High Security and Capacity of Image Steganography for Hiding Human Speech Based on Spatial and Cepstral Domains

Yazen A. Khaleel

Department of Software Engineering, Faculty of Engineering, Koya University,
Danielle Mitterrand Boulevard, Koya KOY45, Kurdistan Region – F.R. Iraq

***Abstract*—A new technique of hiding a speech signal clip inside a digital color image is proposed in this paper to improve steganography security and loading capacity. The suggested technique of image steganography is achieved using both spatial and cepstral domains, where the Mel-frequency cepstral coefficients (MFCCs) are adopted, as very efficient features of the speech signal. The presented technique in this paper contributes to improving the image steganography features through two approaches. First is to support the hiding capacity by the usage of the extracted MFCCs features and pitches extracted from the speech signal and embed them inside the cover color image rather than directly hiding the whole samples of the digitized speech signal. Second is to improve the data security by hiding the secret data (MFCCs features) anywhere in the host image rather than directly using the least significant bits substitution of the cover image. At the recovering side, the proposed approach recovers these hidden features and using them to reconstruct the speech waveform again by inverting the steps of MFCCs extraction to recover an approximated vocal tract response and combine it with recovered pitch based excitation signal. The results show a peak signal to noise ratio of 52.4 dB of the stego-image, which reflect a very good quality and a reduction ratio of embedded data to about (6%–25%). In addition, the results show a speech reconstruction degree of about 94.24% correlation with the original speech signal.**

***Index Terms*—Image steganography, Mel-frequency cepstral coefficients, Speech reconstruction.**

I. Introduction

Steganography is the science of hiding covert information in a cover public media without attracting attention. Modern steganography methods use the characteristics of digital media using them as carriers (covers) to hold secret information. Covers can be different types including text, speech/audio, image, and video. Thus, the sender embeds secret data in a digital cover file using a key to generate a stego-file, in which that an observer cannot feel the existence of the hidden message (Cox, et al., 2008).

The name of the steganography method depends on the type of cover media file used for hiding the secret data (such as image steganography, audio steganography, and video steganography) (Saroj, and Dewangan, 2018).

Many state of art algorithms have been suffering from the capacity storage area of the host image, low security, and robustness. This paper proposes a new technique in image steganography type, in which a secret speech signal is to be hidden inside a digital color image as a cover media. The most two common challenges that facing any steganography techniques are: How can increase the capacity of the host cover image to embed as much as possible secret data and in the same time, and how the unauthorized persons cannot distinguish the presence of hidden message.

This work contributes to develop in these both two challenges mentioned above. As it does not depend on directly hiding all the speech signal samples inside the digital image, it rather extracts some important features from the speech signal and embeds them inside the digital image.

On the authorized person side that receives the stego-image, the algorithm will extract the confidential information (features) from the transmitted stego-image. Then, these returned features will be used to reconstruct the speech clip again by reversing the steps of getting these features and then merging the result with an excitation frequency (Pitch) to get back the speech clip again. These features are called Mel-frequency cepstral coefficients (MFCCs). The usage of MFCCs features as data to be hidden rather than all the samples of the speech signal themselves. This means decreasing the amount of the data to be embedded. The reduction ratio depends on the time width of the frames chosen during the MFCCs extraction process (will be discussed later). The method of hiding the extracting MFCCs features inside the cover digital image will exploit the all bits in the three digital color image components (Red, Green, and Blue) rather than using only the least significant bits (LSBs) or the bits in the image edges as traditionally done

ARO-The Scientific Journal of Koya University
Volume VIII, No.1 (2020), Article ID: ARO.10670, 12 pages
DOI: 10.14500/aro.10670
Received: 28 April 2020; Accepted: 22 June 2020
Regular research paper: Published: 28 June 2020
Corresponding author's email: yazen.adnan@koyauniversity.org

in most image steganography approaches. This technique will increase the hiding capacity of the cover image.

