Hindawi
Security and Communication Networks
Volume 2022, Article ID 9496748, 10 pages
https://doi.org/10.1155/2022/9496748

Research Article

GRACED: A Novel Fragile Watermarking for Speech Based on Endpoint Detection

Shuyun Zhou,[1] Meixin Song,[1] Qing Qian,[1] Wenjing Liao,[1] and Xiaofeng Gong[2]

[1]*School of Information, Guizhou University of Finance and Economics, Guiyang 550025, China*
[2]*Guizhou Science and Technology Information Center, No. 16 Kexue Road of Nanming District, Guiyang 550002, China*

Correspondence should be addressed to Qing Qian; qqian2018_p@163.com

Received 21 August 2022; Accepted 21 September 2022; Published 4 October 2022

Academic Editor: Hanzhou Wu

Content authentication and tampering detection of multimedia is a vital application by using digital watermarking. In this paper, we propose a novel fraGile wateRmArking of speeCh based on Endpoint Detection (namely GRACED) to verify the integrity of speech. Firstly, speech signal is framed word by word and each speech frame includes one intact nonsilence word. Subsequently, feature fusion is adopted to generate the fragile watermark which will be embedded into the coefficients of hybrid domain of discrete wavelet transform (DWT) and singular value decomposition (SVD). Finally, the tampering detection is accomplished without using any synchronous code to detect kinds of attacks. Several experiments are executed in order to quantify the performance of the proposed method. Experimental evaluation and comparisons with other schemes demonstrate that the signal-to-noise ratio of the proposed method is high with a favorable imperceptibility. Additionally, the tampering localization of various malicious attacks can be achieved without using synchronous code and the proposed scheme even can determine the attack types.

1. Introduction

With the advent of the era of big data, the relationship between big data and multimedia security has become closer; speech plays an important role in our life, such as military, courtrooms, and dissemination of policies [1–3]. In some case, speech content contains private information that can be used for judicial expertise. However, it could threaten the national security due to the digital multimedia can be manipulated easily by various software and the content of speech may also be modified or tampered by attackers during the transmission or storage [4, 5]. The "terminal-network-cloud" architecture based on big data brings more challenges to speech content authentication. Therefore, it is essential to evaluate the integrity and authenticate the content of speech.

Generally, there are two technologies to achieve content authentication including content-based identification and information hiding. The first one is perceptual hash function [6–8] and the second one is digital watermarking [9, 10]. Hashing technology produces hash sequence as hash value or message digest. The generated sequence will be stored in cloud and compared with its reconstructed hash sequence to verify the integrity of speech content. A reliable speech perceptual hash authentication algorithm [11] by using the static and dynamic characteristics of speech based on the coefficients of Mel frequency inverted spectrum is introduced. In the process of tampering detection, the hamming distance between the reconstructed hash sequence and the stored hash sequence is calculated to verify the authenticity. In order to achieve content authentication of encrypted speech in the cloud, an efficient encrypted speech authentication method [12] based on uniform sub-band spectrum variance and perceptual hashing is proposed. The reconstructed authentication digest and the original hashing sequence stored in the cloud are matched by hamming distance algorithm to achieve tampering detection. A robust hash method is introduced which is based on MFCC (Mel-Frequency Cepstrum Coefficients) and PCA (Principal Component Analysis) to verify the integrity and authenticate of speech content [13]. Experimental results show that the BER (Bit Error Rate) between the hash value of the

original audio and the tampered audio is low for perceptual manipulations. However, the solutions proposed in the above studies are segmented by using fixed-length framing. Meanwhile, original hash sequence needs to be stored in the cloud with more storage consumption.

On the other hand, digital watermarking is an essential technology to realize content authentication which embeds secure message into speech without noticeable perceptual distortion. Integer Wavelet Transform and Non-negative Matrix Factorization can be used to verify the content authentication of speech [14]. In authentication process, the tampered region can be located by comparing the reconstructed perceptual hashing with the extracted perceptual hashing version. Experiments demonstrate that the proposed scheme is sensitive to malicious tampering of encrypted speech. Two fragile watermarking schemes are proposed [15] by using LSB (Least Significant Bit) in hybrid domain of DCT (Discrete Cosine Transform) and the DST (Discrete Sine Transform). The proposed schemes are sensitive than LSB method in spatial domain but limited to tampering detection. The combination of modifying the least significant digits and G723.1 coding can be used to achieve the speech content authentication and tamper recovery [16]. In order to recover the tampered area, the compressed signal is generated by using G.723 coding and embedded into original speech. An audio watermarking algorithm is proposed in [17]. In this algorithm, watermarks are generated by compressed data of GBT (Graph Based Transform), and then the watermarks are embedded into the coefficients of LSFs (Line Spectral Frequencies) via the combination of LP (Linear Prediction) and DM-QIM (Dither Modulation-Quantization Index Modulation). A secured watermarking algorithm based on chaotic is introduced in [18]. The embedding information is the compressed data of DCT of the secret audio, and then the information is embedded into random sequences of matrixes of singular value via the combination of DWT and SVD (Singular Value Decomposition). The uniform sub-band spectral variance and spectral entropy are fused into fusion features by feature fusion, and the zero-one data of the watermark is determined by comparing the value of each fusion feature with the average value [19]. In [20], the speech is encrypted firstly, and then the G723.1 compression algorithm is used to compress the speech frame data. Finally, the compressed data is embedded into the LSBs of encrypted speech. Therefore, the embedded information can realize the integrity authentication and tampering recovery of the speech content. An audio watermarking scheme in the compressed domain is designed in [21]. In this scheme, the Huffman data of each MP3 frame is used to carry watermark. Experiments present good results in relation to inaudibility, robustness, and capacity rate. A novel blind digital audio watermarking scheme has been proposed in the wavelet and cosine transforms domain [22]. In order to achieve tampering detection and copyright protection, hash sequence is generated with SHA-512 to authenticate the integrity, and image is embedded to protect the copyright. A blind speech watermarking algorithm on a frame-by-frame basis is presented in [23]. The method perceptually manipulates the vector norms drawn from the FFT (Fast Fourier Transform) coefficients firstly and then modifies the speech signal through the combination of DPQIM (Downward Progressive Quantization Index Modulation) and BCIA (Boundary Constrained Iterative Adjustment) according to the watermark bits. A robust dual-domain twofold encrypted image-in-audio watermarking scheme is introduced [24]. Initially, the encrypted binary image is obtained. Then, the encrypted image and the host audio signal are decomposed by the hybrid of DTCWT (Dual-Tree Complex Wavelet Transforms), STFT (Short-Time Fourier Transform) and SVD. Finally, the singular value of encrypted image is embedded in the singular value of host audio signal. Taking the advantage of LWT (Lifting Wavelet Transform) and DCT, the encrypted watermark was embedded into the selected coefficient to ensure the stability of the watermark [25]. In addition, to improve the robustness of watermark, cyclic coding is introduced to correct the errors. In the tampering detection process, the extracted watermark and original watermark are compared to locate the tampered area.

