Secure multimedia exchange using voice biometric based security system for intellectual protection

Ghada Alhudhud, Duaa Alsaeed, Reem Alsaeed

Informaiton Technology Department/College of Computer and Information Sciences King Saud University, Riyadh, Saudi Arabia

E-mail: galhudhud@ksu.edu.sa, dalsaeed@ksu.edu.sa, 438203280@student.ksu.edu.sa

Abstract

In light of the digital transformation requirements, there are increasing demand of the significance of secure exchange of multimedia content across various digital channels. Among the commonly exchanged media are video and audio content. Multimedia secure exchange implies content protection and secure retrieval. The proposed method explores providing biometric authentication for multimedia content. Biometric authentication is based on creating a complex watermark with both voice biometric, and secret image. The voice biometric is analyzed using Mel Frequency Cepstral Coefficient ($MFCC$) for feature extraction. $MFCC$ coefficients are processed by the Gaussian mixture model (GMM) to produce the means matrix μs, covariance arrays of the Gaussians for all clusters in the voice signal, and negative log energy that represents the entropy. Accordingly, a unique voiceprint template is created. Hence, the watermarking will be obtained by embedding both the biometric template and the secret image in a given multimdia file. The embedding process will be performing lifting wavelet transform (LWT). LWT categorizes video frame content into special frequency bands such as low/low LL, Low/High LH, High/Low HL, and High/High HH particular frequency bands. The proposed method requires a) Creating the complex watermark composed of voiceprint template and secret image. b) Embedding the complex watermark using singular value decomposition (SVD) of the second level of LWT into the user video using the LL band of the cover video frames and the HH band of watermark video frames. For testing the methodology, five videos with different proprieties were chosen and the university logo was used as the secret image watermarking. In addition, a dataset was created during this research; containing voice recordings samples for male and female participants between the ages of 15 to 60 years old. The recorded audio clips were for given phrases in Arabic and English languages and the average of each clip is about 2.5 seconds. The total number of play counts in this experiment is 260 as we have five videos tested with 13 audio clips for accepted reads ($False/True$), and another 13 audio clips for rejected reads ($False/True$). Results showed in low False Accepted Rate of (9.2%), low False Rejection Rate of (12%), high True Acceptance Rate of (95.3%), and high True Rejection Rate of (92%). Based on seven matrices evaluations, we found points to improve the performance and accuracy of biometric authentication systems for video content protection. The verification process successful in distinguishing the original video with the original watermark from the tampered video using different voiceprints.Expermineation results exhibit the uniqueness of the complex watermark and verified the proposed method withstand different media processing attacksbased on the various voiceprints.

2010 Mathematics Subject Classification. **55N31**. 54H30, 68M11.
Keywords. biometric-based watermark, multimedia watermarking, image watermark, secure multimedia exchange.

Advanced Studies: Euro-Tbilisi Mathematical Journal 16, supplement issue 2 (2023), pp. 175–199.
DOI: 10.32513/asetmj/1932200823213

Tbilisi Centre for Mathematical Sciences.

Received by the editors: 15 November 2022.
Accepted for publication: 15 January 2023.

1 Introduction

In view of the requirements for digital transformation, there is a significant demand for secure multimedia exchange over various communication channels. Multimedia can be represented as audio, video, and animation in addition to traditional media [i.e. text, graphics, drawings, images, etc.]. Meanwhile, multimedia can be generally defined as a field concerned with the computer-controlled integration of text, graphics, drawings, still and moving images [video], animation, audio, and any other media where every type of information can digitally be represented, stored, transmitted and processed. Among the commonly exchanged media is the video and audio content. Video content exchange adds an exponential growth content delivery performance and effort as it requires to provide a reliable proof of authorization and security. However, the main issue with the video is that it is often vulnerable to various video tampering attacks and protected content. Cruz at al [9] video tampering can be defined as "a process of malicious alteration of video content, so as to conceal an object, an event or change the meaning; conveyed by the imagery in the video." In depth, analysis of many researches in recent years related to digital video manipulation and review of different type of tampering are reviewed in [12]. Multimedia content protection and ownership rights in multimedia are significant problems in multimedia processing and transmission. Multimedia content protection during transmission can be realized by cryptographic methods which secure the information content of multimedia data by encryption. Cryptographic techniques transform data into an unreadable form using a certain transformation based on the key. The encryption can be based on symmetric cryptography, where one key is used for encryption and decryption, or asymmetric cryptography, where one key is used for encryption and one key for decryption. However, cryptographic methods cannot protect the information content of multimedia after the decryption. After the decryption in the receiver, the information content is not protected anymore, and data may be copied easily and without quality degradation [30]. Song et al [29] introduced multimedia encryption and watermarking, a comprehensive survey of contemporary multimedia encryption and watermarking techniques which enable a secure exchange of multimedia intellectual property. Furthermore, multimedia cryptography provides an overview of modern cryptography, focusing on the current image, video, speech, and audio encryption techniques. In addition, digital watermarking evaluates various multimedia watermarking concepts and methods, including digital watermarking techniques for binary images. Most studies of video tampering detection and protection have focused on different techniques of watermarking, authentication, and encryption. In contrast, little attention has been paid to biometric techniques in this regard. Biometric systems substantiate the identity of an individual based on the distinguishing physiological (face, fingerprints, iris, hand geometry, retina, etc.) and behavioral characteristics (voice, signature, gait, keystroke, etc.) which are also known as biometric modalities or traits. Biometric systems are more convenient for use compared with traditional authentication methods. Examples of traditional authentication methods are knowledge-based passwords, PIN (Personal Identification Number), or a token-based (e.g., ID card) authentication. The remaining of this paper is organized as follows: review of related work in section 2, problem statement in section 3, while section 4 gives a background information on architecture of biometric based security systems and attacks on biometric systems. The proposed methodology is explained in section 5. It discusses in details the performance of proposed complex watermarking technique. Finally, section 6 concludes this paper.

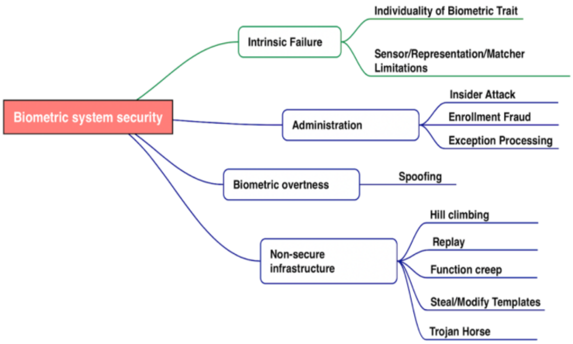

FIGURE 1. Categorizing biometric system security

2 Related work

The most crucial goal of information security is the integrity of exchanged content. In recent years, tampering and altering of digital video have become easier with the rapid development of computer technologies such as digital video editing tools. However, these rapid changes are having a severe effect on preserve the integrity of video content. To date, the researches have founded techniques for video media content protection in face recognition and fingerprint recognition, iris recognition, and signature recognition. Up to our knowledge, no previous study has investigated voice recognition for video media content protection. Cox et. al. [8] analyzed both the advantages and disadvantages of biometric technologies: fingerprint, voice, face, and iris. The results illustrated that the effectiveness of voice biometrics is attractive to investigate because of its prevalence in daily human communication. Voice allows incremental authentication protocols, and conversational biometrics provides inexpensive hardware and is easily deployable over existing ubiquitous communications infrastructure with higher accuracy and flexibility. Voice is, therefore, very suitable for pervasive security management. Voice recognition-based biometric is ”the technology by which sounds, words or phrases spoken by humans are converted into electrical signals, and such signals are transformed into coding patterns to which meaning has been assigned” [16] , [26].

