Detection of Impostor and Tampered Segments in Audio by using an Intelligent System

Zeeshan Mubeen[1]

[1]Department of Computer Science and IT, The University of Lahore, Lahore 54000, Pakistan

Email: zeeshan_mubeen@ymail.com

Mehtab Afzal[1]

[1]Department of Computer Science and IT, The University of Lahore, Lahore 54000, Pakistan

Email: mehtab.afzal@cs.uol.edu.pk

Zulfiqar Ali[2] (Corresponding Author)

[2]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 0HL, United Kingdom

Email: z.ali@essex.ac.uk

Suleman Khan[3]

[3]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8QH, United Kingdom

Email: Suleman.khan@northumbria.ac.uk

Muhammad Imran[4]

[4]College of Applied Computer Science, King Saud University, Riyadh 11451, Saudi Arabia

Email: cimran@ksu.edu.sa

ABSTRACT

The transmission of audio data via the Internet of Things makes such data vulnerable to tampering. Moreover, the availability of sophisticated tampering tools has allowed mobsters to change the context of audio data by altering their segments. Tampered audio may result in unpleasant situations for any member of society. To avoid such circumstances, a new audio forgery detection system is proposed in this study. This system can be deployed in edge devices to identify impostors and tampering in audio data. The proposed system is implemented using state-of-the-art mel-frequency cepstral coefficient features. Meanwhile, a Gaussian mixture model is used to train and validate the system. To evaluate the proposed system, a dataset of tampered audios is created by mixing recordings from two different speakers. The performance of the proposed system in authenticating genuine audio is between 92.50% and 100%, and that in detecting forged audio is between 99.90% and 100%.

Keywords: Audio forgery, splicing, audio forensic, edge computing, machine learning, voice activity detection, audio authentication.

1 INTRODUCTION

Significant improvements in multimedia technology and the enhanced functionality of editing tools have made the forgery of multimedia content an easy task. Approximately 300 million multimedia contents circulate every day on social media to disseminate information. Editing these contents may create adverse circumstances in a person's life. Similar to growing concerns about security in other areas [1], the security of multimedia contents also requires attention. Although watermark-based blind forgery detection techniques have been proposed in the literature [2], they are inapplicable to detecting forgery in audio data because watermarks distort audio signals. Moreover, the recognition rate should not be degraded in case of noise attacks [3].

Electronic proofs are among the important applications of digital multimedia contents in judicial authentication. Therefore, when audio forensics is used in court forensics, the legitimacy, authenticity, and relevance of audio signals must be verified. Moreover, an efficient audio authentication system is highly required for accurate and reliable tampering detection given the considerable use of audio signals in our daily life.

Audio can be tampered using different operations, such as the removal, replacement, and replication (copy–move) of audio segments [4]; and the tampered audio can be used for many purposes. For example, prerecorded audio conversations can be tampered to gain unauthorized access to services, illegally obtain money, or used information as an alibi in the court of law.

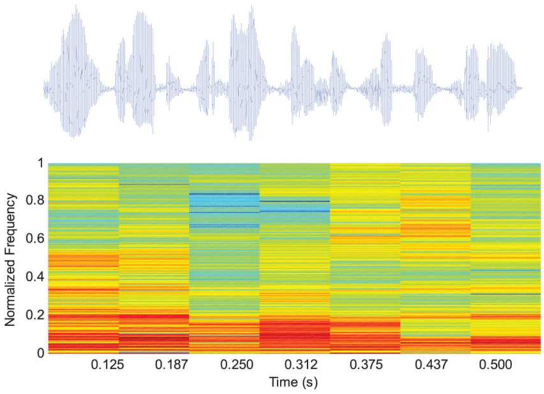


Fig. 1: Forged audio generated by splicing and its spectrogram, which shows no irregularities and discontinuities of tampering [5].

One of the common techniques for generating forged audio signals is splicing. This method generates forged audio by mixing words from different audio tracks [5]. The detection of splicing is a more challenging task compared with detecting copy–move forgery because audio signals from different environments and microphones are involved in the former [6]. By contrast, the contents of a recording are copied and moved to another place within the same audio in copy–move forgery. Thus, no complications due to different microphones and recording backgrounds will appear. One of the basic techniques for detecting forgery is the visual inspection of the spectrogram of forged audio [7]. In this approach, audio tampering can be detected by analyzing irregularities and discontinuities in the spectrogram. However, this approach is unreliable due to the availability of sophisticated tools that can easily cover up tampering without leaving any trace, as shown in Fig. 1.

Several solutions for forgery detection and audio authentication using different approaches have been reported in the literature. In [8], microphones and background environments were classified to detect forgery by determining the origin of an audio streaming. In [9], the existence of more than one microphone in audio was determined to detect forgery. Before extracting the mel-frequency cepstral coefficients (MFCCs), a relative spectral transformation (RASTA) filter was applied to the audio to cancel the channel effect. Then, the result was passed to the Gaussian mixture model (GMM) for classification. In [10], the verification of frame offsets was used to detect insertion, deletion, substitution, and splicing in MP3 audio. Environmental signatures and noise level estimations were also used successfully to detect splicing. RASTA–MFCC was implemented with GMM for audio source authentication and splicing detection by using environmental signatures in [11, 12]. Meanwhile, global and local noise estimations were used in [13] to detect splicing in audio. In addition, reverberation, which is an important parameter for describing the acoustics of a room, was used in [14] to detect forgery in audio.

The following questions arise from the preceding discussion.

- If the environment of audio tracks used for splicing is the same, is detecting forgery using the techniques based on environment classification possible?
- Similarly, if audio tracks recorded with the same microphone are used for splicing, will microphone classification techniques be good at detecting forgery?

- Moreover, if splicing is performed using tracks recorded in the same room and with the same equipment, then the environment signatures and reverberations of all the segments of forged audio will be the same. Thus, will the techniques based on environment signatures and reverberations be able to capture the forgery?

