

RESEARCH ARTICLE

An Attack-Independent Audio Forgery Detection Technique Based on Cochleagram Images of Segments With Dynamic Threshold

BESTE USTUBIOGLU 

Department of Computer Engineering, Karadeniz Technical University, 61080 Trabzon, Turkey

e-mail: bustubioglu@ktu.edu.tr


ABSTRACT Thanks to advanced audio editing software, speech recordings can be tampered with very quickly. If the speech recordings are used as forensic evidence, adding the audio recordings together, cutting them, and changing their content are legally unacceptable and constitute a crime. Audio copy-move forgery is the most common forgery to change the content of the speech. Audio copy-move forgery is performed by copying a segment in the audio and pasting it anywhere in the same audio. This study proposes a robust and new method based on cochleagram images to detect audio copy-move forgery. The proposed method uses cochleagram images of the voiced parts of the audio to detect forgery clues in the input audio file. For this purpose, the audio file is first split into voiced parts using a pitch-based Voice Activity Detection (VAD) method. Each audio part is then converted into a cochleagram image. Structural similarity index measure (SSIM) is used to calculate the similarity between cochleagram images. After calculating the SSIM values between the cochleagram images, the proposed forgery localization algorithm is performed. In this algorithm, the SSIM values among the cochleagram images are first sorted in descending order. The length ratio between these pairs of segments is calculated to determine which values in this descending order are duplicated segment pairs. If this ratio exceeds the specified percentage rate, these segment pairs are marked as forged segments. Finally, the proposed audio copy-move forgery detection method is evaluated against the state-of-the-art approaches with two Copy-Move Forgery Detection (CMFD) database and forged databases created from TIMIT and the Arabic Speech Corpus database. For Copy-Move Forged Datasets, 95% Precision, 98% Recall and 97% F-score were obtained. The experimental results show that the proposed method is *significantly* more robust against post-processing operations than other studies.

INDEX TERMS Cochleagram, forgery detection, copy-move forgery, SSIM.

I. INTRODUCTION

The rapid developments in audio technology allow the production, processing, storage, and distribution of audio data very quickly. In this case, it brings some concerns and sometimes problems. The most important of these problems is the reliability of the audio file. The integrity of the audio data ensures reliability. Integrity is protecting the content of information against threats of unauthorized alteration, deletion, or destruction in any way. Thanks to the ease of use of advanced audio editing software, audio files' integrity can be

damaged, and attackers can perform audio forgery quite easily. For instance, a speech recording with the content "Paul was there the night of the incident, but Nick was not there" can be faked by an attacker before being presented to the court as evidence in a criminal case. For this, the word "not" in this record is pasted between "was" and "there" in the sentence. Thus, the content of the audio recording is changed to "Paul was not there the night of the incident, but Nick was there" and the meaning of the sentence is entirely different, and forged audio recordings are created. In this case, this forged record could ultimately reverse the course of the case. From this point of view, it is essential to research the authenticity of the speech recordings, especially in judicial cases. In this

The associate editor coordinating the review of this manuscript and approving it for publication was Jiachen Yang .

way, audio copy-move forgery is generated by copying one or more segments in the audio and pasting them into another part of the same audio. An instance of audio copy-move forgery is presented in Figure 1. Figure 1(a) shows the audio taken from the TIMIT database (“sa1.wav”), while Figure 1(b) shows the forged audio obtained by copying the second segment of this audio and pasting it on the first segment. Red frames indicate the copied and pasted segments.

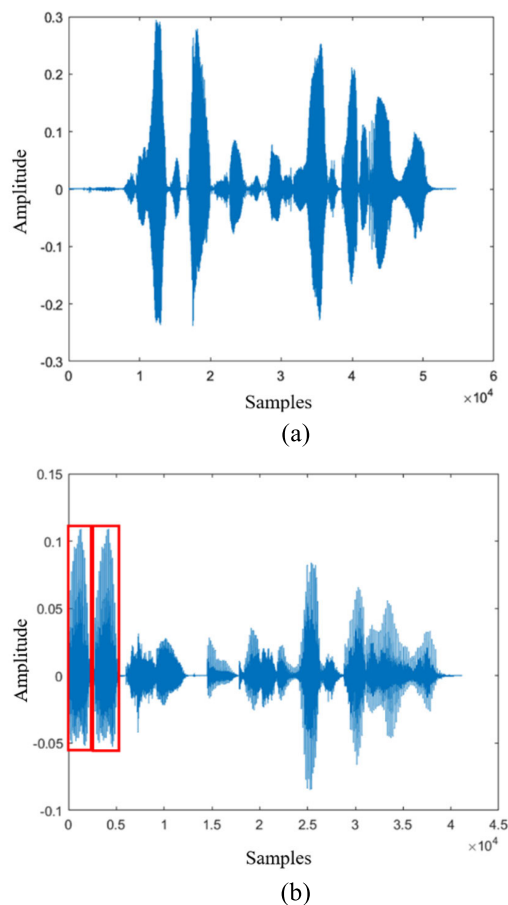


FIGURE 1. a) Original audio (b) forged audio obtained with copy-move forgery.

It is challenging to detect forgery because the duplicated segments are taken from the same speech. This is because the basic speech-related features such as amplitude, length, noise, and frequency will not change for the duplicate segments of the same speech. Another challenge arises when duplicated segments are selected from short-time segments. Most audio forgery detection methods split the speech into voiced parts with VAD-based methods. However, many proposed VAD methods cannot obtain accurate results if the segment is of short time. In addition, the attackers apply post-processing operations such as median filtering, compression, and noise addition to the forged audio after generating the forged audio to remove clues of forgery.

Digital audio authentication methods can be broadly categorized into two types: Active authentication and passive

authentication methods. Active methods [1], [2], [3], [4] necessitate additional information such as watermark and signature. In contrast, passive methods, which are more prevalent, can detect audio forgery by utilizing the features extracted from the audio files without any supplementary information. This is why we focus on the pertinent passive methods to detect audio copy-move forgeries. Passive methods can be further classified into two categories as depicted in Table 1: Methods to detect *Audio copy move forgery* and *Audio splicing forgery*.

Audio splicing is a widely used method of forgery, involving the use of two or more recordings to create forged audio. Pan et al. [31] detected audio splicing forgery by identifying anomalous differences in local noise levels in audio. Chen et al. used discrete wavelet packet decomposition and analyzed the singular points of audio to detect audio forgery operations over time, such as delete, insert, replace, and splice [32]. Gupta et al. [33] proposed a content-based audio copy detection method, where they compared the query fingerprints and the test fingerprints to detect the copies based on the number of matching fingerprints. Several other studies have also been conducted to detect splicing in audio, including the identification of different environments [34], [35], microphone classification [36], [37], and speaker identification [38], [39], [40].

Audio copy-move forgery detection methods can be split into Window-based, VAD-based, and Spectrogram-based.

In Window-based methods, the speech file is divided into non-overlapping or overlapping windows of equal length. Then, robust audio features are obtained from the windows of the audio. Similarity computations are realized between the obtained features of the windows. The windows that are the most similar are marked as duplicated windows. The first study in this area was proposed by Xiao et al. [5]. In their proposed study, the audio file is split into T-period windows. Fast convolutional algorithms are used to calculate the similarity between windows. The windows with similarity above a certain threshold are marked as duplicated windows. Su et al. [6] suggested two Sliding Window (SW) strategies to detect the audio copy-move forgeries within and between voiced segments. They extracted features from the windows with the Constant Q Cepstral Coefficients (CQCC) feature. The similarity between different short windows was calculated using Pearson correlation coefficients. Su et al. [7] used constant Q spectral sketches (CQSS) and the combination of a customized genetic algorithm (GA) and support vector machine to detect duplicated segments. For this purpose, they first averaged the logarithm of the squared-magnitude constant Q transform to extract the CQSS features.

Afterward, they used GA to optimize the extracted CQSS features. In the final stage, they classified the features optimized with SVM and marked the forged segments.

In VAD-based methods, voiced parts of the speech file are extracted using VAD or YAAPT methods. The similarities of the features obtained from the voiced parts are calculated. Voiced parts with high similarity are detected as duplicated segments. Yan et al. [8] extracted the pitch sequence from

the speech file. Average difference (AD) and Pearson correlation coefficient (PCC) methods were used to compute the similarity of the pitch sequence. Wang et al. [9] presented a method to detect audio copy-move forgery with the singular value decomposition (SVD) transform and the discrete cosine transform (DCT). In their method, after the speech file was split into voiced parts with a VAD method, DCT coefficients were obtained from each voice part. Afterward, they obtained eigenvectors by applying the SVD to the square matrix of these coefficients. They used the Euclidean distance (ED) in the similarity calculation. Imran et al. [10] suggested an audio copy-move forgery detection method with the local binary pattern (LBP) method.

In this method, like other VAD-based methods, they first extract the voiced parts from the speech file with their VAD method. Next, they generated the feature vector by generating LBP histograms from each voiced part. Similarity calculation of features was performed with Mean Squared Error (MSE) and Energy Ratio (ER) metrics. Xie et al. [11] extracted four separate features, such as pitch, Discrete Fourier transform coefficients (DFT), Mel frequency cepstral coefficients (MFCCs), and gamma tones from each voiced part, to detect copy-move forgery. They evaluated the similarities of these features using PCC and AD methods. The detection results obtained from four features were combined with the C4 decision tree to obtain the final decision. Anh et al. [12] present an approach based on phonetic sequence. Their method obtains phonetic sequences from extracted voiced parts. They calculated the similarity between the different phonetic sequences with the most minor deviations. Huang et al. [13] extracted the DFT coefficients from the voiced parts. They sorted these features in the proposed method to reduce the computational cost. They performed the PCC method to calculate the similarity between the voiced parts. Yan et al. [14] extracted the voiced parts of the speech file with a normalized low-frequency energy ratio. After pitch and formant sequences from the voiced parts were extracted, similarity calculation was performed with the dynamic time warping (DTW) method. Manneppalli et al. [15] extracted the MFCC features of each voice part after obtaining voiced parts from audio. They evaluated the similarity of MFCC features by the Dynamic Time Warping (DTW) distance. Ustubioglu et al. [16] split the audio into voiced parts with the YAAPT method. They obtained Modified Discrete Cosine Transform (MDCT) coefficients from these voiced parts and took the mean of the transpose of the coefficient matrix as the feature. ED was applied to measure the similarities between the features of voiced parts.

