RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt24613

Speech Signal Splicing Detection system based on MFCC and DTW

Venkata Lalitha Narla [a,*], Gulivindala Suresh [a], Mahesh K Singh [a], M. Vinod Kumar [b]

[a] Department of Electronics and Communication Engineering, Aditya University, Surampalem, Andhra Pradesh, India

[b] Department of Electronics and Communication Engineering, Dhanekula Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India.

* Corresponding Author Email: lalithanarla.ece@gmail.com

DOI: https://doi.org/10.54392/irjmt24613

Received: 27-06-2024; Revised: 10-11-2024; Accepted: 19-11-2024; Published: 21-11-2024

Abstract: One of the key forensics topics has been the detection of speech forgeries, mostly using real evidence in court. The transmission of digital speech recording data over several media exposes the data to the risk of being attacked or tampered with. Several people misuse the audio by altering that using editing software, such as Adobe, Audition CC, etc., which results in speech forgeries. So, to overcome these scenarios speech forgery detection method is deployed. A speech forgery detection method for splicing is implemented in this paper. Firstly, voiced segments are identified in the speech signal and calculated Mel Frequency Cepstral Coefficients (MFCC). These coefficients are considered as features and are stored in the database for the registered speakers. Similarly, features are calculated for each voiced segment of test signals and compared those features with the database by using dynamic time warping. This proposed method is tested on 225 original speech signals that are recorded in two different environments using two different microphones. By combining original recordings of two distinct speakers, a forged dataset of 4900 spliced speech signals is developed to test the efficacy of the developed method. An accuracy of 99.39% was attained and is superior to other existing methods.

Keywords: Speech Splicing Detection, Voice Activity Detection (VAD), MFCC, Dynamic time warping (DTW)

1. Introduction

Multimedia forensics involves manipulating multimedia data such as video, image, and audio. The advancements in artificial intelligence led to synthetic audio generation and detection [1]. The availability of tools for the manipulation of multimedia content throws challenges to investigators. Several researchers developed various techniques to address these challenges contributing to this prominent research area. This research paper concentrates on speech signal forgery detection. The speech signal is vital in communication and has a lot of applications [2]. The most widely performed forgeries on audio content are copy-move forgery and splicing. Some content in the audio may be intentionally deleted or added with other content to form a spliced signal or copied the content of the same signal and moved to another place in the same signal to do forgery resulting in a copy-move forgery. Different methods are being developed to detect these forgeries. Based on the state-of-art, forgery detection techniques are classified into two categories [3, 4], active and passive detection [5]. In active forgery detection, digital watermarking [6] or signature [7] is used to detect the forgery. In the passive detection method, forgery is detected by considering the statistical features of audio/speech signals [8]. The various methods of forgery detection are categorized as shown in Figure 1.

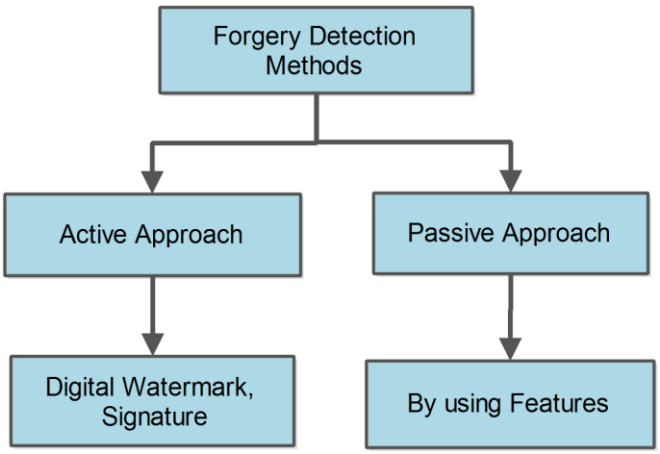

Figure 1. Classification of forgery detection methods

1.1 Active Approaches

With the wide improvement in multimedia technology, digital rights management became vital and several digital watermarking [9] techniques have been developed. Watermarking has been extended to detect

manipulations in audio signal. Audio signal is divided into non-overlapping segments and features are extracted. These features are embedded into least significant bits of another segment [10]. If any segment is tampered then that can be detected by using other segments that carried the features of tested segment and also can be recovered the content. Chen et al. tested their work on 16-bit, 44.1 kHz audio signals. An authentication technique for speech signal using Bessel-Fourier moments is devised to identify and localize the deletion and insertion type of tampering [11]. Imperceptibility, robustness to signal processing operations and tamper localization are tested for this scheme on 16-bit, 44.1 k Hz mono speech signals. Self-embedding speech watermarking is presented in [12] to detect the tampering and for self-recovery. Compressed speech signal and hash bits are used to construct watermark. This watermark is embedded itself into speech signal to detect, localize and to recover the signal. Formant enhancement-based watermarking is developed for tampering detection in the speech signal [13]. The watermark data is inserted by symmetrically changing the pair of LSFs of the corresponding formant. This work is tested on 12 speech signals. The signature content is inserted into singular spectrum analyzed signals by changing the least singular values of those signals for tampering identification [14]. Imperceptibility, robustness and tampering identification of this work is evaluated on twelve Japanese speech database signals.

Audio tampering is detected in [15] by inserting electrical network frequency signal into audio signal. By observing the abnormalities in the ENF signal caused by phase discontinuities due to insertion and deletion of audio clips, audio tampering is detected. The carioca database of 100 unedited phone call recordings are used to test the performance of this method. A high payload blind watermarking scheme is proposed in [16] for secure transmission and tamper detection. Scrambled watermark is embedded in detailed coefficients of wavelet transform of the audio signal. Secret key is generated using SVD and transferred along with the watermarked audio signal. This secret key is used at the receiver end to test whether the received signal is tampered or not. This work is tested on 140 audio signals of various classes with 44.1 kHz sampling frequency and 16-bit quantization levels. Tampered audio recordings are detected by inserting live audio watermarking in systematic way in the audio signal [17]. Nita and Ciobanu selected tic-tac sound as a watermark in this work. Random sequence of delay are introduced between successive tics and by that pattern audio is authenticated. This work is tested on 100 original recordings and 100 tampered signals. Speech watermarking is proposed to detect the speech tampering by modifying the line spectral frequencies which are derived from linear prediction analysis along with dither modulation-QIM [18]. Line spectral frequencies are chosen to embed the watermark because these frequencies are less sensitive to noise and this work is evaluated on 20 kHz, 16-bit Japanese twelve speech signals.

