International Journal on Applications in Information and Communication Engineering
Volume 1: Issue 5: June 2015, pp 42-45. www.aetsjournal.com ISSN (Online) : 2394-6237

Environment and Speaker Recognition System Using MFCC

Nisha K.C

Abstract— Environment recognition and speaker recognition from digital audio for forensic application is a growing area of interest. A significant amount of research can be found in the area of speech recognition or enhancement speaker recognition and authentication of audio. However, a very few researches can be found in the area of environment recognition for digital audio forensics, where foreground human speech is present in the environment recordings, which is a forensics scenario. This paper aims to recognize the environment sound and the speaker sound from an audio sample that contains both the environment sound and speaker sound or from an audio sample of an environment sound with some foreground speech on it. For doing this the feature which is selected to extract from the audio sample is Mel Frequency Cepstral Coefficients (MFCC). The classifier used is K Nearest Neighbour (KNN).

Keywords— digital audio forensics , environment recordings , speaker sound , foreground speech , MFCC, KNN.

I. INTRODUCTION

Digital forensics is the branch of science that deals with different types of scientific techniques for the collection, collection, identification, validation, documentation, presentation, interpretation, and analysis of the digital evidence that are obtained from various digital sources for the facilitation or for the reconstruction of events, usually a criminal event [1]. There are various branches in the field of digital forensics: audio forensics, image forensics, multimedia forensics, video forensics, etc.

This paper mainly deals with digital audio forensics. Audio forensics is the branch of digital forensic science that is related to the acquisition, analysis, and evaluation of sound recordings or audio which were used to present in a court of law or some other official venue as admissible evidence in some criminal cases. The evidence for the audio forensic may be obtained as a part of the official inquiry into an accident, accusation of slander, fraud, or some other incidents or may come from a criminal investigation by law enforcement.

The environment of a speaker can be determined by examining the characteristic background sounds present in the environment, like the sound of trains in the background in a railway station, sound of crowd in a noisy street, sound of fan in a room etc.

But the problem is that, the identification of the environment of a speaker does not provide any information about the actual and exact position (x,y) of the speaker or sound sources and also location cannot be identified [2]. But through speech recognition methods, if a speaker cannot be verified properly or if we have to make sure that the speaker has spoken from a place such as a courtyard or a certain room or at home, the analysis of acoustic characteristics of the speaker's environment will indicate his or her location. So, using this 'indirect' channel, speaker verification or speaker identification can also be attained.

In the area of speech recognition or speaker recognition and authentication of audio, now-a-days many research have been going on. However, we can notice only a very few researches in the area of environment recognition for digital audio forensics, where some foreground human speech will be present in the sound recordings of the environment. This situation is very much complicated.

In this paper, we present an environment sound and speaker recognition system using Mel Frequency Cepstral Coefficients (MFCC) and K Nearest Neighbour (KNN) as the classifier. i.e., In this paper two recognition processes are done, 1) environment sound recognition and 2) speaker recognition. So the aim of the project is to recognize the environment sound and the speaker sound from an audio sample that contains both the two sounds or an audio sample of an environment sound with some foreground speech on it.

II. RELATED WORKS

Digital audio forensics is the field of science that deals with the audio files that are left over in any type of crime spot. But less attention is given to perform the environment recognition from files where some foreground speech is present, which is the exact forensics application. So in [3] the authors, had discussed about such a condition of environment recognition. In this paper, the authors had used a combination of full set of MPEG-7 audio features and MFCC to improve the accuracy. In the paper [4], the task of recognizing and learning different types of environments for mobile robot which is done using audio information is discussed. In this paper, classifiers GMM and SVM were used. The results show that SVM have the highest accuracy rate over the other two methods. The paper [5] discuss about speaker recognition using GTCC, which results in better recognition performance when compared with MFCC.

III. PROPOSED SYSTEM

In this section, the process of environment sound recognition and speaker recognition from a single audio file that contains both environment sound and foreground speech is addressed. The proposed method can be divided into four phases.

Nisha. K. C., P.G. Scholar, Department of ECE, KMEA Engineering College, Ernakulam, Kerala, India. (Email: kcnisha39@yahoo.com)

I. Database capture (via portable recording devices).
II. Feature extraction.
III. Training.
IV. Testing the classifier.

The general block diagram of the proposed scheme is shown below.