Many articles consider the image steganography to hide the human speech signal as the secret information. Saroj and Dewangan (2018) presented a method to hide the audio secret data in image with multilevel protection using LSB techniques. Their technique implements hybrid audio steganography which hides the audio information by encrypting in multiple levels and embed into the variable LSB's of the selected samples based on polynomial expression as a function of audio and image cover file. Sharma (2015) proposed a method works by hiding sequence of speech signal elements in an image by varying the y dimension of the image while keeping x and z dimension of the image as constant. The value of the x dimension is changed by some interval for storing next speech signal elements. For the varied value of x dimension, y dimension is varied across the same interval as above whereas z dimension is constant here the next sequence of speech signal elements is saved.

Nipanikar, Deepthi and Kulkarni (2017) proposed a method for image steganography using sparse representation, and an algorithm named particle swarm optimization (PSO) algorithm for effective selection of the pixels for embedding the secret audio signal in the image. PSO-based pixel selection procedure uses a fitness function that depends on the cost function. Cost function calculates the edge, entropy, and intensity of the pixel for evaluating fitness.

In this work a new method for speech hiding and reconstruction is developed, based on the model of speech production. Section II briefly reviews the speech model parameters, cepstral analysis, and MFCCs that need to be extracted from the speech signal. Section III considers the methodology of the proposed approach. Results, evaluation of the steganography and speech reconstruction is presented in Section IV and a conclusion comes in Section V.

II. Cepstral Analysis and MFCCs

The excitation vocal signal which is generated by the human vocal cords is filtered by the shape of the vocal tract that includes the fauces, tongue, and teeth. This shape specifies what sound comes out. If it can determine the shape accurately, this could give an accurate representation of the phoneme being generated. The form of the vocal tract appears within the envelope of the speech power spectrum. Fig. 1 and equation (1) represent the human voice speech model *s*(*n*).

$$s(n) = e(n) * v(n) \tag{1}$$

Where *e* (*n*) is the glottal excitation signal represents the signal which is produced by the vocal cords. It is periodic pulses with a relatively high frequency in its spectrum E(k). The *v*(*n*) represents as the impulse response of the vocal tract which has low frequency spectrum V(k) compared with the excitation signal frequency. The shape of the vocal tract is unique for every human that gives the person his/her voiceprint. It is referred to as the filter with a relatively smooth frequency response V(k) that specifies what sound comes out and the person voiceprint. Then, the speech features should be extracted from V(k) rather than E(k). Therefore, two components of S(k) are combined (convolution) together and should be separated into two components E(k) and V(k). The Fourier transform of *s*(*n*) gives S(k), as shown in equation (2) below:

$$S(k) = E(k) .V(k) \tag{2}$$

If the components were combined in the convolution, no clear results would be got after using a filter as these two components in the frequency domain (spectrum) are multiplied (nonlinearly combined).

Now, it is necessary to transform the *Spectrum* into a new domain called "Cepstrum" or new frequency domain called "Quefrency domain" that represents a transformation on speech signal with two important properties:

1) The two signal components will be separated.
2) The components will be linearly combined (summation of components).

By taking the logarithm of the absolute part of the spectrum, the real cepstrum domain is generated that achieves the above two properties. Equation (3) and Fig. 2 abbreviate the computation of the real cepstrum.

$$C_s(\omega) = \log|S(k)|$$

$$= \log|E(k).V(k)|$$

$$= \log E(k) + logV(k)$$

$$= C_e(k) + C_v(k) \tag{3}$$

The purpose of real cepstrum is to resolve the two convolved parts of the speech *e*(*n*) and *v*(*n*), into two additive components as in equation (4):

$$c_s(n) = c_e(n) + c_v(n) \tag{4}$$

Using a low time lifter (lifter is equivalent to filter in frequency domain) to select the vocal tract component C_v (*k*) and eliminate the excitation one C_e (*k*).

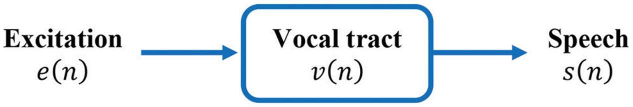


Fig. 1. Basic speech model.

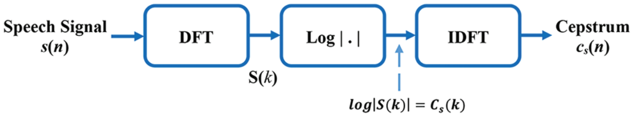


Fig. 2. Computation of the real cepstrum.