2.1 Biometric Based Security Systems

Technologies based on voice biometric security imply defining the traits that can be used to identify individuals, known as authentication. Unlike other biometrics, voice-based authentication technologies can efficiently address this challenge without the need for any special biometrical devices simply using a microphone to capture a sample of the user’s voice. Voice authentication is the analysis of utterances by matching them to a stored voice model template. A voiceprint is the extraction of the unique features in each voice signal, such as pitch and unique modulation. It is a practical biometric system to protect sensitive information, fraudulent acts, and even protect sensitive and critical information [21].

1 illustrates the various causes of biometric system security issues categorized into four main

classes. two of the main classes are intrinsic failure and administration. Intrinsic failure are failures in authentication due to limitation in the microphone used in biometric system, feature extraction approach, voice print template preparation, or matching technique. On the other hand, extrinsic failures are failures due to external attacks by the unauthorized user who targets to tamper the voiceprint template or attempts to acquire the biometric traits of the authorized user and use them to create a copy of that biometric trait. It could be feasible with the existence of security holes in the biometric system design. Jain et al [34] categorize the different types of attacks targeting the biometric attack into:

- The attacks on the user interface (i.e., could not distinguish between the fake and original biometric sample).
- The attack made during data transmission by intercepting the poorly encrypted.
- Biometric data transmitted through insecure communication channels between system modules.
- Attack considered to damage the biometric template database by replacing the templates stored in the system database with the desired template to gain unauthorized access.

2.2 Architecture of biometric system

Biometric recognition is the field of identifying the identity of a person using his/her physiological and behavioral traits. Commonly-used biometric modalities or traits include fingerprint, face, iris, hand geometry, voice, palmprint, handwritten signatures, and gait (see 2 . Automated personal authentication; is based upon prime physical biometric characters or behavioral biometric characters. Physical characteristics; are based on the analysis of the invariable physiological characteristics of a person, such as a face shape and geometry, Fingerprints, the iris of the eye, and Palm, hand, or finger geometry. Behavioral characteristics; are based on the analysis of a person's behavioral characteristics. These characteristics include signature recognition, keystroke dynamics, gait recognition, and voice recognition. Accordingly, the biometric trait is defined as the inherent feature of the person to be recognized [28].

Biometric-based personal authentication methods may produce in two different modes: identification and verification modes [8]. In identification mode, system carries out a one-to-many matching to set up an individual's identity. The main purpose of identification is to prove the answer to the question: "Who am I?". In verification mode, system carries out a one-to-one matching to set up an individual's identity. The main purpose of verification is to prove the answer to the question "Am I who I say I am"?. The biometric system consists of four main modules which are sensor module, feature extractor module, matcher module, and decision module [33]. To collect and process raw biometric data and convert it into some useful information, the block diagram of biometric system is shown in 3. In biometric authentication, the data for the person's characteristics is compared with that person's biometric "template" to determine resemblance. A biometric template is a digital reference of distinct characteristics extracted from a biometric sample or models; templates used during the biometric authentication or verification process.

- Sensor Module This module is the first step in any biometric system for acquiring data by the sensor and it scans the user's raw biometric data to convert it into digital form.

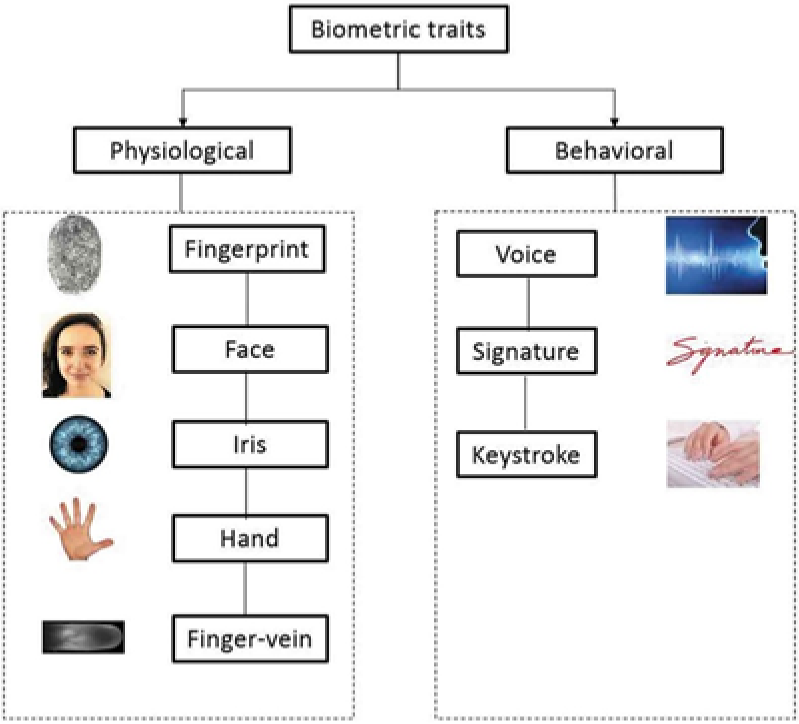

FIGURE 2. Examples of the physiological and behavioral body traits that can be used for biometric recognition [9]

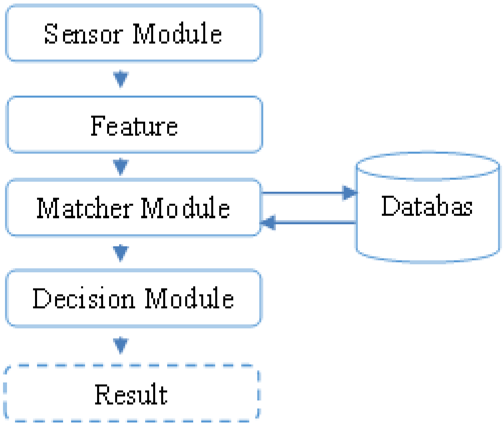

FIGURE 3. The block diagram of biometric system.

- Feature Extractor Module This module is observed and processes the raw data captured by the sensor and generates a biometric template to extract a set of salient or discriminatory features.
- Matcher Module This module in which the features extracted during recognition is compared the input sample with the templates that stored templates in the database to use the matching algorithm and generate matching scores.
- Decision Module This module accepts or rejects the user based on predefined security threshold. If match score is greater than predefined security threshold it will accept the user otherwise reject it.

2.3 Attacks on biometric systems

Biometric systems offer major advantages over traditional methods, but they are vulnerable to attacks [8]. As shown in 4, there are several possible attack points in the biometric system as shown in [8]:

- Attack at the sensor: A fake biometric trait to the sensor by an attacker to bypass recognition systems using synthetic fingerprints, the facial image of a person, and voice recorded.
- Replay attack: This channel is in between the sensor module and the feature extractor module. It is blocked to keep the biometric trait and stored somewhere. The previously stored biometric trait is replayed to the feature extractor to bypass the sensor.
- Attack on feature extractor module: An attacker pressurize the feature extractor module to produce the feature values accepted by the intruder instead of producing the feature values produced from the original data acquired from the sensor.
- Attack on the channel between the feature extractor and matcher: An attacker blocks the communication channel between the feature extractor and matcher modules and keeps the feature values of the actual user. These values can be replayed to the matcher later on.
- Attack on matcher module: It attacked to produce the high matching score as selected by the attacker to bypass the biometric authentication system regardless of the values received from the input trait set.
- Attack on the database: An attacker compromises the security of the database by adding new templates, modifying existing templates, and removing existing templates. It is not an easy task to attack a system database.so, To make a successful attack on the system database, some knowledge of the inner working of the system must be needed.
- Attack on the communication channel between the system database and the matcher: An attacker modifies or tampers the contents of the transmitted template. An attacker blocks the channel to keep, replace or alter the biometric template.
- Attack on the channel between the matcher module and decision module: An attacker may tamper with the match score, transmitted through a communication channel between the matcher module and the decision module. It tampers the match score to change the matcher module's original decision (accept or reject).

