Subrata Mondal, Shiladitya Pujari, Tapas Sangiri / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 6, November- December 2012, pp.123-129

Environmental Natural Sound Detection And Classification Using Content-Based Retrieval (CBR) And MFCC

Subrata Mondal[1], Shiladitya Pujari[2], Tapas Sangiri[3]

[1]Technical Assistant, Department of Computer Science,
Bankura Unnayani Institute of Engineering, Bankura, West Bengal, India
[2]Assistant Professor, Department of IT,
University Institute of Technology, Burdwan University, Burdwan, West Bengal, India
[3]Assistant Professor, Department of IT,
Bankura Unnayani Institute of Engineering, Bankura, West Bengal, India

ABSTRACT

This paper deals with the extraction and classification of environmental natural sounds with the help of content-based retrieval method. Environmental sounds are extremely unpredictable and are very much hard to classify and store in cluster forms according to its feature contents. Environmental sounds provide many contextual clues that enable us to recognize important aspects of our surroundings environment. This paper presented the techniques that allow computer system to extract and classify features from predefined classes of sounds in the environment.

Keywords-content-based retrieval, feature extraction, environmental sound, Mel frequency Cepstral coefficiant.

I. INTRODUCTION

The development of high speed computer has developed a tremendous opportunity to handle large amount of data from different media. The proliferation of world-wide- web has farther accelerated its growth and this trend is expected to continue at an even faster pace in the coming years. Even in the past decade textual data was the major form of data that was handled by the computer. But above advancement in technology has led to uses of non textual data such as video, audio and image which are known as multimedia data and the database deals with this type of data is called multimedia database. This database has much more capability than traditional database. One of the best features of multimedia database is content-based retrieval (CBR). In content base retrieval process the major advantage is that it can search data by its content rather textual indexing. This paper deals with the classification of environmental sounds classification with the help of content-based retrieval and Mel frequency cepstral coefficient (MFCC). Environmental sounds are extremely unpredictable so it's really hard to classify and store in cluster forms according to its feature contents. Environmental sounds provide many contextual clues or characteristics that enable us to recognize important aspects of our surroundings environment. We have presented the different phases of the entire process such as capturing the environmental audio, pre- processing of audio data, feature extraction, training and lastly the testing phase.

The major steps involved in the entire method are as follows:

- Extraction of features for classifying highly diversified natural sounds.
- Making clusters according to their feature similarity.
- Finding a match for a particular sound query from the cluster.

The audio files used for testing purpose are of various categories in sample rate, bit rate and no of channels, but files has been digitized to similar type, i.e. 22k Hz, 16 bit, mono channel. Then all the sounds normalized to -6 db and DC offset corrected to 0%. This reprocessing of sound data helps considering low level descriptor in the sound files are similar, hence we can proceed without any drawback in primary classification methods.

II. RELATED WORKS

A brief survey of previous works relevant to our work is presented here. Parekh R [1] proposed a statistical method of classification based on Zero crossing rate (ZCR) and Root means square (RMS) value of audio file. For this purpose sounds has been categorized in three semantic classes. 1. Aircraft sound, 2.Bird sound, 3.Car sound. First and third are inorganic in nature whereas second is organic in nature. The author tried to extract samples, amplitudes and frequencies of those sounds. To generate the feature vector, ZCR and RMS methods are used. Statistical measure is applied to these features σZ, μZ, μR, σR, T', R' and values are calculated.

Goldhor R. S. *et al* [2] presented a method of classification technique based on dimensional Cepstral Co-efficient. They segmented the sound wave into tokens of sound according to signal power Histogram. From the histogram statistics, a threshold –power level was calculated to separate portions of

the signal stream containing the sound from portions containing only background noise. Result is derived from a high SNR recording. As feature vector they used two-dimensional cepstral-coefficients. They calculated cepstral Coefficient for each token by using Hamming window technique. They also used two types of cepstral coefficients; one is linear cepstral coefficient and another is the Mel frequency cepstral coefficient.

