Multimed Tools Appl (2018) 77:1009–1040
DOI 10.1007/s11042-016-4277-2

Digital multimedia audio forensics: past, present and future

Mohammed Zakariah[1] · **Muhammad Khurram Khan**[2] · **Hafiz Malik**[3]

Received: 12 July 2016 / Revised: 8 December 2016 / Accepted: 15 December 2016 /
Published online: 9 January 2017

Abstract Digital audio forensics is used for a variety of applications ranging from authenticating audio files to link an audio recording to the acquisition device (e.g., microphone), and also linking to the acoustic environment in which the audio recording was made, and identifying traces of coding or transcoding. This survey paper provides an overview of the current state-of-the-art (SOA) in digital audio forensics and highlights some open research problems and future challenges in this active area of research. The paper categorizes the audio file analysis into container and content-based analysis in order to detect the authenticity of the file. Existing SOA, in audio forensics, is discussed based on both container and content-based analysis. The importance of this research topic has encouraged many researchers to contribute in this area; yet, further scopes are available to help researchers and readers expand the body of knowledge. The ultimate goal of this paper is to introduce all information on audio forensics and encourage researchers to solve the unanswered questions. Our survey paper would contribute to this critical research area, which has addressed many serious cases in the past, and help solve many more cases in the future by using advanced techniques with more accurate results.

✉ Muhammad Khurram Khan
mkhurram@ksu.edu.sa

Mohammed Zakariah
mzakariah@ksu.edu.sa

Hafiz Malik
hafiz@umich.edu

[1] College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

[2] Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh, Kingdom of Saudi Arabia

[3] Department of Electrical and Computer Engineering, University of Michigan-Dearborn, Dearborn, MI, USA

Keywords Digital forensics · Audio authentication · Speech intelligibility · Environment detection · Microphone identification · Transcoding detection

1 Introduction

Digital audio forensic analysis consists of the acquisition, analysis and evaluation of audio recordings admissible to a court of law as evidence or for forensic investigations. Digital multimedia forensic analysis is commonly used to determine the authenticity and verify the integrity of the evidence submitted to court involving civil or criminal proceedings. The main objective of the audio forensic analysis process is to achieve one or more of the following tasks:

1. Integrity verification aims to answer the question of "whether the query audio (the audio in question) has been tampered with since its creation or not?"
2. Forensic audio enhancement aims to improve speech intelligibility and the audibility of low-level voice, and
3. Speaker identification aims to identify the talker in the query audio, and

Existing forensics analysis techniques are used for a variety of tasks raging from improving speech intelligibility to extraction and recognition of the background sources/speaker, speech recognition, speaker recognition can improve reliability, and so on. It is therefore important to highlight that *some modifications* (or *processing of the evidence*) are admissible in the court of law. For example, speech enhancement techniques should prove that the original audio files have not been changed even though they have been enhanced by applying some techniques to increase the intelligibility and audibility of the test audio recording. To ensure the admissibility of the findings of the forensic audio enhancement process, the forensic analyst or expert witness must demonstrate that the enhanced copy of the query audio has identical content to the original.

Authenticity and integrity verification for a digital audio recording – audio forensics analysis hereafter – is a complex forensic science. The objective of the audio forensic analysis is to establish as far as possible that the recording is a true 'acoustic representation' of events made at through a specific acquisition system and at a specific location. It has been demonstrated that acoustic environment, acquisition system, and encoding process leave artifacts in the resulting audio. These artifacts are also used for authenticity and integrity verification of the digital audio. For example, Malik et al. in [71] have proposed audio forgery detection techniques that rely on acoustic environment signature. Variations in the acoustic environment signature, estimated from test audio recording, are used for forgery detection. Acquisition system leaves its signature artifacts in the resulting audio, which can be used for authenticity verification of the evidence. As these differ from those of another source, calculating the difference in the noise signal can detect the authenticity of the audio file. Efforts have been made to use ENF-based signature - resulted from coupling the electrical power line frequency to the digital recording system [14] to authenticate the digital recording. The ENF-based approaches may not always be applicable if well-designed audio equipment (for example, condensers or piezoelectric microphones) or battery-operated devices are used to capture the recording. Post-processing techniques such as denoising, compression, filtering, transcoding, etc., are commonly used to suppress traces

of tampering. Efforts have been made to detect post-processing such as single and double compression [4] .

This survey provides an overview of the existing state-of-the-art in the area of audio forensics. It also provides a new classification of existing audio forensics methods.

The main contribution of this survey paper includes:

a) Classification of the digital audio forensic analysis methods
b) Overview of forensic audio enhancement based methods
c) Detailed overview of existing acoustic environment identification based methods
d) Discussion on current state of the microphone identification methods
e) Illustration of transcoding and codec identification methods
f) Discussion on open research issues and future research directions

The remainder of this paper is organized as follows: Section 2 provides a brief discussion on background and history of audio forensics. Section 3 focuses on digital audio authentication based on container- and content-based analysis, which includes compression, time frequency, ENF, enhancement and environment. In section 4 we discuss various techniques for the source detection of audio files. Section 5 provides a detailed discussion on transcoding identification from audio recording, Section 6 discusses existing codec identification methods, and Section 7 overviews existing double-compression detection-based methods. Section 8 provides discussion on open challenges and future directions and conclusion of our findings are discussed in Section 9.

2 History and examples of audio forensic investigations

The initial examination of audio files for forensic detection took place in the 1950s after the invention of live recording systems outside the recording studio. In the early 1960s, the Federal Bureau of Investigation in the United States started developing experts in audio forensics to improve the speech intelligibility, enhancement and authentication of recorded files [26].

2.1 Audio forensics and the law

A case that directly dealt with recorded conversations in the United States was in 1958, namely the ruling in *United States v. McKeever* (169 F. Supp. 426, 430, S.D.N.Y. 1958). For the first time, the judge in the McKeever case was asked to determine the legal admissibility of the conversation recorded that involved the defendant. The judge allowed the written transcript to be presented in court [39]. However, for the recording to be accepted in court, six specific requirements needed to be fulfilled. The following are the specific requirements of audio authenticity in the McKeever case: Audio files, Device operator, authentic recording, *no alteration* to the recording, preservation of the record, and identifying the speaker.

2.2 Methodology for interpreting authenticity

Several types of observations are required for determining the authenticity of audio evidence. The examiner needs to perform visual, physical, electrical and acoustic tests that include: Document history should be carefully reviewed; Recording capability should be checked;

Type of recording and its format should be checked; Medium of recording should not be changed; Entire audio has to be carefully listened; and Continuous recording should be done without interruptions [26, 29, 51].

2.3 Magnetic signature and waveform observations

The magnetic development technique is used to examine magnetic signals. If the evidence is a physical audio tape, then it is compared with the reference signature of the recording carried out with the same recording device. The audio spectrogram is an excellent example of a device that can detect signal irregularities, as shown in Fig. 1.

The insertion of a word at an appropriate place to change the meaning of the sentence can be detected using spectral analysis of the query audio recording, e.g., spectrogram analysis. Shown in Fig. 2 is the plot of spectrogram of tampered audio. Abrupt change in the background noise can be observed from Fig. 2. Other alterations include making a duplicate copy and inserting a specific segment of the speech signal. These methods work fine as long as inserted segment is recorded in a different acoustic environment, but these methods are unable to detect insertions if the inserted segment is made in the same acoustic environments where target recording is made.

3 Classification of audio authentication methods

Digital audio integrity verification methods can be broadly divided into (i) container-based authentication and (ii) content-based authentication. The file structure and metadata of the audio file and its description come under container-based analysis, while the actual bits and bytes of the audio file are related to content-based analysis. Because these actually make up the file, the acoustic events can be further reproduced for future use. Renaming the file may not actually affect the quality of the file contents but it may damage the media support or wrapper. This would raise doubts about its authenticity and may make some types of analyses

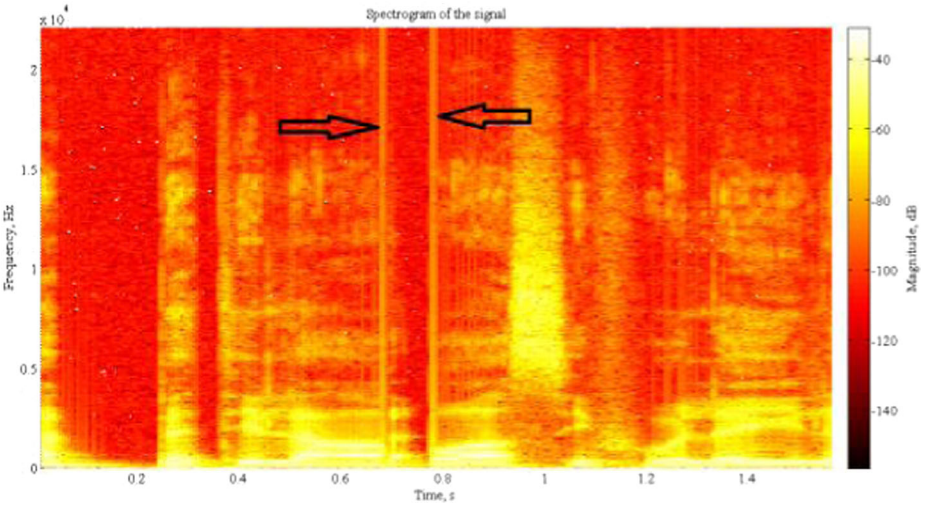


Fig. 1 Spectrogram representation [40]

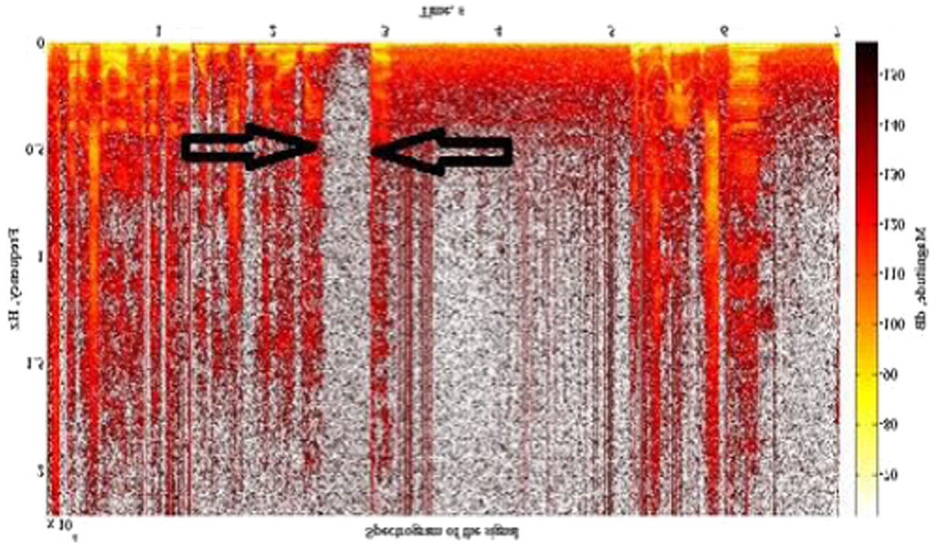


Fig. 2 Spectrogram indicating an abrupt change in the background noise [40]

inconclusive. Furthermore, content analysis-based methods are further divided into two main branches: global and local analyses.

3.1 Container analysis

To gain familiarity with the audio file, experts should perform some container analysis. Container analysis consists of HASH calculations, MAC, and file format analysis as shown in Fig. 3.

- ***Hash-based analysis*:** When the file is received for examination in the lab, first it has to be hashed. To ensure the file has not been tampered, or check if the history of the hash is not changed, certain measures should be taken. A unique character string is derived from the bits and bytes of the audio file and calculated by a mathematically derived hash function. These can be useful to verify that no modifications have occurred to a file from the moment of its HASH calculation is done to the next instance of HASH calculation.

