Journal of ELECTRICAL ENGINEERING, VOL. 62, NO. 4, 2011, 199–205

ENVIRONMENT RECOGNITION FOR DIGITAL AUDIO FORENSICS USING MPEG–7 AND MEL CEPSTRAL FEATURES

Ghulam MUHAMMAD [*] — **Khalid ALGHATHBAR** [**]

Environment recognition from digital audio for forensics application is a growing area of interest. However, compared to other branches of audio forensics, it is a less researched one. Especially less attention has been given to detect environment from files where foreground speech is present, which is a forensics scenario. In this paper, we perform several experiments focusing on the problems of environment recognition from audio particularly for forensics application. Experimental results show that the task is easier when audio files contain only environmental sound than when they contain both foreground speech and background environment. We propose a full set of MPEG-7 audio features combined with mel frequency cepstral coefficients (MFCCs) to improve the accuracy. In the experiments, the proposed approach significantly increases the recognition accuracy of environment sound even in the presence of high amount of foreground human speech.

K e y w o r d s: audio forensics, environment recognition, MPEG-7 audio, MFCC

1 INTRODUCTION

Digital forensics can be defined as the collection of scientific techniques for the preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events, usually of a criminal nature [1]. There are several areas of digital forensics: image forensics, audio forensics, video forensics, multimedia forensics, *etc.*

In this paper, we concentrate on digital audio forensics. Digital audio forensics is to provide evidence from left over audio files contained in audio / video media in the crime spot. This type of forensics can be categorized into four different classes according to its nature: (a) speaker identification / verification / recognition, to find the answer of 'who', (b) speech recognition / enhancement, to find the answer of 'what', (c) environment detection, to find the answer of 'where' or 'situation', and (d) source authentication, to find the answer of 'how'.

A significant amount of research can be found in the area of speech recognition or enhancement [2, 17, 18], speaker recognition [3, 19, 20], and authentication of audio [4]. 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 environment recordings. There are many difficulties while dealing with recognition of environment from audio. Unlike speech or speaker recognition cases, different environment sounds may have similar characteristics (crowded shopping mall and crowded restaurant, and quiet office room and quiet bank, *etc*). Most of the works found in the literature related to environment recognition dealing with audio files that contain only environmental sound. The problem arises when there is foreground speech in the files, which should be the actual case for forensics application. Consider a scenario where a kidnapper is making negotiation through some audio media. In this case, the audio file naturally contains kidnappers' or victim's speech for most of the part, which is not 'relevant' to environment recognition. We present in this paper several experiments on environment recognition for digital audio forensics. Ten different environments and two types of audio files: One containing environment sound only and the other with both foreground speech and background environment sound, are used in the experiments. At first, we find the recognition accuracy of the environments using audio files that contain only corresponding environmental sound. In the second type of experiment, we mix human speech, male and female, to the environmental sounds and repeat the experiment to find the complexity level of detecting environment in the presence of foreground human speech. Finally, we propose a method to improve the recognition accuracy of environment in presence of foreground speech. In the proposed method, we apply a full use of MPEG-7 audio features coupled with MFCC (Mel frequency cepstral coefficient) to represent the environmental sounds, and an HMM (hidden Markov model) based classifier, with separate modeling for each environment sound (class) and human speech, to recognize the environments. The rest of the paper is organized as follows. Section 2 gives a review of related past works; Section 3 describes the data used in the experiments. Section 4 presents the proposed approach to recognize environment sound in presence of human speech. In this section, the experimental results and discussion are also given. Finally, Section 5 draws some conclusions with future direction.

[*] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, PO Box: 51178, Riyadh 11543, Saudi Arabia, ghulam@ksu.edu.sa; [**] Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia,