Lalitha *et al.,* proposed an active approach of speech tampering detection method using hash insertion [6]. Here multiple watermarks are embedded using QIM into DCT coefficients of each voiced segment and then hash bits are inserted. Based on the hash bits, copy-move speech tampering is detected and identified the location. This work is tested on VoxForge database. Perceptual speech hash values are generated using gamma tone filter and Gaussian matrix [19]. These hash values are embedded using QIM into sparse features of stationary wavelet transform coefficients of the speech signal to detect and localize the tampering. Shi et al., Tested their work for four kinds of tampering attacks viz., mute, substitute, pitch changing and mixed attacks. Lalitha et al., proposed an active approach of audio tampering detection schemes [7]. Preprocessed watermark is embedded using QIM into transformed audio signal coefficients to provide authenticity. Hash is generated and inserted into watermarked signal to identity the tampering. This work is tested on five classes of audio signals and three types of tampering. The state-of-art active approaches has commendable performance in detecting the audio forgeries and are tested on different datasets.

At the outset, active approaches entail embedding specific features into audio signals. High payload schemes use scrambled watermarks and secret keys to identify tampering, while live watermarking with tic-tac sounds offers real-time detection. Despite their effectiveness, these active techniques demand significant computational resources for embedding and processing watermarks, limiting their feasibility for real-time applications. This drawback has prompted exploration into passive approaches, which detect tampering using inherent signal characteristics, offering faster and more practical solutions.

1.2 Passive Approaches

Audio forgery detection and localization in time domain are proposed in [20, 21]. In original signals, there is a strong correlation between adjacent samples and this correlation will be degraded because of forgery operations. By analyzing these singular points after applying wavelet packet decomposition, forgery has been detected and located. Chen *et al.,* tested their work on 500 audio files for deletion and insertion forgeries and also observed detection rates by varying sampling rates. Based on the correlation of the magnitude between test and reference audio frames audio splicing was detected in [22, 23]. Zhao *et al.,* Proposed an environmental signature-based audio splicing detection framework. The performance of this work is tested on the TIMIT database and recorded audio in four different environments. This work is also compared with previous

works and the detection rate is reported. It is unable to detect small-sized splicing. Here, audio is divided into 50% overlapping frames and calculated psychoacoustic features, namely the critical band spectral estimation, equal loudness hearing curve and intensity loudness power law of hearing. These features are used to train the model using GMM. This work is evaluated for both text dependently and text independently. The logarithmic spectrum of STFT and channel response is combined to construct multiple features [24]. Here channel response is calculated with the GMM model. The correlation of multiple features for test frame and reference frame is calculated to detect the tampered frames. Rouniyar *et al.,* tested their work for different frame sizes.

Audio tampering can be detected based on the reverberation time analysis [25]. If two audio signals are recorded in two different acoustic properties of rooms and merged then reverberation time will tell the characteristics of the room. It can be decide whether the test signal is tampered or not by observing those acoustic characteristics. Ciobanu et al., tested their work on speech signals recorded in five different room environments. Meng *et al.,* [26] focused on heterogeneous splicing detection means audios are recorded in different environments. Intially, they determined the length of each syllable by detecting the endpoints using the spectral entropy method and then calculated the variance of each syllable's background noise. Finally, compared this variance of each syllable background noise and decided whether the given audio is spliced or not. Three different lengths of 150 original audios and 100 spliced audios are created to test the performance of the method. Edited audio signals can be detected by analyzing its metadata information [27]. Audio recording device model, recording date, audio editing date, modification date, editing software, codec information, and URL address are considered to detect the tampering audio.

Ustubioglu *et.al* proposed a key point-based approach to detect and localize audio copy-move forgeries using the Mel spectrogram representation of audio [28]. The method involves creating a Mel spectrogram, extracting SIFT key points from each RGB channel, and matching them to identify forged regions. A post-processing stage is used to refine the detection and mark forged segments. Experimental results on pitch-based datasets (TIMIT and Arabic Speech Corpus) show that the method is robust against post-processing operations like noise addition, filtering, and compression. Zhaopin *et.al* introduced a robust audio copy-move forgery detection (CMFD) method, SW-CQCC, which uses sliding window (SW) strategies and constant Q cepstral coefficients (CQCC) to identify forgeries, especially short slices within voiced segments [29]. By combining CQCC with Pearson correlation coefficients, the method effectively detects similarities between forged segments. Experimental evaluations on English and Chinese corpora demonstrate that SW-CQCC is highly effective and robust for detecting short forgeries, making it valuable for audio forensic applications.

Many passive techniques suffer from detecting small splices and methods like reverberation analysis are less effective when audio segments are recorded in similar acoustic environments. Variance based techniques struggle with consistent background noise or high-noise recordings. Hence, this motivated us to develop a robust method for splicing in the noise environment. In this paper, a passive type of speech splicing detection method is proposed. Voiced segments in the speech signals are identified using Voice Activity Detection (VAD) module. MFCC features are calculated for each voiced segment and stored in the database. While testing the signal, these features are compared with the test signal features with DTW. A decision will be taken based on the DTW values whether that test signal is an original signal or a forged signal.

1.2.1 Contributions to the work:

1. This is a passive approach to detect forgery.
2. It detects the splicing type of speech forgery detection and also localizes the forged segment.
3. MFCC features are calculated only for voiced segments.
4. Speech recording database is prepared under two environments viz., Noisy Environment and Noise Free Environment.

The organization of the paper is as follows: The preliminary background such as VAD, MFCC and DTW are elucidated in Section II. The suggested method for detecting audio forgery splicing is explained in Section III. The experimental findings are elaborated in Section III. The concluding remarks with future directions were presented in Section IV.

2. Materials and Methods

2.1 Voice Activity Detection Module

Voice Activity Detection (VAD) is also called Active Speech Detection [5]. VAD plays a major role in voice-based applications [30-32]. The start and end points of the voiced segments can be detected using the VAD mechanism. To find the start and end points, the entire signal is divided into small size non-overlapping frames and then amplitude of each frame is calculated using the below Eq. (1).

$$F_{amp} = \sum_{i=1}^{n} |a_i| \quad (1)$$

Where a_i is the amplitude of the i[th] sample, n is the number of frames.

Threshold is calculated using the Eq. (2) to identify the voiced and unvoiced segments of the signal.

$$Th = 3\%\ of\ \left(max(F_{amp}) - min(F_{amp})\right) + min(F_{amp}) \quad (2)$$

If the frame amplitude is above the threshold value, then it is recognized as a voiced frame otherwise the frame is recognized as a non-voiced frame [33]. Original signal, threshold and identified voiced segments can observe in the Figure 2.