The mentioned algorithms for authenticating the integrity of speech content based on hashing need to consume storage, the generated watermark is nonblind; most of the algorithms for authenticating speech content integrity through digital watermarking take fixed-length framing to implement watermark embedding. Embedding watermarks can affect the audibility of speech. Therefore, in order to solve above problems, we propose an efficient speech content authentication scheme based on endpoint detection. The main contributions are listed as follows.

(1) The watermark generating and embedding are focused on nonsilence segment of speech by GRACED. It can better guarantee speech audibility by effectively reducing the amount of watermark and reducing interference with silent frames.

(2) For desynchronization attacks, the misaligned location can be synchronized without extra synchronous codes in GRACED. Meanwhile, the implementation of synchronization does not require bit-by-bit search.

(3) According to the continuity of numbers, attack types can be determined in GRACED.

This paper is organized as follows: Section 2 illustrates the proposed authentication scheme. Section 3 introduces the experimental results of the proposed scheme. The conclusions are described in Section 4.

2. Proposed Method

In this section, we mainly present the proposed method GRACED. Three subsections are written to describe GRACED in detail. Subsection 2.1 introduces the proposed framing method. Subsection 2.2 describes the watermark generation and embedding principle of the proposed method. Subsection 2.3 gives a more specific explanation about the content authentication.

2.1. Speech Framing. For attackers, the purpose of malicious attack is to change content of speech signal. Obviously, a specific word can change the meaning of a message, and the modification of an entire word is more meaningful than the modification of random sampling points. Apparently, whether a speech word is tampered is attracted more concerned than the nonspeech segment. It is well-known that speech endpoint detection refers to the operation of determining the starting point and ending point of every speech segment. Therefore, speech endpoint detection is used to dynamically obtain speech segment instead of using the traditional fixed length frame in this paper. The segmentation method based on endpoint detection technology is illustrated in this paper and the details are described in the following steps.

(1) The speech signal S is first broken into frames. Each frame is denoted as $S_i = \{s_i(m)|1, 2, \ldots, M\}$ which contains M samples.

(2) For each speech frame S_i, the spectral centroid is calculated by using the following equation:

$$C_i = \frac{\sum_{m=1}^{M} (m+1) Y_i(m)}{\sum_{m=1}^{M} Y_i(m)}. \quad (1)$$

Here, C_i is the spectral centroid of the i th frame, Y_i is the coefficients of Discrete Fourier Transform of S_i, m is a variable from one to M, and M is the length of Y_i.

(3) Calculating the short-term energy of S_i according to the following equation:

$$E_i = \frac{1}{M} \sum_{m=1}^{M} \left| s_i(m) \right|^2. \quad (2)$$

Here, m is the sequence number of S_i, $s_i(m)$ is the amplitude of the m th sampling point of S_i, M is the length of S_i, and E_i is the short-time energy of the i th frame.

(4) Calculating two thresholds of the spectral centroid sequence C and the energy sequence E, respectively, as follows.

(i) Computing the histograms of the spectral centroid sequence and the energy sequence and denoted as H_1 and H_2, respectively.

(ii) Selecting two local maxima of the histogram H_1 and denoted as h_1 and h_2.

(iii) Calculating the threshold value of spectral centroid sequence using the following equation:

$$T_1 = \frac{\gamma_1 \cdot h_1 + h_2}{\gamma_1 + 1}. \quad (3)$$

Here, γ_1 is a user-defined parameter.

(iv) Selecting two local maxima of the histogram H_2 and denoted as h_1' and h_2'.

(v) Calculating the threshold value of energy sequence using the following equation:

$$T_2 = \frac{\gamma_2 \cdot h_1' + h_2'}{\gamma_2 + 1}, \quad (4)$$

Here, γ_2 is a user-defined parameter.

(vi) After calculating the threshold T_1 and T_2, the beginning point and ending point of each speech word can be calculated by

$$fg_i = \begin{cases} 1, & \text{if } C_i > T_1 \text{ and } E_i > T_2, \\ 0, & \text{others.} \end{cases} \quad (5)$$

Here, fg_i is the flag of the i th frame. The i th frame belongs to speech segment if the value of fg_i is one, otherwise, the i th frame belongs to nonspeech segment.

According to (5), the result can be shown as Figure 1. In this figure, Figure 1(a) is the waveform of an original speech, the red lines represent the start positions of each speech segment, and the blue lines represent the end positions of each word. Meanwhile, Figure 1(b) is the flags after endpoint detection for the speech. Here, the value of ordinate is used to indicate whether the sampling point belongs to the speech segment. It can be seen that speech signal can be divided into speech segments and nonspeech segments.

Therefore, the speech signal can be divided into N speech segments which are denoted as $\{S_1, S_2, \ldots, S_n, \ldots, S_N\}$.

2.2. Watermark Generation and Embedding Algorithm. Figure 2 illustrates the overall architecture of the watermark generation and embedding process. The detailed introduction of each step is shown below:

2.2.1. Speech Framing. In the speech division processing, we adopt an endpoint detection algorithm to divide the original speech S into N frames which frame includes one speech word and denotes as S_n.

2.2.2. Watermark Generation. The watermark generation process includes six steps: feature extraction, feature fusion, feature watermark generation, frame number watermark generation, watermark connection, and watermark encryption.

(i) Feature extraction. In this step, 2-level discrete wavelet transform is performed on each frame signal S_n to obtain the detail component α_1 and the approximation component α_2 firstly. Then, three features of approximation component are extracted and denoted as $\{f_1, f_2, f_3\}$. f_1 presents the mean value of short-time Fourier transform coefficient. f_2 denotes the mean value of mel spectrum frequency cepstrum coefficient. f_3 is the mean value of the energy of root mean square.

(ii) Feature fusion. In order to reduce the amount of watermark and improve the robustness, the extracted features are merged as $F = \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3$. β_1, β_2, and β_3 are fusion coefficients which satisfy $\sum_{i=1}^{3} \beta_i = 1$.