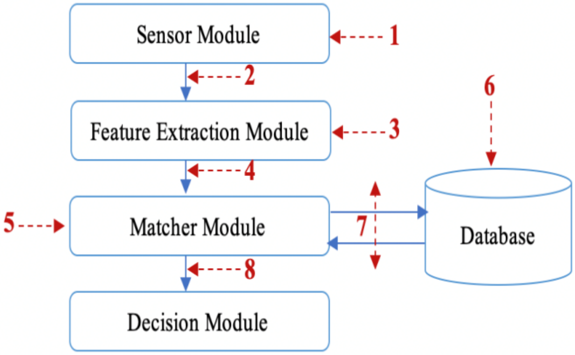

Figure 4. Attack points on biometric system: Red arrows illustrate attack points.

3 Problem statementand objectives

The use of biometric traits as an authentication technology has become widespread from door access to media content protection due to the need for better security in many fields. Much of the researches has been found techniques for media content protection in face recognition and fingerprint recognition. Until recently, no previous study has investigated voice print for video content protection. Particularly, this research paper will examine four main research questions:

1. How effective are the existing researches of securing video with different biometric watermarking?
2. How do the voice biometric-based watermarking techniques for preserving video contents from tampering by unauthorized users, correcting modified versions, and preventing the bitrate increase?
3. How does it examine the voiceprint verification techniques?
4. How effective the proposed approach against the currently available biometric-based video watermarking?

To keep data pure and trustworthy by protecting video data from intentional or accidental changes. So the goal of this thesis is to build a model to ensure that the video content is accurate and reliable and is not changed by unauthorized users or hackers.

The objectives of this research are to determine the following:

- Exploring currently available securing video with biometric watermarking.
- Identifying the gaps in the biometric based watermarking techniques for preserving video contents from tampering by unauthorized users, correcting modified versions, prevention of the bitrate increase.
- Examining the voice print/ verification techniques.

- Validating the proposed approach against the currently available biometric based video watermarking.

4 Background

4.1 Watermarking

It is an approach to protect images or videos by adding an invisible structure to the video content that can be used to authenticate it. If the video is copied and distributed, the watermark is distributed along with the it. Many approaches are available for protecting media content across a network such as: encryption, authentication, and time stamping. One approach to protect video content is to add an invisible structure to the video frames that can be used to authenticate it [4]. Watermark can be visible, fragile, semi-fragile, robust, spatial, image-adaptive or blind watermarking [**?**]. Demands in the sense that some are required to have applications must provide copyright protection and do so compromise with authentication, and vice versa [5]. Characteristic of watermark [11]:

1. Payload data: The numbers of bits encoded within the specified time by the digital watermark.
2. Robustness: Embedded watermark survival from geometric attacks against any image.
3. Security: Define the watermark's ability to be undetectable unauthorized access.
4. Imperceptibility: applies to the sum of resemblance between the image and the watermark. It is also, known as redundancies in perception.
5. Costs: Costs associated with watermarking embedding and extraction.
6. Fragility: sensitivity to any minor attempt to modify.

4.2 Biometric-based watermarking

Kant and Chaudhary [14] proposed a multimodal biometric image watermarking scheme through an integrity verification method using the hidden thumbnail feature vectors for secure authentication of multimodal biometrics data, face and fingerprint, respectively. However, some studies mainly focus on fingerprint, face recognition and iris of protected content video multimedia. In recent years, fingerprint-based authentication systems have been widely accepted in multimedia protection. As a kind of biological feature commonly owned by human beings, fingerprints have enough inter-user differences and individual stability. Maghsoudi and Tappert [17] provided the architecture for joint fingerprinting and decryption holds promise for a better compromise between practicality and security for emerging digital content. The scheme secures a multicast of anti-collision fingerprinted video in streaming applications to improve the quality of the reconstructed sequences at the decoder's side without introducing extra communication overhead [10]. Rane [22] provided a video scrambling and fingerprint embedding with Gaussian watermarks method for digital right protection. Fingerprinting algorithm extracts robust, discriminant, and compact fingerprints for video copy detection quickly and reliably [24].

Face recognition method is very common in video content protection. A morphing-based faces privacy protection method, independence from the video encoding used, and focus on its robustness, reversibility, and security properties [19]. A safe protection scheme of interesting facial region relies on security in video encryption algorithms [6]. In [2], a novel scheme was set to face verification issues in the scrambled domain; to make feature extraction from scrambled face images robust, a biased random subspace-sampling scheme is applied to construct fuzzy decision trees from randomly selected features and fuzzy forest decision using fuzzy memberships obtained from combining all fuzzy tree decisions. However, there is a little difference between different people, and the structures of all faces are similar; even the structures and shapes of facial organs are similar. Such characteristics are not much enough to detect human beings from human faces. In addition, the shape of the face is precarious; observation, angle, and age are all leading factors. In signature recognition [3] a system combines recent advances in biometric scanning technologies to offer a public key infrastructure solution to the issues posed by the growth in digital content sharing over broadband networks.

4.3 Audio Watermarking Using Discrete Wavelet Transform

In Discrete Wavelet Transform (DWT) domain, Attari et. al. [1] proposed audio watermarking scheme based on spread spectrum to achieved high robustness against different types of attacks re-sampling, dequantizing, gaussian noise addition MP3 compression, and low pass filter without significant perceptual distortion. In lifting wavelet transform (LWT) domain, Reynold [23] investigated schemes with 3-level LWT decomposition of the host audio, proposed watermark system embeds three types of watermarks within designated sub bands in a frame synchronous manner. According to a survey by [18], it was found that watermark designed for multimedia files integrity, content originality and ownership authentication needs to be enhanced.

4.4 Biometric Verification Approaches

Commonly, a biometric verification approache is used to ensure secure authentication to overcome the above listed intrinsic failures in the biometric-based security systems. In [32] Uludag and Jain presented a detailed survey of various biometric template protection schemes. They discussed their strengths and weaknesses in light of the security and accuracy dilemma. There are two approaches to deal with this issue, including feature transformation and biometric cryptosystem.

A popular clustering algorithm is known as K-means, which will follow an iterative approach to update the parameters of each cluster. More specifically, it will compute each cluster's means (or centroids) and then calculate their distance to each of the data points. The latter is then labeled as part of the cluster is identified by their closest centroid. This process is repeated until some convergence criterion is met, for example, when we see no further changes in the cluster assignments. A critical characteristic of K-means is that it is a hard-clustering method, which means that it will associate each point to one and only one cluster. A limitation to this approach is that no uncertainty measure or probability tells us how much a data point is associated with a specific cluster. So what about using soft clustering instead of a hard one? It is mean what Gaussian Mixture Models, or simply GMMs, attempt to do. In the next section, discuss this method further.