DOI: 10.2478/v10187-011-0032-0, ISSN 1335-3632

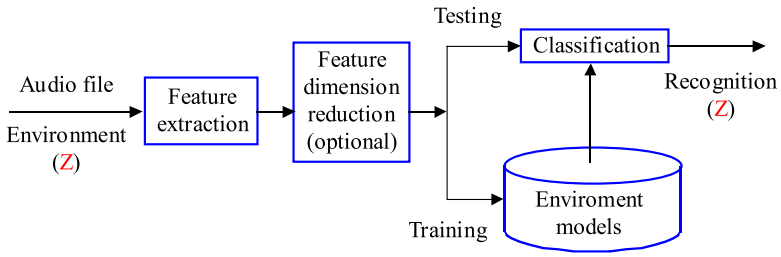

Fig. 1. Block diagram of environment recognition from audio files

2 RELATED WORKS

A block diagram of environment recognition from audio file is given in Fig. 1. The procedure is divided into two main blocks: feature extraction and classification. In the feature extraction block, input audio stream is represented into some suitable feature form. Feature dimension reduction is an optional block that minimizes the dimension of feature vector without losing too much information. In a training phase, different types of environment are modeled using the features. In a testing phase, input features are compared against each model to classify the matched environment. This section describes different types of feature parameters and classification methods used in the literature for environment recognition from audio.

2.1 Feature extraction

MFCCs are the most frequently used features, which are applied not only in environment recognition but also in speech and speaker recognition applications [5]. Eronen et al presented a comprehensive evaluation of a computer and human performance in audio-based context recognition [6]. In their work, they used several time-domain and spectral-domain features in addition to MFCC. Time-domain features included zero- crossing rate and short-time average energy, while spectral-domain features included band energy of logarithmically spaced subbands, spectral centroid, spectral roll-off, and spectral flux. They also used linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), and delta coefficients of MFCC. MFCC and LPC features were also used in [7]. Selina et al [8] introduced matching pursuit (MP) technique [9] in environmental sounds recognition. MP provides a way to extract features that can describe sounds where other audio features (*eg* MFCC) fail. In their MP technique, they used Gabor function based time- frequency dictionaries. It was claimed that features with Gabor properties could provide a flexible representation of time and frequency localization of unstructured sounds like environment. MFCCs and spectral centroid were used as features in [10], while [11] used MFCC features only. Unlike [6, 8, 11, 7, 12, and 13], they used forensics application like audio files, where both ambient (environmental) sound and human speech were present. However, they selected only those segments that were quieter than the average power in an audio file for the experiments. MPEG-7 audio low level descriptors as features were used on environmental sound classification in [12]. The three features they used are: audio spectrum centroid, audio spectrum spread, and audio spectrum flatness. Ntalampiras et al [13] used MFCC along with MPEG-7 features to classify urban environments. They exploited a partial use of MPEG-7 low level descriptors that include audio waveform, audio power, audio spectrum centroid, audio spectrum spread, audio spectrum flatness, harmonic ration, upper limit of harmonicity, and audio fundamental frequency. However, neither [12] nor [13] used full MPEG-7 descriptors, or combination of MFCC and MPEG-7 descriptors.

From the above discussion, we can find that almost all the works used MFCC, some spectral features, or a partial MPEG-7 Audio features.

2.2 Feature dimension reduction

The reduction of feature dimension is applied to reduce the dimension as well as the correlation between feature parameters. Eronen et al [6] used PCA (principal component analysis), ICA (independent component analysis), and LDA (linear discriminated analysis), while Malkin and Waibel [10], and Ntalampiras *et al* [13] applied PCA. Zeng [7] used greedy method to reduce the number of feature dimension.

2.3 Classifier

While HMMs are widely used in the applications, k-NN (k-nearest neighbors) classifier is also applied due to its simplicity [15]. In [6], Eronen et al used two types of classifiers separately: k-NN ($k = 1$), and HMM with number of states and number of mixtures within each state varying from 1 to 4 (and 5), respectively. Selina et al [8] applied k-NN ($k = 1$), and GMM with 5 mixtures.

Malkin and Waibel [10] introduced linear autoencoding neural networks for classifying environment. The autoencoder is a standard feed-forward neural network with a linear transform function. A hybrid autoencoder and GMM was used in their experiments.

A new environmental sound classification architecture that fuses SVM and k-NN were proposed in [12]. For SVM, they used three different types of kernel functions: linear kernel, polynomial, and radial basis kernel. In [16], authors found 96.6 % accuracy for SVM, 94.3 %

for HMM, and 93.4 % for GMM, using forward selection of features.

From classification point of view, we see that HMM or GMM is the most widely used classifier, while some other hybrid classifiers also exist.

2.4 Findings

Twenty four different contexts were grouped into six higher level classes in [6]. Nature and outdoors were recognized with highest accuracy (96–97 %), and library (quiet place) with the lowest (35 %). If the segment length was below 20 second, the performance dropped quickly.