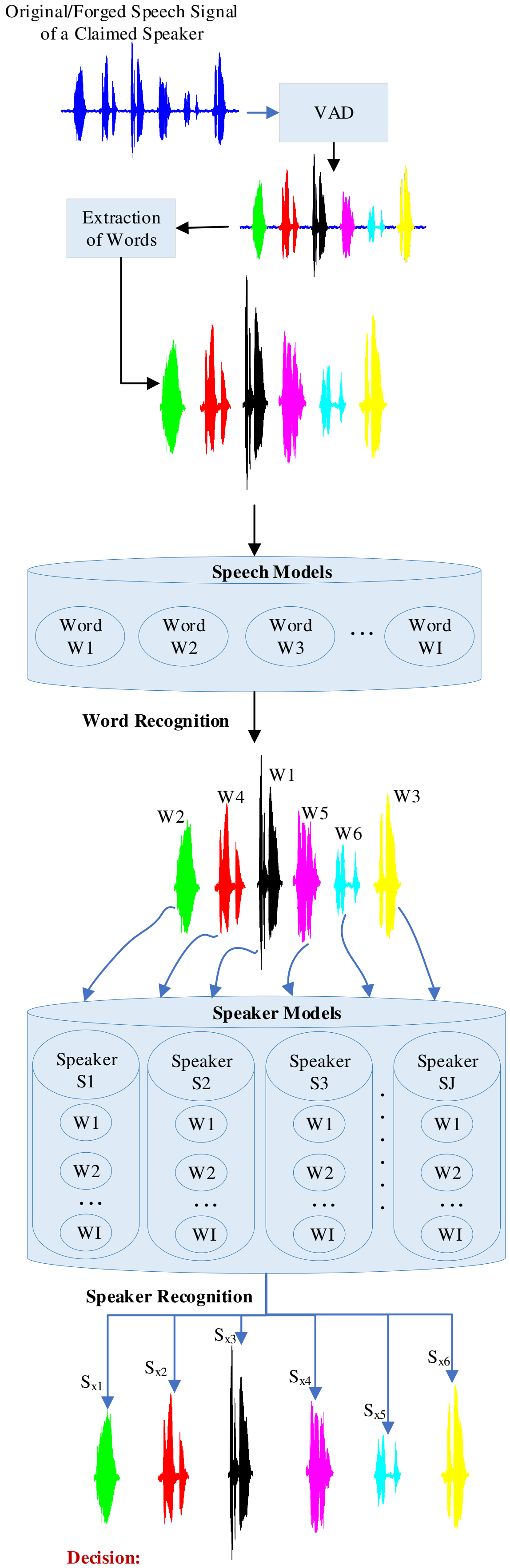
To answer these questions, a forgery detection system that can capture an impostor and tampered segments in audio is designed and developed in the current study. This system can authenticate audio if all the segments belong to the claimed speaker. However, if any of the segments does not belong to the claimed speaker, then the system detects the tampered segments and does not authenticate the audio.

Audio can be collected through the same/different Internet of things (IoT) and edge devices deployed at the same/different locations [15-18]; therefore, the system is evaluated using audio samples generated through the splicing of (i) tracks recorded in the same environment by using the same equipment, (ii) tracks recorded in the same environment by using different equipment, and (iii) tracks recorded in different environments by using different equipment. In each case, the system is evaluated using several sentences produced from words in a specific dictionary. One of the positive aspects of the system is that it is trained with the original audio of speakers. Forged audio is not used in the testing. The system is trained by generating speech models of each word in the dictionary and speaker models of every registered speaker. This study is a continuation of the work presented in [5], wherein forged tracks were used to train an authentication system. This previous system was capable of authenticating audio. However, in contrast with the proposed forgery detection system, it was unable to detect tampered segments in audio.

The remainder of this paper is organized as follows. Section 2 describes the proposed forgery detection system. It presents the speech database used to generate forged audio, the method for generating forged audio, the creation and optimization of speech and speaker models, and the process for authenticating audio. Section 3 provides the experimental setup and results, including the description of recording equipment and the generation of forged sentences. Section 4 presents the experimental results and an analysis of the reliable decisions of the developed system. It also highlights the findings of the study and compares the developed system with existing systems. Lastly, Section 5 provides some conclusions.

2 PROPOSED FORGERY DETECTION SYSTEM

The block diagram of the proposed forgery detection system is shown in Fig. 2. This system takes an original or forged speech signal as input and processes it to determine whether it is a tampered or original signal from the claimed speaker. The complete system is based on the following major components: the selection of an appropriate speech database with diverse characteristics, the generation of forged speech signals using a voice activity detection (VAD) module with multiple



Decision:

If all the identified speakers are the same as the claimed speaker, then the signal is original; otherwise, it is forged.

In case of a forged signal, the words spoken by the impostor are the forged parts of the input signal.

Fig. 2: Block diagram of the proposed forgery detection system.

measures, the computation of speech and speaker models by using feature extraction and machine learning approaches, the optimization of the generated models, and finally, the authentication of audio and the detection of forged segments in case of tampered audio. The design and implementation of these components are described in the subsequent section.

2.1 Speech Database Selection

Most databases are released by the Linguistic Data Consortium and the European Language Resource Association for academic and commercial uses. A database should be selected on the basis of its diversity in terms of recording equipment and environment. Moreover, it must have a good number of speakers with recordings in multiple sessions. Therefore, many databases are analyzed to select a good one for this study.

Some databases are good in certain aspects but lacking in other characteristics. For example, in a speech database in Castilian Spanish called AHUMADA, a variety of texts are recorded using different microphones and sound cards in six sessions. However, all the recordings are performed in one environment. This database is not useful for our study because a signal can be tampered using recordings from different environments. Similarly, the speech database POLYCAST is recorded in two different environments with six sessions per speaker; however, the recording equipment exhibits no variety (i.e., the microphones and sound cards are the same). This database is unsuitable for this study because forgery can be realized using recordings from different equipment. Likewise, the TIMIT speech corpus contains recordings of 10 phonetically rich sentences recorded by 630 speakers in 8 dialects of American English. One of the reasons for not using the TIMIT corpus is the difficulty in detecting the end points of spoken words. Given that it contains running speech (sentence), end point detection becomes a challenging task and demands a sophisticated VAD module that is beyond the scope of the current study. Other limitations of the TIMIT corpus include the lack of variety in recording equipment and environments, which is crucial for a forgery detection system.

For this study, a database with a good number of speakers recorded in different environments and sessions by using various recording equipment is necessary. A comparison of speech databases in terms of the number of speakers and dialects and the variety of text, recording equipment, and environment was presented in [19]. The King Saud University Arabic Speech Database (KSUD) meets all the aforementioned requirements. It has recordings of a good number of speakers with different ethnicities, a variety of text, four different recording equipment, and three environments. These diverse characteristics of KSUD justify its use in the current study. KSUD was also used in the development of forgery detection systems in [5, 6].

2.2 Tampering Using the Speech of a Genuine Speaker and an Impostor

Although some datasets, such as ASVspoof 2017 and ASVspoof 2019, are available, the primary objective of these datasets is to encounter a replay attack, which differs from the underlying problem of this study. In the current work, we are developing a system for detecting the presence of an impostor in audio, and to the best of our knowledge, no dataset exists to build such systems. Therefore, a dataset of forged signals is developed using KSUD in the current study.