4.5 Gaussians mixture model (GMM)

GMM is defined as a linear combination of multiple Gaussian distributions that has multiple modes.Each Gaussian k in the mixture is comprised of the following parameters: a) μ to represent the mean, b) σ to represent covariance. The overall μs and σs for the Gaussian mixture provides a multivariate to include a mixing probability π that defines how big or small the Gaussian function will be. The below graphical representation illustrates the Gaussian mixture parameters for three clusters; a mix of three Gaussians. GMM has been implemented for recognition blocks. $GMMs$ are

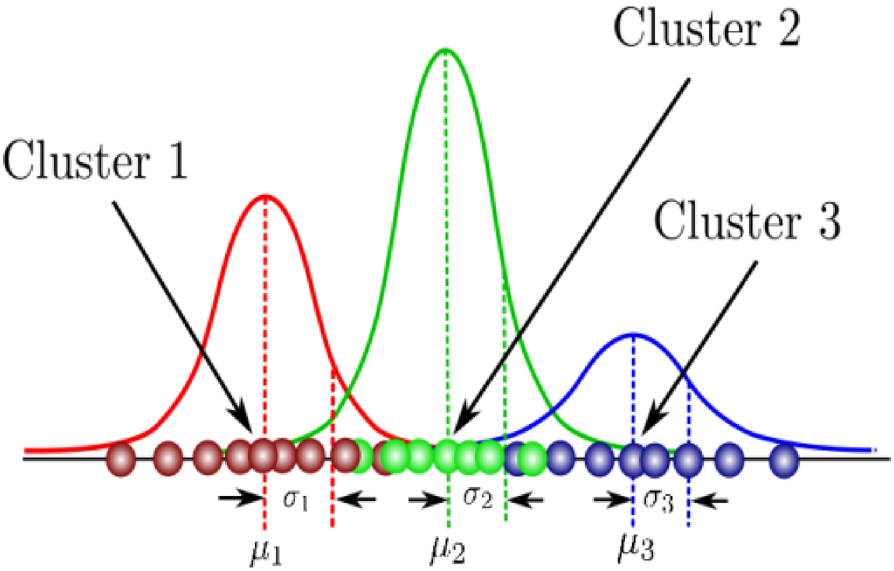

Figure 5. Graphical representation of Gaussian Mixture Parameters [25]

commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system at the core of the speaker recognition system where the μ represents the features- matrix and the σ-matrix, Gaussian probability density functions (PDFs) [27], [13].

The set of μ matrices and σ matrix for these Gaussians are used to ensure that each Gaussian fits the data points are belonging to each cluster, commonly known as maximum likelihood. According to Central Limit Theorem, the sum of independent samples from any distribution with finite mean and variance converges to the normal distribution as the sample size goes to infinity. Hence, GMM employs Machine Learning (ML) to distinguish the different voice signals through an unsupervised learning approach that clusters the voice data with some common characteristics.

4.6 Feature extraction using Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficient $MFCC$ is based on the known variation of the human perception for sensitivity at appropriate frequencies critical bandwidths .it is a compact representation of the spectrum when a waveform is represented by a summation of a possibly infinite number of sinusoids of an audio signal. It is are suitable for understanding humans and the frequency at which humans speak. One advantage of MFCC is no data-specific component and highly uncorrelated parameterization. MFCC coefficients contain information about the rate changes in the different spectrum bands. 2 represents sample of Mel Frequency Cepstral Coefficient (MFCC), Voice print MFCC coefficients; each column represents one extracted coefficient, each row represents the coefficients ts of on frame.

Cepstrum is typically derived by computing Discrete Cosine Transform of (symmetric) log power spectrum of a frame of speech; here, the log power spectrum [curve] is treated as a signal. So, the cepstral coefficients are measures of similarity between a sequence that represents the log power spectrum and cosine waves of different frequencies. The cepstral coefficients capture the rate with which the values of this sequence vary.

The first cepstral coefficient is the dot product of the log power spectrum with the periodic cosine wave. One period begins at the origin (f = 0) in the frequency domain and ends at f=2*Pi radians (or equivalently, sampling frequency). An illustration: the log power spectrum of a vowel has high energy in the low-frequency area (near f = 0) and low energy in the high-frequency area (near f = Pi). In other words, the slope of the log power spectrum curve in the range [0,Pi] has a negative slope in the case of vowels. Since this variation of the log power spectrum is similar to that of the cosine wave mentioned above, the first cepstral coefficient of a vowel speech frame will positively value. GMM based descriptors; defined in the previous section, are extracted using Mel Frequency Cepstral Coefficient (MFCC) that groups the voice data into clusters. Each cluster has its Gaussian distribution, and the corresponding Gaussian features record, such as σ array, and μ array, negative log energy (entropy) for each cluster.

Mel Frequency Cepstral Coefficient (MFCC) features extraction technique extracts using 20 filters linearly spaced in Mel scale from speech frame of 20ms keeping 50% overlap with adjacent frames. One advantage of MFCC is no data-specific component and highly uncorrelated parameterization. It contains information about the rate changes in the different spectrum bands. The MFCC feature extraction process is a 6-steps process shown in 6.

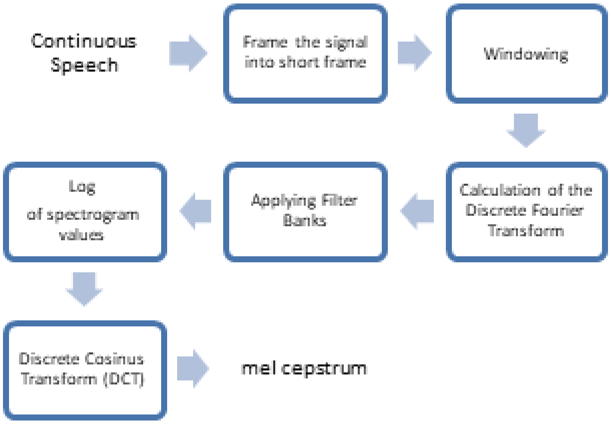
Continuous Speech
Frame the signal into short frame
Windowing
Calculation of the Discrete Fourier Transform
Applying Filter Banks
Log of spectrogram values
Discrete Cosinus Transform (DCT)
mel cepstrum

Figure 6. The MFCC feature extraction process

The MFCC feature extraction process; of input spoken utterances, returns the following voice biometric descriptors: log-energy, delta, and delta-delta the Mel-frequency cepstral coefficients over time. The MFCC coefficients are returned as $LXMXN$ array where:

- L is the number of frames in the spoken utterances signal
- M is the number of coefficients returned per frame.
- N is the number of channels.in the input signal

. These features are stored as voiceprint that will be added to the watermatk

4.7 Multimodal biometric methods

Multimodal of fingerprint and iris recognition in [15] provide framework combination of symmetric and asymmetric key; systems based on fingerprint, iris biometric and a layered encryption and decryption scheme to permit novel uses of protected video multimedia. A new approach of iris recognition is introduced in [20]; a fractional Fourier transform (FRFT), namely, dual parameter fractional Fourier transform, is biometrically encoded bitstream followed by the generation of the keys used in the encryption method in multimedia encryption. Multimodal biometric are much more reliable for building up a safer authentication system. Corcoran and Cucos [7] proposed a digital pattern generated by a methodical fusion of features extracted from iris image and fingerprint image; experimental results indicate that embedding of the digital tag in the image or audio does not tamper the perceptual transparency and is also robust to signal processing attacks.