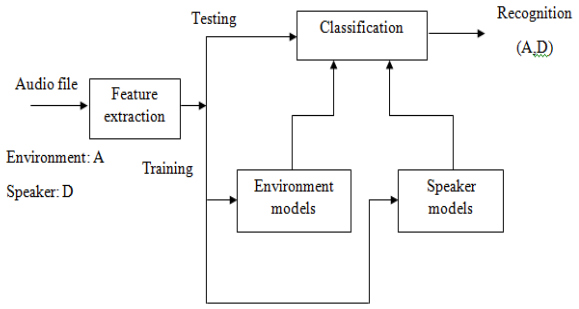

Figure 1. Block diagram of the proposed system

The extraction of the best parametric representation of audio signals is an important task to produce a better performance in the recognition process. For the phases after feature extraction, an important fact that depends is the efficiency of the feature extraction phase, because the behaviour of the feature extraction phase affects the next phase.

MFCC is the acoustic feature which is extracted. For each audio sample (both environment and speaker sound), MFCC feature is extracted and these extracted features are given to the classifier on which the training is done. KNN is used as the classifier. So using the KNN classifier, environment models and speaker models are separately created. During the testing process, audio sample containing both environment sound and speaker sound is given to the classifier and the classifier will correctly recognize which environment it is and who the speaker is.

Mel Frequency Cepstral Coefficients (MFCC)

Various steps to calculate the MFCC of an audio sample is given in detail below.

Step 1:Pre-emphasis

The first step in MFCC extraction is passing the audio sample (both environment sound and speaker sound) into a pre-emphasis filter. Pre-emphasis filter is a high pass filter which will compensate for the high-frequency part of the input audio that is suppressed during the sound production mechanism from the human vocal chord. Also, another function of the filter is to amplify the high-frequency formants.

. The main function of a pre-emphasis filter is to compensate for the problem of spectral tilt. When the audio signal s(n) is sent to the pre-emphasis filter, the filter output is given by:

$$s_2(n) = s(n) - a*s(n-1) \quad (1)$$

where $s_2(n)$ is the output signal and the value of a is usually between 0.9 and 1.0.

Step 2: Frame blocking

Every audio signal will be constantly changing and for doing any operation on such a changing audio signal is very difficult. This is the main reason for framing the signal. If the frame is shorter, then there will not be enough samples to get a reliable spectral estimate, and if the frame is longer the signal changes too much throughout the frame, which also does not give a good result.

So the input audio signal should be segmented into frames of 20~30 ms with an overlap of 1/3~1/2 of the frame size.

Step 3: Hamming windowing

The next step is windowing. Each frame of the audio sample must be multiplied with a hamming window. This is done to keep the continuity of the first and the last points in each frame of the audio signal. If the signal in a frame is denoted as s(n), where n = 0,…N-1, then the signal obtained after the windowing is s(n)*w(n) , where w(n) is the Hamming window defined by:

$$w(n, a) = (1 - a) - a\cos(2\pi n/(N-1)), \quad 0 \leqq n \leqq N-1 \quad (2)$$

where a=0.54

Step 4: Discrete Fourier Transform (DFT)

In order to obtain the magnitude frequency response of each frame after the windowing, discrete fourier transform is done. The discrete fourier transform of each frame is calculated using the following equation.

$$F(j\omega) = \sum_{k=0}^{N-1} f[k]e^{-j\omega kT} \quad (3)$$

Step 5: Triangular Bandpass Filters or Mel Filter Bank

In this step, the magnitude frequency response obtained after performing the discrete fourier transform of each audio frame is multiplied by a set of 20-40 triangular bandpass filters, known as the mel filters, to get the log energy of each triangular bandpass filter. The position of these mel filters are equally spaced along the mel scale. The mel scale conversion formula is explained in eqn 4 given below

$$\text{Mel}(f) = 2595 \log10(1+(f/700)) \quad (4)$$

Step 6: Log Energy

Truly speaking, mel filter bank is not a filter, it will only weight each point of the spectrum with filter weights. To compensate for the increasing band-widths of the mel filter bank, log of each energy coefficient is calculated.