It was shown in [8] that MFCC and MP-based features provided a complementary effect for one another for classifying the classes. The average accuracy for 14 classes was obtained as 83.9 %.

The autoencoder and GMM achieved 77.9 % and 77.57 % accuracy, respectively, in the experiments reported in [10], while a hybrid system between them provided 80.05 % accuracy. MFCC and 11-state HMMs gave 91.5 % average accuracy for 14 classes in [11]. Office, football match, beach, and laundry classes achieved 100 % accuracy, while street class gave 75 %.

A hybrid SVM/kNN system with three MPEG-7 features achieved 85.1 % accuracy averaged over 12 classes [12]. This result was obtained with radial basis kernel in SVM. The same three features with HMM classifier gave 83.2 % accuracy. Ntalampiras et al [13] found that MPEG-7 features reached 75.3 % recognition rate while MFCCs achieved only 64.1 % in their experiments on urban environmental sound classification. They did not use combination of MFCC and MPEG-7 features.

2.5 Environment detection for audio forensics

In an interesting work, which was claimed to be the first approach for digital media forensics to determine the used microphone and the environments of recorded audio signals, Kraetzer et al [14] extracted 63 statistical features from audio signals. Seven of the features were time domain: empirical variance, covariance, entropy, LSB ratio, LSB flipping rate, mean of samples, and median of samples. Besides these temporal features, they used 28 mel-cepstral features and 18 filtered mel-cepstral features. They applied k-NN and Na?ve Bayes classifiers to evaluate microphone and environmental classification for digital audio forensics. In the experiments, they mixed human speech, music sound, and pure tone with the environmental sound to recognize. They reported highest 41.54 % accuracy obtained by Na?ve Bayes classifier with 10 fold cross validation, while 26.49 % as its best by simple k-means clustering. They did not use HMM or GMM for classification.

3 DATA

We recorded audio signals from ten different environments: restaurant, crowded street, quiet street, shopping mall, car with open window, car with closed window, corridor of our university campus, office room, desert, and park. All of the scenes were from Riyadh city.

Sounds were recorded with an IC recorder (ICD-UX71F/UX81F/UX91F). Sampling rate was set to 22.05 kHz, and quantization was 16 bit. Each recording consisted of 30 seconds. There were 200 recordings for each scene. Scene recordings were made at different times and different locations. For example with office room, some recordings were made at Faculty Member X's office room in the morning and in the afternoon; some were made at Y's office room at different times, and so on.

Some Arabic utterances of ten seconds and 20 seconds of lengths from three male speakers and three female speakers were added (overlapped) to all the recordings at random position, keeping in mind that the utterances fit within 30 second length. For simplicity, we fixed three 10 second utterances and added any of the three fixed utterances from any of the speakers to the environment sounds. For 20 seconds of utterances, two 10 second utterances were added with little pause between them. The utterances were added at signal to noise ratio of 15 dB, where environment sound was represented as noise. We used a different set of five male and five female utterances of 20 seconds each for training, which will be described later.

4 THE PROPOSED APPROACH

4.1 Feature parameters

Different types of feature sets are investigated in the experiments. Each audio file are divided into 25 ms frames with 50 % overlap. The most common features in the field of speech processing are MFCCs. MFCCs are fast to extract and proved to be efficient in speech speaker recognition applications. These features are designed to mimic human auditory perception by using filter bank with Mel-scaled frequency. In our experiments, two different dimensions of MFCCs are extracted. The first one is of 13 dimension including 12 MFCCs and log energy from the raw signal. The second one is of 26 dimensions which include those 13 values plus their delta coefficients.

In our approach, we apply MPEG-7 Audio features for environment recognition from audio files. Though [13] utilized partial MPEG-7 features with seven dimensions, we exploit a full advantage of MPEG-7 features in this work. MPEG-7 Audio describes audio content using low-level characteristics, structure, models, etc. The objective of MPEG-7 Audio is to provide a fast and efficient searching, indexing, and retrieval of information from audio files. There are 17 temporal and spectral low-level descriptors (or features) in MPEG-7 Audio. They can be divided into scalar and vector types. Scalar type returns scalar value such as power or fundamental frequency, while vector type returns, for example, spectrum flatness calculated for each band in a frame. In the following we describe, in brief, MPEG-7 Audio low-level descriptors.