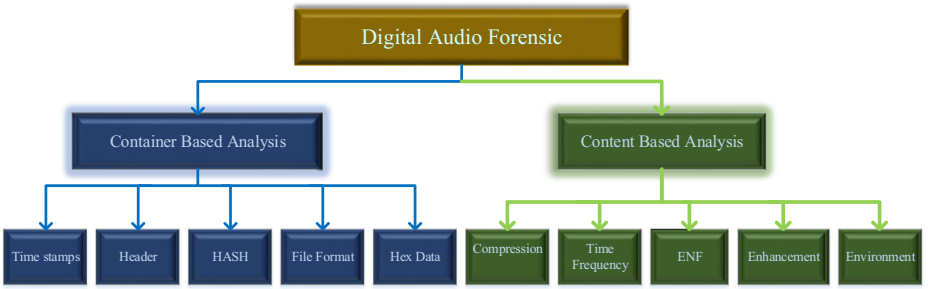


Fig. 3 Scope of the paper

- ***MAC time stamps:*** The date and time of creating the file and its modifications as well as its last access time can be detected by using MAC time stamps. The interlock of the digital system is used to generate the original MAC time stamps but this can be altered by using a copy/transfer operation to another media or through editing operations.
- ***File format:*** The detailed description of the file should be documented for future reference, and a procedural review should be carried out with relevant information for future analysis (file format, codec, sample rate, bit depth etc.). Although this is a simple task, but care must be taken. For example, while the file extension may indicate WAV, several compressions format store audio in WAV files such as Microsoft ADPCM, DVI/IMA ADPCM and A/μ-Law.
- ***Header:*** The examiner can detect a change in the file from the original to the extended version with the help of a hexadecimal reader and the header information of the file format. The file format should match the file name extension (such as RIFF or hex 52 49 46 46 indicating WAV, hex 49 44 33 indicating MP3, hex 30 26 B2 indicating WMA, etc.). Depending on the device and brand, there may be information about the model, serial number, firmware version, time, date and length of the recording (as determined by the internal clock settings). It is useful to note the time stamps and compare them to the date and time claimed by the recordists as to when the file was made [28].
- ***Hex data:*** The raw digital data of the file may contain useful information that can be examined in a hexadecimal reader with an ASCII character viewer. Block addresses of audio information, titles of external software (if present), post-processing operations and other useful information may be displayed [27].

3.2 Content analysis-based audio authentication

Content analysis is the core for the digital audio forensic analysis process and it relies on the actual content of the audio recording to detect traces of tampering, post-processing and anti-forensic processing operation. The majority of existing audio forensic methods use the actual content of test recordings for authentication and integrity verification. The existing state-of-the-art on content-based audio forensics can be broadly classified into the following categories:

3.2.1 The ENF

The Electronic Network Frequency (ENF) is one of the most reliable and robust audio forensic analysis methods, especially for the recordings which are made using the devices powered using mains. For forensic analysis, the ENF method relies on the traces of the ENF present in the recording. For integrity verification or forgery localization, the ENF signature estimated from the input recording is compared with the reference frequency database provided by the power supply company. Figures 4, 5, and 6 show the general block diagram of the ENF extraction process.

In [57], the authors presented a technique to detect audio editing that gave favorable results. The idea was to find abrupt changes in the power grid signal that gives accurate visual characteristics. The editing point and type of editing are determined by visual aid. Automatic discrimination between the original and edited audio file is determined by the use of decision features. The Discrete Fourier transform (DFT) method improves the accuracy of the phase and resolution of the visual characteristics. The audio authentication is detected based on the

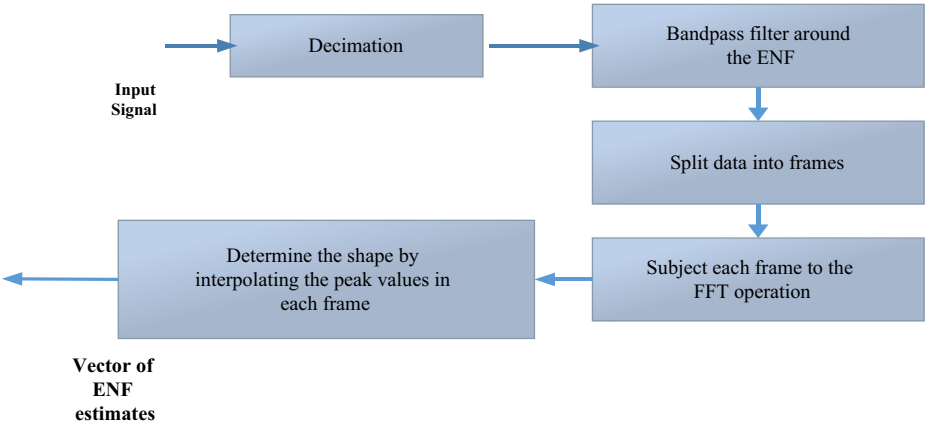


Fig. 4 Block diagram of the ENF extraction process from an input audio recording [18]

max offset for cross correlation (MOCC) between the reference and the extracted signal. ENF signals are extracted from query audio signals and these signals are partitioned into blocks for forgery detection. Both the extracted ENF and the reference ENF signal max offset for cross correlation (MOCC) are calculated block by block. Before calculating the MOCC, an enhancement scheme is introduced to improve the quality of the ENF signal. Both the edited region and the type of editing are detected by taking this approach [38] into consideration.

In [14], the authors have proposed a method for the design and implementation of ENF analysis for audio forensic detection. The conditioning module of the signal was proposed

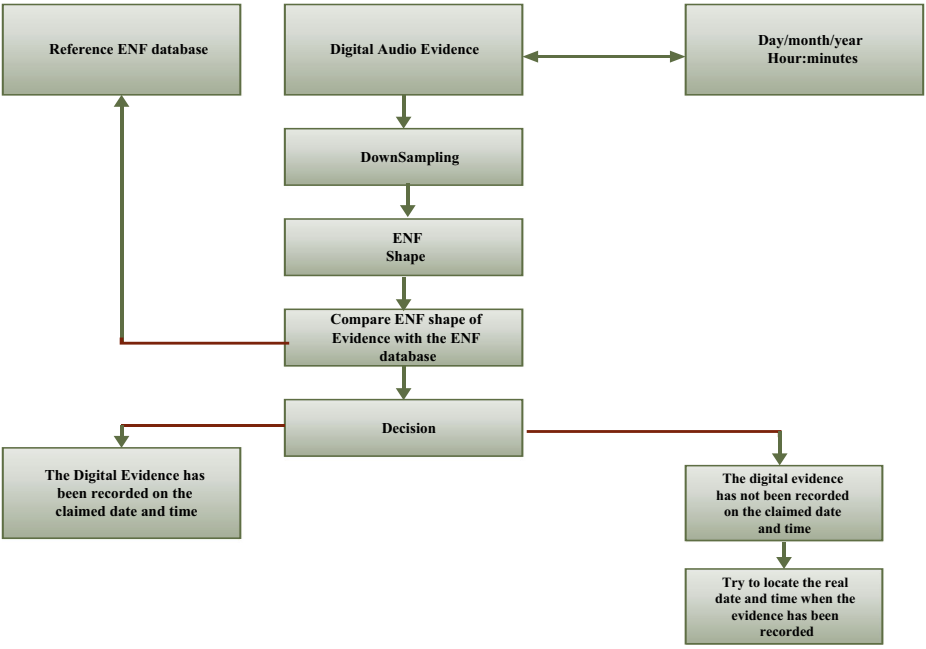


Fig. 5 Automatic system for identifying the date and time when a digital recording was created [15]

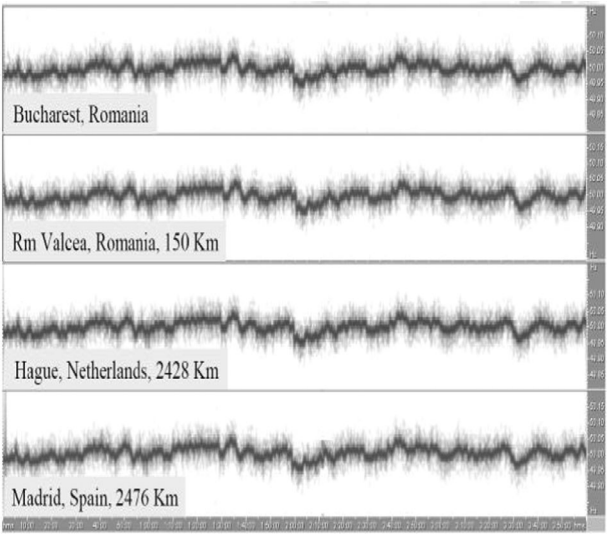


Fig. 6 Comparison of the ENF variation in different cities across Europe, according to [15]

as a prototype. The ENF tracks from the signals are extracted by measuring and analyzing the signal using digital signal processing. To obtain the best possible ENF, emphasis was placed on the frequency precision of the used STFT analysis for a given time instance. Finally, the ENF track logs were popped from the system, which is used in the forensic ENF analysis.

In [15], the authors have discussed the importance of the ENF as a means to detect the integrity and authenticity of digital evidence for forensic analysis. Synchronized recordings made in different locations are compared with the same network to check and establish the stability of the ENF over graphical distances. The experiments were carried out on real cases where the ENF criterion is used to investigate the audio and video files created with secrete surveillance systems. The result for the ENF criteria is to detect the time and place where the editing took place and to cross check whether it was made at the time claimed.

Finally, in [16], if the shape is consistent with A/C mains power, the examiner can prepare the file and begin checking against ENF databases, beginning with the date/time/grid most likely to contain a match.

Three methods to extract the ENF are reported:

1. *Computing time/frequency domain spectrograms* and then visually comparing the questioned samples with the database ENF.
2. *Computing the Fast Fourier Transform(FFT) for short time windows which are in the frequency domain. Here, the maximum magnitude values around 50 Hz are extracted* and the questioned samples are compared with the database ENF.
3. *Time domain analysis* consists of zero-crosses measurement and the questioned samples are compared with the database ENF.

In [6], if the questioned recording and databases have a highly correlated ENF signal, then there are momentary amplitude spikes present in one file and not in the other, these are likely

because of the result of local voltage fluctuations or power surges rather than tampering (Table 1).

3.2.2 Authentication using acoustic environment signature

A typical audio recording consists of a number of acoustic signals including direct source signal, indirect or reflected signals, secondary sources, and ambient noises. These indirect or reflected signals (also known as reservations), secondary sources, and ambient noises are used to characterize an acoustic environment where the recording is made. Acoustic reverberations are caused by the shape and composition of the room, which results in the temporal and spectral smearing of the recorded sound. Secondary audio source activities cause background noise. The challenging task is to extract the acoustic cues from the audio recording. Dynamic acoustic environment identification (AEI) can be calculated depending on the estimated reverberation and background noise. Audio recording authentication and real time crime localization are AEI applications. Literature on modeling and estimating ratio of reverberation and blind reverberation could also be found in [56, 60].

In [8, 23, 42] authors have proposed model-driven approaches to measure the acoustic reverberation parameter using a maximum likelihood framework for automatic acoustic environment parameter estimation and then used them for AEI. In [32], the authors proposed a technique to classify and identify the environment within several known categories of recording environments. Audio signal characteristics are extracted using mel-frequency cepstral coefficients (MFCCs) and time-based features are used to classify the different environments. To classify the room in which these recordings are performed, clustering algorithms are applied to these extracted features. The highest accuracy to classify the recording environment in 10 possible rooms is 41.6% but the performance needs much improvement. The recording environment along with the microphone type can play a major role in successful classification. The results are good for rooms with reverberation and noisy environments with classification accuracy of 75.9%.