In Spectrogram-based methods, audio data is converted into a spectrogram image. In the feature extraction phase, unlike VAD and window-based methods, since the input data is image instead of audio, these methods use image feature extraction methods instead of audio feature extraction methods. After feature extraction, the markings obtained on the spectrogram image are projected into the audio, and duplicated segments are marked on the audio.

Ustubioglu et al. used deep learning with the Mel spectrogram to detect forged segments for the first time in the literature [17]. Their proposed Convolutional Neural Network (CNN) architecture classified the suspicious Mel spectrogram images into original and forged classes. Ustubioglu et al. [18] used the Scale-invariant feature transform (SIFT) method to extract key points on the Mel-spectrogram. The obtained key points from each channel were matched via feature vectors and the image sub-blocks whose key points are determined to be the center were labeled as forged blocks. Ustubioglu et al. [19] used super-resolution spectrogram images to visualize the suspicious audio. They extracted key points and their feature from the spectrogram image with the Binary Robust Independent Elementary Features (BRIEF) method. The Ordering Points To Identify the Clustering Structure (OPTICS) method was used to match the corresponding descriptors with the clustering approach. Then, the method marks the corresponding duplicated segments in the speech file based on the location of the key points in these clusters on the spectrogram image.

This article proposes a new attack-independent audio forgery detection method. The proposed method uses cochleagram images of the voiced parts of the audio to detect forgery clues in the suspicious audio file. For this purpose, the audio file is first split into voiced parts using a pitch-based VAD method. Each audio part is then converted into a cochleagram image. SSIM method is used to calculate the similarity between cochleagram images. After calculating the SSIM values between the cochleagram images, the proposed forgery localization algorithm is performed. In this algorithm, the SSIM values among the cochleagram images are first sorted in descending order. The length ratio between these pairs of segments is calculated to determine which values in this descending order are duplicated segment pairs. If this ratio exceeds the specified percentage rate, these segment pairs are marked as forged segments. In experimental studies, comparisons were made with the state-of-the-art methods in the literature. Since there is no common audio copy-move forgery dataset in the literature, two new datasets created by us have been used.

The remainder of the article is arranged as follows: The contribution of our study is presented in Our Motivation. The details of the proposed audio copy-move forgery method are given in Materials & Methods. Results illustrate the presentation and analysis of the experimental methodology and results. Last, the study is concluded in the Conclusion section, after the Discussion section.

II. OUR MOTIVATION

As is common in the detection of audio copy-move forgery in the literature, the speech file is first split into windows or voice parts. Robust features are then obtained from these windows or voiced parts with audio feature extraction methods to generate an approach robust to post-processing operations. The copied and pasted segments are marked according to the similarity of the obtained features. Each method determines

TABLE 1. Summary of research differences between the existing work and this study.

Authors	Input	Similarity method and Threshold	Database
Su et al. (2020) [6]	Audio	MSE compared with the static threshold, PCCs	Their own dataset :CMFD dataset, 500 authentic recordings and 500 forged recordings for testing.
Huang et al. (2020) [13]	Audio	Compared of each segment Compared with the static threshold	--
Yan et al. (2019) [14]	Audio	Dynamic time warping Compared with the static threshold	Their own dataset: 4000 different words from the TIMIT dataset and the WSJ audio database
Xie et al. (2018) [11]	Audio	C4.5, decision tree, PCCs, and AD Compared with the static threshold	Their own dataset: 1000 copy-move forgery and 1000 audio files
Imran et al. (2017) [10]	Audio	MSE compared with each other Energy ratio used to compare histograms with the static threshold	Their own dataset: created from King Saud University Arabic Speech Database
Wang et al. (2017)[9]	Audio	Distance between any two singular vectors calculated and compared with the static threshold	Their own dataset: 100 normal and 100 copy-move audio files
Yan et al. (2015) [8]	Audio	PCCs and AD compared with the static threshold	Their own dataset: 1000 tampered audios
This study	Spectrogram image of segment	SSIM compared with the dynamic threshold	Their own dataset: TIMIT and Arabic Speech Corpus and CMFD dataset [30]

a static threshold according to its database in the similarity calculation. If their method is given an audio outside their database as an input file, it will fail. For this reason, these studies reported the results of their proposed methods in their databases. Table 1 summarizes research differences between the existing works and this study.

In our previous studies Ustubioglu et al. [17], [18], [19], unlike these studies in the literature, audio data was converted to an image (Mel-spectrogram, High-resolution spectrogram), and the image was taken as input data instead of audio. Since audio is represented by an image, unlike other studies, image feature extraction methods were used instead of audio feature extraction methods. In this proposed study, unlike our previous studies, the spectrogram image was not created from the entire audio, the voiced parts obtained from the audio were converted into spectrogram images. Thus, a more detailed image is obtained by extracting the spectrogram image corresponding to each voiced part instead of extracting the spectrogram image corresponding to the entire audio.

Our motivation can be summarized as follows:

- For the first time in the literature, the possible clues of audio copy-move forgery were investigated using cochleagram images obtained from the voiced parts of the audio file.
- Thanks to the robustness of the cochleagram images extracted from the audio parts against attacks, the proposed audio copy-move forgery detection method is an attack-independent method.
- With the proposed forgery localization algorithm, instead of the static threshold used in similarity calculations, a dynamic threshold was obtained by using the length ratio information between the voiced parts in the audio.
- Experimental results prove that the proposed method shows superior performance in the detection of audio

copy-move forgery compared to other studies in the literature on the audio forgery dataset created from TIMIT and Arabic Speech Corpus dataset and CMFD dataset.

III. MATERIALS & METHODS

We propose a robust audio copy-move forgery method in this study. This section presents all the details of the proposed method as shown in Fig 2. As seen in the Fig. 2, the proposed method consists of 4 phases: *Separating all the voiced segments*, *Extraction of cochleagram images from voiced segments*, *Similarity calculation between cochleagram images*, and *Forgery localization*. In the first step, we use the pitch-based VAD technique [16] to extract all the voiced segments roughly. Next, each voiced segment is converted to a cochleagram image. Then, we calculate the similarity of two cochleagram images with SSIM. SSIM values calculated between all images are saved. At the forgery localization stage, we mark the location of the duplicated segments according to the SSIM values with our proposed localization algorithm. The recorded SSIM values are sorted in descending order in the proposed localization algorithm and kept in a matrix. Starting from the matrix's first value, the largest SSIM value, the length ratio between the segment pairs giving this value is checked because the length of the copied and pasted segment will be mainly preserved during the forged audio creation phase. Finally, all pairs of segments preserved in the length ratio are marked on the audio as duplicated segments. The following subsections will detail each step of the proposed audio copy-move forgery method.

A. SEPARATING ALL THE VOICED SEGMENTS

The pitch-based Vad algorithm proposed in [16] is used to separate the voiced segments of the audio file. Pitch is a

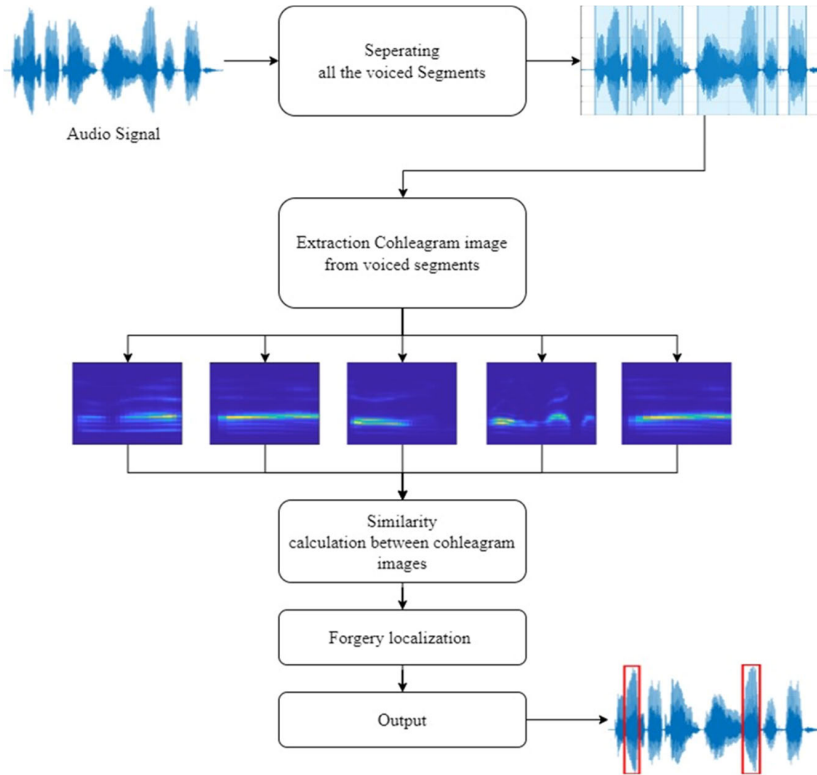


FIGURE 2. The framework of the proposed audio copy-move forgery method.

concept that mentions to the fundamental frequency and gives the vibration frequency of the vocal utterance. Even if a person tells the same voiced parts twice, the pitch sequences of the voiced parts will be different from each other [20]. YAAPT is a popular pitch-tracking method. The main phases of this method are preprocessing, pitch track estimation using spectral information, pitch candidate estimation, and final pitch determination using dynamic programming. After obtaining the pitch sequence with YAAPT, the frequency values greater than zero are saved as voiced segments in the speech. Figure 3(a) shows a sample forged speech file. Duplicated segments in the forged audio are shown in red. The pitch sequence extracted from the speech in Fig. 3(a) is given in Fig. 3(b). The blue lines in Fig. 3(c) represent the voice segment boundaries obtained according to pitch sequences in the audio. As can be seen from the figure, the forged speech is divided into five voiced segments according to the pitch sequence obtained. The second and fifth segments are copy-pasted forged segments.