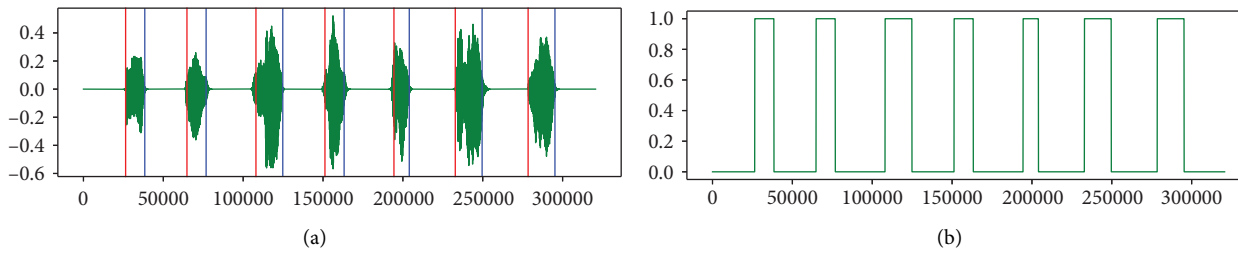

Figure 1: The result of speech framing. (a) Original speech and its endpoints; (b) flags of endpoint detection.

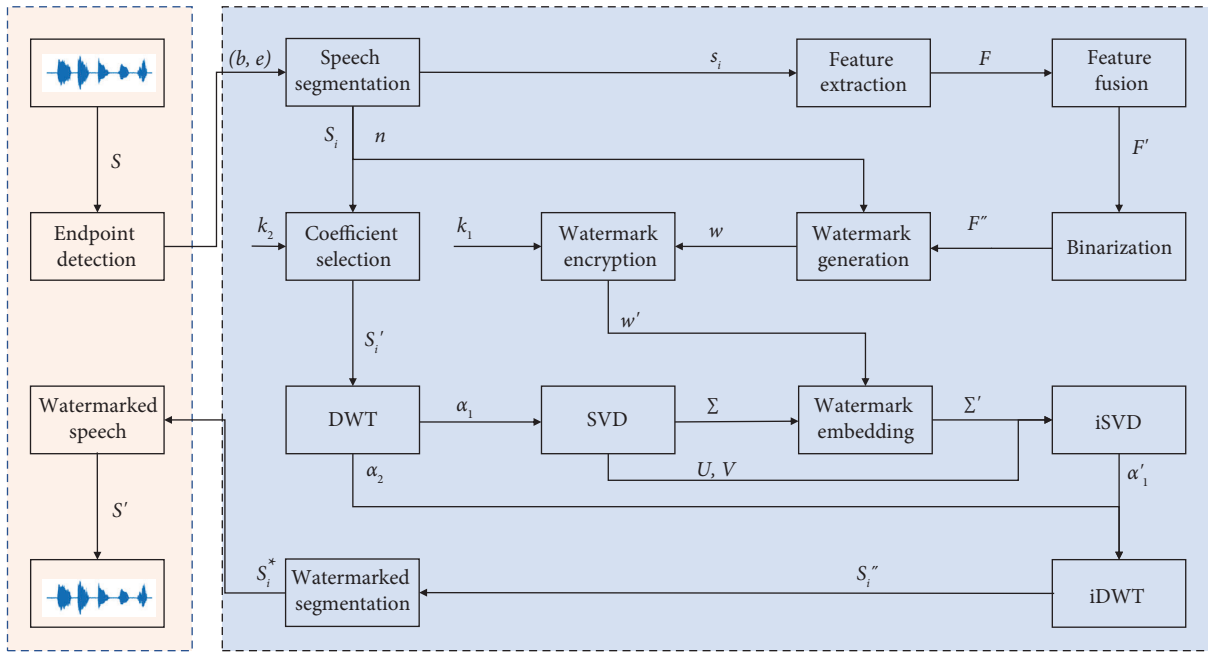

Figure 2: The block diagram of watermark generation and embedding.

(iii) Feature watermark generation. According to the fusion feature F, the perception hashing can be used to generate feature watermark as.

$$w_{n,k}^1 = \begin{cases} 0, & f_{k+1} < f_k, \\ 1, & f_{k+1} \geq f_k. \end{cases} \tag{6}$$

The feature watermark of the n th speech frame is denoted as $\omega_n^1 = \{\omega_{n,1}^1, \omega_{n,2}^1, \ldots, \omega_{n,k}^1, \ldots, \omega_{n,\lambda_1}^1\}$.

(iv) Frame number watermark generation. Each frame number n is converted into binary bits to produce the frame number watermark ω_n^2 with the length of λ_2. Here, $\omega_n^2 = \{\omega_{n,1}^2, \omega_{n,2}^2, \ldots, \omega_{n,k}^2, \ldots, \omega_{n,\lambda_2}^2\}$.

$$\omega_{n,k}^2 = \left\lfloor \left(\frac{n}{2^k} \right) \right\rfloor \%2, \quad k = 1, 2, \ldots, \lambda_2. \tag{7}$$

(v) Watermark connection. The feature watermark ω_n^1 and frame number watermark ω_n^2 are combined as the watermark of the n th speech frame shown as follows:

$$\begin{aligned} w_n &= \{\omega_{n,1}^1, \ldots, \omega_{n,\lambda_1}^1, \omega_{n,1}^2, \ldots, \omega_{n,\lambda_2}^2\} \\ &= \{\omega_{n,1}, \ldots, \omega_{n,k}, \ldots, \omega_{n,\lambda}\} \\ &\quad (\lambda = \lambda_1 + \lambda_2). \end{aligned} \tag{8}$$

(vi) Watermark encryption. A group of pseudorandom sequence $G = \{g_k, 1 \leq k \leq \lambda\}$ is obtained using the logical regression function shown as follows:

$$g_k = \rho \cdot g_{k-1} \cdot (1 - g_{k-1}), \quad 3.5699 \leq \rho \leq 4. \tag{9}$$

Subsequently, the produced sequence G is sorted in ascending order. Then the watermark w_n can be encrypted by using the index of the sorted pseudorandom sequence to disturb the position.

2.2.3. Watermark Embedding. In order to verify the integrity of a speech, the generated watermarks are embedded into speech signal. The steps of watermark embedding are illustrated as follows.

(i) Speech division. Based on the Sec. 1, the speech signal is divided into N words. Each word S_n expresses one frame.

(ii) Position selection. In order to improve the security of GRACED, partial sampling points from the n th speech frame S_n are selected to carry watermark by a secret key k_1.

(iii) DWT transformation. DWT is performed on the selected sampling points. After that, the detail component α_1 and the approximation component α_2 can be obtained.