In [48], thirteen features were extracted by applying principle component analysis on the top ranked 30 MPEG-7 descriptors. To complete the feature sets for the proposed method, the previously 13 extracted features were appended with MFCC features. These features are classified by using Gaussian mixture models (GMMs). Ten different environment sounds

Table 1 Summary of techniques based on the ENF

Ref. No.	Method	Objective	Features	Results
[57]	Discrete Fourier Transform (DFT) method	Find the abrupt changes in the power grid signal	Decision feature	94% accuracy
[38]	Block by block calculation of the extracted ENF and reference ENF	Audio authenticity	max offset for cross correlation (MOCC)	Both the edited region and the type of editing are detected
[14]	Design and implementation of ENF analysis	Audio forensic detection	-	ENF track logs are determined
[15]	Short time windows measurement	Detect the time and place where editing took place	Time domain	-

were used to evaluate the proposed method. The results obtained with the proposed method clearly show a significant improvement in the performance of recognition compared with MFCCs or full MPEG-7 descriptor-based methods. For example, restaurant environment has achieved the maximum accuracy in the result with MFCC and full MPEG-7 while the proposed method gave 90%, 94% and 96% accuracy respectively when used with MPEG-7 based features in conjunction with MFCC's. In [43], inverse filtering is used to estimate the reverberation component from the audio recording. Altogether, 48 dimensional feature vectors were used to capture the traces of reverberation combining both MFCCs and logarithmic mel-spectral coefficients. To classify the features, a multi-class support vector machine (SVM) classifier was used for AEI. The recording environment is accurately identified by the proposed method for both regular and AEI. The performance of the proposed scheme is evaluated by using a dataset consisting of 284 speech recordings as shown in Fig. 7.

The average classification accuracies with (and without) de-reverberation-based identification systems for microphone M1, M2, M3 and M4 are 94%(84%), 92%(86%), 93%(86%) and 92%(86%), respectively. These results indicate that de-reverberation does improve classification accuracy. In the proposed method of [23], authors removed the speech leakage noise signals, which were not detected by traditional methods. These speech leakages signals, although having a low signal-to-noise ratio (SNR), still influence the environment detection and cannot be used for audio forensics. This system is a two-step approach. In the first stage, speech signal is processed and initial noise estimation was calculated by using a spectral subtraction-based method. Second, multiband-based spectral subtraction was used to remove the speech leakage from the initial noise estimates. To check the accuracy of this method, five different environments were used: (i) office, (ii) small office, (iii) room, (iv) stairs and (v) outside. The same speech signal was recorded by using the same device. The results show that this method is better than existing speech enhancement algorithms as shown in Figs. 8 and 9. In [71], authors have presented a method to estimate the amount of reverberation by spectral subtraction and the background noise based on nonlinear filtering and particle filtering techniques. The experiment was carried out with a dataset of two human speakers made in eight acoustic environments with four commercial-grade microphones. The effectiveness of the proposed method was checked in various experimental settings such as microphone independent, semi- and full blind AEI and robustness to MP3 compressions with Temporal Derivative-based Spectrum and Mel-Cepstrum (TDSM)-based features. More than 2240

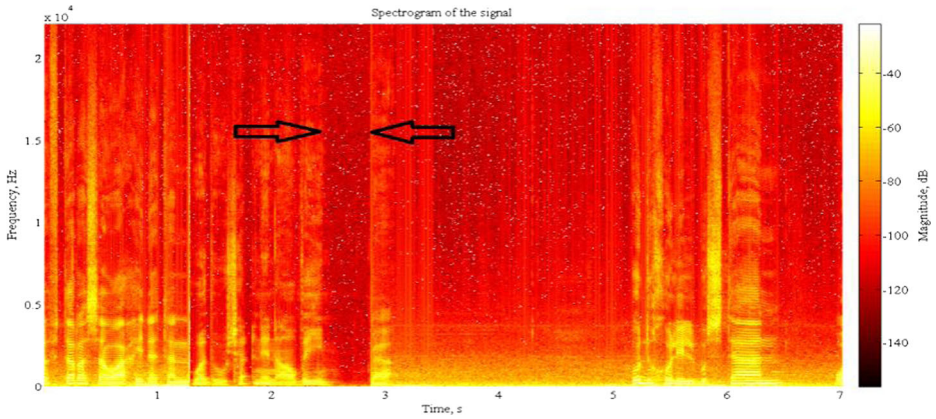


Fig. 7 Irregularity caused by an insertion and recording in different environments [18]

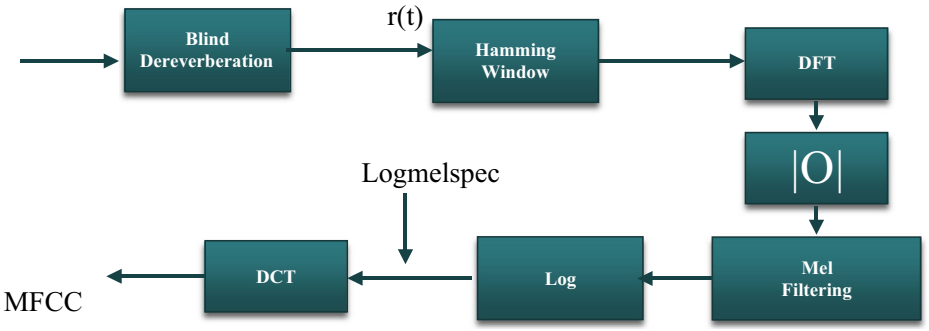


Fig. 8 Flowchart of the feature extraction substage [43]

speech recordings (including both male and female speakers) were used to evaluate the performance of the proposed method [41].

3.2.2.1 Performance evaluation on real-world data The dataset used to conduct the experiments contains 660 audio files, recorded in four different languages (i.e. Arabic, Bahasa Indonesia, Chinese, and English). The duration of each file is approximately three minutes, while the first minute in each audio file is silence. We recorded 264 files in Arabic by two speakers, 132 files by a non-native English speaker, 132 files by a Chinese speaker, and 132 files by an Indonesian speaker. These files were recorded using 22 different microphones in six different acoustic environments, namely: soundproof room (quite room), classroom, laboratory, staircases, parking area, and garden. Seventy-two (12 × 6) sessions were recorded using these microphones in each acoustic environment. Moreover, for each session, a person read a predefined text while sitting approximately 30 cm far from the microphone. Each recording manually aligned to remove starting and ending silence regions. The collected dataset (hereafter referred to as the Digital Multimedia Forensics Dataset – DMFDB) is available online [22]. One common form of tampering in digital audio signals is known as splicing, where sections from one audio is inserted to another audio, it is a new method that can be applied to detect a common form of tampering in digital audio signals known as splicing. This experimental study investigates effectiveness of the acoustic environment signature for splice

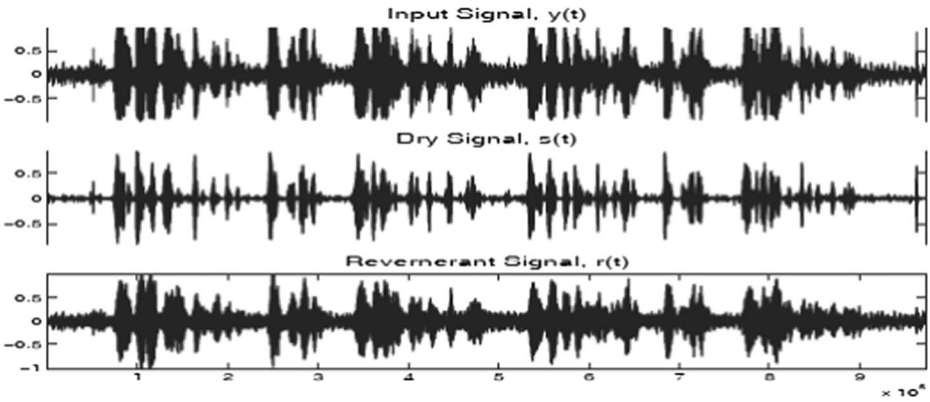


Fig. 9 Top panel: plot of the test recording ***y(t)***, middle panel: estimated ***s(t)***, bottom panel: estimated ***r(t)***blind de-reverberation subsystem [43]

detection and localization. Recently, Zhao et al. [72] proposed an audio splice detection method based on acoustic environment cues. This experiment evaluates the effectiveness of Zhao et al.'s method [72] on the proposed dataset. To this end, the magnitude of the acoustic channel impulse response and ambient noise is used for modeling the intrinsic acoustic environment signature and for splicing detection and splicing location identification. The motivation behind considering the combination of acoustic channel impulse response and ambient noise for audio splicing detection is that acoustic reverberations can be used for acoustic environment identification, that is, to determine where a recording was made. In some of our recent work [20, 41, 43, 70, 71], we showed that acoustic reverberation and ambient noise can be used for acoustic environment identification. One of the limitations of these methods is that they cannot be used for splicing location identification. To address the limitations of reverberation-based methods [20, 41, 43, 70, 71], the magnitude of the channel impulse response is used for audio splicing detection and localization. One of the advantages of the proposed approach is that it does not make any assumptions. In addition, the method in [72] is robust to lossy compression attack. Here, we exploit artifacts introduced at the time of the recording as the intrinsic signature and use it for audio recording integrity authentication. Both the acoustic channel impulse response and the ambient noise are jointly considered to achieve this objective. To this end, each input audio is divided into overlapping frames. For each frame, the magnitude of the channel impulse response and ambient noise is jointly estimated using spectrum classification techniques. The similarity between the estimated signatures from the query frame and the reference frame is computed, which is used to determine whether the query frame is a spliced frame or not. More specifically, a spliced frame is detected and localized if its similarity score with the reference frame is less than the threshold or not. A refining step is further considered to reduce detection and localization errors. Figures 10–12 show the experimental results for audio recordings made with T.Tbone microphones. The title of each sub-figure is the audio name. The points marked in red stars represent the ground truth. It can be observed from Figs. 10–12 that our method can detect the presence of splicing frames for most cases (e.g., Fig. 10 (a-e), Fig. 11 (a, b, c, e), Fig. 12 (a, b, d, f)). The rest also resulted in some false negatives, as shown in Fig. 11 (d & f) and Fig. 12 (e). It was observed that such false negatives could be attributed to the small forgery locations in the test audio. Figure 11 (d & f), and Fig. 12(e) show that only a few frames have been modified in the tampered audio. In this case, it is difficult to obtain reliable signature estimation, which indicates that this method is not very successful for tampered audio with small insertions. It is also observed through extensive experimentation that the larger the insertion in the tampered audio, the easier it is to be detected. Overall, the proposed algorithm resulted in a detection performance of 90% on the developed database. For most cases, the proposed algorithm was able to successfully detect the forgery locations with a very high confidence (Table 2).

3.2.3 Authentication using microphone signature (e.g., source identification)

The device used to record the audio contents usually records other signals also such as signatures, which provide evidence of the ownership of the file and location of the recording. The use of standard security approaches to address this problem (e.g., by digitally signing content within devices right after signal acquisition [9] or watermarking techniques [50]) requires the modification of devices and workflows, and thus this is not always applicable.

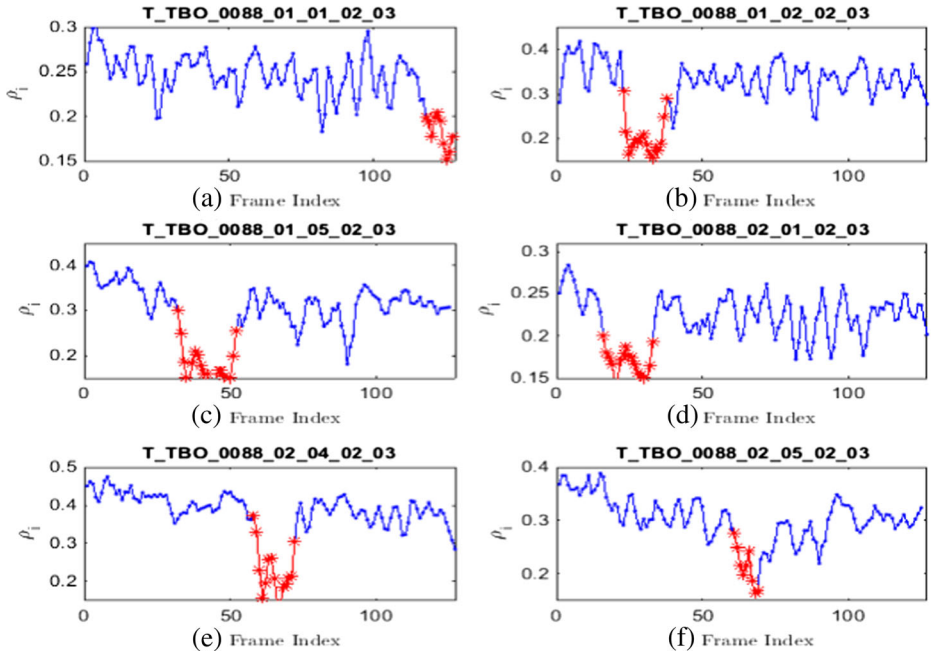


Fig. 10 Forgery detection and localization results

Alternatively, it is possible to rely on passive methods, which exploit the subtle footprints, which are inevitably left by signal processing operators, including acquisition and coding. For instance, in the case of audio signals, the microphones used for content acquisition introduces a characteristic trace, which can be detected [7, 13, 18, 33]. However, at the same time, for somebody to convincingly distort the original meaning of recorded audio material, it is often necessary to introduce content that was recorded elsewhere using a different device. Hence, microphone classification can be used to identify such inconsistencies, i.e., to detect that several microphones were used within a content item that pretends to be one continuous original recording, thus indicating a trace of tampering. It can be seen in Fig. 13.