In their paper, Toyoda Y. *et al* [3] presented the uses of neural network to classify certain types sound such as crashing sound, clapping sound, bell sound, switch typing sound etc All these sounds were classified into three categories; burst sound, repeated sound and continuous sound. For this purpose they fetched the environmental sounds from a sound database RWCP-OB. For further improvement, they developed multistage neural network. In this work, a three layer perception neural network is used. In the first step of recognition they used long time power pattern. This classifies the sound as single impact. Second stage of classification was done by combination of short time power pattern and frequency at power peak.

Hattori Y. *et al* [4] proposed an environmental sounds recognition system using LPC - Cepstral coefficients for characterization and artificial neural network as verification method. This system was evaluated using a database containing files of different sound-sources under a variety of recording conditions. Two neural networks were trained with the magnitude of the discrete Fourier transform of the LPC Cepstral matrices. The global percentage of verification was of 96.66%. First they collected a common sound database. A segmentation algorithm is applied to each file. For this purpose they uses 240 point Hamming window with frame overlap of 50%. LPC cepstral coefficient is computed and DFT is measured. For LPC cepstral calculating they used Levinson-Durbin recursion.

Wang J. *et al* [5] focused on repeat recognition in environmental sounds. For this purpose they designed a three module system; 1) Unit of the repetitions; 2) Distance between units; 3) Algorithm to detect repeating part.

For unit extraction they used the concept of a peak in power envelop. They extracted peak power with the help of following:
1.The power is below the threshold T L
2. The interval between the maximum and the neighbor maximum is than the threshold T2.
3. The ratio of the power to the power of the neighbor local minimum is than the threshold T3. Distance of every pair of unit was calculated using dynamic time warping MFCC.

Mel frequency cepstral coefficient (MFCC) was developed by Stevens and was designed to adapt human perception. They developed an algorithm based on matching approximation of repeating parts of the sequence unit. For separating pattern they use the fact that silent segments affect human perception and are recognized as the separators of perceived patterns, determined the length of the silent portion by dynamic process of histogram of length of silent segments using threshold method. Finally distances between every pair of the units are computed by DTW which minimizes the total distance between units by aligning their time series.

Janku L. [6] used MPEG7 low level descriptor for sound recognition with the help of hidden markov model. Author used four low level descriptor defined in mpeg 7. Those are Audio spectrum envelop, audio spectrum centroid, audio spectrum spread and audio spectrum flatness. The testing method described is a three-step procedure:
(a) Receiving audio wave file and taking 512 points short time Fourier transform.
(b) According to the presented spectrogram, we take the associated parameters.
(c) Taking the SVD of flatness and running the Viterbi algorithm to get the sound class candidate.

Chen Z. and Maher R. C. [7] dealt with Automatic off-line classification and recognition of bird vocalizations. They focused on syllable level discrimination of bird sound. For this purpose they developed a set of parameters those are spectral frequencies, frequency differences, track shape, spectral power, and track duration. The parameters include the frequencies, the frequency differences, the relative the shape, and the duration of the spectral peak tracks. These parameters were selected based on several key insights. For the feature extraction they had two choices: MPEG-7 audio features and Mel-scale Frequency Cepstral Coefficients (MFCC), For the classification they had two choices: Maximum Likelihood Hidden Markov Models (ML-HMM) and Entropic Prior HMM (EP-HMM. They have achieved accuracy of 90% with the best and the second best being MPEG-7 features with EP- HMM and MFCC with ML-HMM.

In their paper, Constantine G. *et al* [8] presented a comparison of three audio taxonomy methods for MPEG-7 sound classification. For the low-level descriptors they used, low-dimensional features based on spectral basis descriptors are produced in three stages:

- Normalized audio spectrum envelope.
- Principal component analysis.
- Independent component analysis.

High-level description schemes they used to describe the modeling of audio features, the procedure of audio classification, and retrieval. For classification they tested three approaches:

- The direct approach.
- The hierarchical approach without hints.
- Hierarchical approach with hints.

They suggested that best approach is hierarchical approach with hints, which results in classification accuracy of around 99%. The direct approach produces the second best results, and the hierarchical approach without hints the third best results.

III. SOUND ANALYSIS TECHNIQUE

Undoubtedly the most important part of the entire process is to extract the features from the audio signal. The audio signal feature extraction in a categorization problem. It is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. It is clear from the above discussion of environmental sound identification and verification systems, that the training and test vector needed for the classification problem grows dimension of the given input vector [2-3], feature extraction is necessary. More variety in sound needs more variety in features for a particular genre of sound.