Step 7: Discrete cosine transform (DCT)

In this step, DCT is applied on the output obtained from the triangular bandpass filters to have L mel-scale cepstral coefficients. The formula for DCT is shown next.

$$C_m = S_{k=1}^{N}\cos[m*(k-0.5)*p/N]*E_k, \; m=1,2, ..., L \quad (5)$$

where N is the number of mel scale bandpass filters, L is the number of mel frequency cepstral coefficients.

IV. Experimental Results

In this paper, the database used consists of ten different environment sounds: closed room, open space, beach, car bus road side, office, class room , rain and bridge and four different speaker sounds: two adults (male and female) and two children (male and female).

Sounds were recorded using Nokia Lumia 520 sound recorder. Sampling rate was set to 44.1kHz, and quantization was 16 bits. Each recording was having 10 seconds duration. The audio samples recorded were converted to .wav format.

The speech samples consisted of some English utterances of 10 seconds of two adult speakers and two child speakers. These speech samples were added (overlapped) to all the recordings of the environment sound. These combined audio samples were used as the test signals in order to perform the comparative study.

The audio patterns were divided into train and test data sets to yield statistically reliable results. The environment sounds and speaker sounds are separately trained on the classifier so that there were separate environment models and speaker models. The test samples were made by overlapping the environment sounds and speaker sounds. Since there are 10 number of environment sounds and 4 number of speaker sounds, there will be a combination of 40 test samples.

In the experiments conducted, different types of features were extracted from the audio samples. This is done to investigate the performance of different features extracted. The task of the proposed recognition system in the experiment is to correctly recognize the ten environment sounds and 4 speaker sounds. Four variants of feature vectors are calculated for MFCC, say MFCC, MFCC(13), MFCC(26) and MFCC(39). For all the 40 combination of test samples, all of these types of features were calculated and analyzed.

The figure 2 represents the hit analysis. This graph shows the number of hits obtained while using different features. In this graph, X-axis represents the number of tries and Y-axis represents the number of hits. It is clearly seen that the hit ratio of MFCC is high when compared with the other features and MFCC(13) is having the least value.

The figure 3 represents the miss analysis. This graph shows the number of miss obtained while using different features. In this graph, X-axis represents the number of tries and Y-axis represents the number of miss. It is clearly seen that the miss ratio of MFCC is less when compared with the other features and MFCC(13) is having the highest value.

The figure 4 represents the accuracy analysis. This graph shows the accuracy obtained while using different features. In this graph, X-axis represents the number of tries and Y-axis represents the accuracy in percentage. The maximum value of accuracy is 100%. The accuracy analysis is plotted for different values of input samples and accuracies are calculated from the recognition of correct environment and speaker sound. It is clearly seen that the accuracy is more for MFCC and lesser for MFCC (13).

The figure 5 represents the error rate analysis. This graph shows the error rate while using different features. In this graph, X-axis represents the number of tries and Y-axis represents error rate in percentage. The maximum value of error rate is 100%. The error analysis is plotted with different values of input samples and error rates are calculated from the incorrect recognition of environment and speaker sound. It is clearly seen that the error rate of MFCC is less and that of MFCC(13) is the highest. As a result, less error rate is obtained only when there are more hits and lesser misses.