1. Audio Waveform (scalar): It describes the shape of the signal by calculating the maximum and the minimum of samples in each frame.
2. Audio Power (scalar): It gives temporally smoothed instantaneous power of the signal.
3. Audio Spectrum Envelop (vector): It describes short time power spectrum for each band within a frame of a signal.
4. Audio Spectrum Centroid (scalar): It returns the center of gravity (centroid) of the log-frequency power spectrum of a signal. It points the domination of high or low frequency components in the signal.
5. Audio Spectrum Spread (scalar): It returns the second moment of the log-frequency power spectrum. It demonstrates how much the power spectrum is spread out over the spectrum. It is measured by the root mean square deviation of the spectrum from its centroid. This feature can help differentiate between noise-like or tonal sound/speech.
6. Audio Spectrum Flatness (vector): It describes how much flat a particular frame of a signal is within each frequency band. Low flatness may correspond to tonal sound.
7. Audio Fundamental Frequency (scalar): It returns fundamental frequency (if exists) of the audio.
8. Audio Harmonicity (scalar): It describes the degree of harmonicity of a signal. It returns two values: harmonic ratio and upper limit of harmonicity. Harmonic ration is close to one for a pure periodic signal, and zero for noise signal.
9. Log Attack Time (scalar): This feature may be useful to locate spikes in a signal. It returns the time needed to rise from very low amplitude to very high amplitude.
10. Temporal Centroid (scalar): It returns the centroid of a signal in time domain.
11. Spectral Centroid (scalar): It returns the power-weighted average of the frequency bins in linear power spectrum. In contrast to Audio Spectrum Centroid, it represents the sharpness of a sound.
12. Harmonic Spectral Centroid (scalar):
13. Harmonic Spectral Deviation (scalar):
14. Harmonic Spectral Spread (scalar):
15. Harmonic Spectral Centroid (scalar): The items (l-o) characterizes the harmonic signals, for example, speech in cafeteria or coffee shop, crowded street, *etc.*
16. Audio Spectrum Basis (vector): These are features derived from singular value decomposition of a normalized power spectrum. The dimension of the vector depends on the number of basis functions used.
17. Audio Spectrum Projection (vector): These features are extracted after projection on a spectrum upon a reduced rank basis. The number of vector depends on the value of rank. For audio spectrum basis and audio spectrum projection vectors, we choose the number of basis function equal to five.

This number of basis function is chosen empirically. The filters are spaced logarithmically with 1/4 octave resolution.

Figures 2 and 3 show the discrimination capability of the two selected MPEG-7 Audio features, namely audio spectrum centroid and audio fundamental frequency, for four different types of environment: car with closed window, car with open window, restaurant, and office room. From Fig. 2 we can see that there is similarity using spectrum centroid between restaurant and office room environments, and between the car environments. However, using fundamental frequency, restaurant environment can be separated from office room environment, which can be seen in Fig. 3. These figures demonstrate that different MPEG-7 Audio features have different types of discrimination capabilities. Therefore, we involve all the MPEG-7 Audio features to recognize environment in our method.

We apply some post-processing on the MPEG-7 Audio features to reduce the dimensionality as well as to remove the correlation between the features. After obtaining MPEG-7 features, we apply logarithmic function, followed by discrete cosine transform (DCT), which decorrelates the features. The decorrelated features are projected onto a lower dimension by using PCA. PCA projects the features onto lower dimension space created by the most significant eigenvectors. All the features are mean and variance normalized.

In our experiments, we use the following sets of feature parameters. The numbers inside the parenthesis after the feature names correspond to the dimension of feature vector.

(i) MFCC (13)
(ii) MFCC (26)
(iii) MPEG-7 audio features after PCA (13)
(iv) MFCC (ii) + MPEG-7 (iii), then PCA (26).

4.2 Classifier

We use HMM as classifier in our approach. First GMMs are created using EM (expectation-maximization) algorithm. GMM approximates probability density functions from feature parameters. For each class (environment), single-state HMM (virtually GMM) with varied number of Gaussian mixtures is used. The number of mixtures is varied between one and eight, and then is fixed to five, which gives the optimal result. For human speech, five-state left-to-right HMMs are used. Each state of the HMMs consists of eight Gaussian mixtures (empirically set) with diagonal covariance.

