Journal of Computer Science 6 (11): 1283-1287, 2010
ISSN 1549-3636

Audio Environment Recognition using Zero Crossing Features and MPEG-7 Descriptors

[1]Saleh Al-Zhrani and [2]Mubarak AlQahtani
[1]Department of Computer Information Systems,
Al Imam Muhammad bin Saud University, Riyadh, Saudi Arabia
[2]College of Computer and Information Sciences,
King Saud University, Riyadh, Saudi Arabia

Abstract: Problem statement: This study investigated zero crossing features and selected MPEG-7 audio descriptors for environment sound recognition applications such as audio forensics. **Approach:** The study implemented several experiments focusing on the problems of environment recognition from audio particularly for forensic applications. **Results:** It was investigated the effect of the temporal zero crossing feature as well as selected MPEG-7 audio low level descriptors on environment sound recognition. The performance was evaluated against a varying number of training sounds and samples per training file. **Conclusion/Recommendations:** Experimental results showed that higher recognition accuracy is achieved by increasing the number of training files and by decreasing the number of samples per training file. This study presented an audio environment recognition using zero crossing features and MPEG-7 Descriptors.

Key words: Audio forensics, zero crossing, MPEG-7

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 (Delp *et al*., 2009). There are several areas of digital forensics: image forensics, audio forensics, video forensics and multimedia.

In this study, we focused on digital audio forensics. Digital audio forensics provides evidence from left-over audio files contained in audio/video media at the crime spot. This type of forensic can be categorized into four different classes according to its nature:

- Speaker identification/verification/recognition to find the answer of who
- Speech recognition/enhancement, to find the answer of what
- Environment detection, to find the answer of where or situation and
- Source authentication, to find the answer of how

A significant amount of research can be found in the area of speech recognition or enhancement (Faghihi and Jangjoo, 2005), speaker recognition (Campbell *et al*., 2006) and authentication of audio (Begault *et al*., 2005). However, little research 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 because, unlike speech or speaker recognition case, die rent environment sounds may have similar characteristics.

The study presents several experiments on environment recognition for digital audio forensics: restaurant, office room, fountain, cafeteria, mall, meeting room and corridor. Temporal Zero Crossing (ZC) feature and some selected MPEG-7 audio low level descriptors are used as features. The MPEG-7 descriptors we use are Audio Waveform (AWF), Audio Power (AP), Audio Spectrum Envelop (ASE), Audio Spectrum Centroid (ASC) and Audio Spectrum Spread (ASS). This selection is based on our ongoing research using the Fisher ratio. Two types of experiments on environment recognition are performed by varying (a) the number of training files and (b) the number of samples per training file.

The study is organized as follows. The next part gives a review of related past works; followed by a

Corresponding Author: Saleh Al-Zhrani, Department of Computer Information Systems, Al Imam Muhammad bin Saud University, Riyadh, Saudi Arabia

description of feature extraction with ZC and MPEG-7 audio descriptors, the classifier and the data used in the experiments, followed by the proposed approach to recognize environment sound. In this section, the experimental results and discussion are also given. Finally, conclusions and future direction are presented.

Literature review and current related work: Most of the previous works in environment detection used Mel Frequency Cepstral Coefficients (MFCC) as features, which are applied not only in environment detection but also in speech and speaker recognition applications (Zulkarnain and Nor, 2010) and Hidden Markov Model (HMM) based classification. While HMMs are perhaps the most widely used in different applications, the k-Nearest Neighbor classifier (k-NN) is also applied due to its simplicity (Duda *et al*., 2000). As noted previously, there is not much work done in the particular area targeting forensic applications, but we can mention some related works that impact on this area.

A comprehensive evaluation of a computer and human performance in audio-based context (environment) recognition is presented in (Eronen *et al*., 2006). In their study, Eronen *et al*. (2006) used several time-domain and spectral-domain features in addition to MFCC. Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Linear Discriminated Analysis (LDA) were used to reduce dimensionality of the feature vector. Two types of classifiers were applied separately: k-NN (k = 1) and HMM with number of states and number of mixtures within each state varying from 1-4 (and 5), respectively. Nature and outdoors were recognized with highest accuracy (96-97%), while the library, a quiet place, had the lowest accuracy (35%).

The researcher Chu *et al*. (2008) introduced the Matching Pursuit (MP) technique (Mallat and Zhang, 1993) in environmental sounds recognition. MP provides a way to extract features that can describe sounds where other audio features such as 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 in the background environment. They applied k-NN (k = 1) and GMM with 5 mixtures (Chu *et al*., 2006; 2008). In (Chu *et al*., 2006), they also used Support Vector Machine (SVM) methods with 2° polynomial as classifier and reduced the dimension by applying forward feature selection and backward feature selection procedures.