For the classification task, support vector machine (SVM) is used with a radial basis function (RBF) kernel and parameters (cRBF, RBF), where cRBF is the cost variable of the SVM and RBF is the gamma parameter of the RBF kernel. Each training audio file is represented by a feature vector $\boldsymbol{f}_{training}$. The complete training set goes through a pre-processing step before feeding the RBF-SVM, i.e. normalization between −1 and +1 of each dimension and a feature selection. The feature selection was performed by computing the F-score of each dimension, on PCM-encoded audio files. The features with the highest F-score are then selected, and the RBF-SVM is trained from the feature matrix FF-score training built by aggregating the feature vectors from all known devices. The training set is balanced, i.e. an equal number of feature vectors per device are present. The proposed method performed this type of tampering detection, using a robust microphone classification algorithm. This new application of microphone classification for tampering detection led to a detection with accuracy higher than 95% for PCM, AAC, and MP3-encoded recordings [10]. In [33], the authors of the proposed method have divided the process into two stages. In the first stage, a suitable context model is designed for microphone recording. In the second stage, the required domain knowledge is generated by

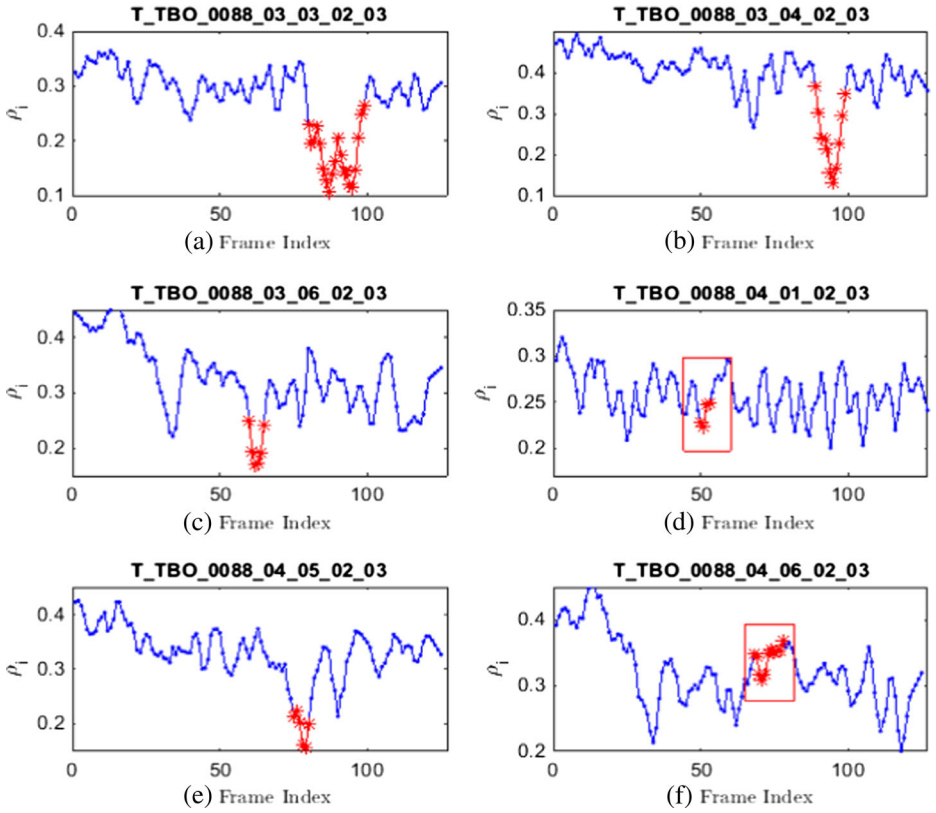


Fig. 11 Forgery detection and localization results using Zhao et al.'s method [71]– small splicing part using Zhao et al.'s [71] method – relatively large splicing

applying the context model. In [34] digital audio forensics is studied to identify the microphone model. The detail definition of microphone model would provide the investigators to prove the similarity among some of the recordings. Authors in [34] performed statistical analysis on the recording, which was gathered from two microphones of the same model as shown in Fig. 14.

The identification of a very suitable classification algorithm is discovered for forensic through this process. Further in [33] authors determined the features, which make considerable changes in the performance of the classification in pattern recognition and microphone detection.

In the proposed method [47] both training and identification phases are used for identification method. During the training phase, the support vector machine (SVM) model is trained together with the reduced noise features and their class information. The classification models, which are produced during phase 1, are used to identify the recorders in the identification phase. The following are the steps for extraction process as shown in Fig. 15: 1. Wiener filters which are used to extract the noise sound from the recorded sound. 2. MIR tool box were used to extract the noise features and then normalized. 3. The interclass standard deviation methods are used to reduce the noise features. Compared to the method of no feature reduction, this method makes 1% improvement for the 11 audio recorders. This feature reduction method is competitive when compared to other well-known methods like PCA, LDA and R-squared as shown in the results.

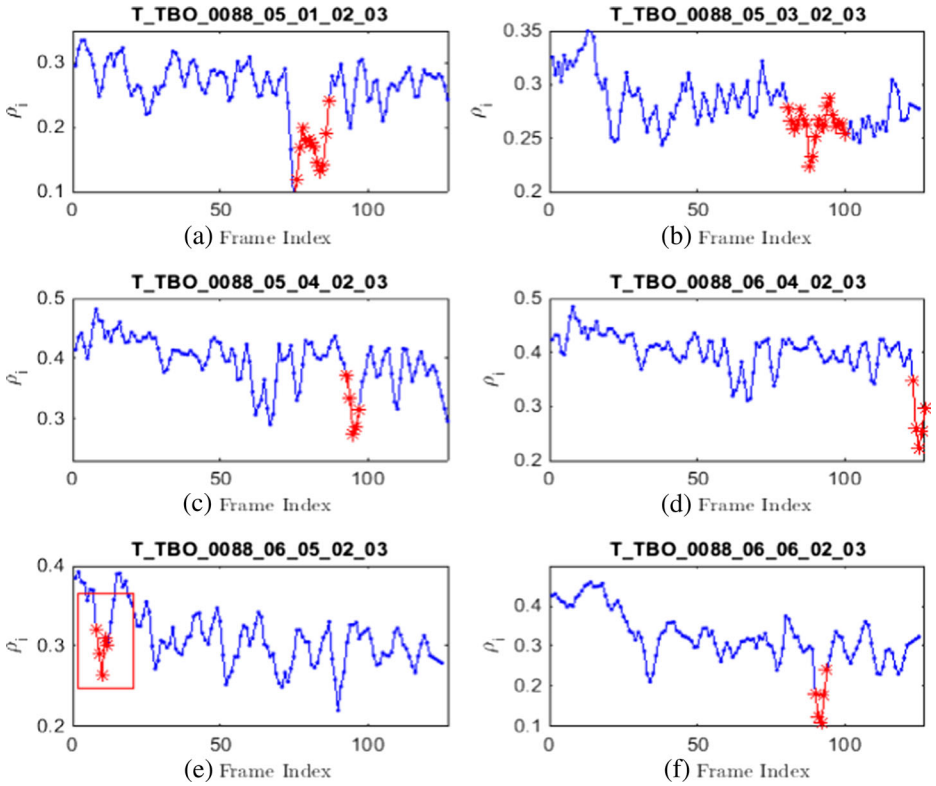


Fig. 12 Forgery detection and localization results using Zhao et al.'s method [71]– small splicing part b

Table 2 Summary of techniques based on the environment detection technique

Ref. No	Objective	Method	Dataset	Features	Classification	Results
[32]	Classify and identify the environment	classify categories of recording environments	Multiple environment features	MFCC	support vector machine (SVM)	75.9%.
[48]	Classify and identify the environment	DCT and PCA techniques	Four Environments	MPEG-7	GMM	--
[43]	Estimate the reverberation component	Inverse filtering	284 speech recordings	MFCCs, logarithmic mel-spectral coefficients	Multi-class support vector machine (SVM)	94% Accuracy
[23]	Speech leakage signal detection	spectral subtraction, multiband-based spectral subtraction	Five different environments			Better results than the traditional methods
[71]	Reverberation and background noise detection,	particle filtering technique	2240 speech recordings	TDSM-based features		Better results

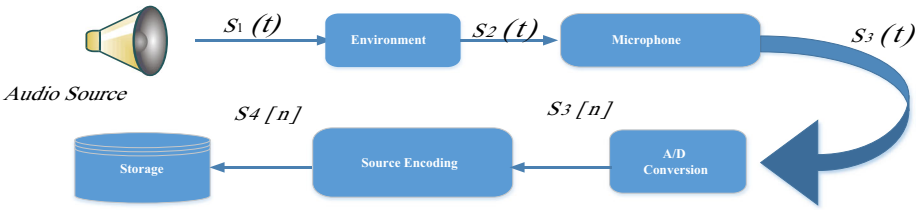


Fig. 13 Mobile recording - process flow [10]

Telephone-based speaker recognition has an important issue regarding the robustness of the environment because speaker verification system users tend to use different handsets in different situations. Accuracy for the recognition degrades when the users have different handsets during recording and verification process. This is a challenging task for the speaker verification system because of the lack of robustness with respect to handset variability. Each handset makes a different degree of distortion in the speech signal. To address this problem, the two-channel compensation approach [68] is introduced to handle the environmental mismatch problem in telephone-based speaker verification system. Probabilistic decision-based neural networks (PDBNNs) deals with both speaker dependence and handset dependence process, while maximum likelihood linear regression (MLLR) deals with handset dependence only. The results, based on 150 speakers of HTIMIT, show that combining MLLR adaptation with handset identification achieves the lowest error rate.