Environmental classes are modeled using environment sound only (no added artificially human speech). One Speech model is developed using male and female utterances without the environment sound. The speech model was obtained using five male and five female utterances of 20 seconds each. These utterances are different than the utterances added to the environment sound.

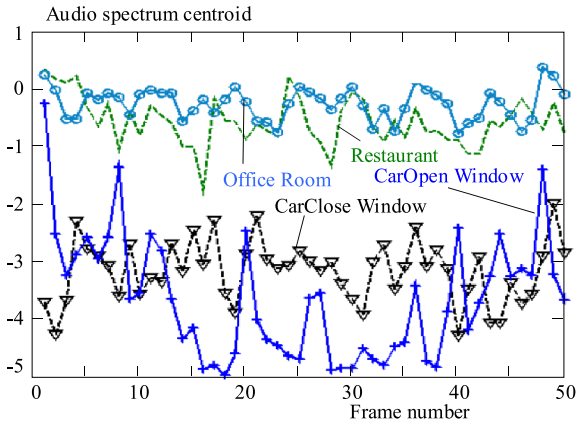

Fig. 2. MPEG-7 Audio feature: Audio Spectrum Centroid of first 50 frames for four different types of environment. Restaurant and office environments have clearly different distributions from car environments.

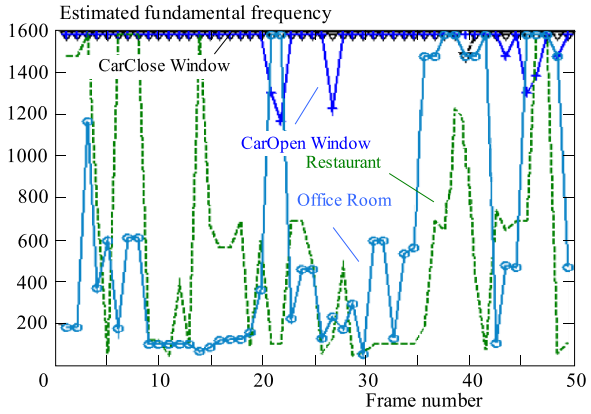

Fig. 3. MPEG-7 Audio feature: Audio Fundamental Frequency of first 50 frames for four different types of environment. Restaurant environment has separable distribution than other environments.

4.3 Experiments

Two categories of experiments were conducted. In the first category, only different types of environment sounds were modeled during training using 100 instances of each type without adding speech. Testing was performed using the rest 100 instances of each type without adding speech, and with 10 second and 20 second added speech, respectively.

In the second category, we modeled each environment sound as well as speech with an extra model (all the utterances corresponded to one speech model).

The MPEG-7 features are calculated using TU-Berlin MPEG-7 audio analyzer [21].

4.4 Results and discussion

Figure 4 shows the recognition accuracies (%) of 10 different environment sounds using four types of feature parameters described in Sect. 4.1. The four bars in each environment class represent accuracies with MFCC (13), MFCC (26), MPEG-7 (13), and MFCC + MPEG-7 (26) features, respectively. From the figure, we can see that the shopping mall environment has the highest accuracy of 92 % using MFCC (13), and it improves to 93 % using MFCC (26). A significant improvement is achieved (96 % accuracy) using MPEG-7 features. However, it improves further to 97 % while using a combined feature set of MFCC and MPEG-7. In case of the park environment, the accuracy is bettered by 11 %, comparing between using MFCC (13) and using combined set. If we look through all the environments, we can easily find out that the accuracy is enhanced with MPEG-7 features, and the best performance is with the combined feature set. This indicates that both the features are complementary to each other, and that MPEG-7 features have upper hand over MFCC for environment recognition.

The least performance is obtained with the desert and the park environments (less than 90 % using combined features). This is because some of the recordings of these environments contain only very low sound without any clear clue. However, some of the recordings contain sound of mild sand storm (desert environment), and sound of gentle breeze (park environment).

Figure 5 gives recognition accuracies of different environment sound in presence of 10 second human foreground speech, and without using speech model. As we can see, the accuracy drops by a large percentage from the case of not adding speech. For example, accuracy falls from 97 % to 84 % using combined feature set for the shopping mall environment. The lowest recognition is again with the park environment.

The accuracy is significantly improved in presence of speech by applying the proposed approach with separate model for each environment and one model of speech, as shown in Fig. 6. Using this separate modeling technique, the performance of the shopping mall environment, for example, jumps to 91 % from 84 % using MFCC + MPEG-7. In fact, accuracies of all the environments are above 80 % using this technique. This result justifies the use of the speech model together with the environmental models for recognition.