(iv) Subsegmentation. The detail component α_1 is divided into λ subsegments ($\lambda = \lambda_1 + \lambda_2$) and denoted as $\alpha_1 = \{Y(1), Y(2), \ldots, Y(j), \ldots, Y(\lambda)\}$.

(v) Bit embedding. In this step, the singular value decomposition is executed on each segment $Y(j)$ to obtain the singular value Σ_j firstly.

Then, the obtained singular value Σ_j is applied to carry one bit watermark using the following equation. Here, $\mu = \lfloor \Sigma_j/\Delta \rfloor$. Δ represents quantization step.

$$\Sigma_j' = \begin{cases} \mu \times \Delta + \dfrac{\Delta}{2}, & \mathrm{mod}(\mu, 2) = w_{n,j}, \\ \mu \times \Delta - \dfrac{\Delta}{2}, & \mathrm{mod}(\mu, 2) \neq w_{n,j}. \end{cases} \tag{10}$$

After that, the inverse singular value decomposition is performed to obtain the watermarked subsegment $Y'(j)$.

This step is repeated until all watermark bits are embedded and obtained the watermarked detail component $\alpha_1' = \{Y'(1), Y'(2), \ldots, Y'(j), \ldots, Y'(\lambda)\}$.

(vi) Inverse transforms. Inverse discrete wavelet transform is performed on the watermarked detail component α_1' and the approximation component α_2 to obtain the watermarked speech subsegment.

2.2.4. Connection. From step 2 to step 3, each speech frame is selected to carry the generated watermark. Then, all watermarked speech frames are connected to acquire the watermarked speech S^*.

2.3. Content Authentication Algorithm

2.3.1. Speech Framing. Based on the speech framing method in Sec. 2.1, the watermarked speech S^* is divided into N frames. Each frame contains one watermarked speech word and denoted as S_n^*.

2.3.2. Feature Watermark Reconstruction. According to step 2 in watermark generation, the reconstructed feature watermark can be calculated and denoted as w_n^1 for the n th frame.

2.3.3. Watermark Extraction. The watermark extraction process is illustrated as follows.

(i) Position selection. For each speech frame S_n^*, partial sampling points are selected by the secret key k_1.

(ii) Frequency domain transformation. Discrete wavelet transform is performed on the selected sampling points to obtain the detail component α_1^* and the approximation component α_2^*. Subsequently, the detail component α_1^* is divided into λ segments ($\lambda = \lambda_1 + \lambda_2$) and singular value decomposition is executed on each segment to acquire the singular value Σ^*.

(iii) Watermark extraction. Based on (11), watermark bits can be calculated one by one.

$$w_{n,k}^* = \begin{cases} 0, & \mathrm{mod}(\mu, 2) = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{11}$$

(vi) Watermark decryption. The logical regression function is performed to generate a group of pseudorandom sequence which is sorted in ascending order. Subsequently, the order index can be used to decrypt the extracted watermark. The decrypted watermark of n th frame is denoted as w_n^*. Then, the feature watermark w_n^{1*} and frame number watermark w_n^{2*} can be separated.

2.3.4. Tampering Location. Calculate the information distance d between the reconstructed feature watermark w_n^1 and the extracted feature watermark w_n^{1*}. The result of tampering detection is defined as

$$T = \begin{cases} 0, & \text{if } d < \text{threshold}, \\ 1, & \text{otherwise.} \end{cases} \tag{12}$$

If $T = 0$ represents the corresponding frame is integrity and the frame number can be recalculated. Otherwise, it means this frame is tampered and the tampered frame number can be calculated by the absence of continuity in the numeric sequence.

3. Experimental Results

In this section, experiments are performed to verify the effectiveness of the designed audio watermarking algorithm. Simulation software is Python 3.9. Additionally, 240 speech signals (including 80 female speech signals, 80 male speech signals, and 80 children speech signals) are selected to evaluate the relative performance. Every speech recording is a 16-bits monaural file in WAVE format.

3.1. The Robustness of Framing. In this paper, endpoint detection technology is used to divide speech. Therefore, the robustness of endpoint detection method directly affects the accuracy of searching speech frames and the accuracy of tampering location. In order to quantify the robustness of framing, the following experiments are performed. Five

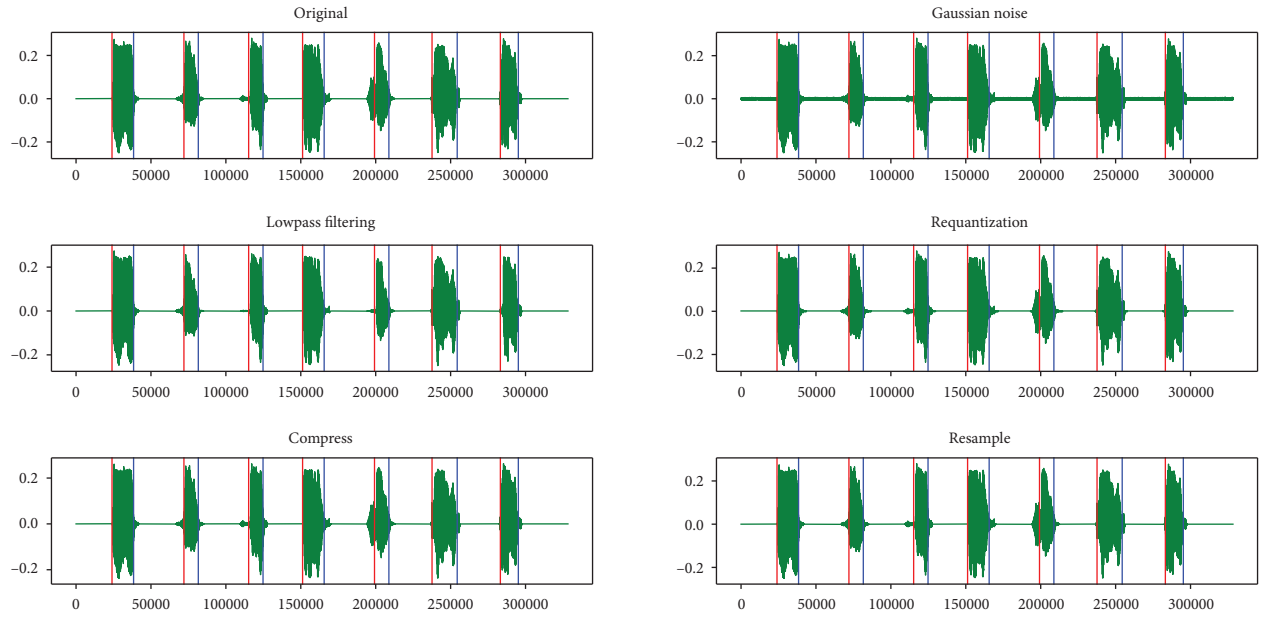

Figure 3: The robustness of speech framing.