3.2.3.1 Performance evaluation using real-world data The effectiveness of the proposed method was tested on the data set recorded using four pairs of microphones. To this end, each recording is segmented into frames of four seconds duration with a 50% overlapping factor. Bicoherence is estimated from each audio segment using the direct (fft-based) approach [49]. The bicoherence is estimated with the following parameter settings: 1) 128- point segment length, 2) 256-point FFT length, 3) no overlap, and 4) Rao-Gabr optimal window for frequency domain smoothing. For each audio segment of a given recording the first four scale-invariant Hu moments, that is, m1,1,m2,0,m2,1, and m3,0 are computed from the bicoherence magnitude spectrum. Shown in Fig. 16 are the scatter plots of scale invariant Hu moments m1,1,m2,0, and m3,0 computed from the bicoherence magnitude spectra of the first (top), second (middle), and third (bottom) recordings made using a pair of Samson R19 dynamic microphonesM1 and M2. Similarly, shown in Figs. 17, 18, and 19 are the scatter plots of scale-invariant Hu moments m1,1,m2,0, and m3,0 computed from the bicoherence magnitude spectra of the first (top), second (middle), and third (bottom) recordings made using

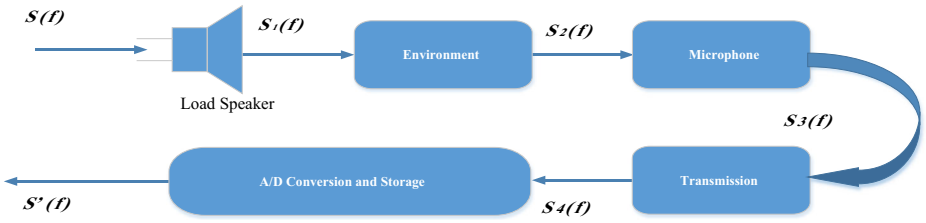


Fig. 14 Recording process pipeline – context model [33]

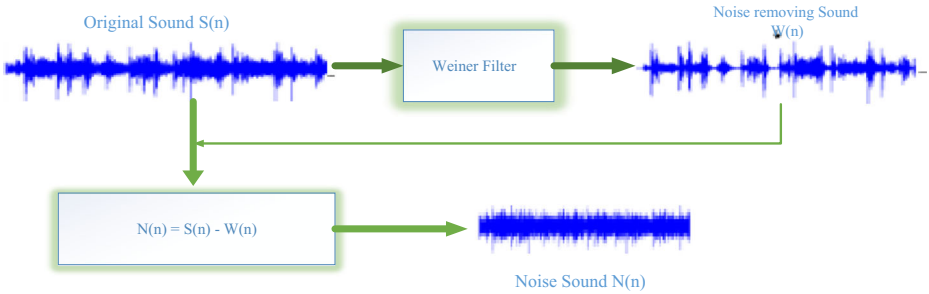


Fig. 15 Process of extracting noise sound by using a Wiener filter [47]

microphone pairs M3 and M4, M5 and M6, and M7 and M8, respectively. It can be observed from Figs. 16, 17, 18, and 19 that there are significant inter- as well as intra-class variations in the frame-based scale-invariant Hu moments. The interclass variations can be used for microphone type identification, whereas, intra-class variation can be used to achieve individual microphone identification. To illustrate, this scatter plots of average scale-invariant Hu moments m1,1,m2,0, and m2,1 computed from the bicoherence magnitude spectra of the first, second, and third recordings using all eight microphones are shown in Fig. 20. Shown in Figs. 17, 18, 19, and 20 are scatter plots of average scale invariant Hu moments m1,1,m2,0, and m3,0 computed from bicoherence magnitude spectra of first, second, and third recordings made using microphone pairs M1 and M2, M4 and M4, M5 and M6, and M7 and M8. The first- and higher-order statistics of the estimated Hu moments are used for microphone identification. To this end, mean, variance, skewness, and kurtosis of the estimated frame-based Hu moments are used for microphone identification. Threshold based multiple hypothesis testing is used for microphone identification which resulted 100% correct classification of 24 recordings for eight classes (microphones) (Table 3).

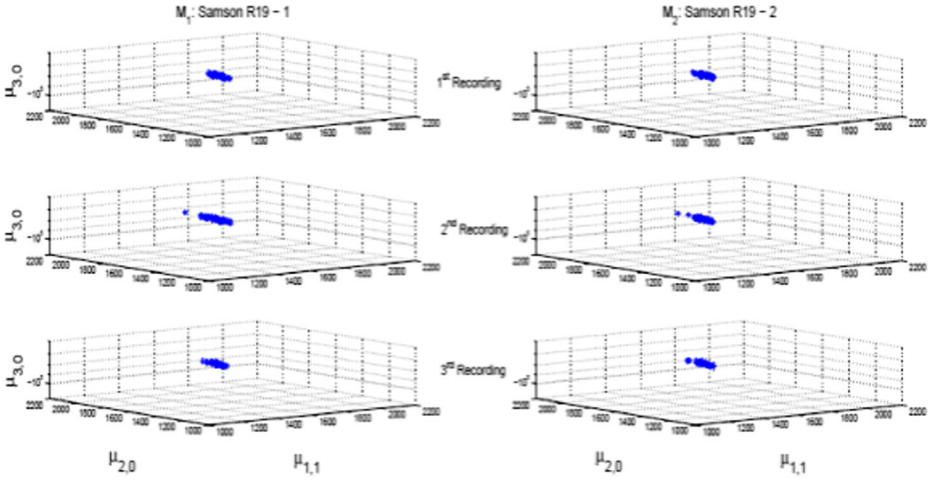


Fig. 16 Shown are scatter plots of scale-invariant Hu moments m1,1,m2,0, and m3,0 computed from the bicoherence magnitude spectra of the first (top), second (middle), and third (bottom) recordings made using a pair of Samson R19 dynamic microphonesM1 and M2

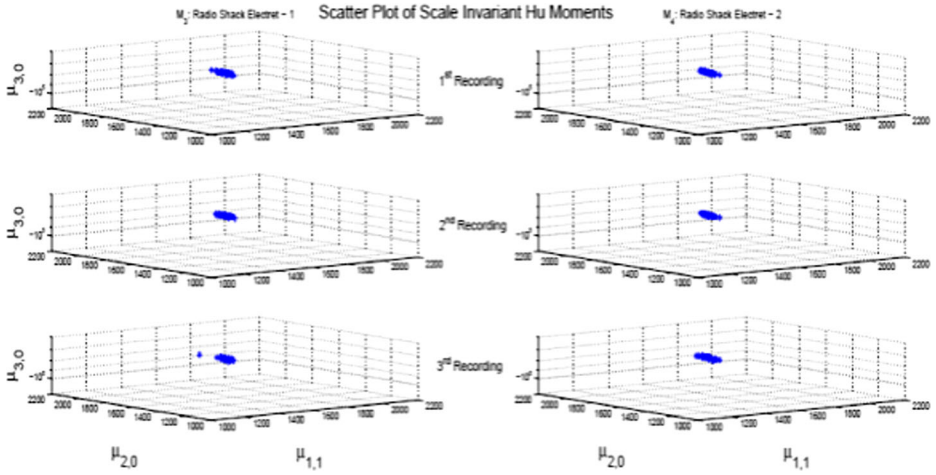


Fig. 17 Shown are scatter plots of scale-invariant Hu moments μ1,1,μ2,0, and μ3,0 computed from the bicoherence magnitude spectra of the *first* (top), *second* (middle), and *third* (bottom) recordings made using a pair of Radio Shack electret microphones *M*3 and *M*4

4 Forensic audio enhancement

Audio enhancement is the process of removing and cleaning unwanted noise from an audio file, which are usually recorded unintelligently. The forensic experts try to remove these noises

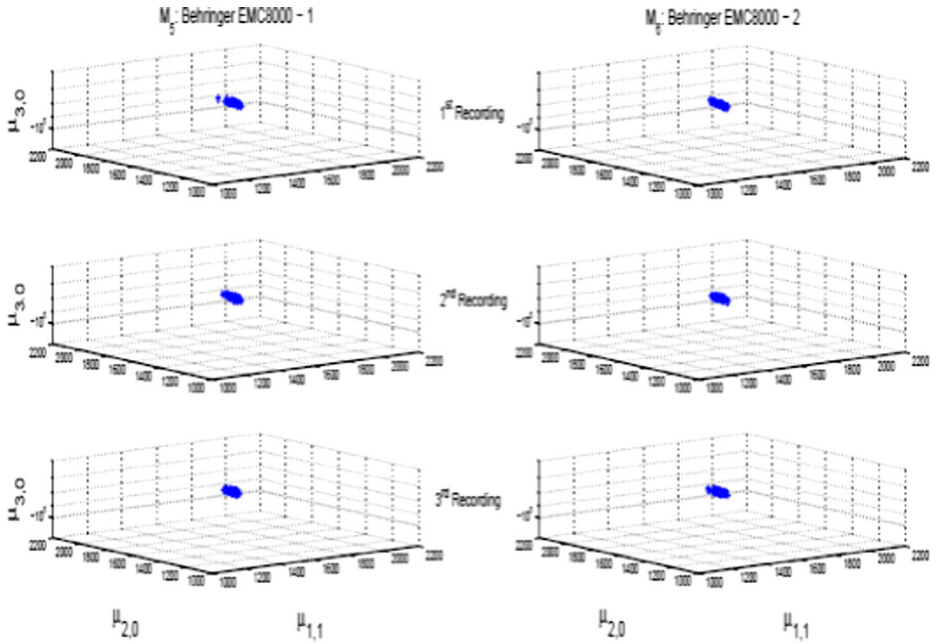


Fig. 18 Shown are scatter plots of scale-invariant Hu moments μ1,1,μ2,0, and μ3,0 computed from the bicoherence magnitude spectra of the *first* (top), *second* (middle), and *third* (bottom) recordings made using a pair of measurement microphones *M*5 and *M*6

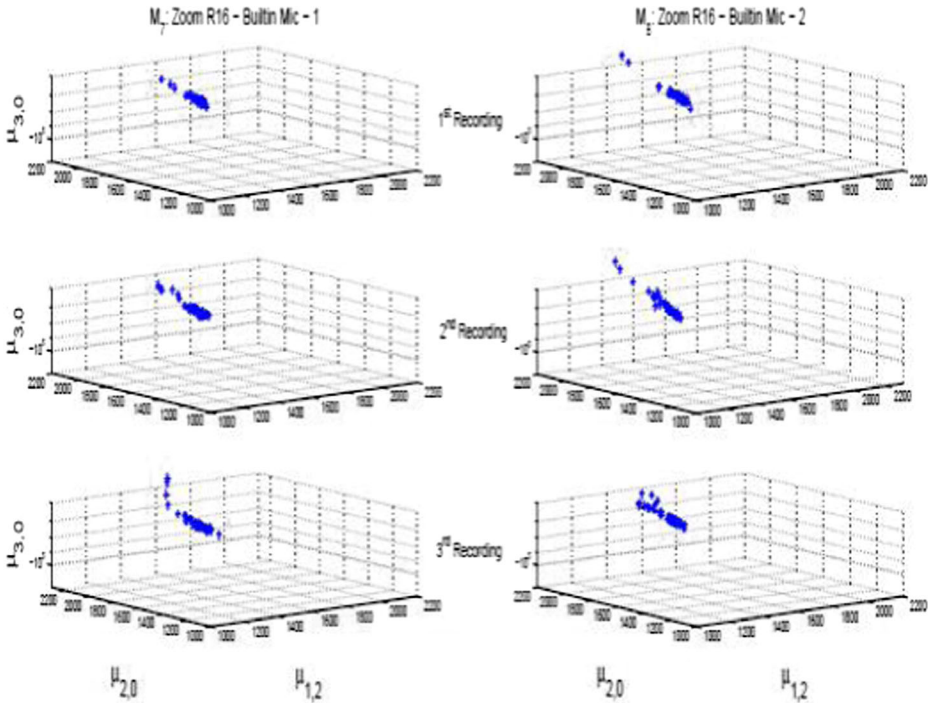


Fig. 19 Shown are scatter plots of scale-invariant Hu moments μ1,1,μ2,0, and μ3,0 computed from the bicoherence magnitude spectra of the *first* (top), *second* (middle), and *third* (bottom) recordings made using a pair of Zoom R16built-in electret microphones*M*7 and *M*8

and enhance the audio file without damaging the original information. Enhancement allows listeners to know "what is said" and prove or disprove the involvement of an individual in a crime. Even though the enhanced file may look worse than the original file, but what actually said is revealed clearly. The key to audio enhancement is to detect the noise problem, because in the tampered file the noise is reengineered in such a way that it becomes a part of the original recording. Therefore, the idea is to detect this noise and extract it from the original recording. The critical listening of the original material is the start of forensic enhancement.