The proposed system must detect a forgery accomplished using the speech signals of more than one speaker recorded in various environments by using different equipment. Therefore, forged signals are generated using the speech signals of two different speakers who are

- recorded in the same or different environments.
- recorded using the same or different microphones.

Both options are crucial for the proposed system because an impostor may use the same or different environments and/or equipment to deceive a system and gain unauthorized access. However, the use of such signals for generating tampered audio makes the task more challenging. For example, a signal recorded in a soundproof room contains nearly no noise, while a signal recorded in office environments may have noise with low or significant intensities. In such situation, the accurate detection of the boundaries of spoken words becomes a challenging task [20]. The accurate identification of spoken words for tampering is crucial, such that the human ears cannot perceive it. Although an improvement in the extraction of boundaries is not under the scope of the current study, the best effort is still exerted to perform the task by implementing a good VAD method [5]. Moreover, if boundaries are not detected accurately in the signals, then they are not included in the experiments. The steps for generating a forged speech signal are illustrated in Fig. 3.

For a tampered signal, two signals from two speakers are selected. One of the signals is from a genuine user of an authorized service, whereas the other signal is from an impostor who wants access by breaching the authentication system. A VAD method based on volume and zero crossing is used to determine the end points of a word. For a speech signal divided into *n* segments, the volume *V* of a segment is the sum of amplitudes of all the samples in it and is defined by Eq. 1. Once the volumes of all the segments are computed in a speech signal, segments whose volumes are higher than an adaptive threshold *adaptiveTh* (given by Eq. 2) are designated as voiced frames; otherwise, they are designated as unvoiced.

$$V = \sum_{i=1}^{n} \left| a_i \right|, \tag{1}$$

$$adaptiveTh = 3\% \text{ of } \left(V_{\max} - V_{\min} \right) + V_{\min}, \tag{2}$$

where a_{i} is the amplitude of the i^{th} sample of a speech segment; and $V_{\max}$ and $V_{\min}$ represent the maximum and minimum volumes in a speech signal, respectively. In general, 10 or more consecutive voiced segments for a speech signal recorded at 16 kHz comprise a word. Zero crossing represents the number of times a speech signal crosses zero amplitude. By using the computed volume and zero crossing, the extracted words are combined to create a tampered signal.

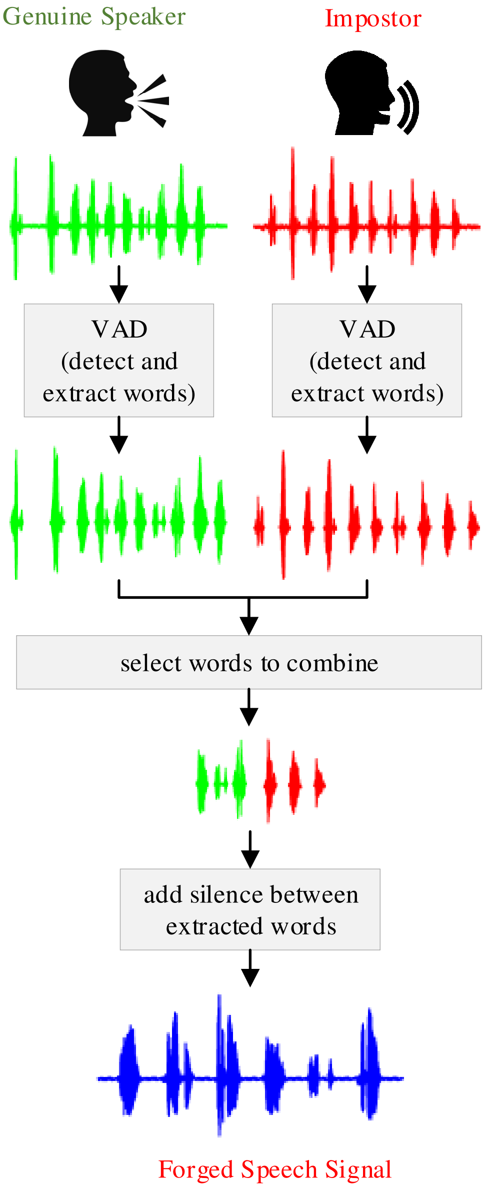


Fig. 3: Generation of a forged speech signal.

2.3 Generation of Speech and Speaker Models by Using Original Speech Signals

An impostor can use any audio for tampering; therefore, tampered audio cannot be used for generating speech and speaker models. Only the original audios of registered speakers (who are legally authorized to access a service) are processed to generate models. Thereafter, these models are used to determine the originality and tampering of audio.

To generate speech models, each spoken word in the dictionary is considered. For example, if a user has to utter a code comprising of digits to gain access, then the dictionary will contain digits from zero to nine. In the current study, the dictionary contains one to nine Arabic digits that are listed in Table 1 with their English translation and international phonetic alphabets.

Table 1: Arabic digits with English translation and international phonetic alphabets

Arabic digits	Translation	International phonetic alphabets
واحد	One	/w/, /a/, /ħ/, /i/, /d/
أثنين	Two	/a/, /th/, /n/, /a/, /y/, /n/
ثلاثة	Three	/th/, /a/, /l/, /ā/, /th/, /a/
أربعة	Four	/a/, /r/, /b/, /ʕ/, /a/
خمسة	Five	/kh/, /a/, /m/, /s/, /a/
ستة	Six	/s/, /i/, /t/, /t/, /a/
سبعة	Seven	/s/, /a/, /ɓ/, /ʕ/, /a/
ثمانية	Eight	/th/, /a/, /m/, /ā/, /n/, /y/, /a/
تسعة	Nine	/t/, /i/, /s/, /ʕ/, /a/

A model of each word is generated. To achieve this, the speech features of a certain word are extracted for all the speakers. We extract MFCC features in this study because of its success in various speech-related applications [21]. The major steps for computing MFCC are as follows: the segmentation of a signal into overlapping frames; the implementation of a hamming window to taper the ends of a segment and avoid spectral leakage during Fourier transform (FT); the application of FT to determine the contribution of each frequency component; and the use of mel-spaced band-pass filters' bank, which is one of the principles of human auditory perception. The latter is the reason why MFCC features are used to simulate the human auditory system and have been widely adopted in different speech-related applications. Lastly, discrete cosine transformation is applied to decorrelate the computed features because it is an important step when passing the features to GMM for the acoustic modeling of words.