5 Proposed method

The proposed method explores providing authentication, confidentiality, and integrity to the user's intellectual multimedia content through creating a complex watermark made of a voice biometric and a secret image. The voice biometric is analyzed using Mel Frequency Cepstral Coefficient ($MFCC$) for feature extraction. $MFCC$ coefficients processed by the Gaussian mixture model (GMM) to produce the means matrix μs, covariance arrays σs of the Gaussians for all clusters in the voice signal, and negative log energy that represents the entropy. Accordingly, a unique voiceprint template was created. Hence, the watermarking will be obtained by embedding both the biometric template and the secret image in the digital assets. The embedding process will be implementing lifting wavelet transform (LWT). LWT categorizes video frame content into low/low LL, Low/High LH, High/Low HL, and High/High HH particular frequency bands. In 7 illustrates the different parts of our proposed method. The proposed method requires the following steps:

1. Creating the dataset,

2. Generating the complex watermark composed of voiceprint template and secret image.

3. Embedding the complex watermark using obtaining singular value decomposition (SVD) of the second level of LWT into the user video using an LL band of the cover video frames and the HH band of watermark video frames.

4. Extracting Video

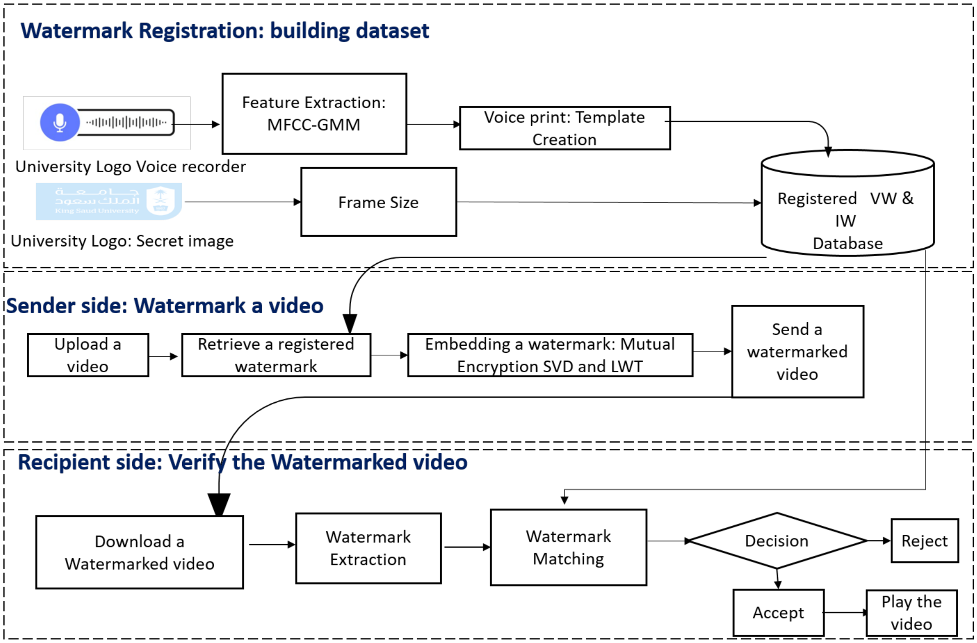
Watermark Registration: building dataset
University Logo Voice recorder
Feature Extraction: MFCC-GMM
Voice print: Template Creation
University Logo: Secret image
Frame Size
Registered VW & IW Database
Sender side: Watermark a video
Upload a video
Retrieve a registered watermark
Embedding a watermark: Mutual Encryption SVD and LWT
Send a watermarked video
Recipient side: Verify the Watermarked video
Download a Watermarked video
Watermark Extraction
Watermark Matching
Decision
Reject
Accept
Play the video

Figure 7. Block diagram of the proposed method process

The proposed method requires the following steps:

1. Creating the dataset: Enrolment
2. Generating the complex watermark composed of voiceprint template and secret image.
3. Embedding the complex watermark using obtaining singular value decomposition (SVD) of the second level of LWT into the user video using an LL band of the cover video frames and the HH band of watermark video frames.
4. Extracting Video

Next, we will discuss each step of the proposed methodology.

5.1 Dataset creation: enrolment by adding voice records to dataset

dataset was created during this research; containing voice recordings samples for 13 male and female participants between the ages of 15 to 60 years old. The recorded audio clips were for given phrases in Arabic and English languages and the average of each clip is about 2.5 seconds. The total number of play counts in this experiment is 260 as we have five videos tested with 13 audio clips for accepted reads ($False/True$), and another 13 audio clips for rejected reads ($False/True$). At the enrolement step, each registered organization will design an authorized member, 8 for recording the organization name, and this will be named VW. Each record is analyzed to extract biometric data to be further classified as key binding and key generation systems

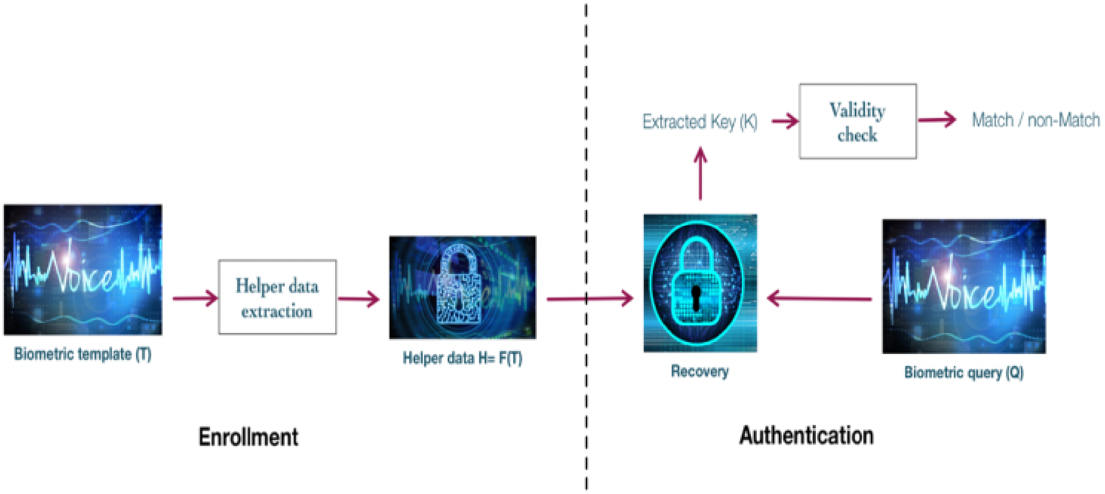

Figure 8. Enrolment of a designated authorizing person

5.2 Input audio signal to be used for biometric based watermarking

The input voice samples recordings of 13 participants in total were used. The majority of the sample is between the age of 15 to 60 years old. There seven males and six females correspondingly. We have recorded an audio clip for the same given phrase in both Arabic or English. The average of each speaker is about 2.5 seconds. As shown in the following table.

Table 1. Audio files information for the used recorded spoken logo at sampling rate=48000

Language	Male/Female	Bit Rate	Total samples	Duration
Arabic	Female	63.6800	118720	2.4733
Arabic	Female	65.4040	132032	2.7507
Arabic	Female	65.8240	92096	1.9187
Arabic	Male	69.1150	101312	2.1107
Arabic	Male	65.7000	115648	2.4093
Arabic	Male	65.1180	97216	2.0253
English	Female	66.7040	149440	3.1133
English	Female	67.1200	162752	3.3907
English	Female	64.6810	122816	2.5587
English	Male	64.4440	114624	2.3880
English	Male	65.3730	121792	2.5373
English	Male	65.0370	144320	3.0067
English	Male	65.9820	155584	3.2413

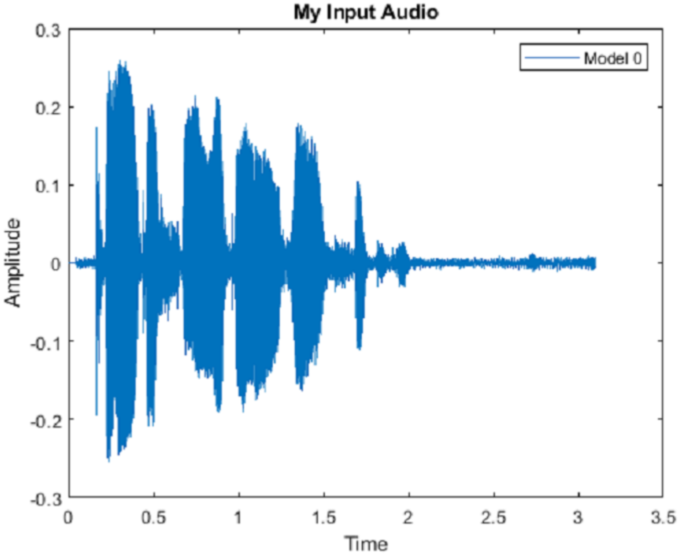

FIGURE 9. The recorded signal for a female 20 years old for 0.2 sec.

