified through mel-scaling for adapting to human perception level. Mel-filterbank consists of n triangular filters whose mel-frequency scale is approximately linear, up to 1 kHz and logarithmic thereafter. In this equation (1), (2) where M is the mel-frequency and f is the spectral frequency.

$$for\ f < 1000Hz, M = f \quad (1)$$
$$for\ f > 1000Hz, M = 2595\ log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

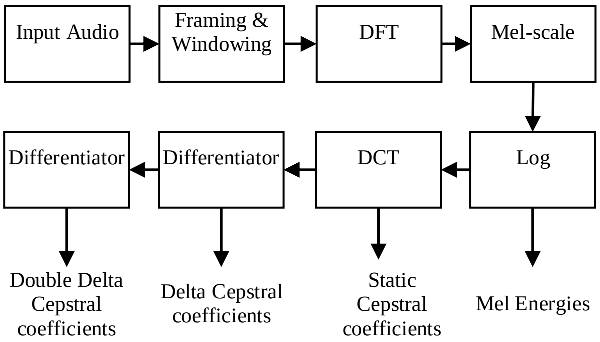

Input Audio
Framing & Windowing
DFT
Mel-scale
Log
DCT
Differentiator
Differentiator
Double Delta Cepstral coefficients
Delta Cepstral coefficients
Static Cepstral coefficients
Mel Energies

**Figure 1.** MFCC Extraction

The calculation of MFCC include the following four steps.

1. Acquired audio sample is framed and windowed for making the short time frame audio locally stationary and free from edge discontinuities.
2. Transform the short time frames into spectrogram frames using DFT.
3. Map the spectral frequency into mel-scale bins.
4. Take the logs of the value of the mel bands.
5. Apply discrete cosine transform (DCT) on the mel bands to derive the cepstral coefficients.
6. Further differentiating the cepstral coefficients results in delta and double delta coefficients.

### Deep Learning

Artificial Intelligence is playing an tremendous role in bringing the capabilities of machines closer to humans. Computer vision is mainly employed in the tasks of audio, image & video data's classification, recognition, recreation, analyzing, etc. Computer vision's growth is maximized with the involvement of deep learning.

- ***Convolutional Neural Network***

CNN is an deep learning classifier which takes two-dimensional matrix as input and learn the network weights and bias from it, for identifying that same class of image during testing. There is almost very less preprocessing present in the CNN. CNN initially reduces the input data as simple as possible without losing the important features.

- ***Architecture of CNN***

The CNN made up of convolution layers with RELU activation function, pooling layers for finding patterns and fully connected layer with softmax function at the end for classification.

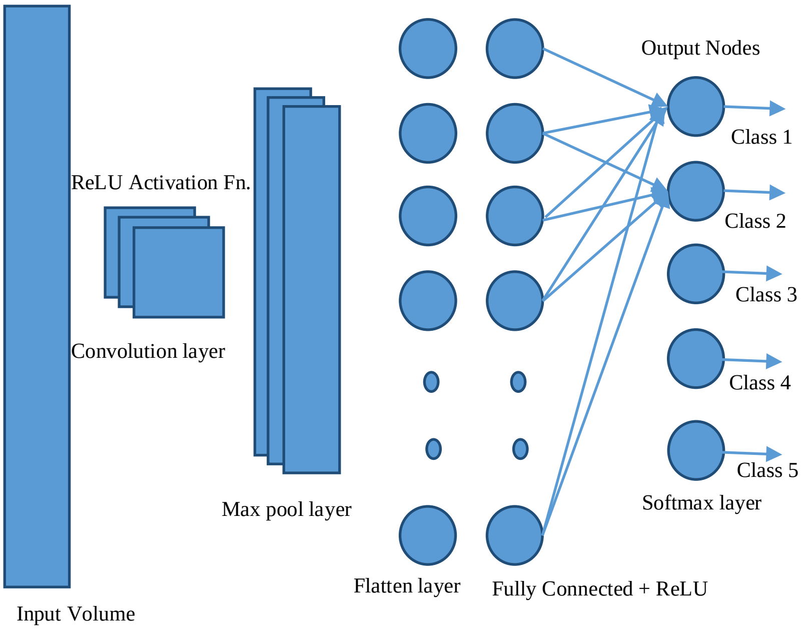
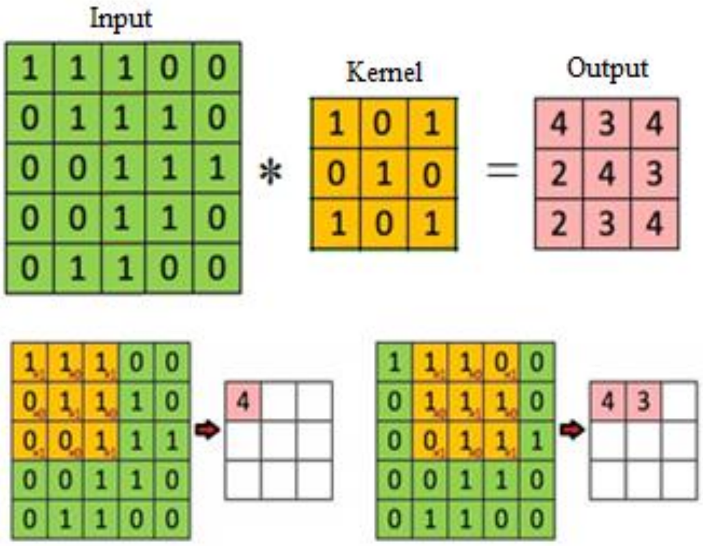