There are various techniques suitable for environmental sound recognition. Recognition of sound (whether it is speech or environmental) is generally consists of two phases:

- The Feature Extraction phase, followed by
- The Classification (using artificial intelligence techniques or any other) phase.

In feature extraction phase, sound is manipulated in order to produce a set of characteristics. Classification phase is to recognize the sound by cataloging the features of existing sounds during training session and then comparing the test sound to this database of features (which is called the testing session). These two phases has been described in the following text.

A. Feature Extraction

Feature extraction mainly of two types. *(a)* Frequency based feature extraction and (*b)* Time-frequency based feature extraction.

Frequency based feature extraction method produces an overall result detailing the contents of the entire signal. According to Cowling [11] Frequency extraction is stationary feature extraction and Time-frequency extraction is non-stationary feature extraction,

Eight popular stationary feature extraction methods are there:

- Frequency extraction (music and speech).
- Homomorphic cepstral coefficients.
- Mel frequency cepstral coefficients (music and speech).
- Linear prediction cepstral (LPC) coefficients.
- Mel frequency LPC coefficients.
- Bark frequency cepstral coefficients.
- Bark frequency LPC coefficients.
- Perceptual linear prediction (PLP) features.

Cepstral frequency based feature extractions are pseudo stationary feature extraction techniques because they split the signal into many small part in time domain. Techniques based on LPC coefficients were based on the idea of a decoder, which is a simulation of the human vocal tract. Since the human vocal tract does not produce environmental sounds, these techniques typically seem to highlight non-unique features in the sound and are therefore not appropriate for recognition of non- speech sounds.

The main non stationary feature extraction techniques are:

- Short-time Fourier transforms (STFT).
- Fast (discrete) wavelet transform (FWT).
- Continuous wavelet transforms (CWT).
- Wigner-Ville distribution (WVD).

These techniques use different algorithms to represent time frequency based representation of signals. Short-time Fourier transforms use standard Fourier transforms of the signal where as wavelet based techniques apply a mother wavelet to a waveform to surmount the resolution issues inherent in STFT. WVD is a bilinear time-frequency distribution.

Moreover there are some other features which are suitable for classification of environmental sounds. These features are described below:

Zero-crossing rate:

The zero-crossing rate is a measure of the number of time the signal value crosses the zero axis. Periodic sounds tend to have small value of it, while noisy sounds tend to have a high value of it. It is computed in each time frame on the signal.

Spectral Features:

Spectral features are single valued features, calculated using the frequency spectrum of a signal. Thus for the time domain signal x (t):

$A(f) = |F[x(t)]|$

Spectral Centroid:

The Spectral Centroid (μ) is bar center of the spectrum. It is a weighted average of the

probability to observe the normalized amplitude. Given A (f) from the equation mentioned above

$$\mu = \int f.p(f)\delta f$$

Where, p (f) = A (f) / ∑A (f)

Spectral spread:

The spectral spread (σ) is a the spectrum around the mean value calculated in following equation

$$\sigma^2 = \int (f -).2p\ (f)\delta f$$

Spectral Roll-off:

The spectral roll-off point (fc) is 95% of the signal energy below this frequency.

IV. DESIGN ARCHITECTURE

In context of multimedia database system major challenge is to classify the environmental sound because they are so unpredictable in nature that they are harder to categorize into smaller number of classes. Classification problem concerns with determining two samples match each other or not under some similarity criterion and to different classes keeping similar type of samples in a class. In case of Content Based Retrieval (CBR) system we must consider one thing that how might people access the sound from database. That may be similar one sound or a group of sounds in terms of some characteristics. For example, class of bird sound, class of sound of car, where system has been previously train on other sound in similar class. To achieve the desired result, we have divided the entire procedure into two sessions - training and testing session.

In order to make an automated classification of environmental sound, main part is to find out the right features of the sound. In this paper, we have selected Mel frequency Cepstral Coefficient (MFCC) to extract features from the sound signal and compare the unknown signal with the existing signals in the multimedia database.