Figure 7 demonstrates accuracies of the environment sounds in the presence of 20 second of foreground speech using the proposed approach. If we compare Figs. 4, 6 and 7, we can see huge performance dip with the increasing amount of speech, which is obvious. However, with the proposed approach, none of the environment sounds have recognition rate less than 65%, which is a huge improvement over the reported result in [14]. Confusion matrix of the environments in the presence of 20 second of foreground speech is presented in Fig. 8. The results are with MFCC + MPEG-7 feature set, and with a separate speech model (corresponds to the fourth bar of Fig. 7). The purpose of the confusion matrix is to analyze the error in terms of confusion between the classes (environments). Figure 8 shows only numbers greater than or equal to five for each confused pair. For example, restaurant sound is confused with mall sound in 10 instances out of total 100 restaurant test instances. On the other hand, mall sound is confused with restaurant sound in

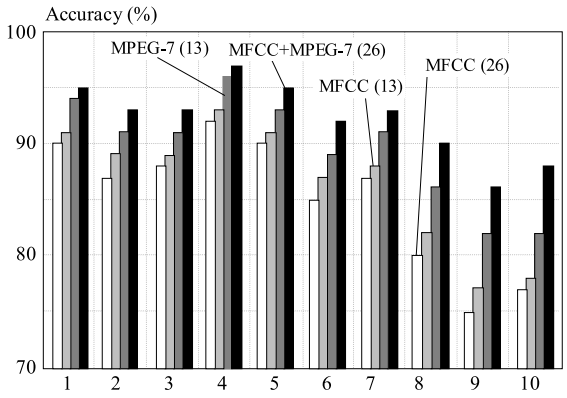

Fig. 4. Recognition accuracy (%) of different environment sounds using different sets of feature parameters. All the audio files contain environment sounds only: 1 – Restaurant, 2 – Crowded Street, 3 – Quiet Street, 4 – Shopping Mall, 5 – Car-Open Window, 6 – Car-Closed Window, 7 – Corridor, 8 – Office Room, 9 – Desert, 10 – Park

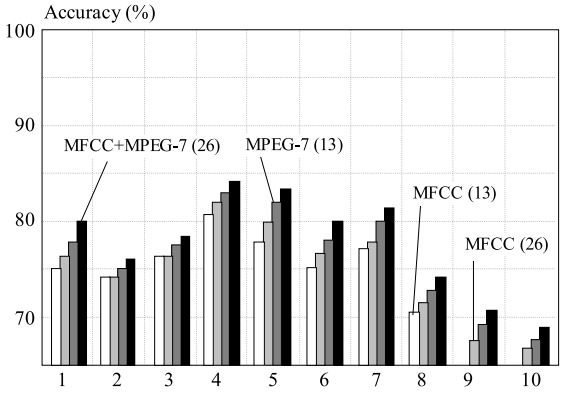

Fig. 5. Recognition accuracy (%) of different environment sounds using different sets of feature parameters. All the audio files contain environment sounds and 10 second of human speech. This result is using environment models only (no added speech): 1 – Restaurant, 2 – Crowded Street, 3 – Quiet Street, 4 – Shopping Mall, 5 – Car-Open Window, 6 – Car-Closed Window, 7 – Corridor, 8 – Office Room, 9 – Desert, 10 – Park

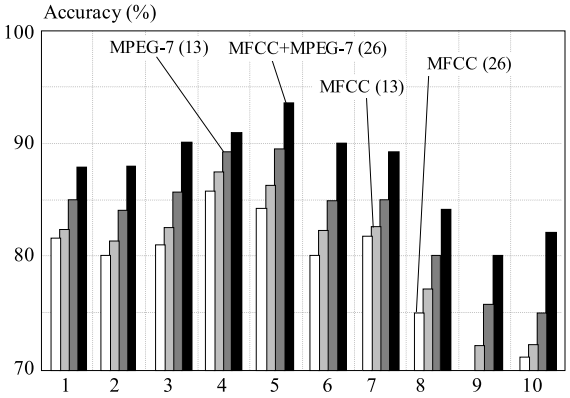

Fig. 6. Recognition accuracy (%) of different environment sounds using different sets of feature parameters. All the audio files contain environment sounds and 10 second of human speech. This result is using environment models and speech model (the proposed approach): 1 – Restaurant, 2 – Crowded Street, 3 – Quiet Street, 4 – Shopping Mall, 5 – Car-Open Window, 6 – Car-Closed Window, 7 – Corridor, 8 – Office Room, 9 – Desert, 10 – Park