Sixty-four dimensional MFCC, plus the spectral centroid were used as features in (Malkin and Waibel, 2005). They used forensic-application-like audio files, where both ambient, i.e., 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. They introduced linear auto encoding neural networks for classifying the environment. A hybrid autoencoder and GMM was used in their experiments and 80.05% average accuracy was obtained.

Wang *et al*. (2006) used three MPEG-7 audio low level descriptors as features in their study on environmental sound classification. They proposed a hybrid SVM and k-NN classifier in their study. For SVM, they used three different types of kernel functions: linear kernel, polynomial kernel and radial basis kernel. The system with 3 MPEG-7 features achieved 85.1% accuracy averaged over 12 classes. Ntalampiras *et al*. (2008) used MFCC along with MPEG-7 features to classify urban environments. They exploited a full use of MPEG-7 low level descriptors, namely, audio waveform, audio power, audio spectrum centroid, audio spectrum spread, audio spectrum flatness, harmonic ration, upper limit of harmonicity and audio fundamental frequency.

To detect the used microphone and the background environments of audio recordings, the researcher Kraetzer *et al*. (2007) 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 Naive Bayes classifiers to evaluate microphone and environmental classification. Their study reported that the highest 41.54% accuracy was obtained by Naïve Bayes classifiers with 10 fold cross validation, while 26.49% was the highest accuracy achieved by simple k-means clustering. They did not use HMM or GMM for classification.

MATERIALS AND METHODS

Feature extraction:

Zero crossing: Zero-crossing is a commonly used term in electronics, mathematics and image processing. In mathematical terms, a “zero-crossing” is a point where the sign of a function changes (e.g., from positive to negative), represented by a crossing of the axis (zero value) in the graph of the function. Zero crossing features are good for extracting sound from environment if we increase number of training files and decrease the number of sample (Johnston and Gulrajani, 2002). Mean

value is subtracted from each signal. Frame length is 512 samples with overlapping 256 samples.

Selected MPEG-7 audio descriptor: MPEG-7 Audio describes audio content using low-level characteristics, structure, models. The objective of MPEG-7 Audio is to provide fast and efficient searching, indexing, retrieval of information from audio les. The characteristics can be divided into scalar and vector types. Scalar types returns scalar values such as power or fundamental frequency, while vector types returns, for example, spectrum flatness calculated for each band in a frame. In the following we briefly describe each characteristic, or descriptor, used. Though (Ntalampiras *et al*., 2008) utilized a partial MPEG-7 feature with seven dimensions, we exploit the full advantage of MPEG-7 features in this study. MPEG-7 Audio low-level descriptors:

- Audio Waveform (AWF): It describes the shape of the signal by calculating the maximum and the minimum of samples in each frame. The maximum and minimum of the waveform are denoted by AWF_max and AWF_min, respectively
- Audio Power (AP): It gives temporally smoothed instantaneous power of the signal.
- Audio Spectrum Envelop (ASE): It describes short time power spectrum for each band within a frame of a signal
- Audio Spectrum Centroid (ASC): It returns the center of gravity (centroid) of the log-frequency power spectrum of a signal. It points out the dominant high or low frequency components in the signal
- Audio Spectrum Spread (ASS): 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 to differentiate between noise-like or tonal sound and speech

Classifier: We used the k-Nearest Neighbor algorithm (k-NN) as classifier. k is the most important parameter in a text categorization system based on k-NN. In the classification process, the k documents nearest to the test document in the training set are determined first. Then, the predication can be made according to the category distribution among this k nearest neighbors. k-NN is one of the most popular algorithms for text categorization (Manning and Schutze, 1999). Many researchers have found that the k-NN algorithm achieves very good performance in their experiments on different data sets (Yang and Liu, 1999; Hirzallah, 2007; Baoli *et al*., 2002). The nearest neighbors are defined in terms of Euclidean distance. The Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler and is given by the Pythagorean formula:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (p_1 - q_1)^2}$$

Data: We recorded audio signals from seven different scenarios: restaurant, office room fountain cafeteria, mall, meeting room and corridor. The duration for each environment is half hour (30 min). Each environment file is separated into many files with fixed number of samples d, for example restaurant environment file 1 from sample 1 to sample d, file 2 from sample d+1 to double 2d, similarly file 10 from (9*d+1) to (10*d). Sounds were recorded with an IC recorder (ICD-UX71F/UX81F/UX91F). Sampling rate was set to 22.05 kHz and quantization was 16 bit.