B. EXTRACTION COCHLEAGRAM IMAGES FROM VOICED SEGMENTS

After the speech file is divided into voiced segments, each voiced part will be represented by a cochleagram. The cochleagram representation of a speech corresponds to the frequency components in the time-frequency image of the speech. These frequency components are based on the frequency selectivity of the human cochlea and are modeled

with a gammatone filter as given in equation (1) [21]

$$h(t) = At^{j-1}e^{-2\pi Bt} \cos(2\pi f_c t + \theta) \quad (1)$$

where A is the amplitude, j is the order of the filter, B is the band-width of the filter, f_c is the center frequency of the filter, θ is the phase, and t is the time.

The equivalent rectangular bandwidth (ERB), a psychoacoustic measure of the auditory filter width at each point along the cochlea, is used to identify the bandwidth of each cochlea filter in [21]. The proposed method uses the ERB filter model as given in [22], which was shown to generate the best results in [23]. After filtering the signal with the gamma tone filter, the implementation of which can be given in [23] and [24], a representation similar to the spectrogram is obtained by adding the energy in the windowed signal for each frequency channel as:

$$C(g, r) = \sum_{n=0}^{N-1} |\hat{x}(g, n)|w(n), \quad g = 1, 2, \dots, G \quad (2)$$

where $\hat{x}(g, n)$ is the gammatone filtered signal, $C(r)$ is the g^{th} harmonic corresponding to the center frequency f_{cg} for the r^{th} frame, and G is the number of gammatone filters.

With the cochleagram representation, the proposed algorithm uses a 64-channel gamma tone filter bank with center frequencies distributed from 50 to 8000 Hz. This filter bank is a standard cochlear filtering model obtained from psychophysical studies of the auditory periphery. Each audio segment is passed through a 64-channel gammatone filter bank. The short-term energy of each vocal segment is

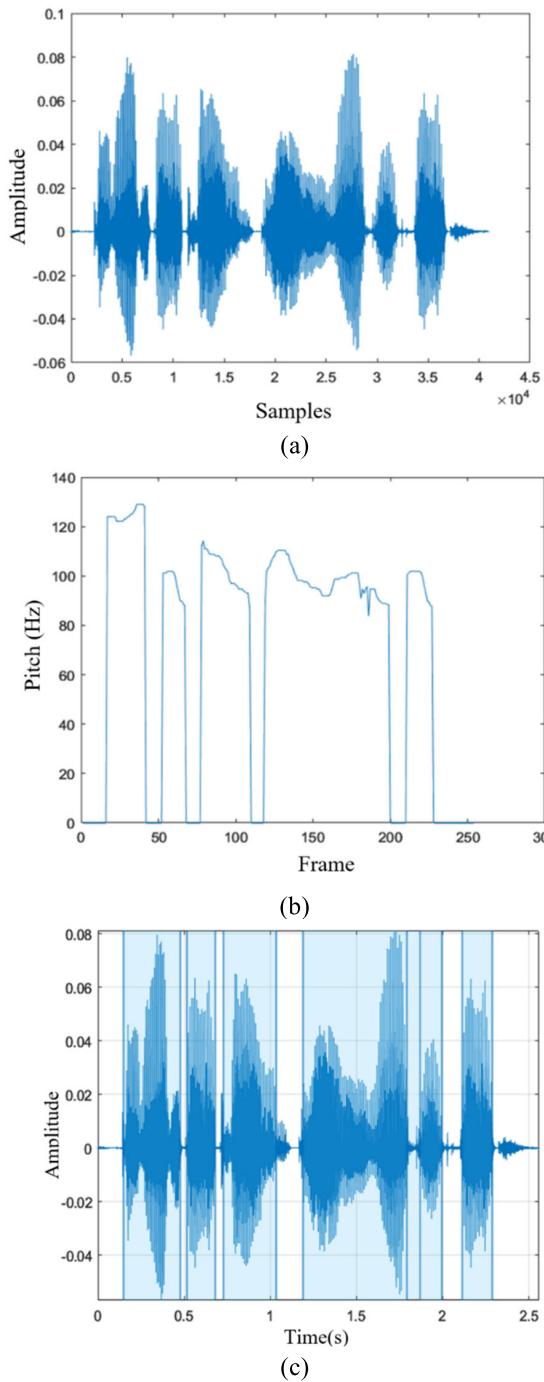


FIGURE 3. (a)The waveform of the forged audio (b)The pitch sequence of the forged audio (c)The voiced segment boundaries of the forged audio.

then calculated to obtain the cochleagram. Figure 4 presents cochleagram images obtained from the five segments separated by blue given in Fig. 3(c). As can be seen in the figure, the cochleagram images corresponding to the 2nd and 5th segments are also similar. In the VAD and window-based methods suggested in the literature to detect audio copy-move forgery, after the audio file is divided into voiced parts or windows, 1-D features are extracted from the voiced

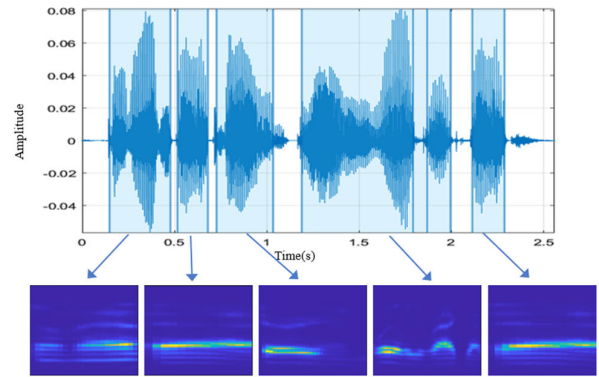


FIGURE 4. Cochleagram images obtained from the voiced parts.

parts or windows. Unlike these studies, the proposed method extracted images from each voiced part instead of 1-D feature extraction. An experimental study has been carried out to show that image extraction is a more efficient method instead of the 1-D feature from each voiced part. For this, both 1-D and 2-D cochleagram features were extracted from the 2nd and 5th segments, which are the duplicated segments, and the similarity between these two segments was calculated by correlation. In this similarity calculation, post-processing operations were also applied to the segments. Table 2 gives the correlation results obtained between the 2nd and 5th segments.

TABLE 2. Correlation values obtained according to the feature size between the 2nd and 5th segments.

	No attack	Median filtering	Noise Addition	Compression
1D feature	0.9286	0.9299	0.9723	0.9596
2D feature	0.9706	0.9714	0.9879	0.9817

As evidenced by the data in Table 2, even when post-processing operations are applied to the 2nd and 5th segments, the correlation value between the 2-D cochleagram features extracted from these segments remains higher than the correlation value between the 1-D cochleagram features. This robustness to post-processing operations is a key strength of the proposed method. In this method, the cochleagram image is extracted from the voiced parts of the speech, further enhancing its resilience. Figure 5 vividly demonstrates the method’s effectiveness, showing the images obtained after post-processing operations such as median filtering, noise addition with 20db, and 32-bit compression to the cochleagram images given in Fig. 4.

As seen in Fig. 5, there is no difference between the no-attack cochleagram images of the segments and the cochleagram images with compression, noise addition, and median filtering. These images show that the cochleagram feature is very robust to post-processing operations.

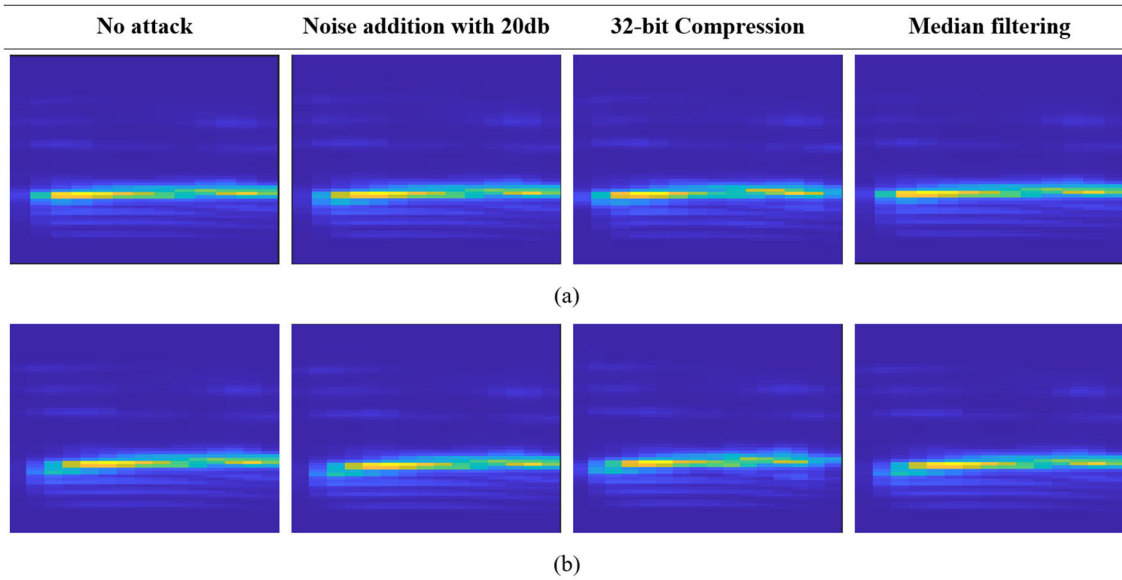


FIGURE 5. Images created as a result of post-processing operations applied to (a) Segment 2 (b) Segment 5.