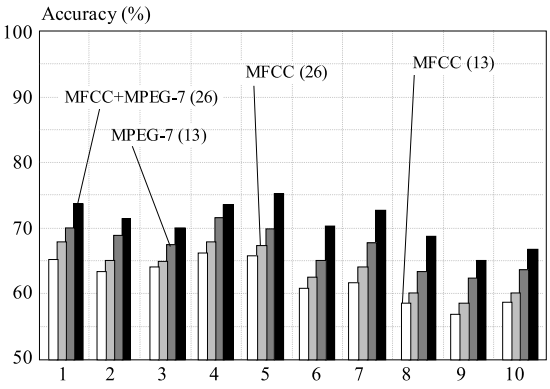

Fig. 7. Recognition accuracy (%) of different environment sounds using different sets of feature parameters. All the audio files contain environment sounds and 20 second of human speech. This result is using environment models and speech model (the proposed approach): 1 – Restaurant, 2 – Crowded Street, 3 – Quiet Street, 4 – Shopping Mall, 5 – Car-Open Window, 6 – Car-Closed Window, 7 – Corridor, 8 – Office Room, 9 – Desert, 10 – Park

Clasified as

	1	2	3	4	5	6	7	8	9	10
1		5		10						
2	11			7	5					
3		9					5	5		
4	7	5								
5		6		5						
6		7						5		
7			6					8		
8			6			5	6		7	
9			8			8				6
10			5				7	5	10	

Fig. 8. Confusion matrix of the environments in presence of 20 second foreground speech. The results are with MFCC + MPEG-7 feature set, and with a separate speech model (corresponds to the fourth bar in Fig. 7.). This figure shows only numbers greater than or equal to five for each confused pair.

seven instances out of total 100 mall test instances. From the matrix, we can find that restaurant sound, crowded street sound, and shopping mall sound are confused between each other. This happens because all these sounds contain some amount of speech (of buyers, sellers, cash persons, pedestrians, *etc*). Corridor, office room, desert, and park sounds are also confused between each other. This is because these sounds are mostly quiet (quiet street is also confused with these sounds).

5 CONCLUSION

We proposed the full use of MPEG-7 Audio features in combination of conventional MFCC features for environment sound recognition for audio forensics. We also used separate modeling for each environment sound and foreground speech. We conducted several experiments with different lengths of foreground speech present in environment sound. The experimental results showed significant improvement in recognition accuracy using combined MPEG-7 Audio features and MFCCs, and speech model. Our future work is to study the effect of different types of other features and classifiers in environment recognition for audio forensics to achieve higher performance.

Acknowledgement

This work is supported by the Research Center, College of Computer and Information Sciences, King Saud University. The authors are grateful for this support.

References

[1] DELP, E.—MEMON, N.—WU, M.: Digital Forensics, IEEE Signal Process. Magazine (March 2009), 14–15.

[2] BROEDERS, A. P. A.: Forensics Speech and Audio Analysis: the State of the Art in 2000 AD, Actas del I Congreso de la Sociedad Espanola de Acustica Forense, March 2000, pp. 13–24.

[3] CAMPBELL, W.—BRADY, K.—CAMPBELL, J.—REYNOLDS, D.—GRANVILLE, R.: Understanding Scores in Forensics Speaker Recognition, ISCA Speaker Recognition Workshop, June 2006, pp. 1–8.

[4] AES, AES43-2000: AES Standard for Forensics Purposes – Criteria for the Authentication of Analog Audio Tape Recordings, Journal of the Audio Engineering Society **48**(3) (June 2000), 204–214.

[5] RABINER, L. R.—JUANG, B. H.: Fundamentals of Speech Recognition, Prentice Hall, Upper-Saddle River, NJ, 1993.

[6] ERONEN, A. J.—PELTONEN, V. T.—TUOMI, J. T.—KLAPURI, A. P.—FAGERLUND, S.—SORSA, T.—LORHO, G.—HUOPANIEMI, J.: Audio-Based Context Recognition, IEEE Trans. Audio, Speech and Language Process. **14**(1) (Jan 2006), 321–329.