RESULTS AND DISCUSSION

The feature extraction and classification used in the experiments are:

- ZC and selected MPEG-7 as features
- k-NN as classifier

Two types of experiments are performed, one with decreasing number of samples per file and the other one with increasing number of file in training. First, we decrease the number of samples with fixed number of training files to six. Second, the same consideration with the number of training files is fifteen.

Six file training: In this case the first six files of each environment are used for training and the last five files for testing. The experiment was run with different number of samples: 1,000,000 and 500,000 for each file. The objective of this experiment is to see the affect of decreasing the number of samples. The results are presented in Fig. 1.

For the ZC feature, the average accuracy for all environments is 20% when the numbers of sample is 1,000,000. When we decrease numbers of sample to 500,000 the average accuracy for all environments is enhanced to 40%. The average accuracy is increased for all features except AP and AWF_min. The ASE feature has the highest average accuracy, followed by ZC.

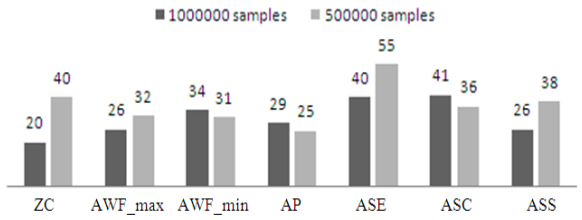

Fig. 1: Recognition accuracy (%) with six file training

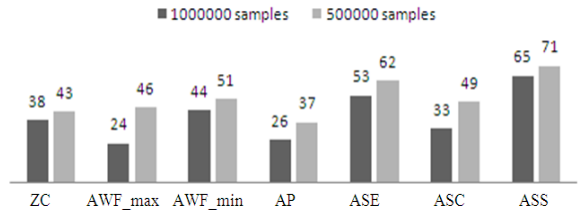

Fig. 2: Recognition accuracy (%) with fifteen file training

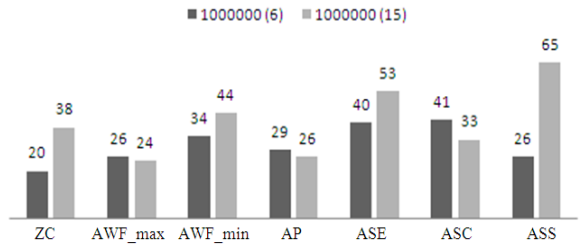

Fig. 3: Comparison between different numbers of training files with fixed sample (1,000,000)

Fifteen file training: The number of training files is increased from six to fifteen files and the same experiment as described previously is repeated. The accuracy is enhanced when the number of file training is increased. The highest accuracy is achieved when the number of sample is 1,000,000. The results are given in Fig. 2.

All features gave an increased overall average accuracy when we decreased the number of samples and increased the number of training files. Figure 3 and 4 give recognition accuracies (%) for fixed samples by varying the number of training files. From the previous diagrams, we can find that by increasing the number of training files, recognition accuracies are increased with all the feature types. However, for the number of samples, the reverse is true. If we decrease the number of samples, the accuracies increase.

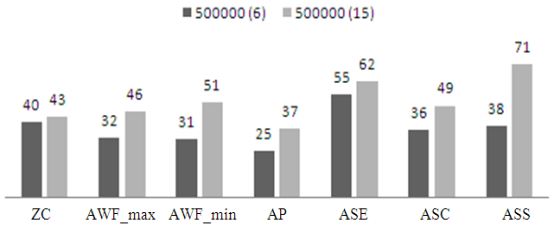

Fig. 4: Comparison between different numbers of training files with fixed sample (500,000)

CONCLUSION

In this study we investigated zero crossing features and selected MPEG-7 audio descriptors for environment sound recognition applications such as audio forensics. The experimental results showed significant improvement in accuracy using MPEG-7 Audio features and zero crossing when we increase the number of training files and decrease the number of samples. The future study is needed to study the effect of other types of features and classifier in environment recognition to achieve yet higher performance. This study provides an attempt to fill the knowledge gap in audio environment recognition using zero crossing features and MPEG-7 descriptors.

REFERENCES

Baoli, L., C. Yuzhong and Y. Shiwen, 2002. Comparative study on automatic categorization methods for Chinese search engines. Proceedings of the 8th Joint International Computer Conference, Nov. 12-14, Zhejiang University Press, Hangzhou, pp: 117-120.