C. SIMILARITY CALCULATION BETWEEN COCHLEAGRAM IMAGES

The proposed method uses the SSIM metric to calculate the similarity between cochleagram images. SSIM, first introduced in [25], is a perceptual metric used to calculate the similarity of two images. The SSIM extracts three features from an image: luminance, contrast, and structure. It uses as visual information the combination of these features. The SSIM value calculated between the two images ranges from -1 to 1 . A value of 1 indicates that the two images whose similarity is calculated are the same, and a value of -1 indicates that these images are very different. Basically, between two images x and y , the structural similarity index identifies a measurement of distance.

$$SSIM_{(x,y)} = l(x,y)^\alpha c(x,y)^\beta s(x,y)^\gamma \quad (3)$$

where α , β , and γ are constants to weights l , c , and s . The latter weights are functions of mean, variance, and covariance of intensities:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2\mu_y^2 + c_1} \quad (4)$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2\sigma_y^2 + c_2} \quad (5)$$

$$s(x,y) = \frac{2\sigma_{xy} + c_2}{2\sigma_x\sigma_y + c_2} \quad (6)$$

The values of the constants c_1 and c_2 are commonly set to 0.01 and 0.03 to continue numerical stability. By substituting Equations (4)-(6) in Equation (3),

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2\sigma_y^2 + c_2)} \quad (7)$$

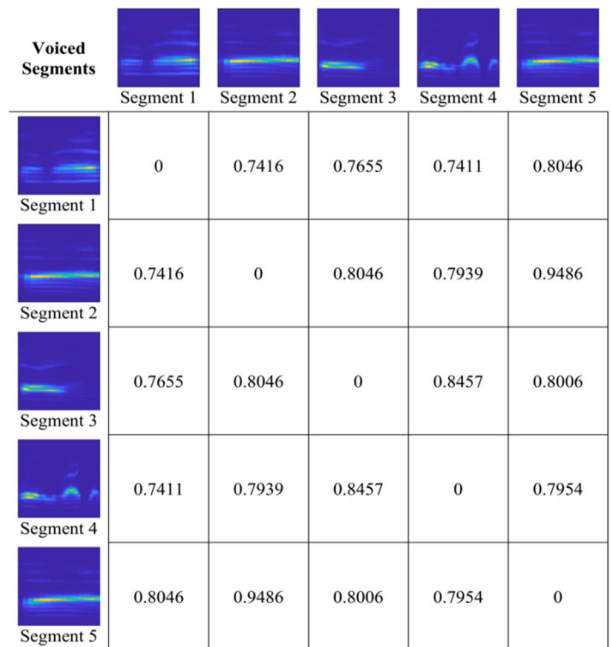


FIGURE 6. SSIM values among the cochleagram images.

In this way, the similarity between the cochleagram images extracted from the speech was calculated. Figure 6 shows the SSIM values among the cochleagram images given in Fig. 4. These SSIM values will form the input for the next stage, the Forgery localization stage.

D. FORGERY LOCALIZATION

In the proposed method, the localization of the forged speech segments will be determined at this stage. In addition to

the SSIM values obtained between the cochleagram images in the previous stage, the length information of the speech segments is also considered. The ratio between the lengths of the segment pairs was used in the localization phase since the size of the copied and pasted segment pairs will be preserved to a certain extent, even if post-processing operations are applied to the speech during the forged audio creation.

Algorithm 1 Forgery Localization

Input: SSIM values between the cochleagram images, the size of segments

Output: Dup_Seg_Pairs

1. Sort the SSIM values in the S vector in descending order
 2. Create the SD vector based on the SSIM values in descending order.
 3. $SD = \{SSIM_1, SSIM_2, \dots, SSIM_n\}$
 4. Save segment pairs giving SSIM values in SD as SP vector.
 5. $SP = \{Seg_{11}, Seg_{12}, \dots, Seg_{xy}\}$
 4. **for** each segment pairs in SP **do**
 5. $R_{XY} = size(segment\ x) / size(segment\ y)$
 6. **if** $R_{XY} > percentage_rate$
 7. Dup_Seg_Pairs = $[x, y]$
 8. **end if**
 9. **end for**
-

For this purpose, in the algorithm given in Algorithm 1, Similarity vector $S = \{SSIM_{11}, SSIM_{12}, \dots, SSIM_{nn}\}$ shows the SSIM values obtained between the cochleagram images in the previous step, where n is the number of voiced segments in the speech. Then the elements of the S are sorted in descending order. SD denotes S sorted in descending order. Thus, the first element in SD will be the element that gives the highest SSIM value among the segment pairs. The length ratio R_{XY} given in Eq. (8) is calculated for the pair of segments from which each SSIM value in SD is obtained. If R_{XY} is greater than $percentage_rate$ the segment pair is marked as a copy-pasted segment. This process is continued until the ratio between the sizes of the segments falls below $percentage_rate$. In the study, the $percentage_rate$ was determined as 80%. All segments before the element that does not satisfy the condition are marked as forged segments.

$$R_{XY} = \frac{size(Segment\ x)}{size(Segment\ y)} \quad (8)$$

In Table 3, an example is given for the proposed localization algorithm using the SSIM values given in Fig. 7. After the SSIM values were sorted in descending order as SD , only segment 2 and segment 5 were marked as forged segments, since the ratio of segment 3's size to segment 4's size (37.97) remained below 80%. Thus, with the proposed localization algorithm, unlike the studies in the literature, for the first time, a static value was not determined for the threshold; a percentage rate was used by using the length information of the segments. When creating forged audio, the segment copied

TABLE 3. The SSIM values and the ratio of the sizes of the segments.

Segment pairs	SD	R_{XY}
Segment 2 and 5	0.9486	86.67
Segment 3 and 4	0.8457	37.97
Segment 2 and 3	0.8046	43.34
Segment 3 and 5	0.8006	50.01
Segment 4 and 5	0.7954	18.99
Segment 2 and 4	0.7939	16.46
Segment 1 and 3	0.7655	76.67
Segment 1 and 4	0.75645	29.11
Segment 1 and 2	0.7416	56.53
Segment 1 and 5	0.7411	65.22

by the attackers can be pasted into multiple audio parts. This forgery is multiple audio copy-move forgery. Moreover, using a dynamic threshold, the proposed method can detect multiple copy-move forgery.

IV. RESULTS

This section will first present the audio copy-move forgery dataset and performance metrics used in the proposed method. After that, effectiveness and robustness tests will be performed to show the effectiveness and robustness of the proposed method. Various experiments will then be performed to demonstrate the effect of cochleagram images extracted from voiced segments on audio copy-move forgery. In the next section, experiments on the determination of the percentage rate used as the dynamic threshold in the forgery localization phase will be given. Another section presents the visual results obtained with the proposed method for the speech recordings from the databases. Finally, the performance evaluation of the proposed and state-of-the-art methods will be given.

A. FORGED AUDIO DATASET

In order to show the results of the proposed copy-move detection method, two databases based on the TIMIT speech database [28] and the Arabic speech corpus database [29] are used. TIMIT speech database contains English-language speeches ranging from two to six seconds in length. The speech format is WAV, with a sample rate of 16 kHz. Arabic Speech Corpus contains spoken utterances. The format of the speech also is WAV. While creating the forged audio file from these databases, the speech was first split into voiced parts (segments) using the proposed VAD method [16]. Afterward, a random voiced part was copied in the speech, and this voiced part was pasted at a random position in the same speech. Each forgery segment is between 0.2 and 0.6 s long. The first created audio copy-move forgery database [16] consists of 368 forged audio files based on the TIMIT database, and the second created forgery database [17] consists of 715 forged audio files based on Arabic Speech Corpus. The

forged databases were created using the public TIMIT and Arabic Speech Corpus databases. TIMIT and Arabic Speech Corpus databases are widely used speech datasets primarily for phoneme recognition and other speech-processing tasks. Since these databases do not contain personal data, they are not directly subject to the General Data Protection Regulation (GDPR), primarily concerned with protecting personal data within the European Union. We also tested the proposed method with the CMFD dataset [30]. This forgery dataset contains 500 authentic recordings and 500 forged recordings for testing.

B. PERFORMANCE METRICS

Accuracy, Precision, Recall (TPR, true-positive rate), and F-score metrics were used in this study to comprehensively compare the proposed method with other methods. Accuracy is the ratio of both the correctly detected forged audios and the correctly detected authentic audios to the total audios. Accuracy is calculated according to the Eq. (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where TP is the number of forged audios that are detected as forged audios; TN is the number of authentic audios that are detected as authentic audios; FP is the number of authentic audios that are detected as forged audios; FN is the number of forged audios that are detected as authentic audios.

While precision shows the ratio of correctly detected forged audios to the total detected forged audios; recall shows the ratio of correctly detected forged audios to all the forged audios. F-score is the weighted average of precision and recall. Precision, Recall, and F-score metrics are given in Eq. (10), (11), and (12), respectively.

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

Accuracy, Recall, F-score and, Precision $\in [0,1]$. A higher F-score shows better overall performance in terms of precision and recall.

C. EFFECTIVENESS AND ROBUSTNESS TEST

After the attackers create the forged audio, the post-processing operations can be applied to the forged audio to remove the clues of forgery. For this reason, audio forgery detection methods must be robust against post-processing operations. In this experiment, five attacks were applied to the forged databases based on TIMIT and Arabic Speech Corpus to show the robustness of the proposed method to post-processing operations. These attack types are compression at 32 kbps and 64 kbps, noise addition at 30 dB and 20 dB, and median filtering. There are 368 forged audio for each of the five different attack types. As a result, 1840 forged audios were

TABLE 4. The result of the proposed method with post-processing terms of accuracy, precision, recall, and F1 score on forged database generated from TIMIT.

Attack Types	Accuracy	Precision	Recall	F-score
No attack	0.9476	0.9548	0.9755	0.9651
Noise addition with 30 dB	0.9456	0.9547	0.9728	0.9637
Noise addition with 20 dB	0.9375	0.9542	0.9620	0.9581
Median filtering	0.9496	0.9549	0.9783	0.9664
32-bit compression	0.9214	0.9466	0.9532	0.9402
64-bit compression	0.9274	0.9536	0.9484	0.9510

created from the TIMIT database. The other dataset from Arabic Speech Corpus, also contains forged audios with additional attacks: median filter, 32 kbps and 64 kbps, 30dB and 20 dB white Gaussian noise. Therefore, this dataset contains $30 \times 5 = 150$ forged audios. These post-processing operations will make it more difficult for audio forensic experts to detect audio copy-move forgery. Table 4 presents the Accuracy, Recall, F-score, and Precision values obtained by our method on the forgery dataset based on TIMIT under five attacks. As can be seen from Table 4, the proposed method is highly robust to compression, median filtering, and noise addition post-processing operations. When the F-score values of the proposed method are examined, they are 0.94 and above. Moreover, the F1-score values obtained under five attacks are quite close. This result shows that the proposed method is an attack-independent method.