GMM is a state-of-the-art machine learning approach, and it has been extensively used in the literature for various scientific domains [22]. The rationale for adopting GMM is to model the extracted features by using a mixture of Gaussian densities. A Gaussian mixture density is a weighted sum of *M* component densities given by

$$p(X \mid \Theta) = \sum_{i=1}^{M} w_i \cdot g(X \mid \mu_i, \Sigma_i), \qquad (3)$$

where μ_i, Σ_i, and w_i are the mean vector, covariance matrix, and weight (prior probability) of the i^{th} Gaussian component, respectively; and $i = 1, 2, 3, \ldots, M$. These parameters are initialized using the *k*-means algorithm and tuned using the well-known expectation–maximization algorithm [23]. The calculated features are represented by the D-dimensional data vector *X*, and the density of each component is given by a D-dimensional Gaussian function with the following form:

$$g\left(X \mid \mu_i, \Sigma_i\right) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (X - \mu_i)^T \Sigma^{-1} (X - \mu_i) \right). \quad (4)$$

The weights of the Gaussian components satisfy the following constraint:

$$\sum_{i=1}^{M} w_i = 1 \quad \text{and} \quad 0 \leq w_i \leq 1 . \quad (5)$$

One model $\Theta = (\mu_i, \Sigma_i)$ of each word is generated and stored. Similarly, models of all the registered users are generated and stored. To improve the accuracy of the forgery system, text-dependent speaker recognition is performed. That is, a model of each speaker of every word in the dictionary is necessary. The next step is to optimize the generated speech and speaker models to ensure that they are providing maximum accuracy.

2.4 Optimization of Speech and Speakers Models

The speech models of all the words (one to nine digits) are generated using 8, 16, 32, and 64 Gaussian mixtures. The models are evaluated by using unknown original signals of all the speakers; these signals are not used during model generation (also known as the enrollment/training phase). The metric for observing the performance of the generated model is given by Eq. 6.

$$\text{accuracy} = \frac{\text{truly detected signals}}{\text{total number of signals}} \times 100 \quad (6)$$

Table 2: Obtained speech recognition accuracies for each digit by using 8, 16, 32, and 64 Gaussian mixtures

Digits	Number of Gaussian mixtures			
	8	16	32	64
One	97%	98%	98%	99%
Two	83%	89%	91%	92%
Three	85%	87%	88%	93%
Four	90%	91%	92%	93%
Five	86%	93%	95%	95%
Six	83%	90%	93%	97%
Seven	75%	77%	85%	85%
Eight	89%	93%	97%	98%
Nine	68%	79%	84%	92%

As shown in Table 2, the accuracy of the digits increases with an increasing number of mixtures. That is, accuracy can still be improved. Therefore, more experiments are conducted to determine the optimal speech models, and the results are provided in Table 3. The obtained results show that the models with 256 mixtures exhibit the best accuracy.

Table 3: Obtained speech recognition accuracies for each digit by using 128 and 256 Gaussian mixtures

Digit	Number of Gaussian mixtures	
	128	256
One	100%	100%
Two	94%	98%
Three	95%	98%
Four	96%	99%
Five	98%	99%
Six	99%	99%
Seven	92%	97%
Eight	99%	100%
Nine	96%	99%

Similarly, different experiments are conducted using 4, 8, 16, 32, 64, and 128 Gaussian mixtures to determine the optimal model for speakers. The obtained results are presented in Table 4.

Table 4: Obtained speaker recognition accuracies for each digit by using 4, 8, 16, and 32 Gaussian mixtures

Digits	Number of Gaussian mixtures			
	4	8	16	32
One	85%	97%	100%	100%
Two	81%	96%	100%	100%
Three	82%	97%	100%	100%
Four	85%	99%	100%	100%
Five	81%	95%	100%	100%
Six	64%	92%	99%	100%
Seven	79%	95%	100%	100%
Eight	82%	97%	100%	100%

Nine	86%	97%	100%	100%

Notably, optimal models of speech recognition are generated using 256 mixtures and those of speaker recognition are obtained with 32 mixtures because of the amount of data provided to GMM. In speech recognition, a model is generated using the features of a digit of all the speakers. In speaker recognition, however, the features of one speaker for a specific digit are used to generate models. The generated models are used to authenticate audio and determine whether it is from a genuine speaker; otherwise, the tampered segments of the audio are identified.

2.5 Audio Authentication and Tampering Detection

When the proposed system receives audio for authentication or the detection of forgery, the first step is to identify the boundary points of every segment, as shown in Fig. 2. Then, the system computes the MFCC features of a segment and compares them with the generated models of all the words to determine the spoken word. Subsequently, the MFCCs are compared with all the speakers' models of that word. In this manner, the system recognizes the speaker who utters the segment. This process is repeated for all the segments of the audio. If all the segments are spoken by the claimed speaker, then the audio is genuine. Otherwise, it is forged, and the system highlights the tampered segments. To evaluate the proposed forgery detection system, the experimental setup and the obtained results are described in the following section.

3 Experimental Results and Analysis

3.1 Experimental Setup

To generate forged audio through splicing, two environments (ENV) are considered. The first is a normal office environment, which is represented by OF. The second is a soundproof room, a quiet environment without background noise, and it is denoted by SR. The pieces of recording equipment (EQP) used in each environment in KSUD are listed in Table 5.

Table 5: Recording equipment used in the office and soundproof room

Equipment	**Office**	**Soundproof Room**
	Make and Model	*Make and Model*
Microphone	SHUR Beta 58A	SHUR PG 42
Microphone	Sony F-V220	SHUR Beta 58A
Sound Mixer	Yamaha MW-12CX	Yamaha MW-12CX
Sound Card	Creative 5.1 Surrounding	Creative 5.1 Surrounding

The audio recording equipment is a combination of a microphone and a sound card/mixture. We use three different recording equipment and two environments to create tampered audio. In KSUD, two different pieces of equipment are selected for the OF environment, namely, Sony F-V220 with Creative 5.1 Surrounding (denoted by MC1) and SHUR Beta 58A with Yamaha MW-12CX (denoted by YM1). Similarly, the selected equipment for the SR environment is SHUR PG 42 with Creative 5.1 Surrounding (denoted by MC2) and SHUR Beta 58A with Yamaha MW-12CX (also denoted by YM1).

The boundary points of digits are easier to determine compared with those of running speech. Therefore, digits from one to nine are used to generate forged audio. The following four sentences are created by mixing two original audios of two different speakers. The numbers of genuine and tampered audios used for the experiments are provided in Table 6.