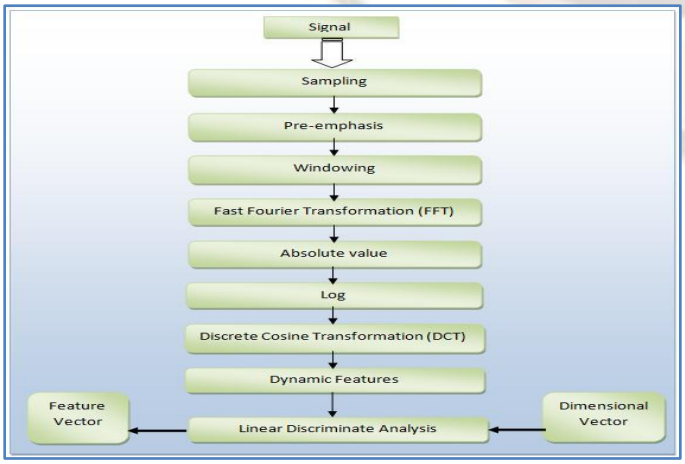

Figure 1: Mel Frequency Cepstral Coefficient pipeline

B. Sampling :

It is the process of converting a continuous signal into a discrete signal. Sampling can be done for signals varying in space, time, or any other dimension, and similar results are obtained in two or more dimensions. For signals that vary with time, let x(t) be a continuous signal to be sampled, and let sampling be performed by measuring the value of the continuous signal every T seconds, which is called the sampling interval. Thus, the sampled signal x[n] given by**:**

$x[n] = x(nT)$, with $n = 0, 1, 2, 3, ...$

The sampling frequency or sampling rate f_s is defined as the number of samples obtained in one second, i.e. $f_s = 1/T$. The sampling rate is measured in hertz or in samples per second.

C. Pre-emphasis :

In processing of electronic audio signals, **pre-emphasis** refers to a system process designed to increase (within a frequency band) the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio (SNR) by minimizing the adverse effects.

D. Windowing :

In signal processing, a **window function** (also known as tapering function) is a mathematical function that is zero-valued outside of some chosen interval. For instance, a function that is constant inside the interval and zero elsewhere is called a rectangular window, which describes the shape of its graphical representation.

Window Terminology**:**

- ***N*** represents the width, in samples, of a discrete-time, symmetrical window function.
- **n** is an integer, with values **0 ≤ n ≤ N-1**.

1) Hamming window:

Hamming window is also called the raised cosine window. The equation and plot for the Hamming window shown below. In a window function there is a zero outside of some chosen interval. For example, a function that is stable inside the interval and zero elsewhere is called a rectangular window that illustrates the shape of its graphical representation, When signal or any other function is multiplied by a window function, the product is also zero valued outside the interval. The windowing is done to avoid problems due to truncation of the signal. Window function has some other applications such as spectral analysis, filter design, audio data compression.

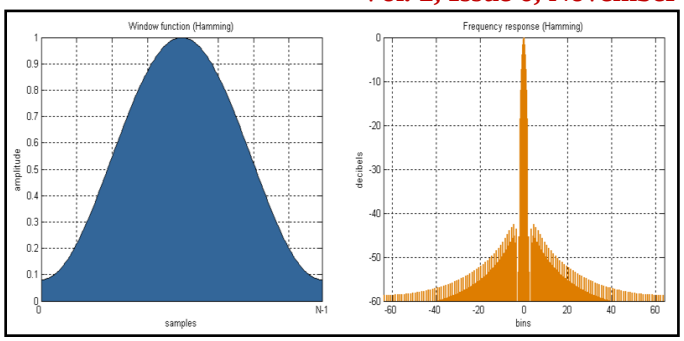

Figure 2: Example of hamming window

The raised cosine with these particular coefficients was proposed by Richard W. Hamming. The window is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window, a raised cosine with simpler coefficients.
Function plotted on the graph bellow :