The same experiment was carried out in the other forged database generated from the Arabic Speech Corpus database. The average results for Accuracy, Recall, F-score and, Precision values of the proposed method on this database are given in Table 5. As can be seen from the table, accuracy and F1-score values over 0.91 were obtained from the proposed method. In addition to the fact that the F1-score values obtained for each attack are quite close, the values obtained as a result of the noise attack are almost equal to the no attack situation. This proves that the proposed method is also quite robust to noise attack.

TABLE 5. The result of the proposed method with post-processing terms of accuracy, precision, recall, and F1 score on forged database generated from arabic speech corpus.

Attack Types	Accuracy	Precision	Recall	F-score
No attack	0.9386	0.8749	1	0.9351
Noise addition with 30 dB	0.9356	0.8746	1	0.9327
Noise addition with 20 dB	0.9412	0.8822	1	0.9425
Median filtering	0.9160	0.8743	0.9312	0.9432
32-bit compression	0.9112	0.8842	0.9434	0.9132
64-bit compression	0.9182	0.8942	0.9334	0.9232

The same experiment was also carried out in the CMFD database. The average results for Accuracy, Recall, F-score, and Precision values of the proposed method on this database are presented in Table 6. As can be seen from the table, the highest accuracy and F1-score value is 0.96 with the proposed method in the CMFD database. However, even if post-processing operations are applied to the speech recordings, the lowest F1-score value is 0.92. These results show that the proposed method is also quite robust to post-processing operations on the speech recordings in the CMFD database.

TABLE 6. The result of the proposed method with post-processing terms of accuracy, precision, recall, and F1 score on the CMFD database.

Attack Types	Accuracy	Precision	Recall	F-score
No attack	0.96	0.97	0.95	0.96
Noise addition with 30 dB	0.95	0.96	0.93	0.94
Noise addition with 20 dB	0.94	0.96	0.91	0.93
Compression	0.936	0.96	0.90	0.92

D. EFFECT OF COCHLEAGRAM IMAGES ON AUDIO COPY-MOVE FORGERY DETECTION

We compared the cochleagram features with the state-of-the-art speech features used in speech recognition to analyze the effect of cochleagram features on audio copy-move forgery detection as a first experiment. These features are spectrogram (Sgram) [27] and Constant Q Transform (CQT) [26]. Table 7 shows the Accuracy, Precision, Recall, and F-score values obtained by our method with the Sgram feature instead of the cochleagram on the forgery database based on TIMIT.

TABLE 7. The accuracy, precision, recall, and F-score values obtained by the proposed method with Sgram feature instead of cochleagram on the forgery database based on TIMIT.

Attack Types	Accuracy	Precision	Recall	F-score
No attack	0.9335	0.9467	0.9647	0.9556
Noise addition with 30db	0.8770	0.9424	0.8886	0.9147
Noise addition with 20db	0.8569	0.9407	0.8614	0.8993
Median filtering	0.8246	0.9377	0.8179	0.8737
32-bit compression	0.8952	0.9438	0.9130	0.9282
64-bit compression	0.9133	0.9452	0.9375	0.9413

As seen from Table 7, using the Sgram feature, the minimum accuracy value was 0.82, and the minimum F-score value was 0.87. These minimum values were obtained in the median filtering attack. At the same time, it is seen that the difference between the minimum values and the no-attack state is significant. These results show that the Sgram feature is not robust to the median filtering attack. As another experiment, the results of the proposed method

obtained by using CQT instead of cochleagram with these four metrics are shown in Table 8. When the results in Table 8 are examined, it is seen that there is not much difference between both accuracy and F1-score values in cases with and without attacks. For example, while the accuracy value of the proposed method is 0.89 in the no-attack situation, it is 0.88 in the case of adding 20db noise. Although this indicates that the CQT feature is attack-independent, it is not a highly accurate feature as these values remain below 0.90.

TABLE 8. The accuracy, precision, recall, and F-score values obtained by the proposed method with the CQT feature instead of cochleagram on the forgery database based on TIMIT.

Attack Types	Accuracy	Precision	Recall	F-score
No attack	0.8992	0.9569	0.9049	0.9302
Noise addition with 30dB	0.873	0.9552	0.8696	0.9104
Noise addition with 20dB	0.8831	0.9559	0.8832	0.9181
Median filtering	0.8609	0.9544	0.8533	0.901
32 bit compression	0.879	0.9556	0.8777	0.915
64 bit compression	0.879	0.9556	0.8777	0.915

As a final experiment, F1-score and Accuracy values obtained using Cochleagram, Sgram, and CQT features are given in Table 9 to show the superiority of the Cochleagram feature. The better result between the three features is highlighted in boldface. As can be seen from Table 9, using the Cochleagram feature yielded the highest results in both F-score and accuracy when compared to CQT and Sgram features. The cochleagram feature is superior, especially in median filtering and noise addition. For example, in the median filtering attack, the F-score values obtained from CQT and Spectrogram are 0.90 and 0.87, respectively, while the F-score value obtained with the Cochleagram feature is 0.96.

TABLE 9. The accuracy, and F-score values obtained by the proposed method with CQT, Sgram, and Cohle feature on the forgery database based on TIMIT.

Attack Types	F-score			Accuracy		
	CQT	Sgram	Cohle	CQT	Sgram	Cohle
No attack	0.9302	0.9556	0.9651	0.8992	0.9335	0.9476
Noise addition with 30dB	0.9104	0.9147	0.9637	0.873	0.8770	0.9456
Noise addition with 20dB	0.9181	0.8993	0.9581	0.8831	0.8569	0.9375
Median filtering	0.901	0.8737	0.9664	0.8609	0.8246	0.9496
32-bit compression	0.915	0.9282	0.9402	0.879	0.8952	0.9214
64-bit compression	0.915	0.9413	0.951	0.879	0.9133	0.9274

Our research involved testing a variety of features. For instance, in the case of noise addition with 20db, the F-score value obtained with the Sgram feature is 0.89, while the F-score value of cochleagram feature is 0.95. As the table

shows, the highest F1-score value with the Cochleagram feature was obtained as 0.96 in the no attack, noise addition with 30db and median filtering, while the lowest was 0.94 in the 32-bit compression attack. When the Accuracy values are examined, as with the F-score values, Cohle features gave higher accuracy than CQT and Speg features for all attacks. This comprehensive testing demonstrates the robustness of the cochleagram feature against attacks.

These experiments, which were carried out in the database generated from the TIMIT database, to measure the performance of the proposed method for different features, were also carried out on the forged database generated from the Arabic Speech Corpus database and the CMFD database. Table 10 shows TPR values obtained by using Cochleagram, Sgram, and CQT features.

TABLE 10. The TPR result of the proposed method on forged database generated from the arabic speech corpus.

	CQT	Sgram	Cochleagram
TPR	0.9832	0.9455	0.9985

As can be seen from Table 10, the highest TPR value in the forged database generated from the Arabic speech database, as in the forged database generated from the TIMIT database, was obtained from the cochleagram feature. Table 11 shows the F-scores and accuracy values obtained using the Cochleagram, Sgram, and CQT features from the CMFD database. The highest score between the three features is highlighted in bold. As shown in Table 11, using the cochleagram feature gave the highest results for both F1 score and accuracy compared to the CQT and Sgram features. The CMFD database also shows that cochlear features are remarkably robust to noise attacks compared to other features. This result is similar to the results from the TIMIT database. For example, a 20db noise addition resulted in an F-score of 0.93 for the cochleagram features, 0.88 for the CQT feature, and 0.86 for the Sgram feature. As with the F- scores, the Cohle features were better than the CQT and Speg features for all attacks in terms of accuracy values in the CMFD database. In the proposed audio copy-move forgery detection method, cochleagram images were extracted from the

TABLE 11. The accuracy, and F-score values obtained by the proposed method with CQT, Sgram, and Cohle feature on the CMFD database.

Attack Types	F-score			Accuracy		
	CQT	Sgram	Cohle	CQT	Sgram	Cohle
No attack	0.9202	0.9456	0.96	0.8892	0.9435	0.96
Noise addition with 30dB	0.8904	0.8947	0.94	0.853	0.8670	0.95
Noise addition with 20dB	0.8881	0.8693	0.93	0.8931	0.8479	0.94
Compression	0.89	0.90	0.92	0.878	0.92	0.936

segments because the cochleagram feature is robust against attacks and gives high accuracy.

E. DETERMINATION OF PERCENTAGE RATE IN LOCALIZATION

In the proposed method, the length information of the speech segments is also considered in addition to the SSIM values obtained between the cochleagram images. We experimented with this subsection to determine the percentage_rate value. For this purpose, 2208 forged audios, with and without attacks, were taken in the dataset. The length ratio between the duplicated segments, percentage_rate, and copy-paste segments in these forged audios was calculated. Figure 7 shows the percentage_rate numbers of audios from the dataset. As shown in Fig.7, the number of audios with a 70% percentage is 8, 50 with 75%, 500 with 80%, 950 with 85%, and 750 with 90%, respectively. Because more than 95% of the audio recordings in the dataset have a percentage_rate of 80% or more, the percentage_rate is determined as 80% in the proposed method. Thus, segment pairs with a percentage_rate value above 80% are marked as forged segments.

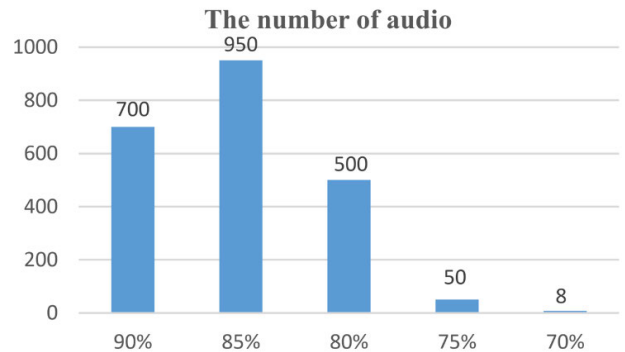


FIGURE 7. The percentage_rate numbers of audios from the dataset.