Sentence 1: [digit1 digit2 digit3 digit4 digit5 digit6]

Sentence 2: [digit2 digit3 digit4 digit5 digit6 digit7]

Sentence 3: [digit3 digit4 digit5 digit6 digit7 digit8]

Sentence 4: [digit4 digit5 digit6 digit7 digit8 digit9]

Table 6: Statistics of genuine and tampered audios

ENV	EQP	Number of Audios		Total
		Genuine	Forged	Audios
OF	MC1	332	3956	4288
	YM	320	3640	3960
SR	MC2	160	3132	3292
	YM	328	3936	4264
OF	*MC1-YM	--	3840	3840
SR	*MC2-YM	--	4372	4372
OF-SR	*MC1-YM	--	3980	3980
Total Audios		*1140*	*26856*	*27996*

* denotes that two different pieces of equipment are mixed to generate forged audio. Therefore, original audio is not possible in this case.

Overall, 27996 audio recordings are used to conduct several experiments for evaluating the developed forgery detection system. The number of forged audios is considerably larger than that of genuine audios. The reason for this result is that a speaker is mixed with many other speakers to generate forged audio, whereas this condition is impossible in the case of genuine audio.

Similar to a real-life scenario, forged audio is unknown to the system. Therefore, forged audio is not used to train the system. It is only used to observe the accuracy and reliability of the system. The developed system is trained using original samples only.

3.2 Tampering of an Original Speaker with Multiple Speakers

To validate the reliability of the developed system, the audio of a genuine speaker is forged with multiple impostors. For example, 115 different impostors are mixed with speaker NS15 to generate forged audio. All the impostors are listed in Fig. 4. That is, only 1 genuine audio of NS15 exists, whereas 115 audios are forged. In such a scenario, identifying the genuine audio in the presence of a large number of tampered audios becomes challenging for the developed system because the voices of some impostors may be extremely similar to that of the claimed speaker. Moreover, the system recognizes the text spoken by the original speaker and the impostor for each audio.

A tampered audio of claimed speaker NS15 is presented to the system for authentication. The contents of this audio are digit2 to digit7. First, the digits are extracted from the audio, and the MFCCs of each digit are computed. Then, the features of these digits are compared with the models of all the generated words, i.e., the models of digits 1 to 9. If the first recognized word is digit2, then it will be compared with the model of digit2 of all the registered speakers to identify the speaker. Suppose that the generated models of speakers 1 to 115 for digit2 are Θ_{21}, Θ_{22}, Θ_{23}, …, Θ_{2k} where $K = 1, 2, 3, \ldots, 115$; and X represents the test utterance of digit2 of an unknown speaker. Then, 115 log-likelihood (LLH) values, log $p(X|\Theta_{2K})$, are computed by comparing the extracted MFCC of X with each model. The decision regarding the identity of the speaker is based on these computed LLH values. All the LLH values of speaker NS15 with all the impostors are presented in Fig. 4.

As shown in Fig. 4, the maximum LLH of digit2 is through NS15, indicating that this digit belongs to the claimed speaker. Similarly, the computed features of each extracted digits are compared with all the registered speakers and the system decides their originality. If all the digits belong to the same speaker, then the system authenticates the claimed speaker and authorizes access to the services. By contrast, if the system determines the existence of more than one speaker in the audio, then the audio is forged. The system also highlights the digits that are not spoken by the claimed speaker.

The LLH values of all the digits of the input audio with every impostor are depicted in Fig. 5. The claimed speaker is shown in green, and the impostors are highlighted in red. Moreover, the speakers identified by the system apart from the claimed speaker are shown in blue. In the case of digit2, the claimed and identified speakers are the same. For digit3, the LLH of S19 (blue) is greater than that of NS15 (green). Therefore, the system determines that this digit is not spoken by the claimed speaker. Similarly, for digits4 and digit 7, the LLH of S19 is greater than that of NS15. A conclusion is drawn that the audio contains more than one speaker. Ultimately, the system will not authenticate the claimed speaker and decline the request to access the required service.

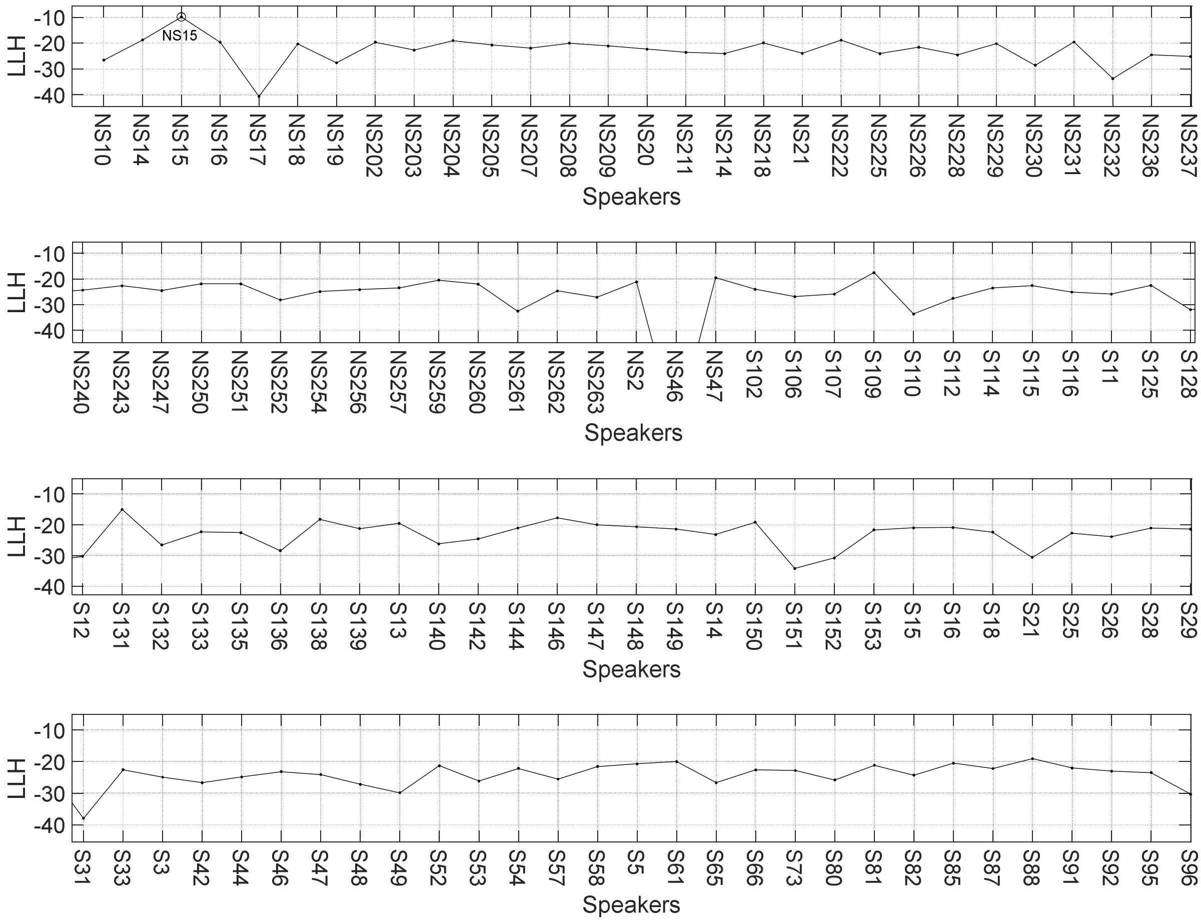


Fig. 4: Computed LLH of digit2 of speaker NS15 with all the impostors.