[7] ZENG, Z.—LI, X.—MA, X.—JI, Q.: Adaptive Context Recognition based on Audio Signal, Proc. 19th International Conference on Pattern Recognition'08, 2008.

[8] SELINA, C.—NARAYANAN, S.—KUO, J.: Environmental Sound Recognition using MP-Based Features, Proc. IEEE International Conference on Acoustics, Speech and Signal Process. (ICASSP08), 2008, pp. 1–4.

[9] MALLAT, S.—ZHANG, Z.: Matching Pursuits with Time-Frequency Dictionaries, IEEE Trans. Signal Processing **41**(12) (1993), . 3397–3415.

[10] MALKIN, R. G.—WAIBEL, A.: Classifying User Environment for Mobile Applications using Linear Autoencoding of Ambient Audio, Proc. IEEE International Conference on Acoustics, Speech and Signal Process. (ICASSP05), 2005, pp. 509–512.

[11] MA, L.—SMITH, D. J.—MILNER, B. P.: Context Awareness using Environmental Noise Classification, Proc. Eurospeech03, 2003, pp. 2237–2240.

[12] WANG, J. C.—WANG, J. F.—HE, K. W.—HSU, C. S.: Environmental Sound Classification using Hybrid SVM/KNN Classifier and MPEG-7 Audio Low-Level Descriptor, Proc. IEEE International Joint Conference on Neural Networks, 2006, pp. 1731–1735.

[13] NTALAMPIRA, S.—POTAMITIS, N.—FAKOTAKIS, N.: Automatic Recognition of Urban Environmental Sounds Events, Proc. CIP2008, 2008, pp. 110–113.

[14] KRAETZER, C.—OERMANN, A.—DITTMANN, J.—LANG, A.: Digital Audio Forensics: a First Practical Evaluation on Microphone and Environmental Classification, Proc. ACM Multi Media and Security, (MMSec'07), 2007, pp. 63–73.

[15] DUDA, R. O.—HART, P. E.—STORK, D. G.: Pattern Classification, 2nd Ed., Willey, New York, 2001.

[16] SELINA, C.—NARAYANAN, S.—JAY KUO—MATARIC, M. J.: Where am I? Scene Recognition for Mobile Robots using Audio Features, Proc. IEEE International Conference on Multimedia Expo06, 2006, pp. 885–888.

[17] MAHER, R. C.: Audio Enhancement using Nonlinear Time-Frequency Filtering, Proc. Audio Engineering Society 26th Conf., Audio Forensics in the Digital Age, Denver, CO, July 2005, pp. 104-112.

[18] MUSIALIK, C.—HATJE, U.: Frequency-Domain Processors for Efficient Removal of Noise and Unwanted Audio Events, Proc. Audio Engineering Society 26th Conf, Audio Forensics in the Digital Age, Denver, CO, July 2005, pp. 65-77.

[19] CHAMPOD, C.—MEUWLY, D.: The Inference of Identity in Forensics Speaker Recognition, Speech Communication **31** (2000), 193–203.

[20] CAMPBELL, J. P. *et al*: Forensics Speaker Recognition: A Need for Caution, IEEE Signal Process. Magazine (March 2009), 95–103.

[21] TU-Berlin MPEG-7 Audio Analyzer. http://mpeg7lld.nue.tu-berlin.de/.

Received 3 January 2010

Ghulam Muhammad received his Bachelor degree in Computer Science and Engineering in 1997 from Bangladesh University of Engineering and Technology, and M. and PhD degrees in 2003 and 2006, respectively, from Toyohashi University of Technology, Japan. After serving as a JSPS (Japan Society for the Promotion of Science) fellow, he joined as a faculty member in the College of Computer and Information Sciences at King Saud University, Saudi Arabia. His research interests include automatic speech recognition, image processing, and multimedia forensics.

Khalid Alghathbar, PhD, CISSP, CISM, PMP, BS7799 Lead Auditor, is Associate Professor and the Director of the Centre of Excellence in Information Assurance in King Saud University, Riyadh, Saudi Arabia. He is a security advisor for several government agencies. His main research interest is in information security management, policies, biometrics and design. He received his PhD in Information Technology from George Mason University, USA.