F. VISUAL RESULTS OBTAINED WITH THE PROPOSED METHOD

Although the proposed method is an audio forgery detection method, it is image-based. The proposed method first divides the audio recording into words, and images are created from each word. Then, the forged audio segments are detected according to the similarity of the word images. The word images obtained from the proposed method are shown in Fig. 8 for audio files from databases used. The 93_6-7.wav file given in the figure is taken from the CMFD database, the si456_1-7.wav file is taken from the forged audio database created from the TIMIT database, and the ARA NORM 0004_7-3.wav file is taken from the forged audio database created from the Arabic Speech Corpus database. The audio file names give information about the segment numbers repeated in the audio. The bolded segments in Fig. 9 display the copy-pasted segment pairs obtained using the proposed method. As seen in Fig. 8, the number of segment images obtained from the given forged audio is 8 for 93.wav, 8 for

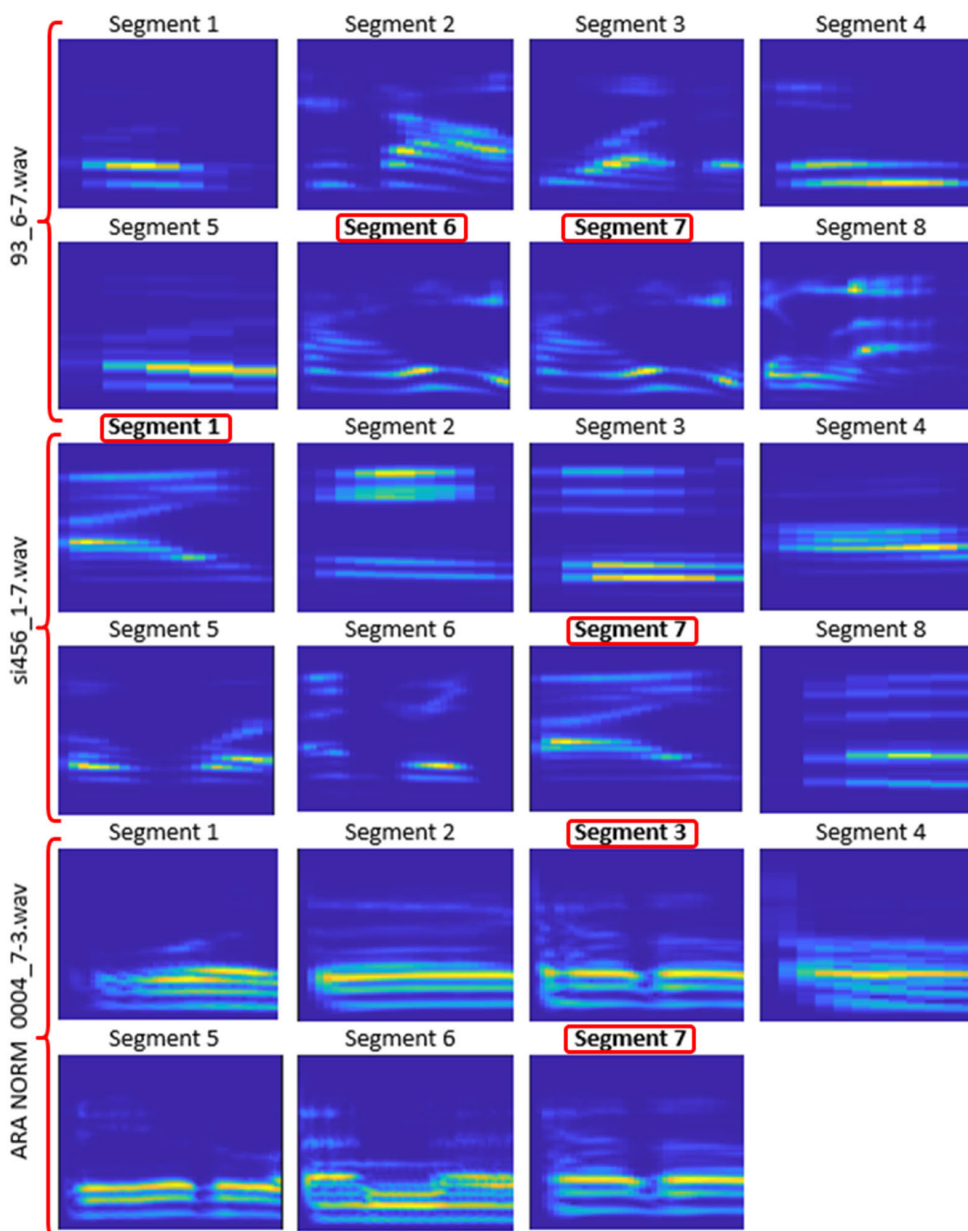


FIGURE 8. The visual results of the proposed method.

si456.wav, and 7 for ARA NORM 0004.wav. The segment numbers found due to detection, and those given in the file-name are the same.

The word images were obtained with all three methods for segment 1 and segment 7, which are duplicated words of the si456.wav audio file, to show that Cochleagram word images are a more effective visual representation than spectrogram and CQT word images as another visual result. Figure 9 shows the CQT, Sgram, and Cochleagram images of segments 1 and 7. As seen from Fig. 9, although Segment 1 and Segment 7 are the same segments, color and shape differences are seen in the CQT and Sgram images created for these words. However, the Cochleagram images are almost identical.

G. PERFORMANCE EVALUATION OF THE PROPOSED METHOD

The comparison of the proposed audio copy-move forgery detection method with other studies in the literature is presented in this part. These other studies are labeled as Lbp [10], Dft [13], Formant [14], DCT-SVD [9], MelSIFT [18], SW-CQCC [6], and CQSS [7]. The codes of all the compared methods except the CQSS method were written in Matlab because only the authors who proposed the CQSS method shared its codes in their study. CQSS method All the tests were done on a personal computer with Intel(R) Core(TM) i7-1165G7 CPU @2.80 GHz, 16 GB of DDR4 RAM, and Windows 10 operating system. The parameters of compared

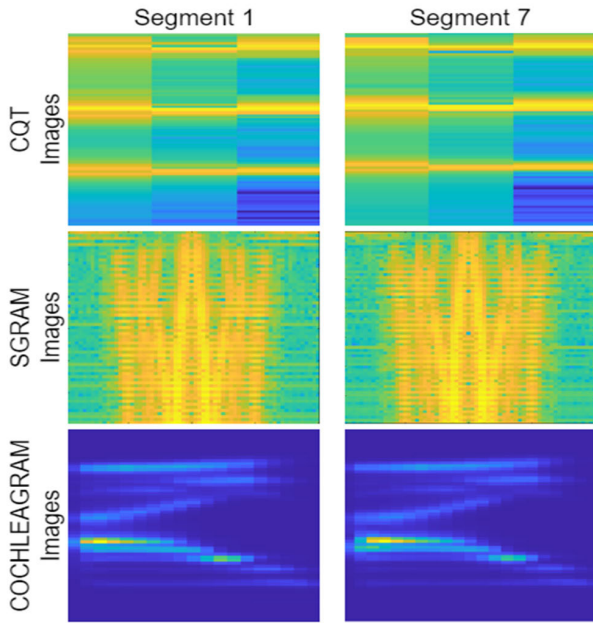


FIGURE 9. CQT, Sgram, Cochleagram Images of duplicated segments.

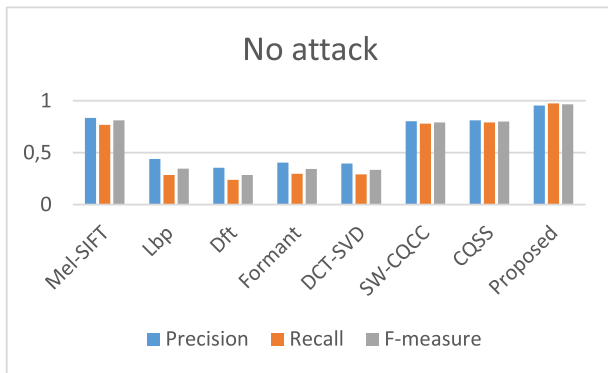


FIGURE 10. Detection results using Precision, Recall, and F-score on audios without additional attacks.

methods are empirically set as below after a series of experiments. In DFT the threshold of PCCs is 0.90. For LBP, the threshold of MSE is 50. In Formant, the threshold for DTW distance is 700. For SW-CQCC, the thresholds are 0.90. This part is split into subparts according to the results of the proposed method obtained with no attack types and each attack type.

H. COMPARISON OF THE PROPOSED METHOD AND OTHER STUDIES WITH NO-ATTACK AUDIOS

In the first experiment, the proposed audio copy-move forgery method was compared with other studies in the literature in case of not applying any attack to the audios. Figure 10 shows the average Precision, Recall, and F-score results from MelSIFT, Lbp, Dft, Formant, DCT-SVD, SW-CQCC, CQSS, and the proposed method. As can be seen Fig. 10, the proposed method is significantly better than MelSIFT, Lbp, Dft,

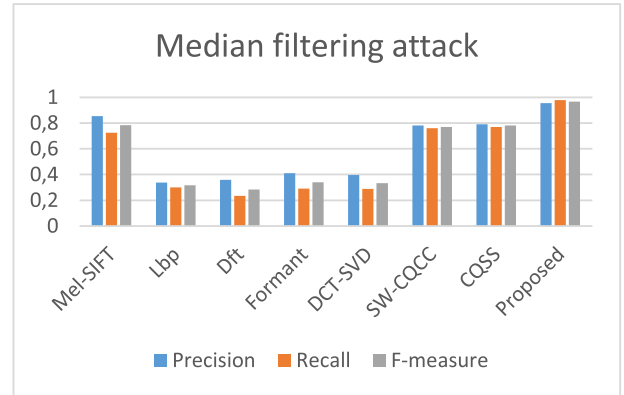


FIGURE 11. Detection results using Precision, Recall, and F-score on audios under median filtering attacks.