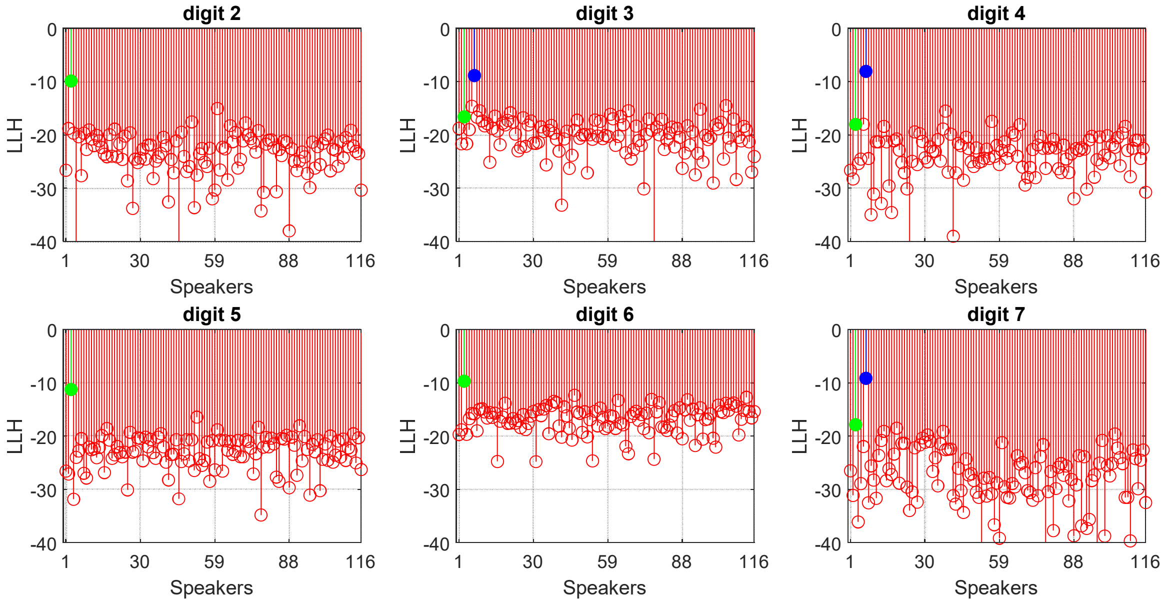


Fig 5: Computed LLH values of digit2 to digit7 of speaker NS15 with 115 impostors.

3.3 Recognition of Original and Tampered Audios

Various experiments are conducted to observe the performance of the developed forgery detection system by using all the sentences. For each sentence, the recognition rates of both environments (ENV) are determined, i.e., office (OF) and soundproof rooms (SR). Moreover, the recordings of two different pieces of equipment (EQP) are used for each environment to evaluate the system. The recognition rates of the developed system using Sentences 1 to 4 are provided in Tables 7 to 10, respectively. The original audio is considered a positive class, whereas the forged audio is regarded as a negative class. The percentages of true positive (TP), false negative (FN), true negative (TN), and false positive (FP) are listed in each table.

Table 7: Recognition rates of original and tampered audios using Sentence 1

ENV	EQP	TP	FN	TN	FP
OF	MC1	98.80%	1.20%	99.90%	0.10%
	YM	96.25%	3.75%	100.0%	0.0%
SR	MC2	92.50%	7.50%	100.0%	0.0%
	YM	95.12%	4.88%	100.0%	0.0%

Table 8: Recognition rates of original and tampered audios using Sentence 2

ENV	EQP	TP	FN	TN	FP
OF	MC1	98.80%	1.20%	100.0%	0.0%
	YM	96.25%	3.75%	100.0%	0.0%
SR	MC2	92.50%	7.50%	100.0%	0.0%
	YM	95.12%	4.88%	100.0%	0.0%

Table 9: Recognition rates of original and tampered audios using Sentence 3

ENV	EQP	TP	FN	TN	FP
OF	MC1	100.0%	0.0%	100.0%	0.0%
	YM	96.25%	3.75%	100.0%	0.0%
SR	MC2	92.50%	7.50%	100.0%	0.0%
	YM	92.68%	7.32%	100.0%	0.0%

Table 10: Recognition rates of original and tampered audios using Sentence 4

ENV	EQP	TP	FN	TN	FP
OF	MC1	100.0%	0.0%	100.0%	0.0%
	YM	98.75%	1.25%	100.0%	0.0%
SR	MC2	92.50%	7.50%	100.0%	0.0%
	YM	95.12%	4.88%	99.90%	0.10%

TNs are 100% in most cases, because the audio is recognized as forged if any of its digits does not belong to claimed speakers. However, TPs vary from 92% to 99%. The system is strict in authenticating audio. If any segment of the audio is not recognized as a claimed speaker, then the system will not authenticate the user.

More experiments are performed to analyze the performance of the system when forged audio is generated using the recordings of different environments and equipment. The recognition results of the developed system when forged audio is created by mixing the recordings of three different pieces of equipment, namely, MC1, MC2, and YM, are presented in Table 11. The environment is the same in this case. Meanwhile, the results when forgery is created using audio from different environments recorded using various equipment are provided in Table 12.