5.3 Feature extraction using *MFCC* for audio watermark

The recorded signal 9 is analysed using the *MFCC* to extract the features by clustering the proportions in the audio signal and accordingly providing the pitch contour. Pitch Contour of Input Spoken 10

Recall that Mel Frequency Cepstral Coefficient *MFCC* is based on the known variation of the human perception for sensitivity; the MFCC extracts the appropriate frequencies critical bandwidths. As it was shown, if a cepstral coefficient has a positive value, most spectral energy is concentrated in the low-frequency regions. On the other hand, if a cepstral coefficient has a negative value, most spectral energy is concentrated at high frequencies. Hence, the cepstral coefficient is being used for distinguishing between the different voiceprint templates. Accordingly, the matching will be using the cepstral coefficient as part of the matching process.

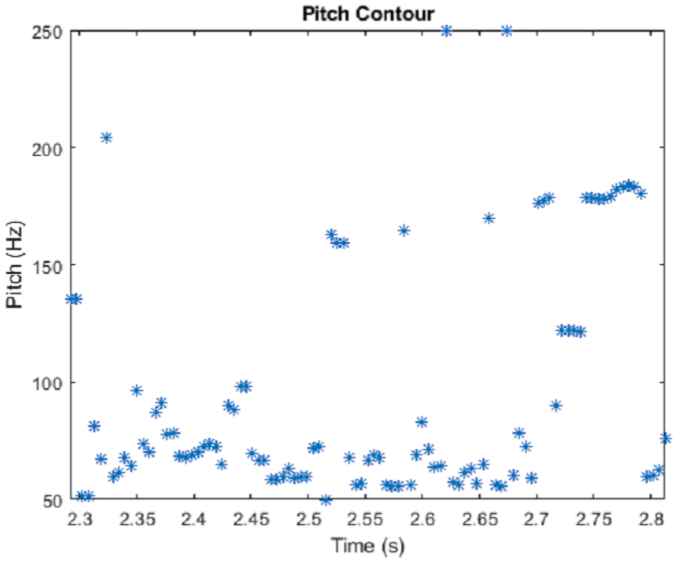

Figure 10. Pitch Contour of Input Spoken signal shown in 9

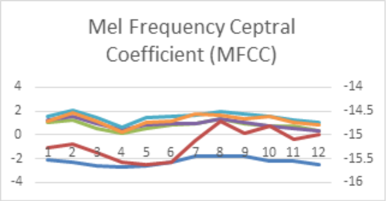

Figure 11. MFCC Results for signal shown in 9

Next step was to create the GMM and Pdf for Audio Watermark, TABLE 2. sample of Mel Frequency Cepstral Coefficient (MFCC), Voice print (1) MFCC coefficients Each row represents the coefficients ts of on frame

Each column represents one extracted coefficient

1	2	3	4	5	6
2.1346751	15.274978	3.2146313	0.1544124	0.2705201	0.3214477
2.3007807	15.198602	3.4968379	0.3647711	0.5178981	0.2372734
2.5666812	15.370973	3.0541843	0.4513928	0.4872558	0.2891993
2.6681539	15.573684	2.7605174	0.306737	0.2795796	0.4904583
2.6369229	15.615577	3.1936598	0.2407337	0.6117264	0.3788296
2.2624644	15.586541	3.1363205	0.0773187	0.5643752	0.4198778
1.835069	15.120995	2.790611	0.0368654	0.7564175	0.093002
1.7584265	14.708819	3.1334548	0.018742	0.5582125	0.2214191
1.7609954	14.955749	2.7414847	0.0838572	0.6517887	0.3645327
2.1613339	14.813353	2.94805	0.0352636	0.8134449	0.0034677
2.2450522	15.089663	3.0255176	0.1981286	0.6704956	0.2230031
2.5391007	15.003218	2.9001741	0.0214046	0.6854187	0.1553873

The results were plotted as the 3D representation of the clusters in each voiceprint using MFCC and also the plots of the Gaussians mixtures in each voice signal being used to extract the coefficients. That refers to 12 & 13. 12 the 3D graph represents Gaussian distributions for a female voiceprint expressing the MFCC coefficients whilst the pdf for All clusters represented by these Gaussians 15. The results exhibit the uniqueness of the watermark based on the various voiceprints being used in experimentations. These differences made the verification process successful in distinguishing the original video with the original watermark from the tampered video using different voiceprints.

5.4 Video Dataset

The Experiment was adopted from the File Examples website as a source dataset of videos. We upload five videos with different properties, as shown below table. **??** shows the information of videos we used. From the table, one can see how the proposed approach is being used with different number of frames and different frames size in these videos.

TABLE 3. Input video Information

Video Type	# of Frames	Length/sec	Frame width	Frame height	Data rate
.mpg	181	06	640	360	1.02 Mbit/s
.mp4	166	06	560	320	2.47 Mbit/s
.mp4	197	07	640	360	214.86 Kbit/s
.mp4	105	07	320	240	663.04 Kbit/s

5.5 Adding the image watermark (IW)

The designated person in each registered organization will add the organization logo as the secret image. This secret image will be used as image watermark and analyze to extract required data to be further classified as key binding and key generation systems.

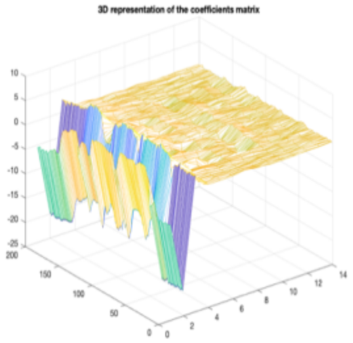

Figure 12. 3D Representation of Coefficients matrix and the gaussian distribution for Video 1, with male voice print with ages 35 years in English language expressing the MFCC coefficients.

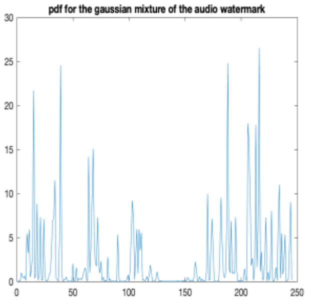

Figure 13. The pdf for All clusters represented by these Gaussians with male voice print of age 35 years in Arabic language expressing the MFCC coefficients

5.6 Embeding the complex watermark using mutual encryption

A mutational encryption technique would implement lifting wavelet transform LWT and singular value decomposition (SVD) in order to embed the complex watermark. SVD of M is given as

$$M_{m\times n} = U_{mxn} \sum_{mxn} V_{mxn}^{T} \tag{1.1}$$

Where the diagonal elements of $\sum_{mxn} V_{mxn}^{T}$ are known as singular values and U, V are called left and right singular vectors. If m=n and the matrix is symmetric, then U, V span the same vector space . Next step is to encrypt user multimedia with a mutation algorithm using Penultimate Least Significant Bit $PLSB$ [31].

$$PLSB = \sum_{i=0}^{n=1} b_i.2^i \tag{1.2}$$

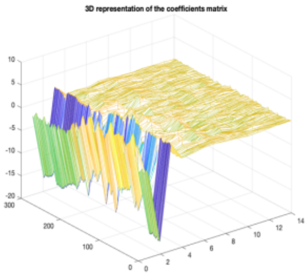

Figure 14. 3D graph represents Gaussian distributions for a male voiceprint with age 20 years in Arabic language expressing the MFCC coefficients.