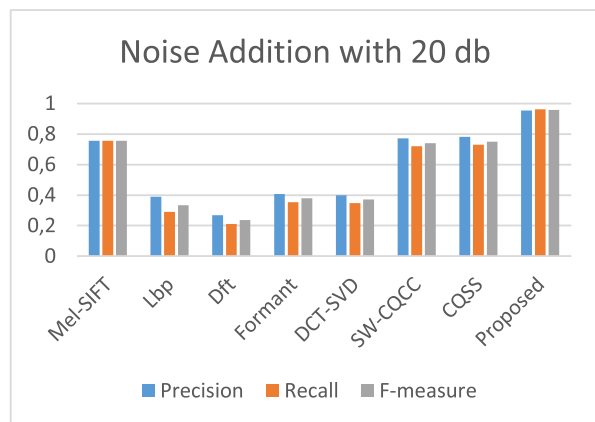
Formant, DCT-SVD, SW-CQCC and CQSS methods. While the F-score value of 0.96 was obtained from the proposed method, the closest F-score value to the proposed method was obtained by the MelSIFT method. The performance of the methods other than the MelSIFT method is quite low.

I. COMPARISON OF THE PROPOSED METHOD AND OTHER STUDIES WITH AUDIOS UNDER MEDIAN FILTERING ATTACKS

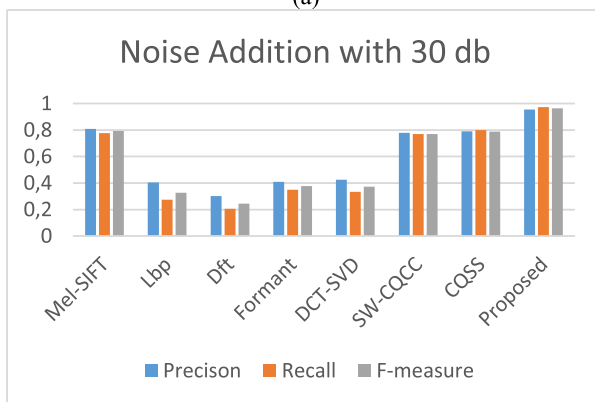
As a second experiment, the audios with the median filtering attack applied were analysed. For this purpose, precision, recall and F-score values were calculated for median filtered audios of both the proposed method and other studies in the literature. The average Precision, Recall, and F-score results from MelSIFT, Lbp, Dft, Formant, DCT-SVD, SW-CQCC, CQSS, and the proposed method are given in Fig. 11. When the results given in Fig. 11 are examined, the proposed method gives the highest F-score value as 0.96, while the DFT method gives the lowest value as 0.28. In this result, it is seen that the proposed method is much more robust to median filtering attack compared to other methods.

J. COMPARISON OF THE PROPOSED METHOD AND OTHER STUDIES IN AUDIOS UNDER NOISE ADDITION ATTACK

In this subsection, an experiment was conducted to show the robustness of the proposed method to noise addition attacks. In this experiment, the audios under noise addition with 20db and 30db were used in the dataset. Precision, Recall, and F-score metrics for the proposed method and other studies were obtained with the audio under noise addition. The performance metric results are shown in Fig.12 (a) for 20db noise and in Fig. 12(b) for 30db noise. As seen from Fig. 12, the Precision, Recall, and F-score values obtained for the proposed method due to adding 30db noise are 0.95, 0.97, and 0.96, respectively. These values are pretty high. Precision, Recall, and F-score values obtained for 20 dB noise addition are 0.95, 0.96, and 0.95, respectively. Although the amount of noise increases, it is seen that the values do not change and



(a)



(b)

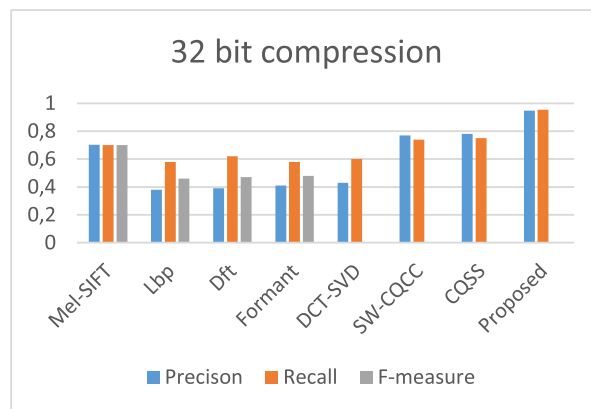
FIGURE 12. Detection results using Precision, Recall and F-score on audios under noise addition attacks (a) Under noise addition with 20 db (b) Under noise addition with 30 db.

are very high. Besides, apart from the proposed method, the highest F-score values obtained for 20db and 30db noise were 0.75 and 0.79, respectively, obtained by the MelSIFT method. The results obtained from this experiment, the proposed audio copy-move forgery detection method is quite robust against noise and gives high accuracy.

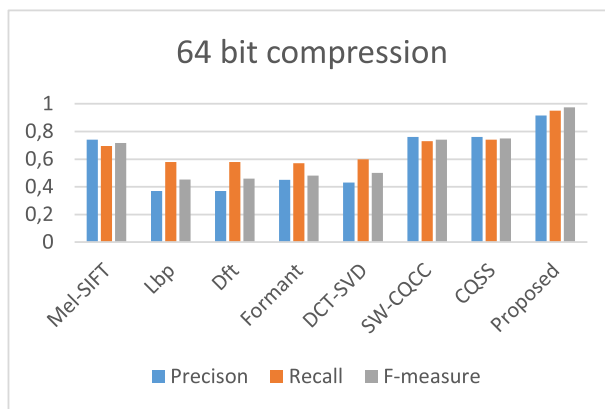
K. COMPARISON OF THE PROPOSED METHOD AND OTHER STUDIES WITH AUDIOS UNDER COMPRESSION ATTACK

This experiment analyzed the robustness of the proposed method and other methods in the literature against compression attacks. For this purpose, 32kbps and 64kbps compressed audios were used in the dataset. The proposed method and other methods have been tested with these audios. As a result of testing the methods with compressed audios, precision, recall and F-measure metrics were obtained for each method. The performance metrics obtained from the methods are given in Fig. 13 (a) for 32 kbps compression and in Fig. 13 (b) for 64 kbps.

As shown in Fig. 13, while the Lbp method gives the lowest average values, the proposed method gives the highest average values for both 32kbps compression and 64kbps compression. The proposed method obtained the F-score



(a)



(b)

FIGURE 13. Detection results using Precision, Recall, and F-score without post-processing audios. (a) Under 32 kbps compression attack (b) Under 64 kbps compression attacks.

values as 0.94 for 32kbps compression and 0.97 for 64kbps compression, respectively. The CQSS method gave the closest F1-measure value to the proposed method. When all the analysis results are evaluated, even if various attacks are applied to the audio files, the proposed method shows a very high performance when compared to other methods in the literature. The forgery detection method is quite robust against noise and gives high accuracy.

L. GENERAL PERFORMANCE OF THE PROPOSED METHOD AND OTHER STUDIES

As a final experiment, the average results of the proposed method and other studies in the literature were analyzed.

Table 12 shows the average precision, recall, and F-score values obtained by different CMFD methods, regarding all post-processing operations. The better result between the nine CMFD methods for each metric is highlighted in boldface. It can be found that for each metric, the proposed method always outperforms the existing methods.

V. DISCUSSION

In this section, we review what the results of the above experiments tell us and further discuss the possible reasons for effectiveness.

TABLE 12. Average metric values (%) obtained by different CMFD methods on the copy-move forged datasets.

Methods	Precision	Recall	F-score
Mel-SIFT	0.84	0.75	0.79
LBP	0.43	0.32	0.34
DFT	0.37	0.25	0.32
Formant	0.40	0.35	0.38
DCT-SVD	0.40	0.32	0.36
SW-CQCC	0.79	0.77	0.78
CQSS	0.80	0.78	0.77
Proposed	0.95	0.98	0.97

Initially, five attacks were applied to the forged databases based on the TIMIT, Arabic Speech Corpus, and CMFD datasets to show the robustness of the proposed method for post-processing operations. The accuracy, Precision, Recall, and F-score values obtained by the proposed method are more significant than 0.92 on the forged database generated from TIMIT, 0.87 on the forged database generated from Arabic Speech Corpus, and 0.90 on the forged database CMFD, respectively. This result shows that the proposed method is quite robust to post-processing operations. At the same time, these high results are preserved in the CMFD database. This result is due to using a dynamic threshold in the proposed method instead of choosing a static threshold specific to our database.

The second experiment concerned the effect of cochleagram images on forgery detection. For this purpose, the cochleagram features were compared with the state-of-the-art speech features (spectrogram (Sgram) and Constant Q Transform (CQT)) used in speech recognition. According to experimental results, using the Cochleagram feature yielded the highest results in both F-score and accuracy when compared to CQT and Sgram features. The superiority of the cochleagram feature is seen especially in median filtering and noise addition. For example, in the median filtering attack, the F-score values obtained from CQT and Spectrogram are 0.90 and 0.87, respectively, while the F-score value obtained with the Cochleagram feature is 0.96 on the forgery database based on TIMIT. The highest TPR value, 0.99 in the forged database generated from the Arabic speech database, was obtained from the cochleagram feature. Compared to other features, the CMFD database also shows that cochlear features are remarkably robust to noise attacks. This result is similar to the results obtained from the TIMIT database. For example, the F-score for the cochleagram features was 0.93, the CQT feature was 0.88 and the Sgram feature was 0.86 when 20 db of noise was added. As with the F-scores, the cochleagram traits outperformed the CQT and Speg traits in the CMFD database for all attacks. This result shows how robust the cochleagram feature is to attacks on all databases.

The third experiment was the determination of the percentage_rate value. To do this, 2208 fake audios, with and without

attacks, were taken from the dataset. The percentage_rate, i.e., the proportion of copied and pasted segments in these fake audio files, was computed. In the proposed method, the percentage_rate is set at 80%. Because more than 95% of the audio recordings in the dataset have a percentage of 80% or more. This way, pairs of segments with a percentage_rate value of more than 80% are marked as fake.