Table 11: Recognition of tampered audio generated using audio from different equipment with Sentence 1

ENV	EQP	TN	FP
OF	MC1-YM	100.0%	0.0%
SR	MC2-YM	100.0%	0.0%

Table 12: Recognition of tampered audio generated using audio from different environments and equipment with Sentence 1

ENV	EQP	TN	FP
OFSR	MC1-YM	100.0%	0.0%

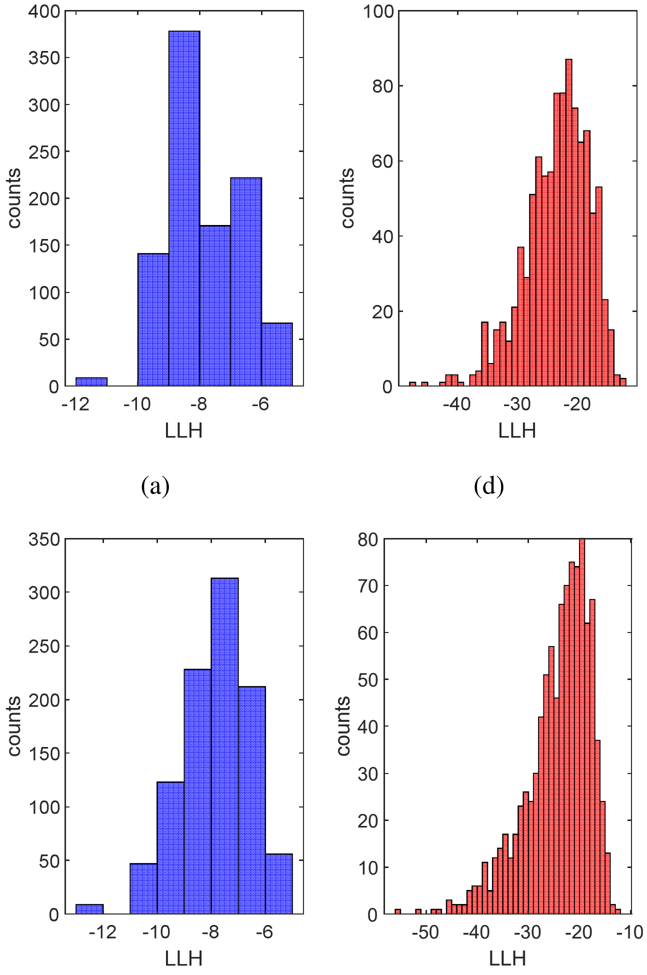
The developed system recognizes the forged audio with an accuracy of 100%. The criterion is the same; that is, if any segment does not belong to the claimed speaker, then the system labels the audio as forged. In Tables 10 and 11, the results are provided using Sentence 1 because the accuracy for the other sentences is the same. The analysis of the developed system with regard to recognizing forged audio is discussed in the subsequent section.

4 Analysis and Comparison

4.1 Analysis of Forged Audio Recognition

Determining the approach used by the developed system to declare audio as forged is crucial, and eventually, the request of a claimed speaker to access a service declines. To analyze the process, the distribution of the LLH values of 989 audios (Sentence 1) is depicted in Fig. 6. In each recording, digit1, digit4, and digit5 are spoken by genuine speakers (shown in blue in Fig. 6). By contrast, digit2, digit3, and digit6 are uttered by impostors (shown in red in Fig. 6).

When a digit of a genuine speaker is compared with its model, the computed LLH is extremely high. For example, approximately 99% of the LLH values of digit1 are found between −10 and −5, with an average value of −7.8. By contrast, when a digit of an impostor is compared with that of the claimed speaker, its LLH value is extremely low and falls within the range of −48 to −12. Although only 20 values are found between −15 and −12, 98% values are between −15 and −48. The same trend is observed for digit3; that is, all the LLH values are within the range of −56 to −12, as shown in Fig. 6(e), with 16 values within the range of [−15 −12]. The gaps between the computed LLH values make the system capable of differentiating between genuine and forged audios.

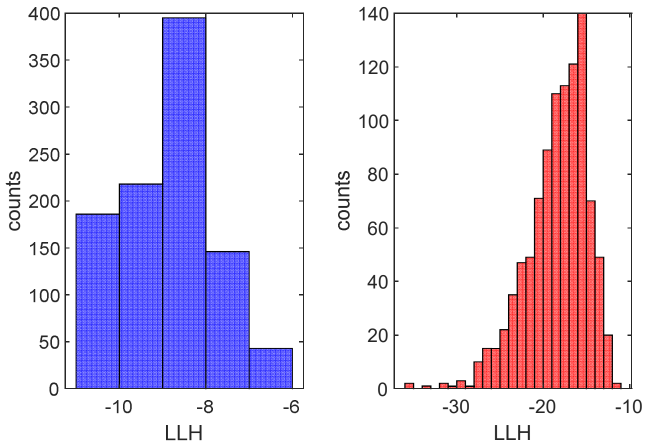


(a)

(d)

(b)

(e)



(c) (f)

Fig. 6: Distribution of LLH values for (a–c) digit1, digit4, and digit5 of genuine speakers, and (d–f) digit2, digit3, and digit 6 of impostors.

The slight overlap that occurs between the LLH values of genuine and forged audios (Fig. 7) is the reason for the decrease in the percentage of TPs. That is, this overlap introduces false rejection. However, false acceptance is virtually zero in all the cases, which is a positive aspect of the developed system. Such false acceptance value indicates that the system does not authenticate forged audio.

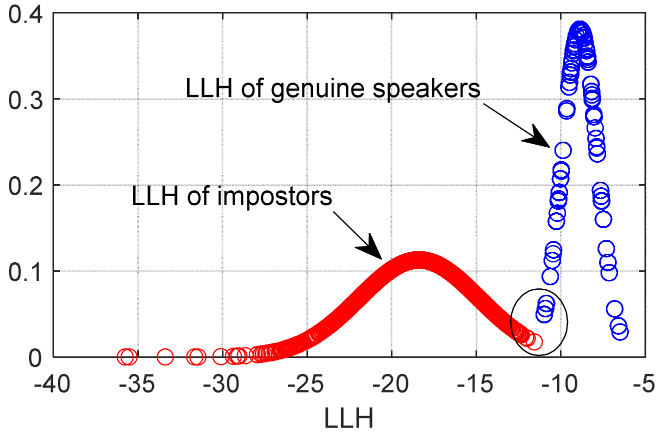


Fig. 7: Probability density function of the LLH values of genuine speakers and impostors for digit5 (genuine speaker) and digit6 (impostor).

4.2 Summary and Comparison

The developed system exhibits good performance for genuine and tampered audios. The maximum obtained results for the authentication of genuine audio and the detection of tampered audio is 100% each. Meanwhile, the minimum results for authentication and forgery detection are 92.50% and 99.90%, respectively. A decrease of 7.50% in authenticating genuine audio is due to the inaccurate detection of boundary points in SR. However, the overall average performance of both types of audio (genuine and forged) is 96.6% and 99.98%, respectively, which are better than those of many existing forgery detection systems.