2.2 MFCC

MFCC is the most dominant technique for feature extraction in the field of speech and speaker recognition [34]. It represents the acoustic signals in the form of cepstral coefficients [35]. The first step is framing and windowing then applied it converts into frequency domain using Discrete Fourier Transform. The mel-frequency warping is used to map the actual frequencies into human being perceived frequencies by using Eq. (3). Later log function is applied and again converted to time domain using inverse DFT to obtain the MFCC features. The block diagram of MFCC feature generation is shown in Figure 3.

$$mel(f) = 1127\ ln\left[1 + \frac{f}{700}\right] \quad (3)$$

2.2. Dynamic Time Warping

Time series analysis employs Dynamic Time Warping (DTW) [8], a technique for contrasting two temporal sequences with potential speed discrepancies. For instance, DTW could still identify patterns in walking even if one person was moving more swiftly than the other or if there were accelerations and decelerations during an observation. With DTW, you may look at any data that can be transformed into a linear series, including temporal video, audio, and graphic data sequences. Automatic speech recognition has been a well-known application for dealing with varying speaking speeds. Other uses include online signature recognition and speaker recognition. It can also be applied in applications requiring partial shape matching. DTW measures the similarity between two sequences using Euclidean distance and given in Eq. (4).

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (4)$$

Where $d(p,q)$ is the DTW distance between p and q which are MFCC features.

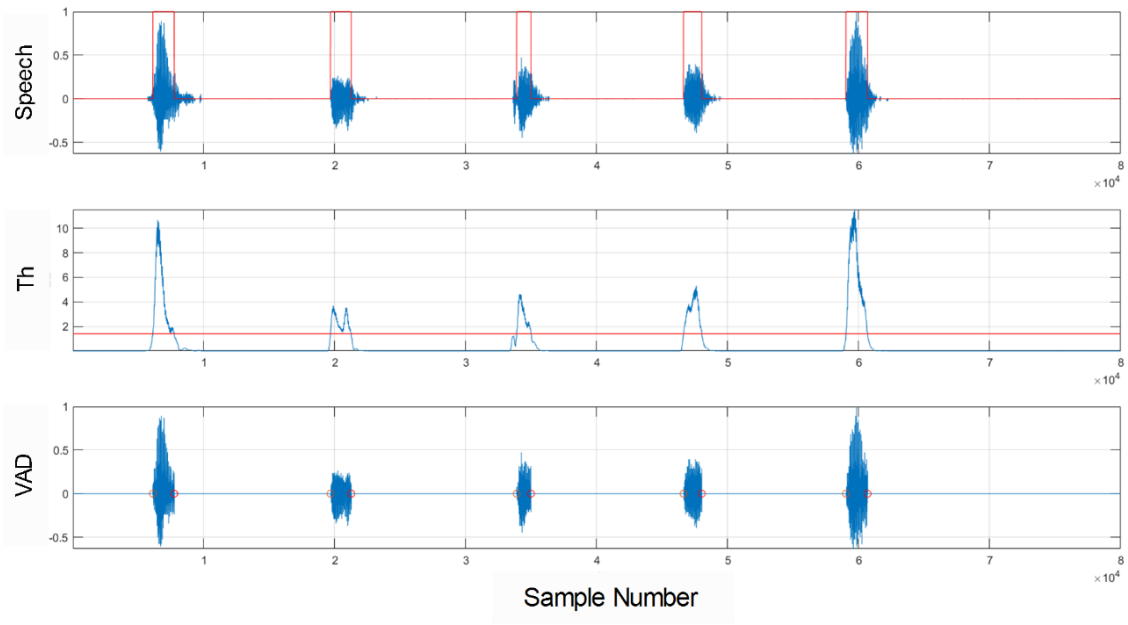

Figure 2. Identification of voiced segments in a speech signal

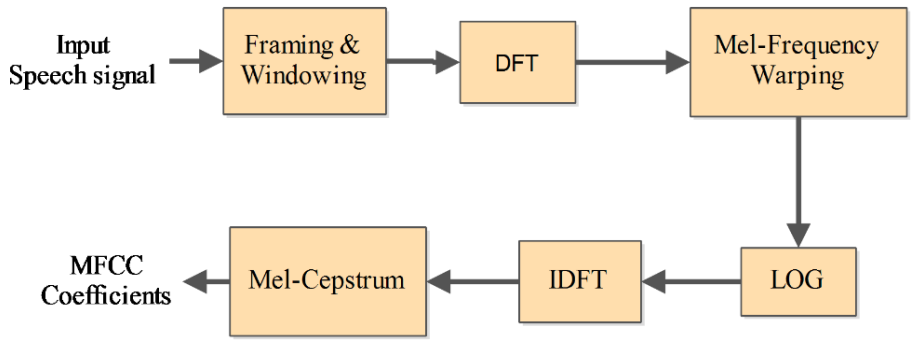

Figure 3. MFCC Block diagram

3. Proposed Method for Audio Forgery Splicing Detection

In this work, VAD is applied on registered speech signals to separate the voiced segments. Features are calculated using MFCC for each recognized voiced segment and stored in the database. The parameters of MFCC are sampling rate: 8 KHz, frame size: 20 mSec, frame stride: 50% overlap, Hamming window with 40 number of filters to get 13 cepstral coefficients. MFCC features characterizes the speaker. This database is used in the testing session to identify the test signal is spliced one or not. In testing session, voiced segments are identified using VAD and then features are calculated. These features are compared with the database of the registered speakers using DTW to conclude whether that segment is spoken by registered speaker or not. If all the voiced segments are spoken by the same speaker then it is concluded that the testing speech is original otherwise test signal is spliced one. The proposed speech splicing detection system block diagram is shown in Figure 4.

The step-by-step procedure for each training and testing sessions are explained here. Training Session: Training all registered speech signals by using the following steps.

Step 1: Read the registered speech signal.

$$\mathcal{R}_s = [\mathcal{S}_1^r, \mathcal{S}_2^r, \quad \cdots \quad \mathcal{S}_N^r] \tag{5}$$

Where N is the number of registered speakers.