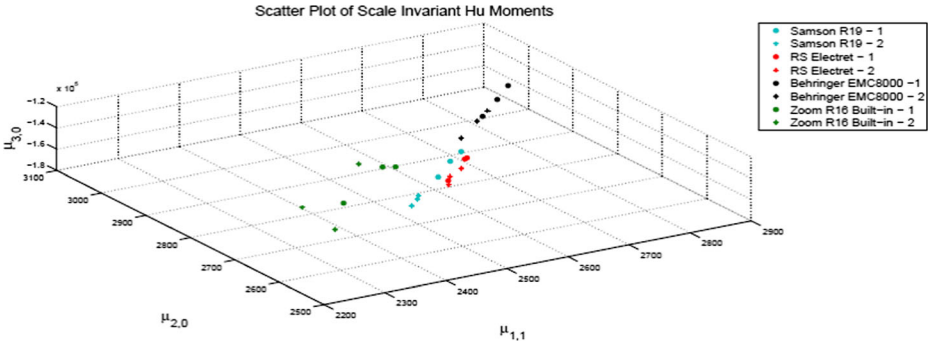


Fig. 20 Shown is scatter plots of average scale-invariant Hu moments μ1,1,μ2,0, and μ3,0 computed from the bicoherence magnitude spectra of *first*, *second*, and *third* recordings made using all eight microphones

Table 3 Summary of the techniques based on source detection techniques

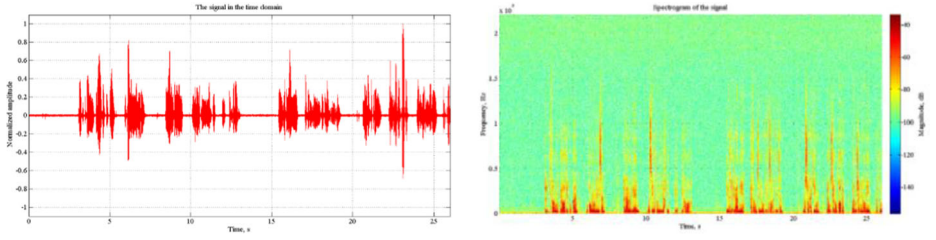
Ref. No	Objective	Methodology	Classification	Accuracy	Results
[10]	Audio tampering detection	Microphone classification	support vector machine (SVM)	95%	-
[47]	Audio recorder identification	Phase 1 is for training and phase 2 for identification	support vector machine (SVM)	1% increase	Feature reduction
[68]	Telephone-based speaker verification	Adaptation/ transformation techniques: PDBNNs, MLLR	-	Error Rate: PDBNN:8.44 MLLR: 6.67	Low error rate
[33]	Classification of microphones	Classify the seven microphones using six classifiers	Logistic regression, support vector machine SVMs, decision trees and nearest neighbor	93%	-

The following are the ultimate goals of forensic audio enhancement: Increase speech intelligibility, Increase the accuracy of transcription, Decrease listener's fatigue, and Reduce SNR. Broadband noise reduction is a common request for forensic audio recording [5, 19, 35, 46, 62]. The digital copy of the original forensic recording is used to apply the noise reduction process to implement several enhancement techniques without making any damage to the original files.

It is desirable to enhance the SNR for an audio recorded file that contains unwanted noise before playback [35, 55, 63]. With the overlap-add procedure [5, 46] all the subsequent frames create the entire output signal as shown in Figs. 21(a) and (b).

4.1 Classification of forensic audio enhancement

To improve the intelligibility of the target speech, time variation could be compensated between the primary and reference inputs but it should be relatively slow. The time drift acceleration and deceleration rate should be under some limit to allow DCAF (drift-compensated adaptive filtering) to track the reference. DCAF can achieve a noticeable interference reduction for rates as large as ±1% per 60 s at a 16 kHz sampling rate [12].



(a) (b)

Fig. 21 **(a)** Sample of enhancement of forensic audio [40]. **(b)** Example of forensic audio enhancement [40]

A landmark-based acoustic fingerprinting technique is proposed to automatically identify and align then subtract the reference sound and bring the speech of interest to the forefront. This two-stage approach reduces the effect of interfering music, television or noise. It identifies and automatically aligns the reference sound. A signal reference cancellation algorithm technique is then applied to subtract the reference sound to bring the speech of interest to the forefront. A small reference music database consisting of 50 songs were used. The database consisted of pop, rock and instrumental music. The sampling rate of the reference music files used in these experiments was 44, 100 Hz, with a bitrates of 16 bits and uncompressed Microsoft WAV files as shown in Fig. 22.

Existing LMS-based two-channel reference cancellation approaches could be applied to robustly cancel the interfering audio and then leave the speech of target speakers largely intact [1]. A spectral subtraction algorithm in the modulation domain has been proposed in [52] to overcome additive noise distortion. Both objective and subjective speech enhancement experiments were carried out to evaluate the proposed approach. To enhance speech quality, a combination of the ModSpecSub and MMSE methods in the STFT magnitude domain were proposed. The fusion method was also evaluated through both subjective and objective speech enhancement experiments. The experimental results show that this approach improves the quality of speech and does not suffer from musical noise, which is typically associated with the spectral subtraction algorithm.

Extracting DOA information and signal phase from the background noise by using this method improved speech separation performance when measured with PESQ, Segmental SDR

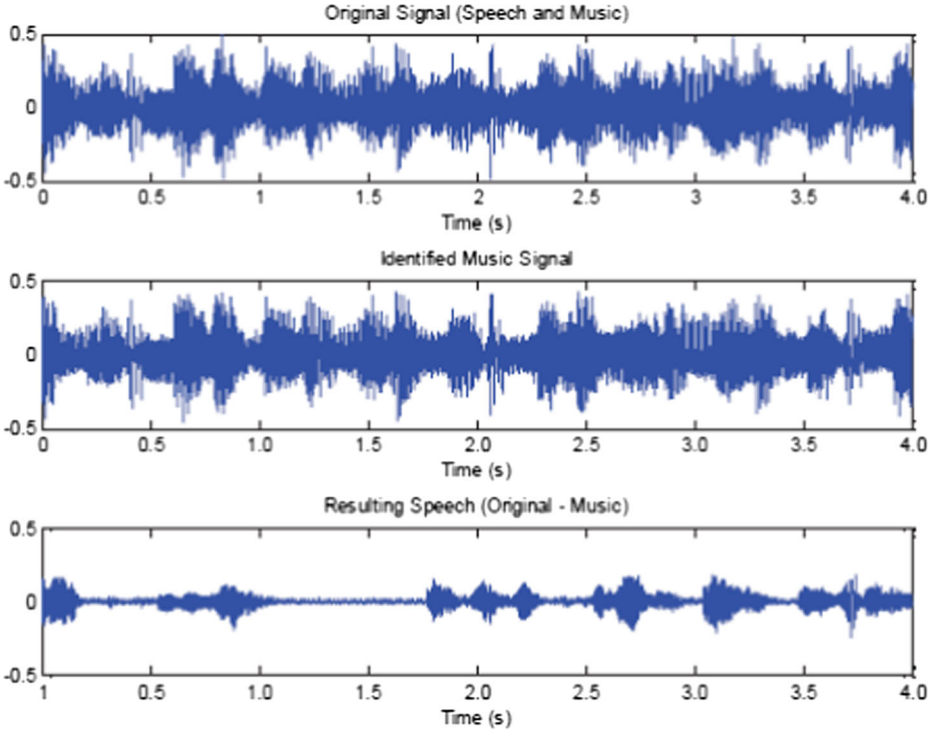


Fig. 22 Waveforms of the original signal, subtracted identified music signal and the result of music signal cancellation [1]

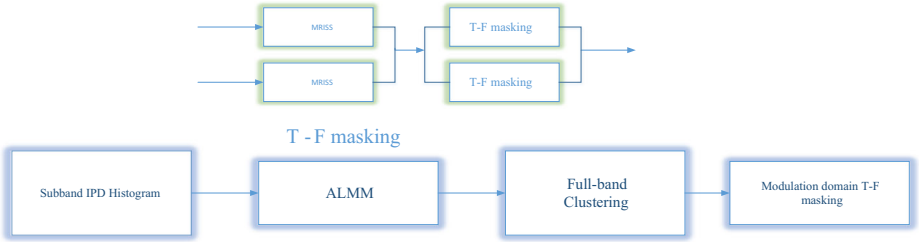


Fig. 23 Block diagram of the presented method [69]

and SIR gain. This method as shown in Fig. 23 also improved robustness through the contribution of MRISS preprocessing, sub-band IPD histograms, ALMM and modulation domain separation. Future studies could aim to investigate the integration of modulation domain separation with other blind separation methods [69] (Table 4).

5 Transcoding identification

Audio files recorded from handheld devices and uploaded on the web do not appear to listeners straight away because of transcoding. Only when transcoding is finished the listeners are able to listen to the file. To explain transcoding, we must understand how digital media are stored. Digital media have a container that stores metadata information about the dimension and duration of the file as well as the number of tracks. Each of these

Table 4 Summary of the techniques based on the audio enhancement technique

Ref. No	Objective	Methodology	Dataset	Results
[12]	Reduce interference, even if the magnitude of the timing drift rate is 1–2%	DCAF	-	Reduce the number of features
[1]	Audio fingerprinting and reference cancellation for improving intelligibility	Landmark-based acoustic fingerprinting: reducing the effect of interfering music, using the signal reference cancellation algorithm technique for reducing reference sounds	Small reference database of music containing 50 songs, sampling rate of 44,100 Hz, bitrates of 16 bits	Reduce reference sounds and bring the speech of interest to the forefront
[52]	to overcome additive noise distortion	combination of the ModSpecSub and MMSE methods	-	Improved speech quality
[69]	Extracting the DOA information and recovering signal phase from the background noise, and improving speech separation	Robust blind speech separation method: pre-processing MRISS performed, DOA-based T-F masking implemented	-	Contributed to the robust performance of the proposed method

tracks is encoded to improve the quality and reducing the file size. These encoded tracks are then stored back in the container. The following undertakings are considered to transcode a digital file: Extract the track from the container, Filter the track, Encode the track, and Multiplexing the new track.

Transcoding is usually carried out to convert a file from one format to another (e.g., converting a DivX AVI file into H.264/AAC in MP4 for delivery to mobile devices, set-top devices and computers). Media is transcoded for a number of reasons: To convert a high quality file into a digitally distributed format to send the file over the Internet, to convert it from a high quality music file library stored in ACC or Vorb is to MP3 files, To convert the file into a suitable format used by the user, To convert the format to save storage space (e.g., transcoding old MPEG2 HDV tapes into H.256)).

Transcoding detection is helpful to know if the downloaded audio file is in its original state. The problem of audio transcoding has been studied in earlier works [3, 25, 61].The quality of the audio file is checked by measuring the bit rate of the MP3 file. However, this check and the results are affected when bit rate is transcoded from a lower bit rate to a higher bit rate. The original lower bit rate of the audio file is detected by analyzing the high frequency spectrum of the audio file. Then, the SVM classifier is used to classify the five classes of bit rates (CBR 128 kbps, 192 kbps, 256 kbps, 320 kbps, VBR-0). Classification accuracy of about 97% was achieved to detect the original compressed bit rate of the file. A dataset of about 2512 different songs were used for this analysis. In the proposed technique, a high frequency spectrum of the audio file is considered to detect the original bit rate of the file by comparing it with the known spectrum patterns of various bit rates. By using the proposed method, 1945 of the 2512 songs tested were correctly classified (97% accuracy). Hence, the high frequency spectrum is a reliable method to determine the true bit rate [11]. In method [25], transcoding was performed by employing frequency domain signals. This approach reduces the memory requirements as well as the computing power, making it suitable for applications with limited computing power such as multimedia recording and handheld devices. An attempt was made to combine the MP2 decoder and MP3 encoder by removing and exchanging certain blocks. The processing power of the MP3 encoder was reduced by 50% by removing the filter bank and psychoacoustic model. Hence, the filter banks of both the MP2 decoder and the MP3 encoder were removed with the psychoacoustic model computation. Because of the relationship between the time and frequency domains, this advantage reduces the processing power and memory requirements. An efficient implementation for converting MPEG-2/MPEG-4 AAC-encoded data into Dolby Digital AC-3 has been described. In [45], the information present in the decoder was exploited to simplify audio transcoding as well as implementation to various algorithms in the encoder. Because of the similarity between standard audio coders, optimization is achieved in transcoding. A case study is proposed to prove the efficiency of these techniques to simplify encoder implementation. A large set of audio files were tested in the proposed method, finding that a significant amount of encoder complexity is reduced with no degradation in audio quality**.** The study reused the bit allocation information in audio transcending by exploiting similarities in sub-band audio coding schemes. This shows that important information can be deduced to reduce encoder complexity, even if the two coders employ different psychoacoustic models. A case study is provided with MPEG AAC/Dolby AC-3 transcoding. However, the proposed algorithms can be extended to other audio transcoding schemes [44] (Table 5).