Begault, D.R., B.M. Brustad and A.M. Stanley, 2005. Tape analysis and authentication using multi-track recorders. AES 26th International Conference, July 7-9, AUDIOFORENSICS, Denver, Colorado, USA., pp: 1-7. http://www.audioforensics.com/PDFs/AuthenticationPaper.pdf

Campbell, W.M., K.J. Brady and J.P. Campbell, R. Granville and D.A. Reynolds, 2006. Understanding scores in forensic speaker recognition. Proceeding of the IEEE Odyssey: The Speaker and Language Recognition Workshop, June 28-30, IEEE Xplore Press, USA., pp: 1-8. DOI: 10.1109/ODYSSEY.2006.248091

Chu, S., S. Narayanan and C.C. Jay Kuo, 2008. Environmental sound recognition using MP-based features. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 31-Apr. 4, IEEE Xplore Press, Las Vegas, NV., pp: 1-4. 10.1109/ICASSP.2008.4517531

Chu, S., S. Narayanan, C.C.J. Kuo and M.J. Mataric, 2006. Where am I? Scene recognition for mobile robots using audio features. Proceedings of IEEE International Conference on Multimedia Expo 06, July 9-12, IEEE Xplore Press, Toronto, Ont., pp: 885-888. DOI: 10.1109/ICME.2006.262661

Delp, E., N. Delp, Memon and M. Wu, 2009. Special issue on forensic analysis of digital evidence. IEEE Sign. Process. Mag., 26: 26-28.

Duda, R.O., P.E. Hart and D.G. Stork, 2000. Pattern Classification. 2nd Edn., Willey- Interscience, New York, ISBN: 10: 0471056693, pp: 654.

Eronen, A.J., V.T. Peltonen, J.T. Tuomi, A.P. Klapuri and S. Fagerlund *et al*., 2006. Audio-based context recognition. IEEE Trans. Audio, Speech Lang. Process., 14: 321-329. DOI: 10.1109/TSA.2005.854103

Faghihi, F. and A. Jangjoo, 2005. Intensity position modulation for free-space laser communication system. Am. J. Applied Sci., 2: 1178-1181. http://scipub.org/fulltext/ajas/ajas271178-1181.pdf

Johnston, P.R. and R.M. Gulrajani, 2002. An Analysis of the zero-crossing method for choosing regularization parameters. SIAM J. Sci. Comput., 24: 428-442. DOI: 10.1137/S1064827500373516

Hirzallah, N., 2007. An authoring tool for As-in-class E-lectures in E-learning systems. Am. J. Applied Sci., 4: 686-692. http://www.scipub.org/fulltext/ajas/ajas49686-692.pdf

Kraetzer, C., A. Oermann, J. Dittmann and A. Lang, 2007. Digital audio forensics: A first practical evaluation on microphone and environmental classification. Proceedings of ACM Multi Media and Security, Sept. 20-21, ACM Press, Dallas, Texas, USA., pp: 63-74. DOI: 10.1145/1288869.1288879

Mallat, S.G. and Z. Zhang, 1993. Matching pursuits with time-frequency dictionaries. IEEE Trans. Sign. Process., 41: 3397-3415. DOI: 10.1109/78.258082

Malkin, R.G. and A. Waibel, 2005. Classifying user environment for mobile applications using linear autoencoding of ambient audio. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 18-23, IEEE Xplore Press, USA., pp: 509-512. DOI: 10.1109/ICASSP.2005.1416352

Manning, C.D. and H. Schutze, 1999. Foundations of Statistical Natural Language Processing. 1st Edn., The MIT Press, Cambridge, ISBN: 10: 0262133601, pp: 620.

Ntalampiras, S., I. Potamitis and N. Fakotakis, 2008. Automatic recognition of urban environmental sounds events. International Association for Pattern Recognition Workshop on Cognitive Information Processing, June 2008, EURASIP, USA., pp: 110-113.

Wang, J.C., J.F. Wang, K.W. He and C.S. Hsu, 2006. Environmental sound classification using hybrid SVM/KNN Classifier and MPEG-7 audio low-level descriptor. Proceedings of IEEE International Joint Conference on Neural Networks, June 26-29, IEEE Xplore Press, Vancouver, BC., pp: 1731-1735. DOI: 10.1109/IJCNN.2006.246644

Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 15-19, ACM Press, Berkeley, California, United States, pp: 42-49. DOI: 10.1145/312624.312647

Zulkarnain, R. and M.J.M. Nor, 2010. Noise control using coconut coir fiber sound absorber with porous layer backing and perforated panel. Am. J. Applied Sci., 7: 260-264. http://www.scipub.org/fulltext/ajas/ajas72260-264.pdf