Step 2: Apply VAD to identify the speech segments in the signal.

$$\mathcal{S}_i^v = VAD(\mathcal{S}_i^r), for\ i = 1,2,3,\dots,N \tag{6}$$

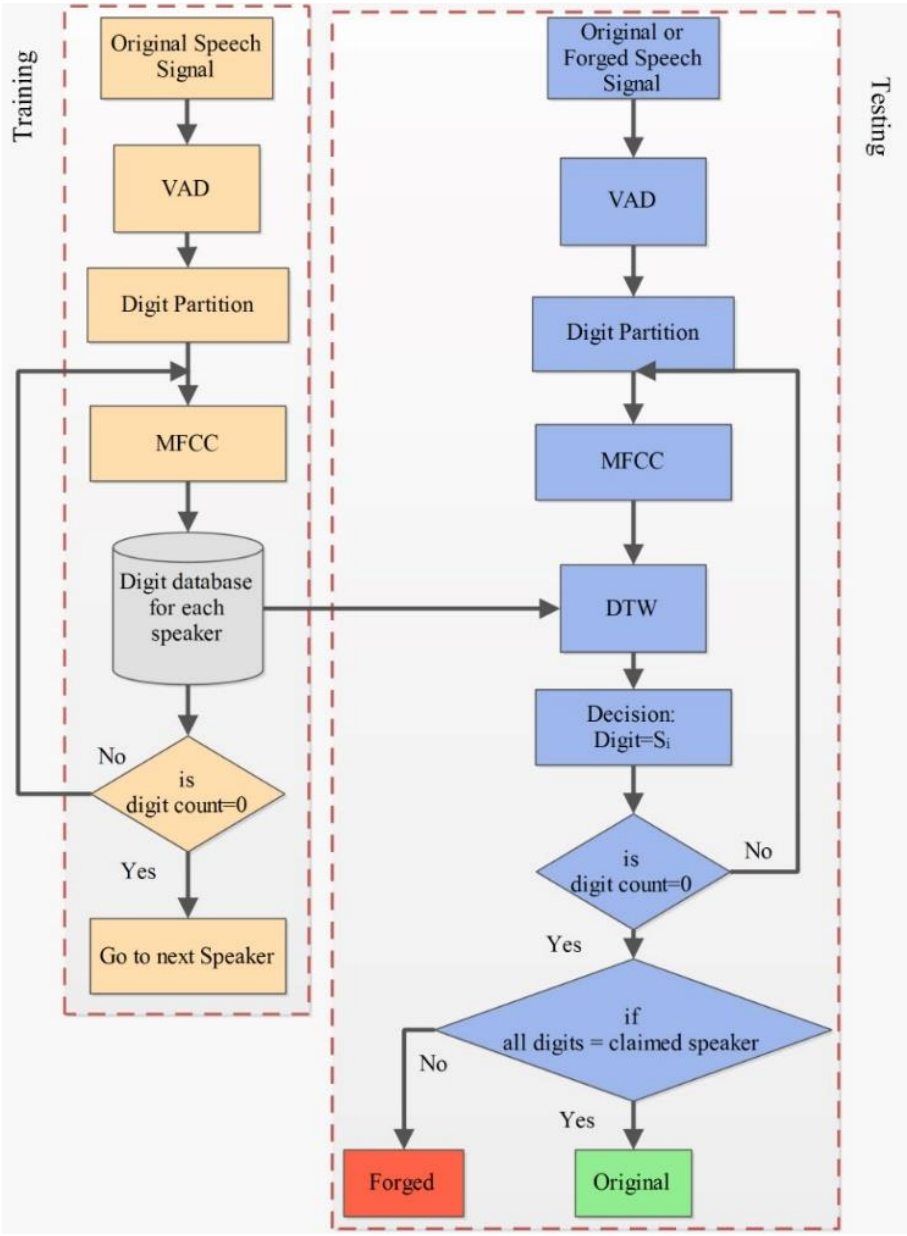

Figure 4. Block diagram of proposed speech splicing detection system

Step 3: Separate the speech segments and calculate MFCC features for each segment.

$$[\mathcal{D}_1^i, \mathcal{D}_2^i, \mathcal{D}_3^i, \quad \dots \quad \mathcal{D}_M^i] = \mathcal{DP}(\mathcal{S}_i^{\mathcal{V}}), for\ i = 1,2,3, \dots, N \quad (7)$$

Where $\mathcal{DP}(\cdot)$the digit partition of the speech is signal and M is the number of digits spoken by the speaker.

$$\mathcal{F}_j^i = MFCC(\mathcal{D}_j^i), for\ i = 1,2,3, \dots, N\ and\ j = 1,2,3, \dots, M \quad (8)$$

Step 4: Store these features in the database under that registered speaker.

$$\mathcal{F}_{db} = \begin{bmatrix} \mathcal{F}_1^1 & \mathcal{F}_2^1 & \dots & \mathcal{F}_M^1 \\ \mathcal{F}_1^2 & \mathcal{F}_2^2 & \dots & \mathcal{F}_M^2 \\ \vdots & & \dots & \vdots \\ \mathcal{F}_1^N & \mathcal{F}_2^N & \dots & \mathcal{F}_M^N \end{bmatrix} \quad (9)$$

Step 5: Repeat Step 4 until all digit counts become zero.

Step 6: Repeat Step 1 to Step 5 for all registered speaker speech signals.

Testing Session: The following steps must follow for checking whether the test signal is original or forged.

Step 1: Read the test speech signal $\mathcal{S}^t$.

Step 2: Apply VAD to identify the voiced and unvoiced segments.

$$\mathcal{TS}^{\mathcal{V}} = VAD(\mathcal{S}^t) \quad (10)$$

Step 3: Separate the voiced segments and calculate the MFCC features for every segment.

$$[\mathcal{D}_1^t, \mathcal{D}_2^t, \mathcal{D}_3^t, \quad \dots \quad \mathcal{D}_M^t] = \mathcal{DP}(\mathcal{TS}^{\mathcal{V}}) \quad (11)$$

$$\mathcal{F}_j^t = MFCC(\mathcal{D}_j^t), for\ j = 1,2,3, \dots, M \quad (12)$$

Step 4: Calculate the distance between these features with registered speaker features by using DTW.

$$dt_j = DTW(\mathcal{F}_j^t, \mathcal{F}_j^i), for\ j = 1,2,3, \dots, M \quad (13)$$

Step 5: The decision of the voiced segment or digit is related to which speaker is based on the DTW values.

$$\mathcal{D}_j^t = \mathcal{D}_j^i, if\ dt_j \leq Threshold$$

$$\mathcal{D}_j^t \neq \mathcal{D}_j^i, if\ dt_j > Threshold \quad (14)$$

Step 6: Repeat Step 4 and Step 5 for the remaining digits or segments.

Step 7: If all digits are related to the claimed speaker, then the decision is tested signal is genuine or original signal. If all digits are not related to one speaker the decision is tested signal is forged signal. \

$$\mathcal{S}^t = \mathcal{S}^i, if\ [\mathcal{D}_1^t, \mathcal{D}_2^t, \mathcal{D}_3^t, \quad \dots \quad \mathcal{D}_M^t] = [\mathcal{D}_1^i, \mathcal{D}_2^i, \mathcal{D}_3^i, \quad \dots \quad \mathcal{D}_M^i] \quad (15)$$

All digits are spoken by speaker $\mathcal{S}^i$ and test speech is original signal.

$$\mathcal{S}^t \neq \mathcal{S}^i, if\ [\mathcal{D}_1^t, \mathcal{D}_2^t, \mathcal{D}_3^t, \quad \dots \quad \mathcal{D}_M^t] \neq [\mathcal{D}_1^i, \mathcal{D}_2^i, \mathcal{D}_3^i, \quad \dots \quad \mathcal{D}_M^i] \quad (16)$$

All digits are not related to same speaker and it is spliced speech signal.