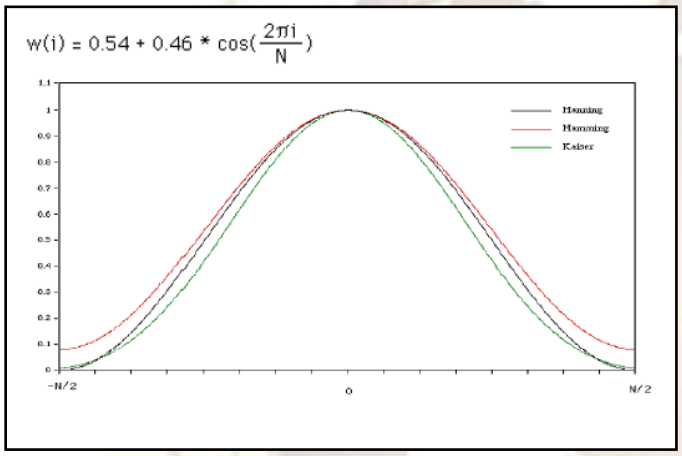

Figure 3: Function is plotted

E. Cepstrum :

Cepstrum name was derived, from the spectrum by reversing the first four spectrums. We can say cepstrum is the Fourier Transformer of the log with phase of the Fourier Transformer frequency bands are not positioned logarithmically. As the frequency bands are positioned logarithmically in the Fourier Transform the MFCC approximates the human system response better than any other system. These coefficients allow better processing of data. In the Mel Frequency Cepstral Coefficients the calculation of the mel is as the real Cepstrum except the Mel cepstrum's frequency is to up a correspondence to the Mel scale. The Mel scale was projected by Stevens, Volkmann and Newman In 1937.The Mel scale is mainly based on the study of observing the pitch or frequency of human. The scale is divided into the units mel. In this test the person started out hearing a frequency of 1000 Hz, and labeled it 1000 mel. Listeners were asked to change the frequency till it reaches to the frequency of the reference frequency. Then this frequency is labeled to 2000 mel. The process is repeated for half of the frequency, then this frequency as 500 and so on. On this basis the normal frequency is mapped into the mel frequency. The mel is normally a linear mapping below 1000 Hz and logarithmically above 1000 Hz.

Transformation of normal frequency into mel frequency and vice versa can be obtained by using the following formulas shown here:

$$m = 1127.01048 \log(1 + \frac{f}{700}) \ldots\ldots\ldots\ldots (1)$$

Where ***m*** is the Mel frequency and f is the normal frequency.

$$f = 700(\frac{e^m}{1127.01048} - 1) \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Here mel frequency is transformed into normal frequency. Figure 4 showing the mapping of normal frequencies to the mel frequencies.

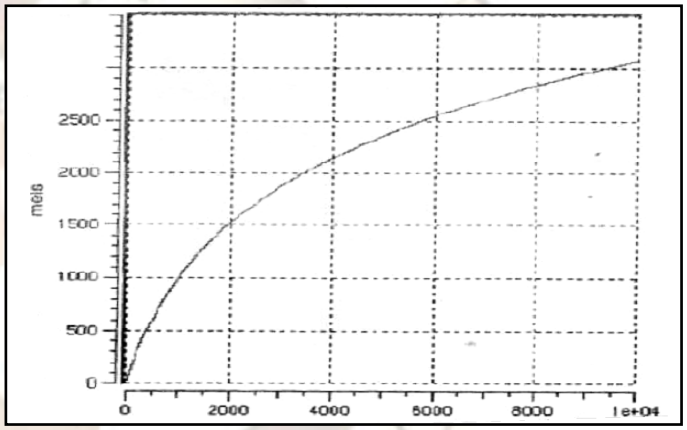

Figure 4: Mapping of normal frequency into Mel frequency

F. Distance measure:

In sound recognition phase, an unknown sound is represented by a sequence of feature vectors (x1, x2 ... xi), and then it is compared with the other sound from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance. The Euclidean distance is the scalar distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem, The formula used to calculate the Euclidean distance is as follows:

The Euclidean distance between two points P = (x1, x2...xn) and Q = (y1, y2… yn)

$$\boldsymbol{d} = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x_i} - \boldsymbol{y_i})^2}$$

The sound signal with the lowest distortion distance is chosen to be identified as the unknown signal.

G. Fast Fourier Transformation :

FFTs are of great importance to a wide variety of applications, from digital signal processing and solving partial differential equations to algorithms for quick multiplication of large integers.

H. Absolute value:

In mathematics, the **absolute value** (or **modulus**) $|a|$ of a real number a is the numerical value of a without its sign. The absolute value of a number may be thought of as its distance from zero.