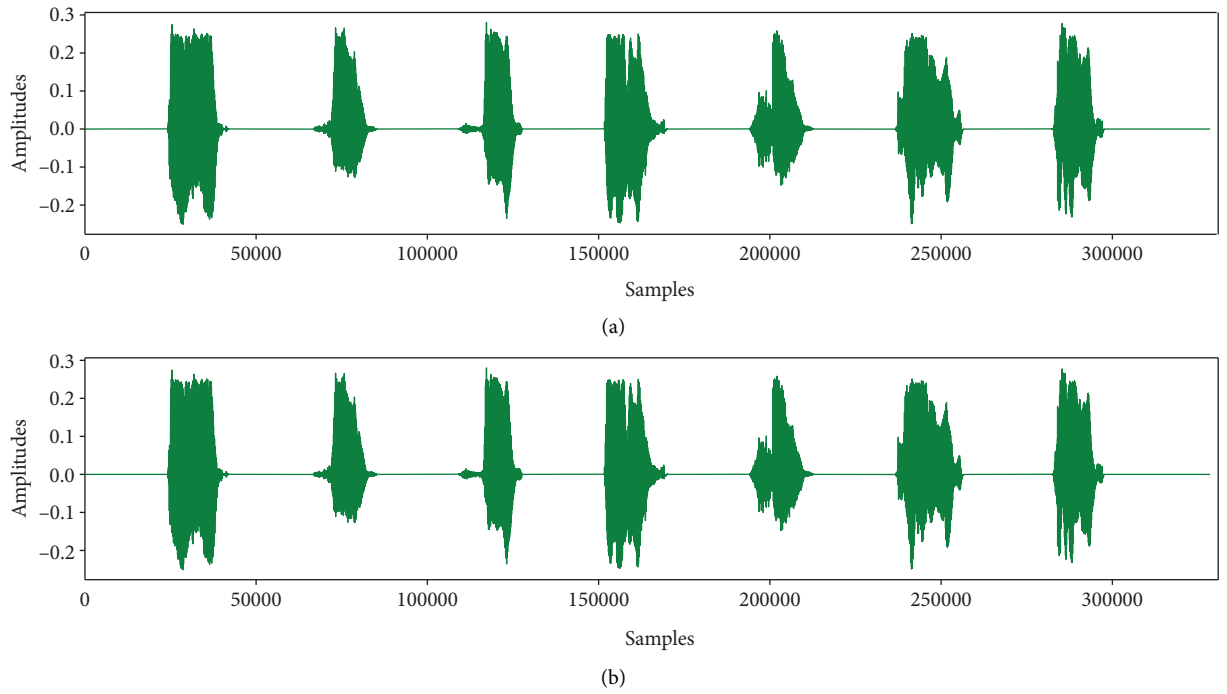

Figure 4: Waveform comparison of original speech and watermarked speech. (a) Original speech; (b) watermarked speech.

common signal processing is used to attack original speech signal such as low-pass filtering, quantization, noise addition, MP3 compression, and resampling. Subsequently, the attacked speech signal is framed according to the endpoint detection method. From Figure 3, it can be seen that attacked speech can be accurately divided into words after above conventional signal processing. Therefore, it is believed that the framing method has good robustness in this paper.

3.2. Inaudibility. Inaudibility usually can be classified into subjective assessment and objective assessment. On the one hand, the waveforms of the original speech and the watermarked speech are shown in Figure 4. It shows that there is no obvious difference between original speech and watermarked speech. On the other hand, SNR value is employed to measure the quality of watermarked speech and the equation is shown as follows. Wherein, x represents the sampling value of original speech sequence, y represents the

sampling value of watermarked speech, and L represents the number of sampling points of speech.

$$\mathrm{SNR} = 10\lg \frac{\sum_{l=1}^{L} x^2(l)}{\sum_{l=1}^{L} (x(l) - y(l))^2}. \quad (13)$$

In this experiment, different algorithms are chosen to evaluate inaudibility using the same speech signal and embedding capacity. From Table 1, it can be seen that the SNR values of GRACED are larger than Ref. [14], Ref. [16], and Ref. [25]. It means that GRACED can achieve the integrity authentication of speech signal with better inaudibility.

3.3. Fragility. Fragility represents that the watermark is sensitive to all kinds of malicious and nonmalicious attack. It means that the embedded watermark will be changed after malicious attacks (such as insertion attack, deletion attack, mute attack, and substitution attack) and common signal processing (such as resampling, low-pass filtering, and compression). The bit error rates (BER) between the generated watermark and extracted watermark can be used to evaluate the fragility. BER can be defined in the following formula:

$$\mathrm{BER} = \frac{\lambda_e}{\lambda}, \quad (14)$$

where λ_e is the number of different bits between the generated watermark and extracted watermark and λ is the total number of watermark bits.

In order to test the fragility of the proposed algorithm, several kinds of typical common signal processing are performed on the watermarked speech. The details are listed as follows.

(1) AWGN: 30 dB while Gaussian noise is added into the watermarked speech.

(2) Low-pass filtering: Low-pass filter with a cutoff frequency of 1.5 kHz is performed on the watermarked speech.

(3) Requantization: The watermarked speech is quantized from 16 bits per sample down to 8 bits per sample and requantized from 8 bits per sample up to 16 bits per sample.

(4) Compression: The format of watermarked speech is changed from WAV to MP3.

(5) Resampling: The sampling rate of watermarked speech is downsampled from 4.8 kHz to 1.6 kHz and then upsampled from 1.6 kHz to 4.8 kHz.

Table 2 shows the fragility of our proposed algorithm using different speech signals. In this table, each BER value is the average of the BER values of 80 speech signals. It can be found that the BER between the reconstructed watermark and the extracted watermark is around 0.5 after common signal processing. Obviously, the error bits are random. Therefore, it is considered that GRACED is very vulnerable to the operation of conventional processing.

Table 1: SNR (dB) between original speech and watermarked speech.

SNR (dB)	Ours	Ref. [14]	Ref. [16]	Ref. [25]
Male	57.43	37.87	29.57	30.37
Female	61.41	33.71	28.05	35.68
Child	64.61	30.51	27.55	39.02
Mean value	61.15	34.03	28.39	35.02

Table 2: BER comparison of different after common signal processing.