The accuracy of the system developed in [8] was 75.9%, and it performed classification of the environment and the microphone to detect audio tampering. In addition to its lower accuracy, this system cannot detect forgery if splicing is performed using audios recorded in the same environment and with the same microphone. In another system [9], tampering was detected via microphone classification, and an accuracy of 95% was obtained. This approach is also inapplicable if forgery is performed using audio recorded with the same microphone. In [12], the accuracy of the system was 100% for detecting splicing in audio. However, such accuracy was provided only to tampered audio, which is also 100% for our proposed system. In addition, one of the limitations of this previous study is that the system requires training with forged audio. In our proposed system, training the system with forged audio is not required. When detecting splicing, the systems developed in [11], [24], and [4] achieved accuracies of 96%, 96.9%, and 85.54%, respectively. The comparison of existing studies with our proposed forgery detection system concludes that our system outperforms existing systems.

5 CONCLUSION

The proposed forgery detection system determines the existence of more than one speaker in audio to identify tampering. In addition, it highlights the tampered segments that do not belong to the claimed speaker, and it can detect forgery introduced by combining recordings from different equipment and environments. The experimental results show that the system is accurate and reliable in decision-making. The acceptance criterion for authentication is strict; that is, the system can reject genuine audio but will not authenticate forged audio. Given its high accuracy, the system can be used reliably in IoT platforms for authenticating audio. The proposed system can be used in various real-world applications, particularly when an impostor intends to deceive a deployed authentication system by tampering audio with a voice similar/close to the genuine speaker. This scenario may occur easily, because voice authentication systems process the entire audio to identify the claimed speaker. Hence, an authentication system cannot recognize that some segments belong to an impostor, and thus, grants access to required services to an unauthorized person. The proposed system can also be used to provide scientific evidence in situations wherein doubt that audio contains more than one speaker exists. Such evidence may be used in courts to support the decision to prove a person guilty. In the future, the proposed system will be implemented using a large dictionary that contains more common words and sentences.

REFERENCES

[1] M. A. Amanullah, R. A. A. Habeeb, F. H. Nasaruddin, A. Gani, E. Ahmed, A. S. M. Nainar*, et al.*, "Deep learning and big data technologies for IoT security," *Computer Communications,* vol. 151, pp. 495-517, 2020.

[2] O. Benrhouma, H. Hermassi, A. A. Abd El-Latif, and S. Belghith, "Chaotic watermark for blind forgery detection in images," *Multimedia Tools and Applications,* vol. 75, pp. 8695-8718, 2016.

[3] N. Wang, Q. Li, A. A. Abd El-Latif, T. Zhang, and X. Niu, "Toward accurate localization and high recognition performance for noisy iris images," *Multimedia Tools and Applications,* vol. 71, pp. 1411-1430, 2014.

[4] S. K. Rouniyar, Y. Yingjuan, and Y. Hu, "Channel Response Based Multi-Feature Audio Splicing Forgery Detection and Localization," presented at the Proceedings of the 2018 International Conference on E-Business, Information Management and Computer Science, Hong Kong, 2018, pp. 46 - 53.

[5] Z. Ali, M. Imran, and M. Alsulaiman, "An Automatic Digital Audio Authentication/Forensics System," *IEEE Access,* vol. 5, pp. 2994-3007, 2017.

[6] M. Imran, Z. Ali, S. T. Bakhsh, and S. Akram, "Blind Detection of Copy-Move Forgery in Digital Audio Forensics," *IEEE Access,* vol. 5, pp. 12843-12855, 2017.

[7] B. E. Koenig and D. S. Lacey, "Forensic Authentication of Digital Audio Recordings," *Journal of Audio Engineering Society,* vol. 57, pp. 662-695, 2009.

[8] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: a first practical evaluation on microphone and environment classification," presented at the Proceedings of the 9th workshop on Multimedia &amp; security, Dallas, Texas, USA, 2007, pp. 63 - 74.

[9] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, 2013, pp. 177-182.

[10] R. Yang, Z. Qu, and J. Huang, "Detecting digital audio forgeries by checking frame offsets," presented at the Proceedings of the 10th ACM workshop on Multimedia and security, Oxford, United Kingdom, 2008, pp. 21-26.

[11] H. Zhao, Y. Chen, R. Wang, and H. Malik, "Audio splicing detection and localization using environmental signature," *Multimedia Tools and Applications,* vol. 76, pp. 13897-13927, 2017.

[12] H. Zhao, Y. Chen, R. Wang, and H. Malik, "Audio source authentication and splicing detection using acoustic environmental signature," presented at the Proceedings of the 2nd ACM workshop on Information hiding and multimedia security, Salzburg, Austria, 2014, pp. 159 - 164.

[13] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 1841-1844.

[14] A. Ciobanu, T. Culda, C. Negrescu, and D. Stanomir, "Analysis of reverberation time blind estimation used in audio forensics," in *2014 11th International Symposium on Electronics and Telecommunications (ISETC)*, 2014, pp. 1-4.

[15] B. Chen, J. Wan, A. Celesti, D. Li, H. Abbas, and Q. Zhang, "Edge Computing in IoT-Based Manufacturing," *IEEE Communications Magazine,* vol. 56, pp. 103-109, 2018.

[16] H. Tang, D. Li, J. Wan, M. Imran, and M. Shoaib, "A Reconfigurable Method for Intelligent Manufacturing Based on Industrial Cloud and Edge Intelligence," *IEEE Internet of Things Journal,* vol. 7, pp. 4248-4259, 2020.

[17] B. Chen, J. Wan, Y. Lan, M. Imran, D. Li, and N. Guizani, "Improving Cognitive Ability of Edge Intelligent IIoT through Machine Learning," *IEEE Network,* vol. 33, pp. 61-67, 2019.

[18] C. Jiang, J. Wan, and H. Abbas, "An Edge Computing Node Deployment Method Based on Improved k-Means Clustering Algorithm for Smart Manufacturing," *IEEE Systems Journal,* pp. 1-11, 2020.

[19] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU Speech Database: Text Selection, Recording and Verification," in *Modelling Symposium (EMS), 2013 European*, 2013, pp. 237-242.

[20] Z. Ali and M. Talha, "Innovative Method for Unsupervised Voice Activity Detection and Classification of Audio Segments," *IEEE Access,* vol. 6, pp. 15494-15504, 2018.

[21] M. Alsulaiman, G. Muhammad, and Z. Ali, "Comparison of voice features for Arabic speech recognition," in *2011 Sixth International Conference on Digital Information Management*, 2011, pp. 90-95.

[22] Z. Ali, M. Imran, S. McClean, N. Khan, and M. Shoaib, "Protection of records and data authentication based on secret shares and watermarking," *Future Generation Computer Systems,* vol. 98, pp. 331-341, 2019.

[23] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review,* vol. 26, pp. 195-239, 1984.

[24] X. Lin and X. Kang, "Supervised audio tampering detection using an autoregressive model," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2142-2146.