**Figure 2.** CNN Architecture

**Convolution Layer—The Kernel:** Figure 3 shows an 2D matrix of dimension 5 (Height) x 5 (Breadth) is given to the CNN along with a 3x3 convolutional kernel. First Convolution Layer captures the Low-Level features. With additional convolutional layers, CNN network adapts to the High-Level features as well, thus giving a complete understanding of that 2D matrix rom the dataset.

**Stride Value:** Stride value denotes how many boxes the filter slide at every step. Kernel/Filter moves from right top to left top with specified Stride Value till it completes the width. Now it hops down one box to the beginning (left) of the 2D matrix and proceed towards right with same Stride Value and this process go on until the entire 2D matrix is travelled.

In Figure 3, the filter travels to 9 different positions since Stride = 1, at each position they performing a dot product operation between filter and corresponding region of the 2D matrix over which the filter is currently lying.



$$\begin{bmatrix}1&1&1&0&0\\0&1&1&1&0\\0&0&1&1&1\\0&0&1&1&0\\0&1&1&0&0\end{bmatrix} * \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix} = \begin{bmatrix}4&3&4\\2&4&3\\2&3&4\end{bmatrix}$$

**Figure 3.** Convolution Operation on a 5x5 Matrix with a 3x3 Kernel

**Zero Padding:** Convolution layer gives either same dimensional output or reduced or increased dimensional output depending on the padding used.

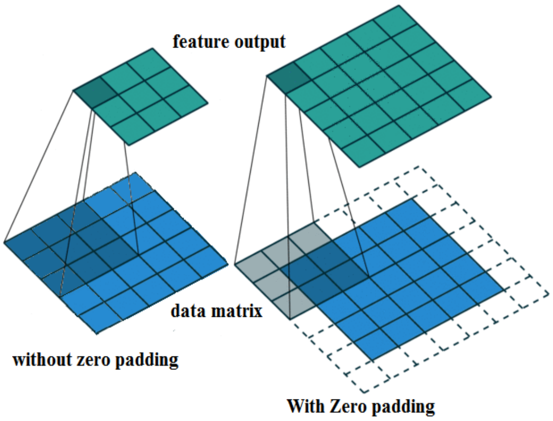


**Figure 4.** Convolution Operation with and without Padding

If we want same dimensional features as that of input matrix, we do zero padding as shown in Figure 3. Here 5x5x1 data matrix is first zero padded and then convolved with 3x3x1 kernel, we get 5x5x1 feature dimension. If we do that same convolution without zero padding then we get 3x3x1 feature dimension.

**Rectified Linear Unit (ReLU) activation function:** The convolution layer output is given to a ReLU activation function, which introduces non-linearity in data.ReLU makes the values ≤ 0 as zero and all positive values will be unchanged.

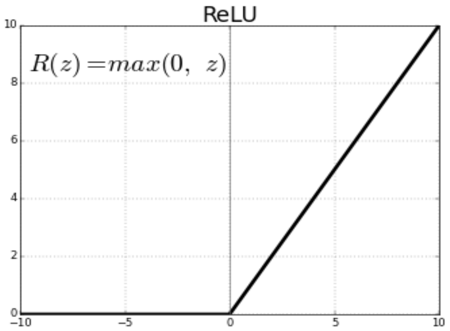


**Figure 5.** ReLU Activation Function

**Pooling Layer:** Pooling layers were used to decrease redundancy present in the data matrix. These layers will down sample the data so that the resource needed for running that program is also reduced. In pooling, a analysis window hovers over the 2D data matrix with an stride value. For max-pooling at each step, the greatest value among the analysis window is chosen. Whereas for average pooling, the average of all values in analysis window is selected.