The last experiment concerns the proposed method's superiority over other methods. Surprisingly, under five types of commonly used post-processing operations, the proposed method significantly outperforms the state-of-the-art CMFD methods on three datasets with post-processing operations. In general, the experimental results demonstrate that the proposed method's average metric values are over 0.95.

The above improvements are due to the fact that the extracted cochleagram features are robust to post-processing operations. Furthermore, in the proposed forgery localization algorithm, a dynamic threshold was obtained using the inter-segment length ratio information. In conclusion, the proposed method seems more accessible and applicable in practice than the existing CMFD methods. However, since the method is segment-based, the proposed method will not give accurate results due to copying and pasting a syllable in a segment. Although it is not a meaningful forgery, it is possible to adapt the proposed method to be syllable-based instead of segment-based, but this would require much processing time. For this reason, improvements in the proposed method are aimed to be made effectively so that forged syllables can also be found.

Improvements to make higher-performance detection of audio copy move forgery possible create the basis of our future studies. To succeed in this aim, approaches that generate a problem-specific deep learning-based network will also be investigated. In addition, other new attack types that can be applied to the speech file are evaluated to aim to be robust to these attack types.

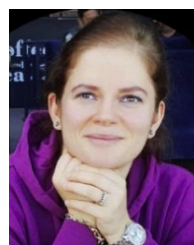
VI. CONCLUSION

This study suggests a new approach to detecting audio copy-move forgery. We observed that 1-D feature extraction and similarity calculation methods are used in the methods suggested in this field because the input data is an audio file. Considering the diversity of 2-D feature extraction and similarity calculation methods, this limits the field of voice forgery detection. On the other hand, a static threshold is used for the similarity threshold in all of the proposed methods. In this case, determining this threshold for each data set requires great experimental effort. For this purpose, we extracted cochleagram images from each audio part in the audio file. We used SSIM to calculate the similarity of cochleagram images. We obtained a dynamic threshold using the length ratio between segment pairs with the proposed forgery localization algorithm to locate the duplicated segments. Finally, we analyzed our method through extensive experiments against the state-of-the-art methods with three datasets. Experimental results show that our audio copy-move forgery detection method gives superior performance in

CMFD and is effective and attack-independent. In the proposed method, the VAD method may incorrectly detect segment boundaries in other data sets. Deep learning-based networks that can localize duplicated segments will be investigated to avoid these false detections.

REFERENCES

- [1] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," in *Proc. 9th workshop Multimedia Secur.*, 2007, pp. 63–74.
- [2] C. Kraetzer, K. Qian, M. Schott, and J. Dittmann, "A context model for microphone forensics and its application in evaluations," *Proc. SPIE*, vol. 7880, pp. 253–267, Jul. 2011.
- [3] H. Malik and H. Farid, "Audio forensics from acoustic reverberation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 1710–1713.
- [4] R. Yang, Z. Qu, and J. Huang, "Detecting digital audio forgeries by checking frame offsets," in *Proc. 10th ACM workshop Multimedia Secur.*, Sep. 2008, pp. 21–26.
- [5] J.-N. Xiao, Y.-Z. Jia, E.-D. Fu, Z. Huang, Y. Li, and S.-P. Shi, "Audio authenticity: Duplicated audio segment detection in waveform audio file," *J. Shanghai Jiaotong Univ., Sci.*, vol. 19, no. 4, pp. 392–397, Aug. 2014.
- [6] Z. Su, M. Li, G. Zhang, Q. Wu, and Y. Wang, "Robust audio copy-move forgery detection on short forged slices using sliding window," *J. Inf. Secur. Appl.*, vol. 75, Jun. 2023, Art. no. 103507.
- [7] Z. Su, M. Li, G. Zhang, Q. Wu, M. Li, W. Zhang, and X. Yao, "Robust audio copy-move forgery detection using constant Q spectral sketches and GA-SVM," *IEEE Trans. Dependable Secure Comput.*, vol. 20, no. 5, pp. 1–15, Oct. 2022.
- [8] Q. Yan, R. Yang, and J. Huang, "Copy-move detection of audio recording with pitch similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1782–1786.
- [9] F. Wang, C. Li, and L. Tian, "An algorithm of detecting audio copy-move forgery based on DCT and SVD," in *Proc. IEEE 17th Int. Conf. Commun. Technol. (ICCT)*, Oct. 2017, pp. 1652–1657.
- [10] 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.
- [11] Z. Xie, W. Lu, X. Liu, Y. Xue, and Y. Yeung, "Copy-move detection of digital audio based on multi-feature decision," *J. Inf. Secur. Appl.*, vol. 43, pp. 37–46, Dec. 2018.
- [12] N. T. Anh, H. T. T. Hang, and G. Chen, "One approach in the time domain in detecting copy-move of speech recordings with the similar magnitude," *Int. J. Eng. Appl. Sci. (IJEAS)*, vol. 6, no. 4, pp. 9–11, Apr. 2019.
- [13] X. Huang, Z. Liu, W. Lu, H. Liu, and S. Xiang, "Fast and effective copy-move detection of digital audio based on auto segment," in *Digital Forensics and Forensic Investigations: Breakthroughs in Research and Practice*. Hershey, PA, USA: IGI Global, 2020, pp. 127–142.
- [14] Q. Yan, R. Yang, and J. Huang, "Robust copy-move detection of speech recording using similarities of pitch and formant," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2331–2341, Sep. 2019.
- [15] K. Manepalli, P. V. Krishna, K. V. Krishna, and K. R. Krishna, "Copy and move detection in audio recordings using dynamic time warping algorithm," *Int. J. Innov. Technol. Exploring Eng.*, vol. 9, no. 2, pp. 2244–2249, Dec. 2019.
- [16] B. Ustubioglu, B. Küçükuğurlu, and G. Ulutas, "Robust copy-move detection in digital audio forensics based on pitch and modified discrete cosine transform," *Multimedia Tools Appl.*, vol. 81, no. 19, pp. 27149–27185, Aug. 2022.
- [17] A. Ustubioglu, B. Ustubioglu, and G. Ulutas, "Mel spectrogram-based audio forgery detection using CNN," *Signal, Image Video Process.*, vol. 17, no. 5, pp. 2211–2219, Jul. 2023.
- [18] B. Ustubioglu, G. Tahaoglu, and G. Ulutas, "Detection of audio copy-move-forgery with novel feature matching on mel spectrogram," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118963.
- [19] B. Ustubioglu, G. Tahaoglu, G. Ulutas, A. Ustubioglu, and M. Kilic, "Audio forgery detection and localization with super-resolution spectrogram and keypoint-based clustering approach," *J. Supercomput.*, vol. 80, no. 1, pp. 486–518, Jan. 2024.
- [20] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *J. Acoust. Soc. Amer.*, vol. 123, no. 6, pp. 4559–4571, Jun. 2008.
- [21] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds. Oxford, U.K.: Pergamon, 1992, pp. 429–446.
- [22] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.
- [23] R. V. Sharan and T. J. Moir, "Cochleagram image feature for improved robustness in sound recognition," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Jul. 2015, pp. 441–444.
- [24] M. Slaney, "Auditory toolbox for Matlab," Interval Res. Corporation, Palo Alto, CA, USA, Tech. Rep. 1998-010, 1998.
- [25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [26] P. Singh, G. Saha, and M. Sahidullah, "Non-linear frequency warping using constant-Q transformation for speech emotion recognition," in *Proc. Int. Conf. Comput. Commun. Inform. (ICCCI)*, Jan. 2021, pp. 1–6.
- [27] E. Sejdic, I. Djurovic, and L. Stankovic, "Quantitative performance analysis of scalogram as instantaneous frequency estimator," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3837–3845, Aug. 2008.
- [28] J. S. Garofolo et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. [Online]. Available: <https://catalog.ldc.upenn.edu/LDC93S1>
- [29] N. Halabi. (2016). *Arabic Speech Corpus. Oxford Text Archive Core Collection*. [Online]. Available: <https://en.arabicspeechcorpus.com/>
- [30] WuQinfang. (2020). *Copy-Move-Forgery-Detection-in-Digital-Audio-Forensics*. [Online]. Available: <https://github.com/WuQinfang/Copy-move-forgery-detection-in-digital-audio-forensics>
- [31] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 1841–1844.
- [32] J. Chen, S. Xiang, H. Huang, and W. Liu, "Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet," *Multimedia Tools Appl.*, vol. 75, no. 4, pp. 2303–2325, Feb. 2016.
- [33] V. Gupta, G. Boulianne, and P. Cardinal, "Content-based audio copy detection using nearest-neighbor mapping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 261–264.
- [34] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and mel-frequency cepstral coefficients," in *Proc. 5th Int. Conf. Digit. Telecommun.*, Athens, Greece, Jun. 2010, pp. 11–16.
- [35] H. Zhao, Y. Chen, R. Wang, and H. Malik, "Audio splicing detection and localization using environmental signature," *Multimedia Tools Appl.*, vol. 76, no. 12, pp. 13897–13927, Jun. 2017.
- [36] C. L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process.*, Sep. 2013, pp. 177–182.
- [37] Q. Liu, A. H. Sung, and M. Qiao, "Detection of double MP3 compression," *Cognit. Comput.*, vol. 2, no. 4, pp. 291–296, Dec. 2010.
- [38] D. Luo, R. Yang, B. Li, and J. Huang, "Detection of double compressed AMR audio using stacked autoencoder," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 2, pp. 432–444, Feb. 2017.
- [39] X. Lin and X. Kang, "Supervised audio tampering detection using an autoregressive model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2142–2146.
- [40] Y. Zhang, S. Dai, W. Song, L. Zhang, and D. Li, "Exposing speech resampling manipulation by local texture analysis on spectrogram images," *Electronics*, vol. 9, no. 1, p. 23, Dec. 2019.



BESTE USTUBIOGLU received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Karadeniz Technical University (KTU), Turkey, in 2010, 2013, and 2018, respectively. She is currently an Assistant Professor with the Department of Computer Engineering, KTU. She works on cybersecurity and machine learning.

...