Table 5 Summary of the techniques based on the transcoding technique

Ref No.	Objective	Technique	Methodology	Results
[25]	Reduce processing power and memory requirements	Relationship between the time and frequency domains	Combine both the MP2 decoder and the MP3 encoder	Memory is reduced
[11]	Measure the bit rate of the MP3 file	Detect the original bit rate of the file by comparing it with the known spectrum patterns of the various bit rates	Analyze the high frequency spectrum of the audio file	Classification accuracy of about 97%
[45]	Reduce encoder complexity	Encoder information is used	Embedding decoder information into the transcoder	Complexity is reduced
[44].	Describe strategies for reusing bit allocation information	Study the reuse of bit allocation information	Exploit similarities in sub-band audio coding schemes	Important information can be deduced to reduce encoder complexity

6 Codec identification

Codecs are used to encode and decode the digital audio. The primary goal of designing a codec is to compress the digital audio file and music file for more compact storage over the internet and transmit the voice communication over cellular network and VoIP networks [58]. Telephony system is another field where identification of codec is essential to know the history of the audio stream [2]. Telephone infrastructure is diversified and non-centralized and no exact and reliable mechanism is available to track the route of the incoming call as the voice signal is travelling through many routes over the network. Because of this inability to verify the origin of incoming calls many malicious activities takes place like voice spam and voice phishing attacks. To determine the quality of the file and its originality, i.e. to check if the low bit rate file is transcoded to high bit rate and then pretending to be of high quality, is another aspect of research which is based on codec identification [65].

For detecting the authenticity of the file, once the above information is extracted automatically then the origin and authenticity could be determined.

Technique In [65], a source dependent technique is focused. The speech media authentication is processed in two steps. The type of speech codec used to generate the signal is determined in the first step, and then media authentication is performed based on the properties detected on the codec in the second step. Tampering detection algorithm is proposed based on the codec detector. The goal of this method is to detect if some alterations are done after encoding and decoding a speech file with a specific codec. Cellular dataset is used to test the algorithm, dataset consists of recordings which are recorded directly from the cell phones [73]. The proposed method uses multiclass classification based on the features, which describe the randomness and chaotic behavior of coded data and support vector machines. The experimental procedure consists of two steps, one for identification of codec among 16 audio codecs, most of the codecs are identified accurately with an average accuracy of 85%. In the second step the transcoding is done to the audio files, which are encoded with other codec, and the technique is to identify the first codec. The experimental results display that the singly coded and transcoded audio codecs can be distinguished from

each other with an accuracy close to 100%, and codec before and after transcoding can be identified at accuracy of about 80% [21].

The accurate identification of decoded speech from the codec is detected in the proposed method without access to its original encoded speech. The technique detects the decoded speech signals with the multidimensional profile, which consists of noise spectra and time domain amplitude histogram from multiple speeches. Comparison is done between these profiles and the reference profile from candidate codecs. Results demonstrate that the proposed technique is highly accurate with 100% correct identification for most of the codecs [24]. In the proposed method, authors have introduced a non-intrusive data driven method for detecting codec in the presence of background noise. During the training phase, a number of speech features are used. The result is demonstrated as the performance of the method on different noise types with wide range of SNRs. The results showed that the proposed method can identify a codec and its bit rate to an accuracy of 92% and are able to detect the presence of a codec with an accuracy of 97% at −5 dB SNR [59]. AMR decompressed audio files are focused in the work [37], which are then used for the purpose of detecting the source file of the recording, which ultimately help in detecting the digital audio forensics (Table 6).

7 Double compression detection

MP3 is typically manipulated by the compression and decompression of audio files for malicious purposes. Because of these manipulations, many studies have been conducted in this field to authenticate audio files, and several solutions have been proposed to detect both double and single compression for multimedia files. The main purpose of manipulating and compressing an audio file is to recompress it at a higher bit rate to pass it off as a high quality track [17]. This technique describes the statistical features extracted from modified discrete cosine transform (MDCT) coefficients and other parameters that may be obtained from compressed audio files. Tampering activities and trace identification are detected because of multiple compressions. Based on the analysis and inherent parameters of compression encoder

Table 6 Summary of techniques based Codec Technique

Ref. No	Objective	Technique	Method	Result
[73]	Tampering detection algorithm is proposed based on the codec detector	source dependent technique	Two step procedure is followed	-
[21]	detect the codec used to encode the given audio file	multiclass classification	Two step procedure, identification of the codec, and identify the first codec after transcoding	First step 100% accuracy and second step 80%
[24]	identification from the decoded speech of the codec used	identifying the traces left from the signal processing	Multidimensional profile with noise spectra	100% accuracy
[59]	codec detection and identification in the presence of background noise	features used to train a CART classifier	non-intrusive data driven method	97% at −5 dB SNR

identification, an algorithm was developed to enhance robustness. A large music database was used to test the effectiveness of this method with about one million compressed audio files. The results achieved substantially contribute to the development of scientific tools for forensic audio analysis [30].

In this technique, the authors tried to localize the presence of double compression in MP3 audio files and uncover possible tampered parts. It detects whether an MP3 audio file is singly compressed or doubly compressed as well as derives the bit rate of the first compression. It also detects the short temporal windows to localize the tampered portions of the MP3 file under analysis. The technique is effective when the bit rate of the second compression is higher than that of the first; however, it has limited performance in the opposite direction. The features are based on a simple histogram [4].

The following techniques are implemented by experts to detect the tampering of audio files. In [31], an audio encoding algorithm was applied to compress and decode audio files. This algorithm employed modified discrete cosine transforms based on frame-offset measurement. The regularity in the audio file was disturbed when the file was modified such as cutting off or pasting a part of the audio recording. By detecting the small value of the spectral components, an additional histogram analysis was performed to enhance robustness. The technique was tested on a database consisting of 15 music tracks with harmonic components and slowly changing audio backgrounds.

This method [54] was based on the statistical patterns extracted from quantized MDCT coefficients and their derivatives. Both up-transcoded and down-transcoded MP3 audio files were detected and the real compression quality was revealed. The false predictions caused by individual characteristics were minimized for diversified audio clips. Reference audio signals were generated by calibrating and recompressing the audio files as well as measuring the difference between signal-based and reference-based features.

In [65], to address the fake quality of MP3 files, authors of the proposed technique observed many more quantized MDCT coefficients with small values in a singly compressed MP3 file than in a fake quality MP3 file, regardless of which bitrates the fake quality MP3 was transcoded from.

In [66], authors of the proposed technique used the SVM classifier to detect the double MP3 compression of an audio file. Then, MDCT coefficients with the distribution of first-digit quantization were used to form the feature vector for classification. In particular, a global method was proposed, where the statistics on the first digits of all quantized MDCT coefficients were taken, and then the computed probability distributions of nine digits were used as features (nine dimensions) for training an SVM.

In [36, 53], authors of the proposed technique detected double MP3 compression, with some statistical features extracted from MDCT, and an SVM was used to classify the extracted features. A set of the statistical features of zero and non-zero MDCT coefficients from the frequency range as well as individual scale bands were adopted. In [64, 67], a forgery detection method for MP3 audio files was proposed (Table 7).

8 Open challenges and future directions

Forensics audio enhancement Although Drift-Compensated Adaptive Filtering schemes can survive the timing drift between two inputs with a good accuracy, but the conventional scheme completely fails even with a small fraction of time drift. When the reference and

Table 7 Summary of the techniques based on audio compression

Ref No.	Method	Database	Objective	Features	Results
[30]	Compression encoder identification algorithm	one million compressed audio files	Tampering detection in compressed audio files	MDCT coefficients	Development of scientific tools for forensic audio analysis
[4]	histogram distance method	-	Localize the presence of double compression	Simple statistical feature	Detection of double compression
[31]	Audio encoding algorithm	15 music tracks	Tampering detection	Modified discrete cosine transforms	minimizes the number of false detections of forgeries
[54]	Calibrating and recompressing	-	Reveal the real compression quality	-	False predictions are minimized

primary inputs are asynchronous then DCAF is suitable. A landmark based acoustic fingerprinting technique is not directly applicable to badly clipped, pre-filtered, or heavily compressed recordings, or to recordings where there is a dynamic 'drift'. MMSE and Kalman Filter can be used in future techniques to enhance the Intelligibility. Fusion of ModSpecSub and MMSE methods can be done in STFT magnitude domain to get further enhanced quality of speech.

Acoustic environment identification Many statistical recognition based techniques were proposed in the past but most of them work only in the raw domains with low accuracy and inability to link the recording and acquisition device in a unique style. The noise spectrum for each time slot is updated by noise removal algorithm but that can also remove the speech signal from the input signal, which can be taken as future research to only remove the interfering noise. Towards the future research, an important point is intra-room classification under room identification, recordings done in different locations in the same room.

Microphone forensics Research for identification of different microphones of the same model could be done by combining other features to increase the discriminatory power, some of the features are discovered but they are not implemented with combination. As a future research, a preprocessing should be done to the audio file to have less influence of environment to enhance the discrimination power of magnitude response of the microphone channel. Microphones usually record audio signals, which are very normal files, and the original signal could be extracted from the recorded file, which can yield distortions introduced by the microphone, this could be used as a feature extraction.

Transcoding identification Decrease in audio quality due to transcoding in loose format is a very serious problem, which occurs because of compression in each generation, it is named as digital generation loss. Reducing the quality of audio file due to transcoding is a challenging task. To re-encode the audio file into any format and for editing it digitally, users make a master copy in a lossless format which takes lots of storage space and also these copies cannot be transcoded into any other format in future without a subsequent loss of quality.

Codec identification Analyzing the input signal only in frequency domain would neglect the possibility of extracting the valuable traces of the signal. Codec's may be limited in the range or set of sample amplitudes that can occur at the output. For example, the output samples from ITU-T G.711 codecs are quantized to set of 256 discrete amplitudes from among the usual 16-bit linear PCM space used for representation in memory. Since because both SS and SR are located in the intermediate network node they have limited influence on the choice of overt codec, due to its fact they are bounded to rely on the codec chosen by overt, non-steganography calling parties or they can interfere with the choice of the overt codec during the signaling phase of the call where codec negotiation is taking place.

9 Conclusion

Audio forensics plays an important role in crime detection as most of the human conversations are done through speech/voice which are then recorded as audio file. Detecting the authenticity of recorded audio files play an important role because many audio tampering software are easily available on the market. Many serious cases have been successfully investigated because of the implementation of the audio forensic detection techniques and some of the cases are listed in this paper for the readers to understand the importance of this subject. Authors have tried to list all the techniques by classifying them in to enhancement, environment, source, transcoding, and codec sections and each section is described by its background, experimental setup, database used and methodology applied to successfully detect the tampering with the latest accuracy results with tabular format and diagrammatical presentation to ease the understanding of the readers and finally open challenges in each section with the future directions to further explore the research for new findings. Many survey papers are written on audio forensics but a combination of these sections is not done in the past.

Acknowledgement "This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-INF2634-02)".