BER	Male	Female	Child
Without attack	0.00	0.00	0.00
AWGN	0.34	0.47	0.50
Low-pass filtering	0.49	0.50	0.50
Requantization	0.49	0.50	0.50
Compression	0.50	0.50	0.51
Resampling	0.56	0.54	0.59

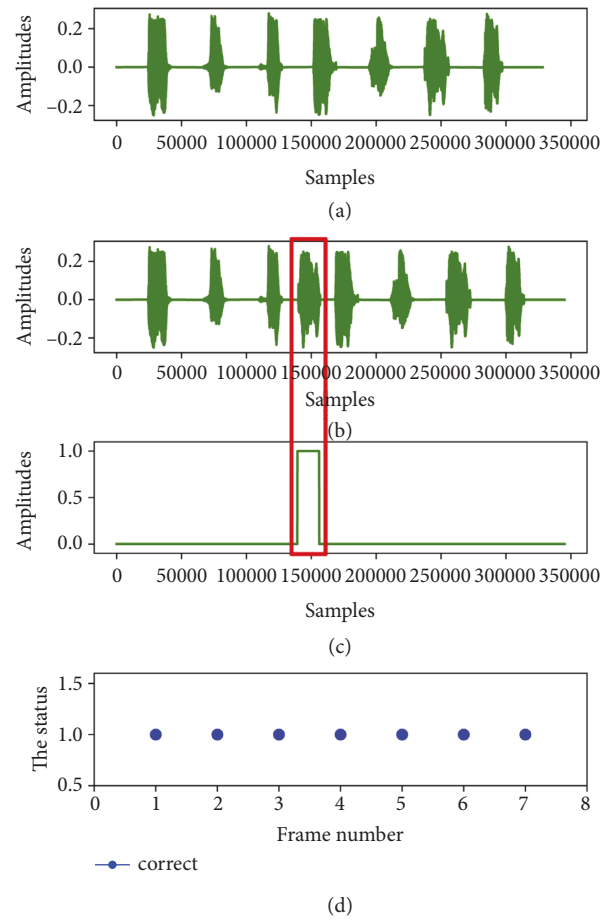

Figure 5: Tampering detection results of insertion attack. (a) Watermarked female speech; (b) tampered speech; (c) the location result of tampered detection; (d) the status of the frame number.

3.4. Tampering Detection and Location. Actually, malicious attacks are executed on speech signals to the purpose of modifying content information, and the modification of an entire word is more meaningful than the modification of

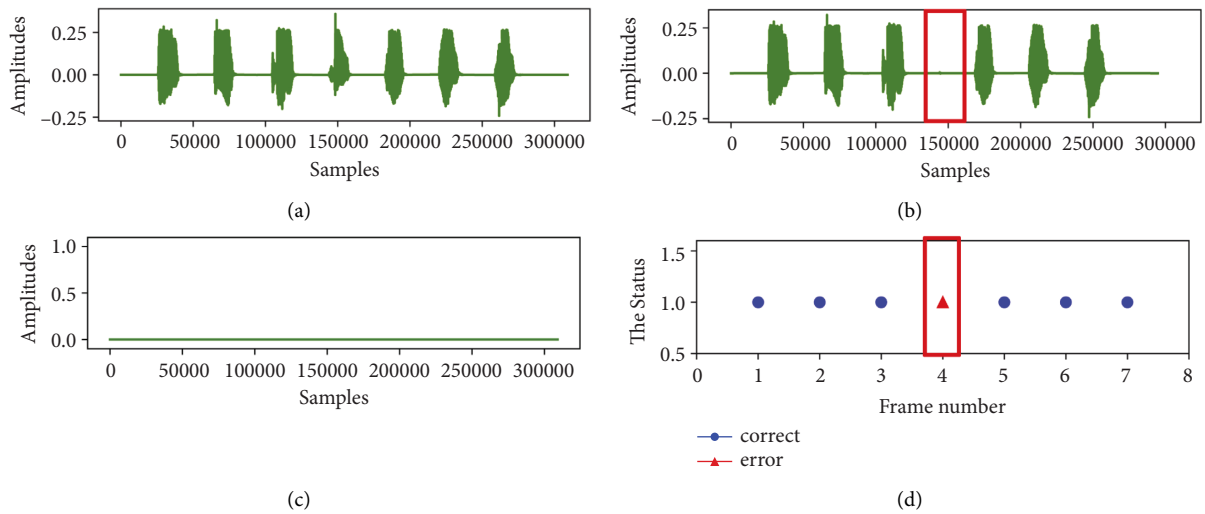

Figure 6: Tampering detection results of deletion attack. (a) Watermarked male speech; (b) tampered speech; (c) the location result of tampered detection; (d) the status of the frame number.

random sampling points. Therefore, a fragile watermark for speech authentication based on endpoint detection is proposed to verify the integrity of speech words. The tampering detection of malicious attacks is mainly focus on entire words rather than the speech segment with fixed length. In order to validate the proposed method, several malicious attacks are operated on the watermarked speech and the details are shown as follows.

3.4.1. Insertion Attack. In this attack, one word is inserted into the watermarked speech signal. Here, one word is inserted after the third word of the watermarked speech. Figure 5 shows the watermarked speech signal (Figure 5(a)), the attacked speech (Figure 5(b)), the location result (Figure 5(c)), and the status of frame number (Figure 5(d)). In the experiment, the watermarked speech is divided into seven frames. However, the attacked speech is divided into eight frames and the fourth frame is classified as tampered frame which is shown in Figure 5(c). Meanwhile, from the intact speech frames, the correct frame number sequence is {1, 2, 3, 4, 5, 6, 7}. Obviously, the extracted frame number sequence is a consecutive one as Figure 5(d). Hence, the attack type is considered as insertion attack. Therefore, it is believed that GRACED can accurately locate the tampering position without using synchronous code and judge the attack type.

3.4.2. Deletion Attack. It represents that one or more speech words are deleted from the watermarked speech. In the deletion experiment, the watermarked speech is shown in Figure 6(a), and the third word is deleted (including 9600 sampling points) as shown in Figure 6(b). According to the content authentication, it can be found that all words in Figure 6(c) are considered as unattacked words and the correct frame numbers are {1, 2, 3, 5, 6, 7} as shown in Figure 6(d). Hence, according to the continuity of frame numbers, the missing frame number can be confirmed. Meanwhile, according to the detection result and the missing frame number, the type of attack can be judged as deletion attack in the experiment.