4. Experimental Results

The proposed method is evaluated on the spliced forged audio database. The experimental set-up consists of personal laptop with i5 processor, 8 GB RAM, 2 GB graphics card and Matlab 2022 for simulation. The steps to create the splicing forged speech database and the performance evaluation are discussed in this section. Further, the efficacy of the proposed method is compared with other existing methods.

4.1 Generation of Splicing Forged Audio Database

To assess the effectiveness of the suggested strategy, a database of spliced fabricated speech is created. The fabricated audio is thought to be produced through the recording of various settings and tools. The VAD module is set up so that no human should be able to judge a falsified recording. Speech signals are recorded in two environments with two different microphones and created original speech database. Voiced segments of the original speech signals are replaced with other speaker voiced segments to create spliced speech signal database. The process of spliced speech signal formation is shown in Figure 5.

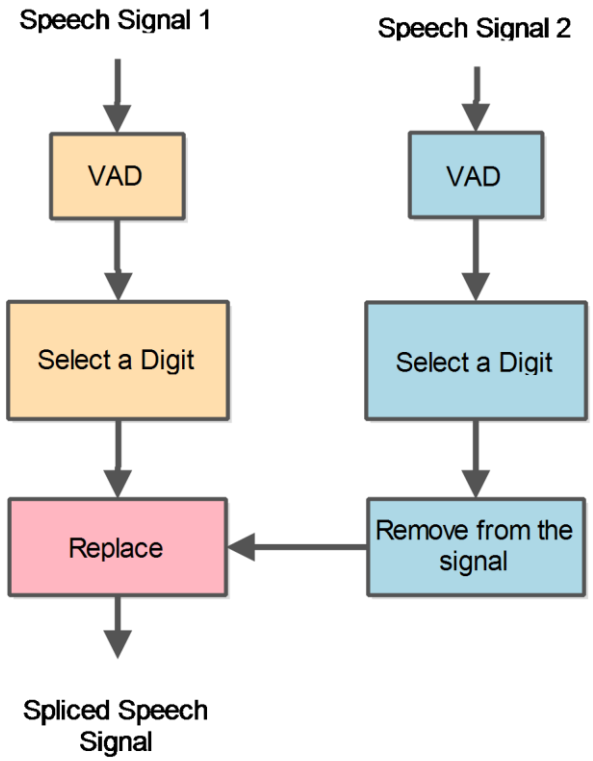

Figure 5. Spliced signal database formation

The proposed system is tested on a total of 225 original signals and 4900 tampered signals in .wav format with 10 seconds in length. Speech signals 225 are recorded in two environments viz., Noisy Environment (NE) and Noise Free Environment (NFE). In each environment, two different microphones (Mic1 and Mic2) are considered to record speech. 25 audio clips are recorded with Mic1 and 20 audio clips are recorded with Mic2 in both environments. Noise environment has been depicted by adding Additive White Gaussian Noise (AWGN) with SNRs ranging from 30dB to 60dB. Each original speech signal is spliced with other original speech signals and created 4900 tampered signals. The experimentation is carried out in 10 scenarios and the same is given in Table 1.

Table 1. characteristics of Actual And Fraudulent Audio

Scenario	Genuine audio	Tampered audio	Total audios
NFE -Mic1	25	600	625
NFE -Mic2	20	380	400
NE-60dB - Mic1	25	600	625
NE-60dB - Mic2	20	380	400
NE-50dB - Mic1	25	600	625
NE-50dB - Mic2	20	380	400
NE-40dB - Mic1	25	600	625
NE-40dB - Mic2	20	380	400
NE-30dB - Mic1	25	600	625
NE-30dB - Mic2	20	380	400
Total Audios	225	4900	5125

4.2 Performance Metrics

If original speech signals are identified as original or genuine speech, then it is termed as True Positive (TP). If spliced speech is identified as a forged speech, then it is termed a True Negative (TN). If the system identified genuine speech as a forged speech, then it is termed False Negative (FN). The system is identified tampered speech as an original, then it is termed as True Negative (TN). The performance of the proposed system is measured with metrics specificity, sensitivity and accuracy.

$$Sensitivity = \frac{TP}{TP+FN} \quad (17)$$

$$Specificity = \frac{TN}{TN+FP} \quad (18)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (19)$$

Sensitivity, specificity, and accuracy are key metrics used to evaluate the performance of a detection system. Sensitivity measures the system's ability to correctly identify true positives (original speech as genuine), ensuring minimal missed detection of genuine signals. Specificity measures the system's ability to correctly identify true negatives (forged speech as tampered), crucial for minimizing false positives. Accuracy represents the overall effectiveness of the system by quantifying the proportion of correct identifications (both genuine and forged) out of all cases, providing a balanced measure of system reliability.

The performance of the proposed system in terms of sensitivity, specificity and accuracy for ten scenarios are shown in Table 2.

Each digit MFCC feature distances are measured using DTW to test speech signals with all registered speech signals. DTW values for five digits of a genuine speech signal are shown in Figure 6. In that Figure, digit 1 is spoken by speaker 8 (S8) because the DTW is less for S8. Similarly, remaining all digits are also related to speaker 8 (S8). From the Figure 6, all digits are related to speaker 8 (S8) then it can be conclude that the speech is genuine speech.

From the Figure 7, it is evident that all the digits are related to speaker 7 (S7) except digit 2 because digit 2 DTW value is less for speaker 3 (S3). So, the test speech signal is not an original signal and it has tampered with S3 signal at digit 2.

Further, the performance of the present method is compared with existing methods in terms of accuracy as shown in Table 3. Forgery detection is performed in [20, 21] using DWT and by analyzing singularity points. Based on the variation in the correlation, deletion and insertion types of tampered regions are identified. An accuracy Rate of 84.20% is achieved for the speech database with a sampling rate of 44.1 KHz [20]. It has been extended and speech analysis is done for different sample rates [21]. The tampering detection accuracy rate is increased by increasing the sampling rate of the speech signal. Multiple features are constructed using channel response of the audio signal and logarithmic spectral characteristics of the tested signal to detect the splicing forgery [24]. A database of size 504 audio clips is considered which are recorded in 21 mobile phones and by 24 speakers and the highest accuracy of 97.6% is achieved. Splicing detection and location identification are done in [22] using acoustic environment signature. The magnitude of the acoustic channel impulse response and ambient noise is considered as environment signature.