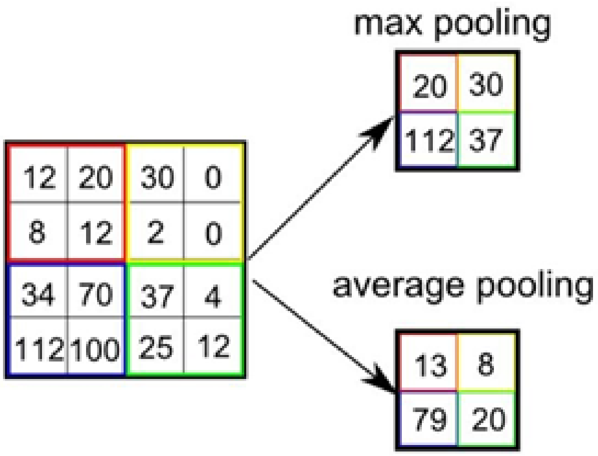


**Figure 6.** Types of Pooling

**Classification—Fully Connected Layer (FC Layer):** Finally, FC layers are added at last to learn from the high-level features derived from the previous layers. At this stage, the 2D data matrix is flatted out and given to a feed forward neurons and backpropagation is utilized to their network weights at each iteration.
Thus, by using the softmax activation function at the last output layer, the input data were classified into their corresponding classes.

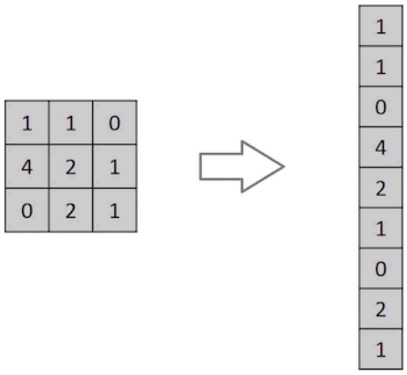


**Figure 7.** Flattening of a 3x3 Image Matrix into a 9x1 Vector

## 4. Experimental Setup

In this proposed system, deep learning-based modelling is employed through CNN classifier for recognizing an event in a specific scene from its acoustic data. The audio data is represented through the MFCC cepstral features. Proposed system comprises of training and testing, while each phase has three stages namely pre-processing, audio feature extraction and finally classification as shown in Figure 8.

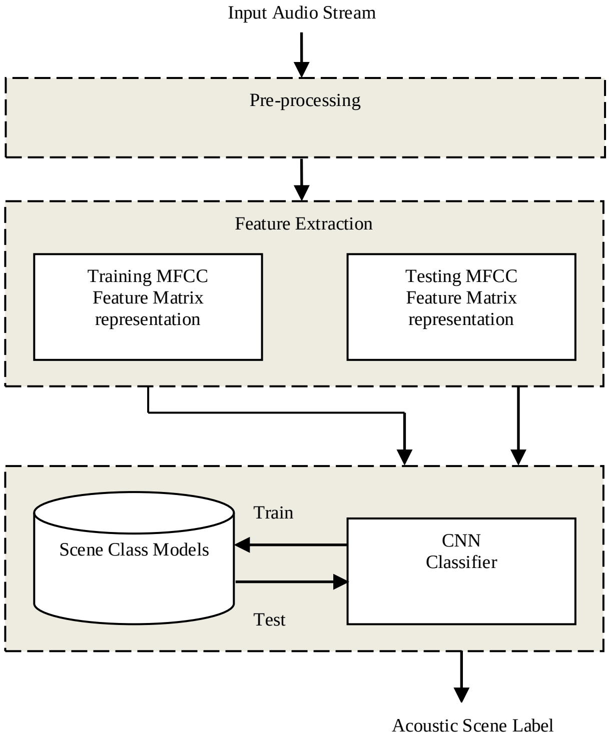


**Figure 8.** Block Diagram of the Proposed Auditory Scene Recognition System

### Database Description

The AURORA database consists of 1000 audio samples which are of duration 2 second, 3 second and 5 second in length and are sampled at 16kHz. Dataset consists of non-speech and non-music sounds from general classes namely 230 kitchen noises, 190 living room noises, 200 laundry sounds, 220 meeting sounds, 160 office sounds as shown in table 1 and belong to events (steel plate clashing, music player playing, washing machine running, flush, background speech, footsteps, typing sound, etc.). Dataset is randomly split in the ratio of 80:20 between the training and unknown testing sets.