3.4.3. Muteness Attack. It represents that one or more words are silenced in the watermarked speech. Obviously, the muteness attack also deletes the content of speech. Therefore, the muteness attack is considered as a kind of deletion attack. In the experiment, the third word contains 12000 sampling points as shown in Figure 7(a). The third word is silenced as an attacked speech as shown in Figure 7(b). According to the content authentication process, all speech frames are considered as integrity in Figure 7(c). However, the same as the deletion attack, the missing frame number can be confirmed according to the continuity of frame numbers in Figure 7(d). Here, in the correct frame number sequence, the intact frames are {1, 2, 4, 5, 6, 7}, the silenced content is the third word. Therefore, the tampering detection result and the missing frame numbers indicate that the type of attack is a muteness attack in the experiment.

3.4.4. Substitution Attack. In this attack, one or more speech words are replaced by another word or random sampling points. In the substitution attack, the fifth word is replaced by random sampling points and the attack speech as shown in Figure 8(b). According to GRACED, the attack speech is divided into seven frames and the fifth frame is judged as a tampered frame. Meanwhile, the extracted correct frame numbers are {1, 2, 3, 4, 6, 7} and the missing frame numbers can be determined. According to the status of frame number and the result of tampering location, the attacked type is regarded as substitution attack.

In this section, four attack experiments are executed. From what has been discussed above, we may safely arrive at the conclusion that GRACED can locate the tampered content of speech and judge the type of attack based on the location result of tampering detection and the status of frame number.

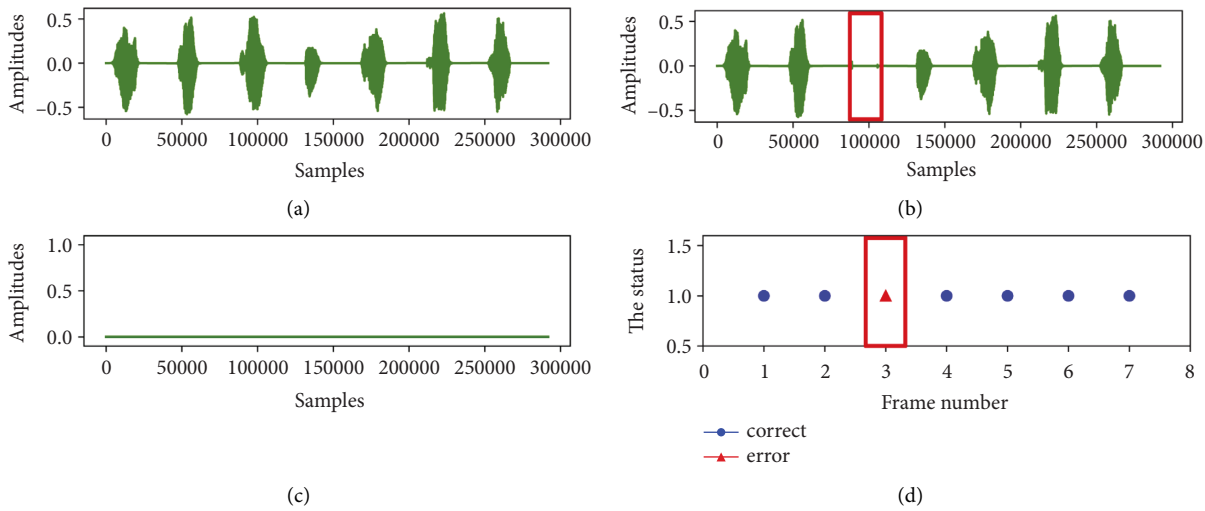

Figure 7: Tampering detection results of muteness attack. (a) Watermarked child speech; (b) tampered speech; (c) the location result of tampered detection; (d) the status of the frame number.

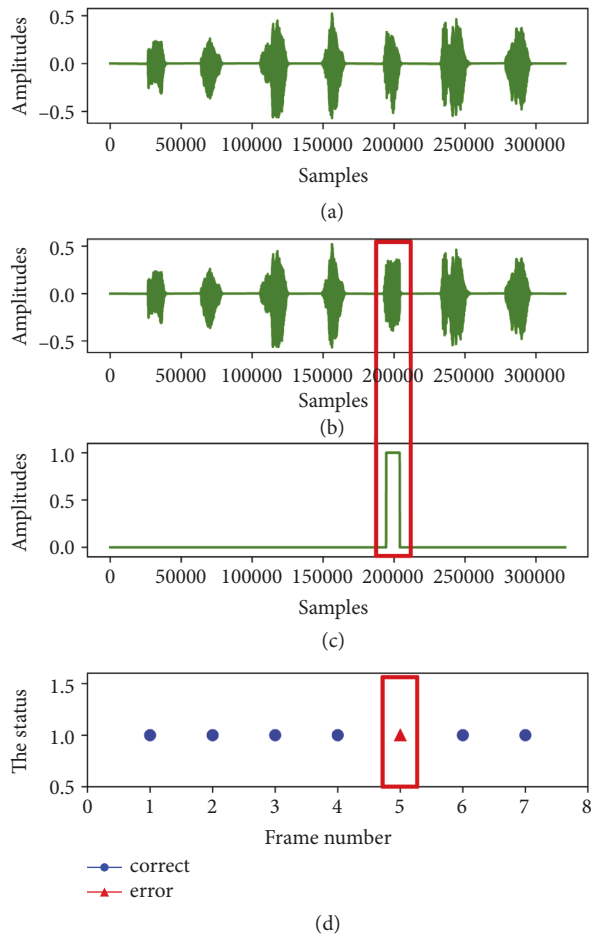

Figure 8: Tampering detection results of substitution attack. (a) Watermarked child speech; (b) tampered speech; (c) the location result of tampered detection; (d) the status of the frame number.

4. Conclusion

In order to realize the integrity authentication and tampered localization of speech content, a content authentication of speech based on endpoint detection GRECED is proposed in this paper. Firstly, speech signal is divided into frames word by word using endpoint detection. For each extracted word, its approximate and detail components can be calculated by discrete wavelet transform. Secondly, feature fusion and perceptual hashing are combined to generate authentication watermark. Finally, the integrity of the speech content is authenticated and tampering localization is achieved by that watermarking. Extensive experiments show that GRECED is sensitive to conventional processing. Meanwhile, the embedded watermark has good imperceptibility. Compared with other algorithms, the signal-to-noise ratio is high, and tampered localization of various malicious attacks can be achieved. Even more, the attack type can be identified by the continuity of frame numbers of those intact speech words.