Table 2. Performance Parameters of the Proposed System

Scenario	TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Accuracy (%)
NFE -Mic1	25	597	3	0	100%	99.50%	99.52%
NFE -Mic2	20	377	3	0	100%	99.21%	99.25%
NE-60dB -Mic1	25	593	7	0	100%	98.83%	98.88%
NE-60dB -Mic2	20	374	6	0	100%	98.42%	98.50%
NE-50dB -Mic1	24	590	7	4	86%	98.83%	98.24%
NE-50dB -Mic2	19	372	5	4	83%	98.67%	97.75%
NE-40dB -Mic1	22	585	10	8	73%	98.32%	97.12%
NE-40dB -Mic2	18	369	8	5	78%	97.88%	96.75%
NE-30dB -Mic1	20	576	16	13	61%	97.30%	95.36%
NE-30dB -Mic2	17	360	15	8	68%	96.00%	94.25%

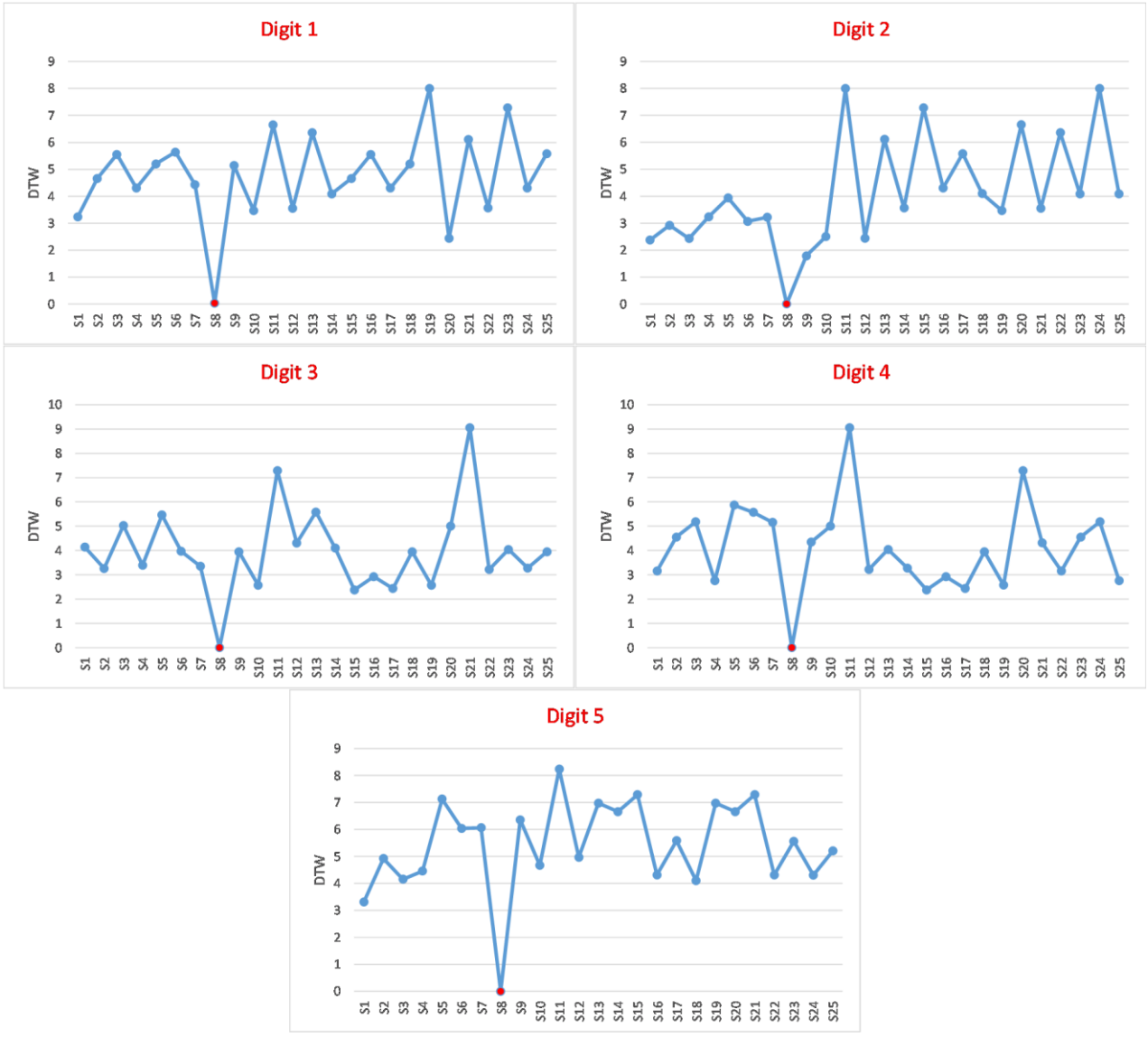

Figure 6. DTW values for five digits of genuine speech

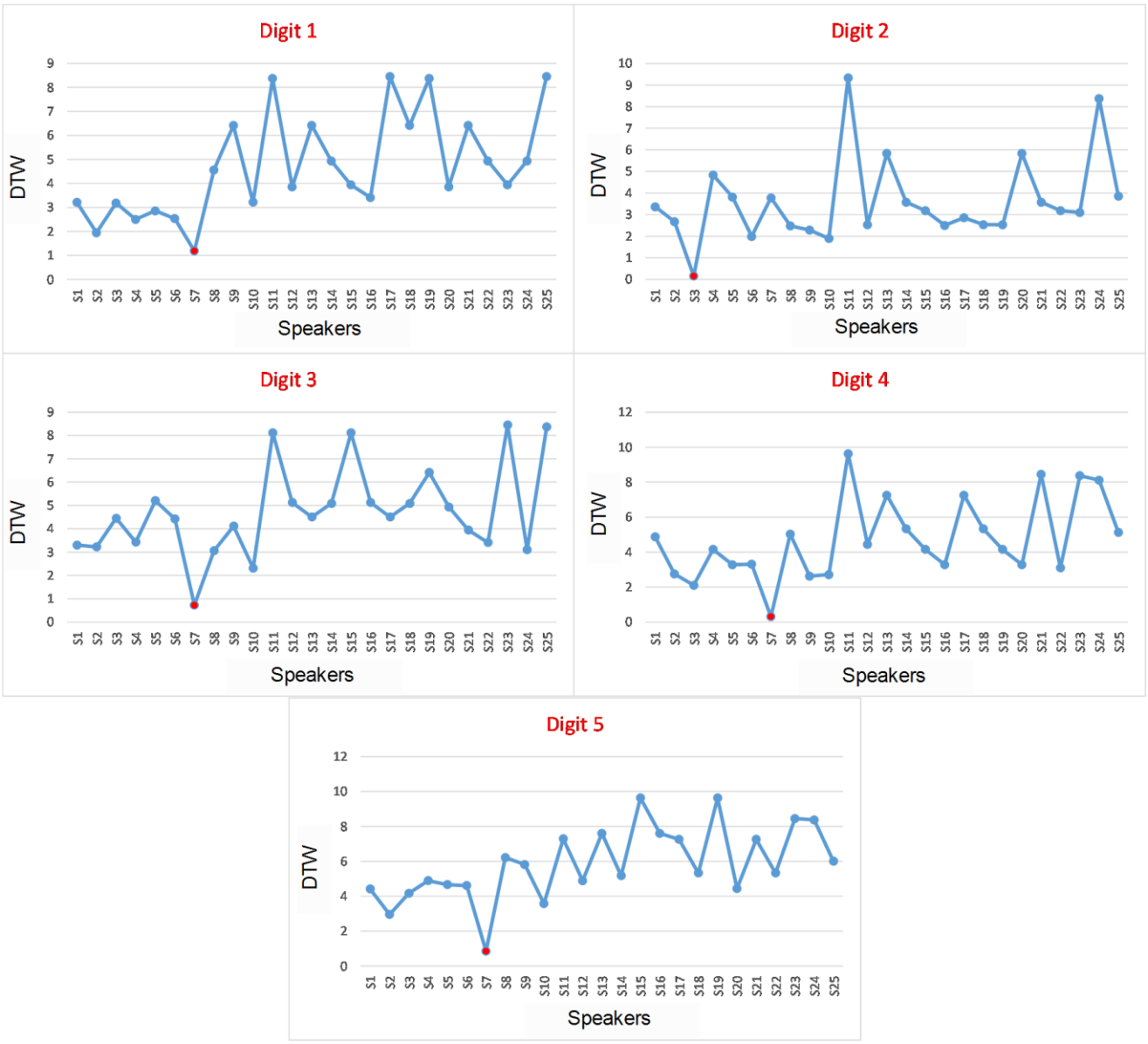

Figure 7. Computed DTW values for five digits of the claimed speaker with remaining all registered speakers

Table 3. Comparison of Previous Work with Proposed method

S.no	Methods	Detection accuracy
1	Singularity points analysis [20]	84.20%
2	Analysis of environmental signature [22]	96%
3	Singularity analysis with wavelet packet [21]	89.5%
4	Channel response based multi-feature [24]	97.6%
5	Proposed Method	99.39%

5. Conclusion and Future Scope

Considering how frequently audio is utilised in daily life nowadays, such as in court cases when it is used as evidence, it is important to know whether the audio has been tampered with or not. In this paper, the investigation of audio tamper detection is done for text-dependent speech with recorded audio in NE and NFE. To detect tampering, the proposed forgery detection method evaluates whether there are several speakers in the audio. Additionally, it identifies forgeries introduced by merging recordings made with various pieces of equipment and surroundings. The modified portions that

don't belong to the stated speaker are displayed. The experimental findings demonstrate the proposed method's decision-making accuracy and dependability. To detect audio tampering MFCC features and the DTW method are used, the MFCC features are compared with the DTW method. The developed method has achieved an average accuracy of 99.39% and 97.11% in a noise-free and noisy environment respectively. This capability recommends its usage in digital audio forensic examination. Future versions of the system will be created using a large vocabulary that includes more commonly used terms and phrases along with the implemented features. The detection performance can be enhanced in the future, taking into consideration tampering in a lengthy speech. An amplitude-based VAD is used to segment speech and there is a scope to explore different VAD systems which can recognise the speech signal at low SNRs. Further, the dataset can be made more realistic by performing smoothing at the boundaries of the spliced signals. This challenging task can be taken up in future work to analyse smoothed spliced signals realistically.

References