Generalizations of the absolute value for real numbers occur in a wide variety of mathematical settings. For example an absolute value is also defined for the complex numbers, the quaternion, ordered rings, fields and vector spaces. The absolute value is closely related to the notions of magnitude, distance, and norm in various mathematical and physical contexts.

I. DCT :

In particular, a DCT is a Fourier-related transform similar to the discrete Fourier transform (DFT), but uses only real numbers. DCTs are equivalent to DFTs of roughly twice the length, operating on real data with even symmetry (since the Fourier transform of a real and even function is real and even), where in some variants the input and/or output data are shifted by half a sample. There are eight standard DCT variants, of which four are commonly used.

J. Linear Discriminate Analysis (LDA):

Linear discriminate analysis (LDA) and the related **Fisher's linear discriminate** are methods used in statistics, pattern recognition and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier or, more commonly, for dimensionality reduction before later classification.

LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempts to express one dependent variable as a linear combination of other features or measurements. In the other two methods, however, the dependent variable is a numerical quantity, while for LDA it is a categorical variable (*i.e.* the class label). Logistic regression is more similar to LDA, as it also explains a categorical variable. These other methods are preferable in applications where it is not reasonable to assume that the independent variables are normally distributed, which is a fundamental assumption of the LDA method.

LDA is also closely related to principal component analysis (PCA) and factor analysis in that both look for linear combinations of variables which best explains the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminate analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made.

V. TRAINING AND TESTING SESSIONS

As stated earlier, due to achieve the best result, the entire sound detection and classification process is divided into two sessions. One is called the training session, during which the system is being trained with known sound signals and features of that known signals are extracted. These extracted features are then stored into the multimedia database as the reference model for future reference. Figure 5 shows the processes during the training session. Form this figure, we can see that the known sound signal is first re-sampled, then re-referenced and normalized. The normalized signal is then passed through the process to extract the features. System is then trained with the help of these features and reference model is built.

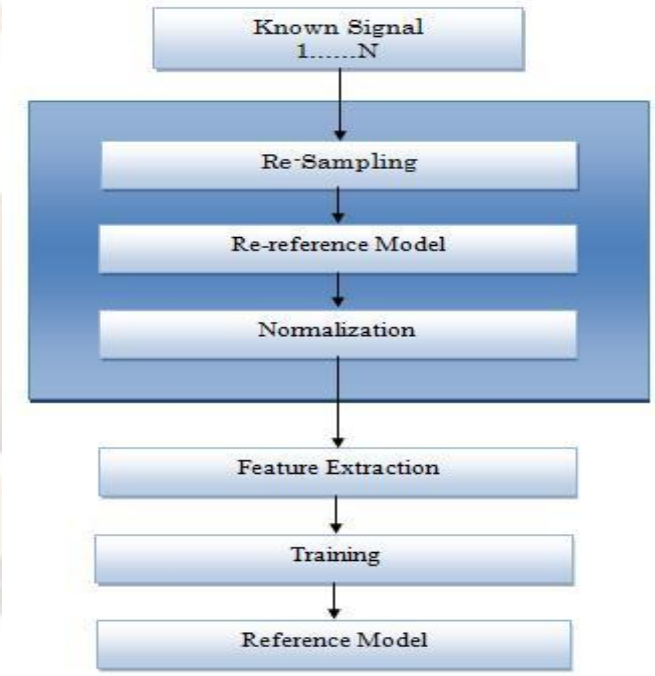

Figure 5: Flow chart of Training Session

Through the training session, the multimedia database is populated by feature vectors of known sound signals so that those known environmental sound can be detected and classified later. In case of unknown signals, signals are newly classified during training session and detection can be done with that classification.

Another session of the detection and classification process is the testing session. During this session, depicted in figure 6, the input signal is sampled, DC offset is determined and then normalization is done. After normalization, features are extracted for the input signal. Then for a particular sound, feature vectors are retrieved from the multimedia database using content-based retrieval (CBR) method. The retrieved features are then matched with the feature vector of the input sound signal.