Data Availability

The experimental data used to support the findings of this study can be obtained from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no financial, affiliations, intellectual property, personal, ideology, and academic conflicts of interest in this paper.

Authors' Contributions

Shuyun Zhou and Meixin Song contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China, under Grant 61902085, the Guizhou Provincial Science and Technology Projects, under Grant no. Qian Ke He Jichu-ZK[2021]YiBan312, and Technological Talents of Guizhou Provincial Science, under Grant no. Qian Jiao He KY Zi[2021]136.

References

[1] W. Lu, L. Li, Y. He, J. Wei, and N. N. Xiong, "RFPS: a robust feature points detection of audio watermarking for against desynchronization attacks in cyber security," *IEEE Access*, vol. 8, pp. 63643–63653, 2020.

[2] R. C. W. Phan, Y. Y. Low, K. S. Wong, and K. Minemura, "Strengthening speech content authentication against tampering," *Speech Communication*, vol. 129, pp. 41–57, 2021.

[3] E. Salah, K. Amine, K. Redouane, and K. Fares, "A Fourier transform based audio watermarking algorithm," *Applied Acoustics*, vol. 172, Article ID 107652, , 7 page, 2021.

[4] S. Helal and N. Salem, "A hybrid watermarking scheme using walsh hadamard transform and SVD," *Procedia Computer Science*, vol. 194, pp. 246–254, 2021.

[5] C. J. Chen, H. N. Huang, S. Y. Tu, C. H. Lin, and S. T. Chen, "Digital audio watermarking using minimum-amplitude scaling on optimized DWT low-frequency coefficients," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2413–2439, 2021.

[6] Y. B. Huang, T. F. Chen, Q. Y. Zhang, Y. Zhang, and S. H. Yan, "Encrypted speech perceptual hashing authentication algorithm based on improved 2d-henon encryption and harmonic product spectrum," *Multimedia Tools and Applications*, vol. 81, no. 18, pp. 1–24, 2022.

[7] Y. Y. Zhang, Y. B. Huang, D. H. Chen, and Q. Y. Zhang, "Long sequence biohashing speech authentication based on biometric fusion and modified logistic measurement Matrix," in *Proceedings of the 2021 International Conference on Computer Engineering and Application*, pp. 426–434, Kunming, China, June 2021.

[8] Y. Huang, T. Chen, X. Pu, and Q. Zhang, "Speech biohashing authentication based on power spectrum," *Journal of Physics: Conference Series*, vol. 2010, Article ID 012058, 2021.

[9] Y. Luo, D. Peng, Y. Sang, and Y. Xiang, "Dual-domain audio watermarking algorithm based on flexible segmentation and adaptive embedding," *IEEE Access*, vol. 7, pp. 10533–10545, 2019.

[10] K. M. Abdelwahab, S. M. Abd El-atty, W. El-Shafai, S. El-Rabaie, F. E. Abd El-Samie, and A. El-Samie, "Efficient SVD-based audio watermarking technique in FRT domain," *Multimedia Tools and Applications*, vol. 79, no. 9-10, pp. 5617–5648, 2020.

[11] L. Li, Y. Li, Z. Wang, X. Li, and G. Shi, "A reliable voice perceptual hash authentication algorithm," in *Proceedings of the International Conference on Mobile Multimedia Communications*, pp. 253–263, Cham Switzerland, November 2021.

[12] Q. Zhang, D. Zhang, and L. Zhou, "An encrypted speech authentication method based on uniform subband spectrumvariance and perceptual hashing," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 28, no. 5, pp. 2467–2482, 2020.

[13] D. Renza, J. Vargas, and D. M. Ballesteros, "Robust speech hashing for digital audio forensics," *Applied Sciences*, vol. 10, no. 1, pp. 249–316, 2019.

[14] C. Shi, H. Wang, Y. Hu, and X. Li, "A novel NMF-based authentication scheme for encrypted speech in cloud computing," *Multimedia Tools and Applications*, vol. 80, no. 17, pp. 25773–25798, 2021.

[15] R. Sripradha and K. Deepa, "A new fragile image-in-audio watermarking scheme for tamper detection," in *Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems*, pp. 767–773, Thoothukudi, India, December 2020.

[16] Q. Qian and Y. Cui, "A fragile watermarking algorithm for speech authentication by modifying least significant digits," in *Proceedings of the 2020 5th International Conference on Computer and Communication System*, pp. 680–684, Shanghai, China, May 2020.

[17] D. V. Vaishnavi, R. Dhanalakshmi, and S. Tripatha, "Digital watermarking and tamper detection in speech signal using blind detection," in *Proceedings of the 13th International Conference on Signal Processing Systems*, vol. 12171, pp. 340–346, Portugal, November 2022.

[18] K. P. Kumar and A. Kanhe, "Secured speech watermarking with DCT compression and chaotic embedding using DWT and SVD," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 10003–10024, 2022.

[19] Q. y. Zhang, D. H. Zhang, and F. J. Xu, "An encrypted speech authentication and tampering recovery method based on perceptual hashing," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 24925–24948, 2021.

[20] Q. Qian, Y. Cui, H. Wang, and M. Deng, "REPAIR: fragile watermarking for encrypted speech authentication with recovery ability," *Telecommunication Systems*, vol. 75, no. 3, pp. 273–289, 2020.

[21] S. Masmoudi, M. Charfeddine, and C. Ben Amar, "A semi-fragile digital audio watermarking scheme for MP3-encoded signals using Huffman data," *Circuits, Systems, and Signal Processing*, vol. 39, no. 6, pp. 3019–3034, 2020.

[22] V. L. Narla, S. Gulivindala, S. R. Chanamallu, and D. P. Gangwar, "BCH encoded robust and blind audio watermarking with tamper detection using hash," *Multimedia Tools and Applications*, vol. 80, no. 21-23, pp. 32925–32945, 2021.

[23] H. T. Hu, H. H. Chou, and T. T. Lee, "Robust blind speech watermarking via fft-based perceptual vector norm modulation with frame self-synchronization," *IEEE Access*, vol. 9, pp. 9916–9925, 2021.

[24] M. S. Islam, N. Naqvi, A. T. Abbasi et al., "Robust dual domain twofold encrypted image-in-audio watermarking based on SVD," *Circuits, Systems, and Signal Processing*, vol. 40, no. 9, pp. 4651–4685, 2021.

[25] Z. J. Yang, "Semi-fragile audio watermarking algorithm based on cyclic codes in LWT-DCT," *Tie Dao Xue Bao*, vol. 40, no. 8, pp. 91–97, 2018.