[1] V.L. Narla, G. Suresh, A.K. Sahu, M. Kollati, (2024) A Watermark Challenge: Synthetic Speech Detection. Multimedia Watermarking, Springer, Singapore. https://doi.org/10.1007/978-981-99-9803-6_5

[2] K.V. Satya, A.K. Gogoi, G. Sahu, (2011). Regressive linear prediction with doublet for speech signals. In 2011 IEEE International Conference on Control System, Computing and Engineering, IEEE, Malaysia. https://doi.org/10.1109/ICCSCE.2011.6190491

[3] G. Suresh, C.S. Rao, Copy move forgery detection through differential excitation component-based texture features. International Journal of Digital Crime and Forensics (IJDCF), 12(3), (2020) 27–44. https://doi.org/10.4018/IJDCF.2020070103

[4] S. Panda, M. Mishra, (2018) Passive techniques of digital image forgery detection: developments and challenges. In Advances in Electronics, Communication and Computing: ETAEERE-2016. Springer, Singapore. https://doi.org/10.1007/978-981-10-4765-7_29

[5] M. Imran, Z. Ali, S.T. Bakhsh, S. AkramBlind detection of copy-move forgery in digital audio forensics. IEEE Access, 5, (2017) 12843-12855. https://doi.org/10.1109/ACCESS.2017.2717842

[6] N.V. Lalitha, C. Srinivasa Rao, P.V.Y. JayaSree, Localization of copy-move forgery in speech signals through watermarking using DCT-QIM. Intl Journal of Electronics and Telecommunications, 65(3), (2019) 527–532. https://doi.org/10.24425/ijet.2019.129809

[7] V.L. Narla, S. Gulivindala, S.R. Chanamallu, D.P. Gangwar, BCH encoded robust and blind audio watermarking with tamper detection using hash. Multimedia Tools and Applications, 80(21–23), (2021) 32925–32945. https://doi.org/10.1007/s11042-021-11370-5

[8] Q. Yan, R. Yang, J. Huang, Robust copy-move detection of speech recording using similarities of pitch and formant. IEEE Transactions on Information Forensics and Security, 14(9), (2019) 2331–2341. https://doi.org/10.1109/TIFS.2019.2895965

[9] G. Suresh, V.L. Narla, D.P. Gangwar, A.K. Sahu, False-Positive-Free SVD Based Audio Watermarking with Integer Wavelet Transform. Circuits, Systems, and Signal Processing, 41(9), (2022) 5108–5133. https://doi.org/10.1007/s00034-022-02023-5

[10] F. Chen, H.J. He, H.X. Wang, (2008) A fragile watermarking scheme for audio detection and recovery. Congress on Image and Signal Processing, IEEE, China. https://doi.org/10.1109/CISP.2008.298

[11] Z. Liu, H. Wang, A novel speech content authentication algorithm based on Bessel–Fourier moments. Digital Signal Processing, 24, (2014) 197–208. https://doi.org/10.1016/j.dsp.2013.09.007

[12] S. Sarreshtedari, M.A. Akhaee, A. Abbasfar, a Watermarking Method for Digital Speech Self-Recovery. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(11), (2015) 1917–1925. https://doi.org/10.1109/TASLP.2015.2456431

[13] S. Wang, R. Miyauchi, M. Unoki, Tampering Detection Scheme for Speech Signals using Formant Enhancement Based Watermarking. Journal of Information Hiding and Multimedia Signal Processing, 6(6), (2015) 1264–1283.