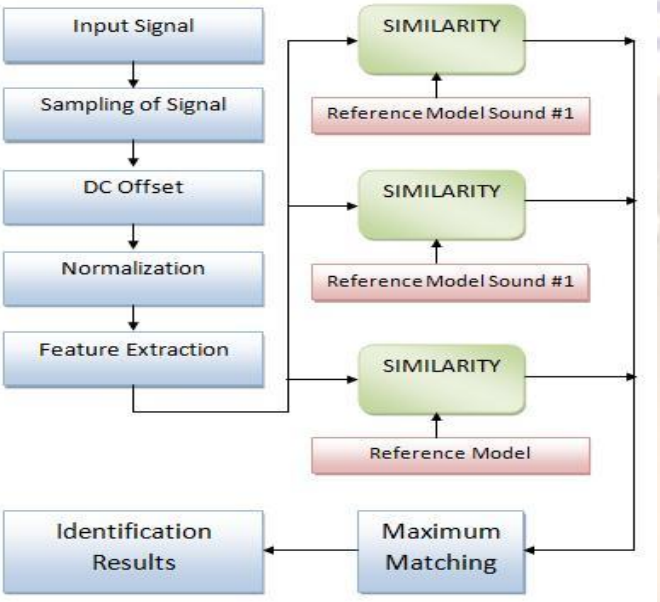

Figure 6: Flowchart of Testing Session

After feature extraction, the matching process is carried on and feature vector of input signal is matched with every reference model of sound signals stored in the database. Whenever a maxim um matching is found, the input signal is said to be detected. This is the entire process of detection and classification of environmental sounds. In both sessions, feature extraction process is done using the MFCC pipeline.

VI. CONCLUSION

This method of environmental sound detection and classification is developed using MFCC pipeline and CBR for extraction of features of a particular sound and retrieval of sound features from the multimedia database respectively. This method can be implemented in the domain of robotics where sound detection and recognition may be possible up to a satisfactory level. If the method will be properly implemented with computer vision, then human-computer interaction process can be developed much. MFCC is undoubtedly more efficient feature extraction method because it is designed by giving emphasis on human perception power. Using more than one features of a sound may obviously improve the performance of the method. Applying clustering technique, accuracy can be boosted. Another good feature available today is Audio spectrum projection provided by MPEG7 specification. Inclusion of this feature may increase the performance measure of the method.

REFERENCES

[1] Ranjan Parekh, "Classification Of Environmental Sounds",ICEMC2 2007, Second International Conference on Embedded Mobile Commu- nication and Computing, ICEMC2 2007.

[2] R. S. Goldhor, "Recognition of Environmental Sounds/' Proceedings of ICASSP, vol.1, pp.149-152, April 1993.

[3] Yoshiyuki Toyoda, Jie Huang, Shuxue Ding,Yong Liu, "Environmental Sound Recognition by Multi layered Neural Networks" ,The Fourth International Conference on Computer School of Education Technology, Jadavpur University, Kolkata 32 and Information Technology (CIT'04) pp. 123-127,2004.

[4] Yuya Hattori, Kazushi Ishihara, Kazunori Komatani, Tetsuya Ogata, Hiroshi G.Okuno/Repeat Recognition for Environmental Sounds", Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication Kurashiki, Okayama Japan September 20-22, pp. 83-88, 2004.

[5] Jia-Ching Wang, Jhing-Fa Wang, Kuok Wai He, and Cheng-Shu Hsu/'Environmental Sound Classification Using Hybrid SVM/KNN Classifier and MPEG-7 Audio Conference on Low-Level Descriptor", 2006 International Joint Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[6] Ladislava Janku, "Artificial Perception: Auditory Decomposition of Mixtures of Environmental Sounds - Combining Information Theoretical and Supervised Pattern Recognition Approaches", P. Van Emde Boas et al. (Eds.): SOFSEM 2004, LNCS 2932, pp. 229-240,2004

[7] Z, Chen, R. C, Maher, '"Semi-automatic classification of bird vocalizations1 Accost Soc Am., Vol. 120, No. 5, pp. 2974-2984, November 2006.

[8] G. Constantine, A, Rizzi, and D, Casali, '"Recognition of musical instruments by generalized min-max classifiers/' in IEEE Workshop on Neural Networks for Signal Processing, pp. 555-564, 2003.