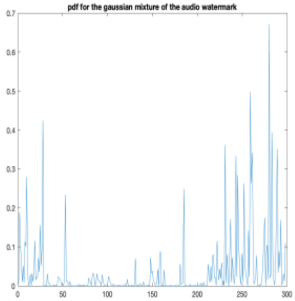

Figure 15. The pdf for All clusters represented by these Gaussians with male voice print of age 20 years in Arabic language expressing the MFCC coefficients.

where b_i denotes the value of the bit with number i and N denotes the number of bits in total.

5.7 Extraction approach

- Apply the second level LWT on the watermarked image: $lwt2$(Watermarked Image)
- Select the sub band LL and HH and apply SVD. [72]
- Select the matrices and mutate using scaling factor α
- Perform Inverse SVD in order to Construct extracted watermark image by performing second level $ILWT$.

5.8 Performance evaluation metrics

The evaluation of a biometric-based security system is based on the evaluation of all components; the recognition system performance, the communication interface, the matching and decision step, and other factors such as ease of use, acquisition speed, and processing speed. However, a method to compare biometric system performance be based on the end decision's accuracy. In this section, a set of evaluation metrics is used. In order to test the accuracy and reliability of biometric-based security of the transmitted videos, the following metrics uses: a) False Accepted Rate [FAR], b) False Rejection Rate [FRR]. False Acceptance Rate (FAR) is defined as the percentage of identification instances in which unauthorised persons are incorrectly accepted. False Rejection Rate (FRR) is defined as the percentage of identification instances in which authorised persons are incorrectly rejected.

$$FAR = FP/(FP + TN) \tag{1.3}$$

$$FRR = FN/(FN + TP) \tag{1.4}$$

But, How do the FAR and FRR impact on each other?. As the number of false acceptances (FAR) goes down, the number of false rejections (FRR) will go up and vice versa 16. The point at which the lines intersect also has a name: the Equal Error Rate (EER). This is where the percentage of false acceptances and false rejections is the same.

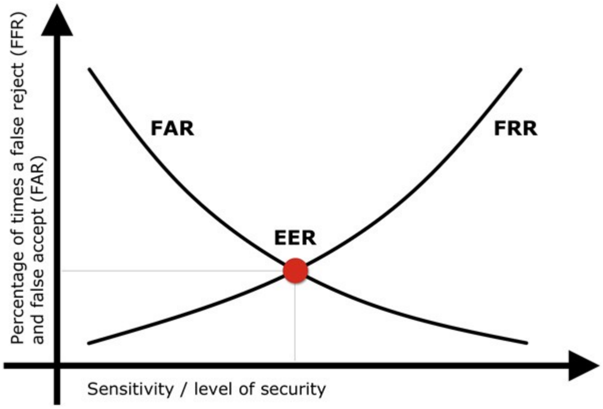

Figure 16. FAR and FRR impact on each other

Now, How does this affect the security level in terms of user experience. If you try to reduce the FAR to the lowest possible level, the FRR is likely to rise sharply. In other words, the more secure your access control, the less user-friendly will be the system, as users are falsely rejected by the system. The same also applies the other way round. As you go for increasing the user user-friendly experience, the system would be likely to be less secure (higher FAR). Accordingly, the FAR and FRR can usually be configured in a security system's software by adjusting the appropriate criteria so that they are more or less strict. We can conclude from the information above that this will result in a system that is more secure (but less user-friendly) or less secure (but more user-friendly). In high critical security ststems priority is given to the security rather than user experience. This is clearly the case in situations where they expect a high level of security.

Applying the above security measures to evaluate the performance of our system, the explanation above, our experiement included a total of 260 runs as we have five videos tested with 13 audio files (illustrated information of voiceprint in table 8) for (False / True) the true reads and (False / True) for rejection reads. In this regard, we have used the below symbols to represent the variables we used to calculate the measures.

For testing the recordings, one participant in this stage is 26 years old and the input video is of MPG extension video duration 8 seconds, 210 frame and resolution 352 × 288 Pixel. The data rate is 1150 Kbps. The frame rate is 25 fps. The data in this stage consists of different voiceprints with the same video in the first stage. In this stage, a varying sample of women of ages. The participant [VPF2] is female in 30 − 40 years of age was selected and another female [VPF3] in 40-50 years of age. Both of them recorded an audio clip for the phrase in the Arabic language. Besides, [VPFE1] – [VPFE3] ages from 15-60 years recording an audio clip of approximately 2 seconds uses with the spoken phrase (King Saud University) in the English language. The same video in this stage tests male participants, men with age ranges between 15-50 years. Results illustrated in table 7 watermarked video2 with 13 audio files: Male/Female, ages between 30-60, language Arabic, and English.

Table 4. Voiceprints participants information

Template Number	Audio Name	Gender	Age	Language
1	VPF1	Female	20-30	Arabic
2	VPF2		30-40	
3	VPF3		40-50	
4	VPM1	Male	30-40	
5	VPM2		15-20	
6	VPM3		20-30	
7	VPFE1	Female	20-30	English
8	VPFE2		30-40	
9	VPFE3		15-20	
10	VPME1	Male	15-20	
11	VPME2		20-30	
12	VPME3		15-20	
13	VPME4		40-50	

The number of wrong verifications of unauthorized claims of identity in a total of 65 runs is six.

Two female participants have aged 27 and 43 years; In the Arabic language. One female with 35 ages years and two male with ages 30 and 45 ages years; in the English language

Table 5. Number of wrong verifications of unauthorized claims of identity

Numbers of runs	Numbers of wrong verifications of unauthorized claims of identity	
Female Arabic (15 runs)	2	
Male Arabic(15 run)	1	
Female English (15 runs)	1	9.2 %
Male English (20 run)	2	
65 runs	6	

The number of wrong verification to reject authorized claims of identity in a total of 65 runs is eight. One female participant has aged 30 years, and two male participants with ages 18 and 33 years; In the Arabic language. Two female participants have aged 20-23 years, and Three male participants have aged between 20-45 years.; In the English language.

Table 6. Number of wrong verification to reject authorized claims of identity

Numbers of runs	Numbers of wrong verifications of authorized claims of identity	
Female Arabic (15 runs)	1	
Male Arabic(15 run)	2	
Female English (15 runs)	2	12%
Male English (20 run)	3	
65 runs	8	

5.9 Performance results discussion

Recall that both FRR and FAR are functions dependent on the variable threshold t, We obtain the result of [FAR] which is 9.2% and of [FRR] which is 12%. It decreases the value of this threshold to make the system tolerant of handling real users' variations and noises, and decrement of the FRR results in increasing FAR. On the other hand, if t is increased to reject some small inter users variations and more secure the system with low FAR, then the FRR raises and evaluated vary in opposite directions. In practice, one has to make a trade-off between these two rates to optimize the parameters based on the targeted application . One should define the decision threshold for a biometric system, so this can be achieved with a rate of FRR and FAR; that is, the percentage of attempts declared does not match the biometric sample within the allowed number of attempts. IN this regard, an attacker would have to try to impersonate vast numbers of different people to gain 90% confidence to be wrongly accepted once within the permissible number of attempts. The values of FAR and FRR are thus dependent on this decision threshold to reduce the system's rate error. The decision conditions must be adjusted according to the desired characteristics for the application considered. High-security applications require a low FAR, which increases the FRR, while Low-security applications are less demanding in terms of FAR.