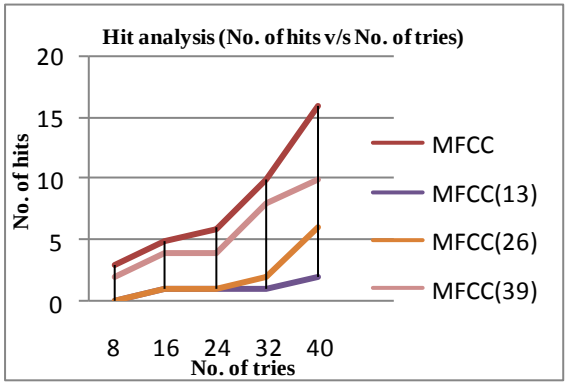

Figure 2: Hit analysis

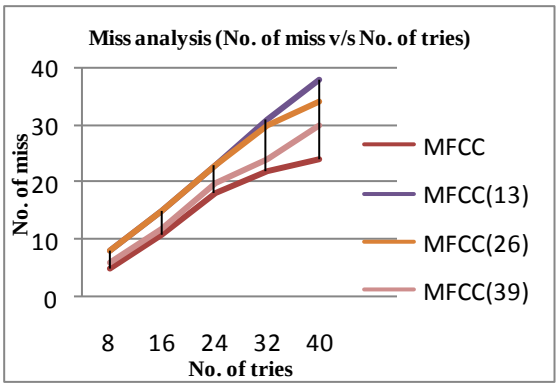

Figure 3: Miss analysis

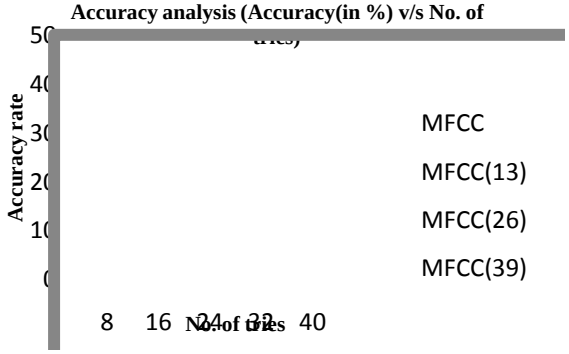

Figure 4: Accuracy analysis

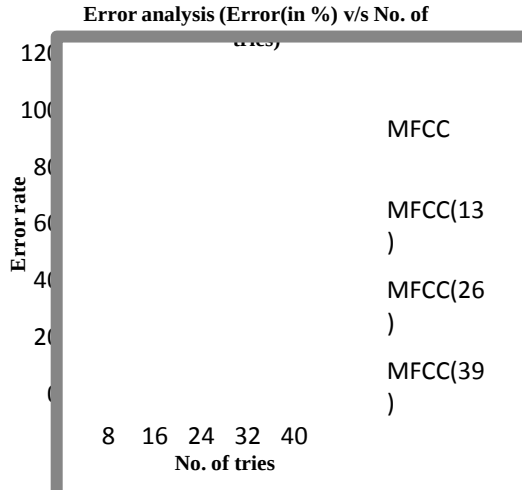

Figure 5: Error analysis

V. CONCLUSIONS

In this paper, an environment & speaker recognition system has been proposed for digital audio forensics application, which is a system that will recognize a given environment sound and speaker sound from an audio sample of environment sound with some foreground speech.

In this paper, MFCC feature was used for the recognition procedure of the proposed system. Also KNN is used as the classifier.

REFERENCES

[1] Delp E., Memon N., and Wu M., "Digital Forensics" ,IEEE Signal Processing Magazine, vol. 3, no. 1, pp. 14-15, 2009.

[2] Andrea Oermann, Andreas Lang, Jana Dittmann Ottovon Guericke, "Verifier Tuple for Audio Forensic to determine speaker environment", in Proc. 7th workshop on multimedia and security, pp. 57-62, August 2005.

[3] Ghulam Muhammad and Khaled Alghathbar, "Environment Recognition for Digital Audio Forensics Using MPEG-7 and Mel Cepstral Features", The International Arab Journal of Information Technology, Vol. 10, No. 1, January 2013.

[4] Selina Chu, Shrikanth Narayanan, and C.C. Jay Kuo, "Content Analysis for Acoustic Environment Classification in Mobile Robots", American Association for Arti_cial Intelligence, vol. 2, 2006.

[5] Rahana Fathima, Raseena P E, "Gammatone Cepstral Coefficient for Speaker Identification", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 2, Special Issue 1, December 2013.

[6] Patterson, R., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., "Complex sounds and auditory images, Auditory Physiology and Perception", (Eds.) Y Cazals, L. Demany, K.. Horner, Pergamon, Oxford, pp. 429-446, 1992.

[7] Gelfand, S. A, "Hearing: an introduction to psychological and physiological acoustics (4th ed.)", New York: Marcel Dekker, 2004.

[8] Brian C. J. Moore, "Perceptual Consequences of Cochlear Damage(1st ed.)", Oxford Medical Publications, Oxford University Press, December14, 1995.