**Table 1.** Acoustic Database Description

| Context | Total amount of database |
|---|---|
| Kitchen | 23 % |
| Living Room | 19 % |
| Laundry | 20 % |
| Meeting | 22 % |
| Office | 16% |

**Preprocessing**

Pre-processing includes audio signal normalization, framing, followed by windowing and then silence removal. The audio signal which is sampled at 16000Hz, is separated into successive frames. Each frame is of duration 20ms and having 320 samples (16000Hz / 1000 = 16 samples per 1 ms since 1 second = 1000 milli second). Such 20ms duration frames were extracted every 10ms. These frame's signal amplitude is normalized to be in the range [−1 1]. Followed by applying Hamming windowing to each extracted audio frame of size 20ms. For each frame, the overall energy mean is calculated and those frames with low energy values are discarded to remove empty audio frames.

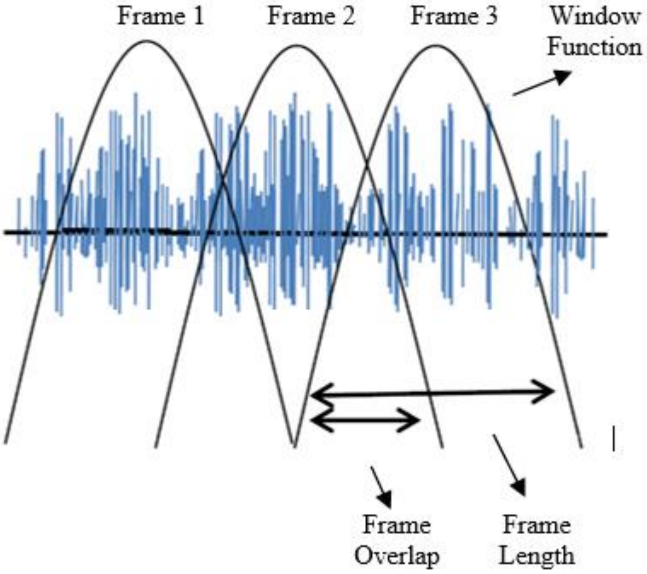


**Figure 9.** Framing & Windowing in Pre-Processing

**Feature Extraction**

Audio frames were analyzed in cepstral domain for deriving most discriminative feature vectors, which were the high-level representation of the input. In this system, frame-level 39 dimensional double delta, delta and static MFCC feature are utilized for effectively distinguishing environmental sounds.

**Modelling using CNN**

The extracted features whose dimension [39 X no. of frames] is taken as input for training. CNN network is trained using known training dataset's feature vector for all five class. After the CNN is trained, the testing data which is newer to the classifier is given one by one to analyse the recognition accuracy.

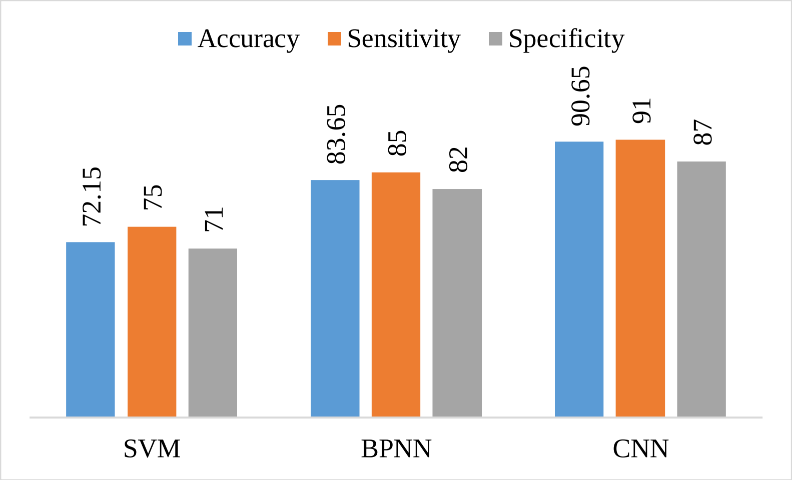
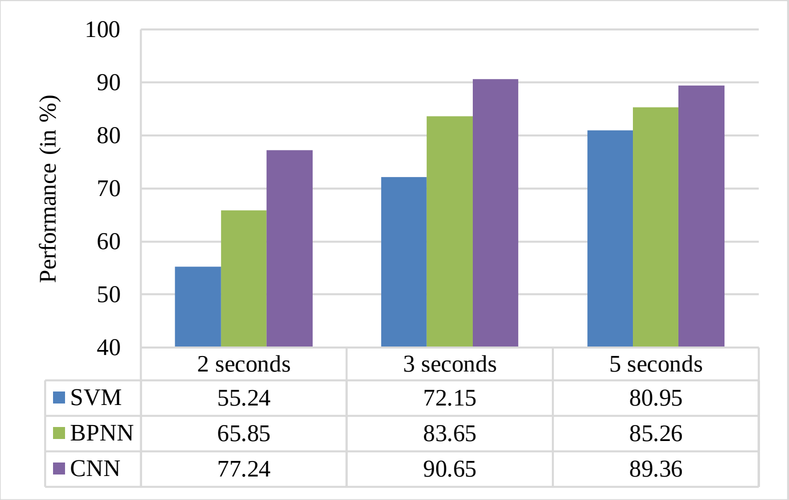