6 Conclusion

This research proposed a model to ensure that the multimedia content exchange is accurate, reliable, and not tampered by unauthorized users or hackers. The model is based on using a complex watermarking for video multimedia files with multimodal protection. The complex watermark comprises a secret watermark image and a voiceprint encrypted template resulted from the spoken text. This complex watermark is then embedded in the media file using singular value decomposition (SVD) of the second level of LWT into the user video using an LL band of the cover video frames and the HH band of watermark video frames. The experiment dataset contained five videos with different proprieties in addition to the secret image to be used for watermarking the media file, and voice samples recordings of 13 participants in total. The majority of the sample ranged between the age of 15 to 60 years old. The recorded audio clips contains the same phrase, could be the institution name in Arabic and English language. The average of each clip is about 2.5 seconds. The total number of play counts in this experiment is 260 as we have five videos tested with 13 audio clips for the accepted reads (False/True), and we have the same five videos being tested with 13 audio clips for rejected reads (False/True). We obtain the result of [FAR] which is 9.2

7 Acknowledgments

This research project was supported by a grant from the “Research Center of the Female Scientific and Medical Colleges”, Deanship of Scientific Research, King Saud University for funding this work through research group No (SMRC -1903).

Bibliography

1. A. A. Attari, A. Asghar, and B. Shirazi. “Robust and Transparent Audio Watermarking based on Spread Spectrum in Wavelet Domain.”.

2. S. Arora and M. P. S. Bhatia. Challenges and opportunities in biometric security: A survey. *Information Security Journal*, pages 1–21, 2021.

3. Russell E. Adams. *Sourcebook of Automatic Identification and Data Collection.* Van Nostrand Reinhold, 1990.

4. S. I. Ao and A International Association of Engineers. *Review of the Fingerprint, Speaker Recognition, Face Recognition and Iris Recognition Based Biometric Identification Technologies.* Newswood Ltd, 2011.

5. A. O. Alaswad, A. H. Montaser, and F. E. Mohamad. *Vulnerabilities of Biometric Authentication ‘Threats and Countermeasures.* [Online. Available, 2014.

6. O. C. Carrasco. Gaussian mixture models explained from intuition to implementation. 2019.

7. P. Corcoran and A. Cucos. Techniques for securing multimedia content in consumer electronic appliances using biometric signatures. *IEEE Transactions on Consumer Electronics*, 51(2):545–551, 2005.

8. I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker. Elsevier, 2007.

9. C. Cruz, R. Reyes-Reyes, M. Nakano, and H. Perez-Meana. Image content authentication system based on semi-fragile watermarking. *p*, pages 306–309, April 2008.

10. T. Ghazali and N. H. Zakaria. Security performance evaluation of biometric lightweight encryption for fingerprint template protection. *International Journal of Advanced Computer Research*, 9:232–241, April 2019.

11. J. Galbally Herrero. Vulnerabilities and attack protection in security systems based on biometric recognition. 2009.

12. H. T. Hu and T. T. Lee. Hybrid blind audio watermarking for proprietary protection, tamper proofing, and self-recovery. *IEEE Access*, 7:80395–18040, 2019.

13. Z. Jalil and H. Aziz. S. bin shahid. In *M*, pages 2010–2010, and A. M. Mirza, “A zero text watermarking algorithm based on non-vowel ASCII characters,” in ICEIT International Conference on Educational and Information Technology, Proceedings vol. 2, 2010. Arif.

14. C. Kant and S. Chaudhary. A watermarking based approach for protection of templates in multimodal biometric system. *Procedia Computer Science*, 167:932–941, 2020.

15. J. Li, Y. Wong, and T. Sim. Towards protecting biometric templates without sacrificing performance. In *2016 23rd International Conference on Pattern Recognition (ICPR*, pages 1029–1034, 2016.

16. G. Q. Liu, X. S. Zheng, Y. L. Zhao, and N. Li. A robust digital video watermark algorithm based on dct domain. In Iccasm International, editor, *Conference on Computer Application and System Modeling, Proceedings vol*, pages 2010–2010, no. Iccasm, pp. 202–205, 2010. 2.

17. J. Maghsoudi and C. C. Tappert. A behavioral biometrics user authentication study using motion data from android smartphones. *Proceedings* -, 2016:184–187, 2017.

18. C. Müller. *Fundamentals Features and Methods*. Springer Berlin, Heidelberg.

19. K. Nandakumar, A. K. Jain, and A. Nagar. Biometric template security. *Eurasip Journal on Advances in Signal Processing*, 2008, 2008.

20. R. E. O. Paderes. A comparative review of biometric security systems. In *Proceedings - 8th International Conference on Bio-Science and Bio-Technology*, pages 8–11. BSBT 2015, 2016.

21. Q. Qian, H. Wang, X. Sun, Y. Cui, H. Wang, and C. Shi. Speech authentication and content recovery scheme for security communication and storage. *p*, pages 635–649, 2018.

22. S. Rane. Standardization of biometric template protection. *IEEE MultiMedia*, 21(4):94–99, 2014.

23. D. Reynolds. *Gaussian Mixture Models*. Encyclopedia of Biometrics. Springer, Boston, MA, In, 2015.

24. E. A. Rúa, E. Maiorana, J. L. A. Castro, and P. Campisi. Feature fusion for template stability in biometric cryptosystems. an application to face biometrics based on eigen-models. In *2012 IEEE First AESS European Conference on Satellite Telecommunications (ESTEL*, pages 1–6, 2012.

25. D. Renza, J. Vargas, and D. M. Ballesteros. Robust speech hashing for digital audio forensics. *Applied Sciences (Switzerland)*, 10:1, 2020.

26. C. Sharma and A. Bhaskar. Proceedings, no. xxxx, Materials Today, 2020.

27. R. Schafer. “springer handbook of speech processing,” 2008. *p*, 161.

28. P. Singh, B. Raman, and N. Agarwal. Toward encrypted video tampering detection and localization based on pob number system over cloud. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2116–2130, September 2018.

29. C. Song, S. Sudirman, M. Merabti, and D. Llewellyn-Jones. Analysis of digital image watermark attacks. *p*, pages 1–5, April 2010.

30. M. Stamp. Wiley.

31. Encompassing intelligent technology and innovation towards the new era of human life. In Ieee Thailand Section and Thai-Nichi Institute of Technolog, editors, *Proceedings of 2019 4th International Conference on Information Technology (InCIT)*, pages 24–25, Thailand, 2019. Institute of Electrical and Electronics Engineers, Bangkok.

32. U. Uludag and A. K. Jain. Multimedia content protection via biometrics-based encryption. In *03. Proceedings (Cat. No.03TH8698) vol. 3, pp. III–237.* 2003 International Conference on Multimedia and Expo. ICME, 2003.

33. Yamila. *Bypassing Biometric Systems with 3D Printing and ‘Enhavced’ Grease Attacks.* Dreamlab Technologies, June 2020.

34. S. Zhang, X. Guo, X. Xu, L. Li, and “a C. C. Chang. video watermark algorithm based on tensor decomposition,” mathematical biosciences and engineering, vol. 16, no. 5. *p*, 3435, 2019.