References

1. Alexander A, Forth O, Tunstall D (2012) Music and noise fingerprinting and reference cancellation applied to forensic audio enhancement. In: Audio engineering society conference: 46th international conference: audio forensics
2. Balasubramaniyan VA, Poonawalla A, Ahamad M, Hunter MT, Traynor P (2010) PinDr0p: using single-ended audio features to determine call provenance. In Proceedings of the 17th ACM conference on computer and communications security, pp 109–120
3. Bang KH, Park Y-C, Youn D-H (2006) A dual audio transcoding algorithm for digital multimedia broadcasting services. In: Audio Engineering Society Convention 120
4. Bianchi T, Rosa AD, Fontani M, Rocciolo G, Piva A (2014) Detection and localization of double compression in MP3 audio tracks. EURASIP J Inf Secur 2014:10
5. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. Acoustics, Speech and Signal Processing, IEEE Transactions on 27:113–120
6. Brixen EB (2007) Techniques for the authentication of digital audio recordings. In: Audio Engineering Society Convention 122
7. Buchholz R, Kraetzer C, Dittmann J (2009) Microphone classification using Fourier coefficients. In: Information hiding, pp 235–246

8. Chaudhary UA, Malik H (2010) Automatic recording environment identification using acoustic features. In: Audio Engineering Society Convention 129
9. Chen N, Xiao H-D, Wan W (2011) Audio hash function based on non-negative matrix factorisation of mel-frequency cepstral coefficients. Information Security, IET 5:19–25
10. Cuccovillo L, Mann S, Tagliasacchi M, Aichroth P (2013) Audio tampering detection via microphone classification. In: Multimedia Signal Processing (MMSP), 2013 I.E. 15th International Workshop on, pp 177–182
11. D'Alessandro B, Shi YQ (2009) MP3 bit rate quality detection through frequency spectrum analysis. In: Proceedings of the 11th ACM workshop on multimedia and security, pp 57–62
12. Ding H, Havelock DI (2010) Drift-compensated adaptive filtering for improving speech intelligibility in cases with asynchronous inputs. EURASIP J Adv Signal Process 2010:12
13. Garcia-Romero D, Espy-Wilson CY (2010) Automatic acquisition device identification from speech recordings. In: Acoustics Speech and Signal Processing (ICASSP), 2010 I.E. International Conference on, pp 1806–1809
14. Gerazov B, Kokolanski Z, Arsov G, Dimcev V (2012) Tracking of electrical network frequency for the purpose of forensic audio authentication. In: Optimization of Electrical and Electronic Equipment (OPTIM), 2012 13th International Conference on, 2012, pp 1164–1169
15. Grigoras C (2007) Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. Forensic Sci Int 167:136–145
16. Grigoras C (2009) Applications of ENF analysis in forensic authentication of digital audio and video recordings. J Audio Eng Soc 57:643–661
17. Grigoras C (2010) Statistical tools for multimedia forensics. In: Audio engineering society conference: 39th international conference: audio forensics: practices and challenges
18. Gupta S, Cho S, Kuo C-C (2012) Current developments and future trends in audio authentication. MultiMedia, IEEE 19:50–59
19. Hatje U, Musialik CM (2005) Frequency-domain processors for efficient removal of noise and unwanted audio events. In: Audio Engineering Society Conference: 26th International Conference: Audio Forensics in the Digital Age
20. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87:1738–1752
21. Hicsonmez S, Sencar HT, Avcibas I (2011) Audio codec identification through payload sampling. In: Information Forensics and Security (WIFS), 2011 I.E. international workshop on, pp 1–6
22. http://cybertechnos.com/datasets
23. Ikram S, Malik H (2010) Digital audio forensics using background noise. In: Multimedia and Expo (ICME), 2010 I.E. International Conference on, pp 106–110
24. Jenner F, Kwasinski A (2012) Highly accurate non-intrusive speech forensics for codec identifications from observed decoded signals. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 I.E. international conference on, pp 1737–1740
25. Ju F-S, Fang C-M (2006) Time-frequency domain fast audio transcoding. In: Multimedia, 2006. ISM'06. Eighth IEEE international symposium on, pp 750–753
26. Koenig BE (1990) Authentication of forensic audio recordings. J Audio Eng Soc 38:3–33
27. Koenig BE, Lacey DS (2009) Forensic authentication of digital audio recordings. J Audio Eng Soc 57: 662–695
28. Koenig BE, Lacey DS (2012) Forensic authenticity analyses of the header data in re-encoded WMA files from small Olympus audio recorders. J Audio Eng Soc 60:255–265
29. Koenig BE, Lacey DS, Killion SA (2007) Forensic enhancement of digital audio recordings. J Audio Eng Soc 55:352–371
30. Korycki R (2014a) Authenticity examination of compressed audio recordings using detection of multiple compression and encoders' identification. Forensic Sci Int 238:33–46
31. Korycki R (2014b) Detection of montage in lossy compressed digital audio recordings. Archives of Acoustics 39:65–72
32. Kraetzer C, Oermann A, Dittmann J, Lang A (2007) Digital audio forensics: a first practical evaluation on microphone and environment classification. In: Proceedings of the 9th workshop on Multimedia & security, pp 63–74
33. C. Kraetzer, K. Qian, M. Schott, and J. Dittmann (2011) A context model for microphone forensics and its application in evaluations. In: IS&T/SPIE Electronic Imaging, pp 78800P–78800P-15
34. Kurniawan F, Rahim MSM, Khalil MS, Khan MK (2016) Statistical-based audio forensic on identical microphones. International Journal of Electrical and Computer Engineering (IJECE) 6:2211–2218
35. Lim JS, Oppenheim AV (1979) Enhancement and bandwidth compression of noisy speech. Proc IEEE 67: 1586–1604
36. Liu Q, Sung AH, Qiao M (2010) Detection of double MP3 compression. Cogn Comput 2:291–296

37. Luo D, Yang R, Huang J (2015) Identification of AMR decompressed audio. Digital Signal Processing 37: 85–91
38. Lv Z, Hu Y, Li C-T, Liu B-B (2013) Audio forensic authentication based on MOCC between ENF and reference signals. In: Signal and Information Processing (ChinaSIP), 2013 I.E. China Summit & International Conference on, pp 427–431
39. Maher R (2009) Audio forensic examination. Signal Processing Magazine, IEEE 26:84–94
40. Maher RC (2010) Overview of audio forensics. In: Intelligent multimedia analysis for security applications. Springer, vol. 282, pp. 127–144
41. Malik H (2013) Acoustic environment identification and its applications to audio forensics. Information Forensics and Security, IEEE Transactions on 8:1827–1837
42. Malik H, Farid H (2010) Audio forensics from acoustic reverberation. In: Acoustics Speech and Signal Processing (ICASSP), 2010 I.E. International Conference on, pp 1710–1713
43. Malik H, Zhao H (2012) Recording environment identification using acoustic reverberation. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 I.E. International Conference on, pp 1833–1836
44. Mansour MF (2009) Strategies for bit allocation reuse in audio transcoding. In: ICASSP, pp 157–160
45. Mansour MF (2012) A transcoding system for audio standards. IEEE transactions on multimedia 14: 1381–1389
46. McAulay R, Malpass M (1980) Speech enhancement using a soft-decision noise suppression filter. Acoustics, Speech and Signal Processing, IEEE Transactions on 28:137–145
47. Moon C-B, Kim H, Kim BM (2014) Audio recorder identification using reduced noise features. In: Ubiquitous information technologies and applications, Springer, pp 35–42
48. Muhammad G, Alotaibi YA, Alsulaiman M, Huda MN (2010) Environment recognition using selected MPEG-7 audio features and Mel-Frequency Cepstral Coefficients. In: Digital Telecommunications (ICDT), 2010 Fifth International Conference on, pp 11–16
49. Nikias CL (1993) Higher-order spectral analysis. In: Engineering in Medicine and Biology Society, 1993. Proceedings of the 15th Annual International Conference of the IEEE. pp 319–319
50. Olanrewaju R, Khalifa O (2012) Digital audio watermarking; techniques and applications, In: Computer and Communication Engineering (ICCCE), 2012 International Conference on,pp 830–835
51. Owen T (1996) AES recommended practice for forensic purposes-managing recorded audio materials intended for examination. J Audio Eng Soc 44(4):275
52. Paliwal K, Wójcicki K, Schwerin B (2010) Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm 52:450–475
53. Qiao M, Sung AH, Liu Q (2010) Revealing real quality of double compressed MP3 audio. In: Proceedings of the international conference on multimedia, pp 1011–1014
54. Qiao M, Sung AH, Liu Q (2013) Improved detection of MP3 double compression using content-independent features. In: Signal Processing, Communication and Computing (ICSPCC), 2013 I.E. international conference on, pp 1–4
55. Rabiner LR, Schafer RW (1978) Digital processing of speech signals, vol 100. Prentice-hall, Englewood Cliffs
56. Ratnam R, Jones DL, Wheeler BC, O'Brien WD Jr, Lansing CR, Feng AS (2003) Blind estimation of reverberation time. The Journal of the Acoustical Society of America 114:2877–2892
57. Rodríguez DPN, Apolinário JA, Biscainho LWP (2010) Audio authenticity: detecting ENF discontinuity with high precision phase analysis. Information Forensics and Security, IEEE Transactions on 5:534–543
58. Shanmugasundaram K, Kharrazi M, Memon N (2004) Nabs: a system for detecting resource abuses via characterization of flow content type. In: Computer security applications conference, 2004. 20th Annual, pp 316–325
59. Sharma D, Naylor PA, Gaubitch ND, Brookes M (2012) Non intrusive codec identification algorithm. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 I.E. international conference on, pp 4477–4480
60. Soulodre GA (2010) About this dereverberation business: A method for extracting reverberation from audio signals. In: Audio Engineering Society Convention 129
61. Takagi K, Miyaji S, Sakazawa S, Takishima Y (2006) Conversion of MP3 to AAC in the compressed domain. In: Multimedia Signal Processing, 2006 I.E. 8th Workshop on, pp 132–135
62. Tsoukalas DE, Mourjopoulos JN, Kokkinakis G (1997) Speech enhancement based on audible noise suppression. Speech and Audio Processing, IEEE Transactions on 5:497–514

63. Weiss M, Aschkenasy E, Parsons T (1975) Study and development of the INTEL technique for improving speech intelligibility. DTIC Document
64. Yang R, Qu Z, Huang J (2008) Detecting digital audio forgeries by checking frame offsets. In Proceedings of the 10th ACM workshop on multimedia and security, pp 21–26
65. Yang R, Shi Y-Q, Huang J (2009) Defeating fake-quality MP3. In: Proceedings of the 11th ACM workshop on multimedia and security, pp 117–124
66. Yang R, Shi YQ, Huang J (2010) Detecting double compression of audio signal. In: IS&T/SPIE electronic imaging, pp 75410 K–75410 K-10
67. Yang R, Qu Z, Huang J (2012) Exposing MP3 audio forgeries using frame offsets. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 8:35
68. Yiu K-K, Mak M-W, Kung S-Y (2003) Environment adaptation for robust speaker verification. In: INTERSPEECH
69. Zhang Y, Zhao Y (2013) Modulation domain blind speech separation in noisy environments. Speech Comm 55:1081–1099
70. Zhao H, Malik H (2012) Audio forensics using acoustic environment traces. In: Statistical Signal Processing Workshop (SSP), 2012 IEEE, 2012, pp 373–376
71. Zhao H, Malik H (2013) Audio recording location identification using acoustic environment signature. Information Forensics and Security, IEEE Transactions on 8:1746–1759
72. Zhao H, Chen Y, Wang R, Malik H (2014) Audio source authentication and splicing detection using acoustic environmental signature. In: Proceedings of the 2nd ACM workshop on Information hiding and multimedia security, pp 159–164
73. Zhou J, Garcia-Romero D, Espy-Wilson CY (2011) Automatic speech codec identification with applications to tampering detection of speech recordings. In proceedings of Interspeech, Florence, Italy, August, 2011, pp. 2533–2536

Mr. Mohammed Zakariah is a Research Assistant of Computer Science department in the College of Computer and Information Sciences, King Saud University, Riyadh, Kingdom of Saudi Arabia. His research interest includes Digital audio Forensics, cloud computing, multimedia, healthcare and social media.

Dr. Muhammad Khurram Khan is currently working as a Full Professor at the Center of Excellence in Information Assurance (CoEIA), King Saud University, Kingdom of Saudi Arabia. He is one of the founding members of CoEIA and has served as the Manager R&D from March 2009 to March 2012. He developed and successfuly managed the research program of CoEIA, which transformed the center as one of the best centers of research excellence in Saudi Arabia as well as in the region. Prof. Khurram is the Editor-in-Chief of a well-esteemed international journal '*Telecommunication Systems*' published by Springer. He has published over 275 research papers in the journals and conferences of international repute. In addition, he is an inventor of 10 US/PCT patents. He has edited 7 books/proceedings published by Springer-Verlag and IEEE. He has secured several national and international research grants in the domain of information security. His research areas of interest are Cybersecurity, digital authentication, biometrics, multimedia security, and technological innovation management. He is a Fellow of the IET (UK), Fellow of the BCS (UK), Fellow of the FTRA (Korea), senior member of the IEEE (USA), a member of the IEEE Technical Committee on Security & Privacy, and a member of the IEEE Cybersecurity community.

Dr. Hafiz Malik is Associate Professor at University of Michigan-Dearborn, his Research Interests are Digital Forensics, Wireless Sensor Network Security, Video Surveillance, Multimedia & Biometric Security, Steganalysis, Multimedia Signal Processing, Adaptive Filtering, Blind Source Separation, Pattern Recognition, and Machine Learning.