**Figure 10.** Performance for Various Classifiers

The duration of the audio dataset is varied and their corresponding results are noted every time. Figure 11 clearly show that CNN classifier gives almost same accuracy for 3 and 5 second duration audio samples.



| | 2 seconds | 3 seconds | 5 seconds |
|---|---|---|---|
| ■ SVM | 55.24 | 72.15 | 80.95 |
| ■ BPNN | 65.85 | 83.65 | 85.26 |
| ■ CNN | 77.24 | 90.65 | 89.36 |

**Figure 11.** Accuracy for Various Sound Durations and Classifiers

## 5. Conclusion

A scene recognition system is proposed on audio input modality and is implemented in Matlab. Proposed system utilizes cepstral MFCC features which discriminative well and reduce the redundancy in the audio data for effectively classifying environmental sounds into predefined classes. Deep learning CNN classifier and two Machine learning classifiers namely SVM and BPNN were employed in recognizing the acoustic events related to the respective acoustic scenes. Test results clearly depicts a higher accuracy of 90.65% using deep CNN while lower accuracy for other machine learning classifiers such as 72% for SVM and 83% for BPNN. Thus, CNN clearly outperforms other state-of-art classifiers in the recognition accuracy.

## References

1. Aucouturier, J.J., Defreville, B., & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. The Journal of the Acoustical Society of America, 122(2), 881-891.
2. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In Proceedings of the 6th international conference on Multimodal interfaces, 205-211.
3. Chen, S.S., Donoho, D.L., & Saunders, M.A. (2001). Atomic decomposition by basis pursuit. SIAM review, 43(1), 129-159.

4. Chu, S., Narayanan, S., & Kuo, C.C.J. (2009). Environmental sound recognition with time–frequency audio features. IEEE Transactions on Audio, Speech, and Language Processing, 17(6), 1142-1158.
5. Costa, Y.M., Oliveira, L.S., Koerich, A.L., Gouyon, F., & Martins, J.G. (2012). Music genre classification using LBP textural features. Signal Processing, 92(11), 2723-2737.
6. De León, P. J.P., & Inesta, J.M. (2002). Musical style identification using self-organising maps. In Second International Conference on Web Delivering of Music, 2002. WEDELMUSIC. Proceedings. 82-89.
7. Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In Machine learning proceedings 1995, 194-202.
8. Ebenezer, S.P., Papandreou-Suppappola, A., & Suppappola, S.B. (2004). Classification of acoustic emissions using modified matching pursuit. EURASIP Journal on Advances in Signal Processing, 3, 1-11.
9. Ellis, D.P., & Lee, K. (2004). Minimal-impact audio-based personal archives. In Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences, 39-47.
10. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. Speech communication, 52(1), 12-40.
11. Lee, H., Grosse, R., Ranganath, R., & Ng, A.Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th annual international conference on machine learning, 609-616.
12. McKay, C., & Fujinaga, I. (2004). Automatic Genre Classification Using Large High-Level Musical Feature Sets. In ISMIR 2004. 525-530.
13. Meng, A., Ahrendt, P., & Larsen, J. (2005). Improving music genre classification by short time feature integration. In Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.5, v-497.
14. Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7), 971-987.
15. Pineau, J., Montemerlo, M., Pollack, M., Roy, N., & Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. Robotics and autonomous systems, 42(3-4), 271-281.
16. Seyerlehner, K., Schedl, M., Pohle, T., & Knees, P. (2010). Using block-level features for genre classification, tag classification and music similarity estimation. Submission to Audio Music Similarity and Retrieval Task of MIREX, 2010.
17. Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9-10), 1062-1087.
18. Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. IEEE Transactions on speech and audio processing, 10(5), 293-302.
19. Wu, M.J., Chen, Z.S., Jang, J.S.R., Ren, J.M., Li, Y.H., & Lu, C.H. (2011). Combining visual and acoustic features for music genre classification. In 10th International Conference on Machine Learning and Applications and Workshops 2, 124-129.
20. Yanco, H.A. (1998). Wheelesley: A robotic wheelchair system: Indoor navigation and user interface. In Assistive technology and artificial intelligence, 256-268.