[14] J. Karnjana, K. Galajit, P. Aimmanee, C. Wutiwiwatchai, M. Unoki, Speech Watermarking Scheme Based on Singular-Spectrum Analysis for Tampering Detection and Identification. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, Malaysia. https://doi.org/10.1109/APSIPA.2017.8282027

[15] P.M.G.I. Reis, J.P.C.L. Da Costa, R.K. Miranda, G. Del Galdo, ESPRIT-Hilbert-Based Audio Tampering Detection with SVM Classifier for Forensic Analysis via Electrical Network Frequency. IEEE Transactions on Information Forensics and Security, 12(4), (2017) 853–864. https://doi.org/10.1109/TIFS.2016.2636095

[16] A. Kaur, M.K. Dutta, High Embedding Capacity and Robust Audio Watermarking for Secure Transmission Using Tamper Detection. ETRI Journal, 40(1), (2018) 133–145. https://doi.org/10.4218/etrij.2017-0092

[17] V.A. Nita, A. Ciobanu, Tic-Tac, (2018) Forgery Time Has Run-Up! Live Acoustic Watermarking for Integrity Check in Forensic Applications. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Canada. https://doi.org/10.1109/ICASSP.2018.8461538

[18] X. Chen, W. Yuan, S. Wang, C. Wang, L. Wang, Speech Watermarking for Tampering Detection Based on Modifications to LSFs. Mathematical Problems in Engineering Mathematical Problems in Engineering, 2019, (2019). https://doi.org/10.1155/2019/7285624

[19] C. Shi, X. Li, H. Wang, A novel integrity authentication algorithm based on perceptual speech hash and learned dictionaries. IEEE Access, 8, (2020) 22249–22265. https://doi.org/10.1109/ACCESS.2020.2970093

[20] J. Chen, S. Xiang, W. Liu, H. Huang, Exposing digital audio forgeries in time domain by using singularity analysis with wavelets. In First ACM workshop on Information hiding and multimedia security, (2013) 149–158. https://doi.org/10.1145/2482513.2482516

[21] J. Chen, S. Xiang, H. Huang, W. Liu, Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet. Multimedia Tools and Applications, 75(4), (2016) 2303–2325. https://doi.org/10.1007/s11042-014-2406-3

[22] H. Zhao, Y. Chen, R. Wang, H. Malik, Audio splicing detection and localization using environmental signature. Multimedia Tools and Applications, 76(12), (2017) 13897–13927. https://doi.org/10.1007/s11042-016-3758-7

[23] H. Zhao, Y. Chen, R. Wang, H. Malik, Audio source authentication and splicing detection using acoustic environmental signature. ACM Information Hiding and Multimedia Security Workshop, 2014, (2014) 159–164. https://doi.org/10.1145/2600918.2600933

[24] S. K. Rouniyar, Y. Yingjuan, Y. Hu, Channel response based multi-feature audio splicing forgery detection and localization. in International Conference on EBusiness, Information Management and Computer Science, Hong Kong, (2018) 46–53. https://doi.org/10.1145/3210506.3210515

[25] A. Ciobanu, V.A. Nita, V. Popa, (2018) Forgery detection based on reverberation time estimation in multiple bands. in 2018 13th International Symposium on Electronics and Telecommunications, ISETC 2018 - Conference Proceedings, Romania. https://doi.org/10.1109/ISETC.2018.8583961

[26] X. Meng, C. Li, L. Tian, Detecting audio splicing forgery algorithm based on local noise level estimation. In 2018 5th international conference on systems and informatics (ICSAI) (2018), IEEE, China. https://doi.org/10.1109/ICSAI.2018.8599318

[27] D.P. Gangwar, A. Pathania, M. DFSS, Authentication of digital audio recording using file's signature and metadata properties. International Journal of Engineering Applied Sciences and Technology, 5(3), (2020)162-165.

[28] B. Ustubioglu, G. Tahaoglu, G. Ulutas, Detection of audio copy-move-forgery with novel feature matching on Mel spectrogram. Expert Systems with Applications, 213, (2023) 118963. https://doi.org/10.1016/j.eswa.2022.118963

[29] Z. Su, M. Li, G. Zhang, Q. Wu, Y. Wang, Robust audio copy-move forgery detection on short forged slices using sliding window. Journal of Information Security and Applications, 75, (2023) 103507. https://doi.org/10.1016/j.jisa.2023.103507

[30] E. Chuangsuwanich, S. Cyphers, J. Glass, S. Teller, (2010) Spoken command of large mobile robots in outdoor environments. IEEE Spoken Language Technology Workshop, IEEE, USA. https://doi.org/10.1109/SLT.2010.5700869

[31] Y. K. Bharath, S. Veena, K. V. Nagalakshmi, M. Darshan, R. Nagapadma, (2016) Development of robust VAD schemes for Voice Operated Switch application in aircrafts: Comparison of real-time VAD schemes which are based on Linear Energy-based Detector, Fuzzy Logic and Artificial Neural Networks. In 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), IEEE, India. https://doi.org/10.1109/ICATCCT.2016.7911990

[32] D.S. Jat, A.S. Limbo, C. Singh, (2019) Voice activity detection-based home automation system for people with special needs. In Intelligent Speech Signal Processing, Academic Press.

[33] Z. Ali, M. Imran, M. Alsulaiman, An automatic digital audio authentication/forensics system. IEEE Access, 5, 2994-3007. https://doi.org/10.1109/ACCESS.2017.2672681

[34] M.K. Singh, Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques. Multimedia Tools and Applications, (2024) 77327–77345. https://doi.org/10.1007/s11042-024-18602-4

[35] A.K.H. Al-Ali, D. Dean, B. Senadji, V. Chandran, G.R. Naik, Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access, 5, (2017) 15400-15413. https://doi.org/10.1109/ACCESS.2017.2728801

Acknowledgment

The authors thank the anonymous reviewers who helped to improve the quality of the paper.

Authors Contribution Statement

Venkata Lalitha Narla: Conceptualization, Writing Draft, Review and Editing; Gulivindala Suresh: Formal analysis, Methodology, Writing – original draft; Mahesh K Singh: Conceptualization and Implementation, Writing Draft and Editing, Writing – original draft; M. Vinod Kumar: Writing Draft, Methodology and Implementation, Review and Editing. All the authors read and approved the final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License