For this reason, the MFCCs are selected as good features that reflect the speech in which these features come from cepstral analysis. The job of MFCCs is to accurately represent the envelope of the speech spectrum (Huang, Acero and Hon, 2001). The motivating idea of using MFCCs is to reduce information about the vocal tract (smoothed spectrum) into a little number of coefficients.

MFCCs are features widely used in speech recognition, speaker identification, and verification. MFCCs are understood to represent the filter (vocal tract). They were presented by Davis and Mermelstein (1980).

This following introduces the MFCCs extraction steps, as shown in Fig. 3:

1. Dividing the signal into series of short overlapped time segments (frames) with width 20–40 ms.
2. Windowing each individual frame (Typically Hamming window).
3. Computing the Fast Fourier Transform for each frame and its power spectrum (periodogram).
4. Applying the Mel-frequency filter bank to the power spectra, sum the energy in each filter.
5. Taking the logarithm of all filter-bank energies.
6. Taking the discrete cosine transform (DCT) of the log filter-bank energies.

The Mel filter bank has a triangular band-pass frequency response, as shown in Fig. 4. The spacing and the bandwidth are determined by a constant Mel frequency interval. The number of Mel spectrum coefficients is 12–40, as shown in Fig. 4. The filter bank comes as vectors. Each vector is mostly zeros, but it is non-zero for a particular section of the spectrum. To calculate filter-bank energies, each filter bank is multiplied with the power spectrum, and then sum of the coefficients is numbers that show how much energy was in each filter bank. The Mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Equation (5) represents the transfer function from linear to Mel frequency scale. Mel filter banks perform the sum of energy that exists in various frequency regions. The first filter is very narrow and gives a sign of how much energy exists close to 0 Hertz. As the frequencies go higher, the filters become wider because it becomes less attentive about variations (Chakroborty, Roy and Saha, 2007).

$$F\left(Mel\,scaled\right)=\left[2595\log\left(1+f\left(Hz\right)/700\right)\right] \qquad (5)$$

Once the filter-bank energies is calculated, the logarithm of them is taken. Human hearing also motivates this. The human does not hear loudness on a linear scale. In general, to double the perceived volume of a sound it must put 8 times as energy amount into it. This means that large variations in energy might not sound all that different if the sound is loud, to start with. This operation makes the extracted features match more closely what humans actually hear.

The last step is converting the log Mel spectrum coefficients back to time domain using the DCT. DCT decorrelates the features, taking the DCT of the log filter-bank energies to get the cepstral coefficients. According to the application and the accuracy required, any desired number of coefficients can be kept and neglecting the others. For example, in case of speech or speaker recognition, only 12 (2nd–13th) of the DCT coefficients are kept as they are enough for this application. In the current research, 26 coefficients per frame will be taken. A short notation can summarize the steps as in equation (6):

$$c_v = DCT(\ \log\left[\ Mel\left(PSD\right)\right]\ ) \qquad (6)$$

Where c_v is the cepstral coefficients (MFCCs) associated with the vocal tract part of speech model, Mel means the Mel-Filter bank, and power spectral density (PSD) is the PSD of the speech signal (magnitude spectrum).

III. Methodology

In general, the proposed technique consists of two main parts. The first one is the analyzing a voice in terms of its pitch (fundamental frequency) and spectral envelope and then extracting the MFCCs features. These extracted features are embedded inside the cover digital color image. The second part is the recovering the hidden data (MFCCs features and the pitches) and then reconstructing (synthesizing) the speech clip.

The proposed technique is implemented using MATLAB R2019a. A speech clip of about 10 s long will be used as an example case through the following steps of the methodology:

A. Preprocessing on the Speech Signal

The first step is the enrollment of the human voice utterance. Some preprocessing steps are required and carried out such as converting the audio file to WAV format, one channel, and sampled with a sampling frequency (*fs*). For example, if the speech clip is with 10 s as assumed above and sampled with 48 kHz sampling rate then the total number of samples (*N*) is 480,000.

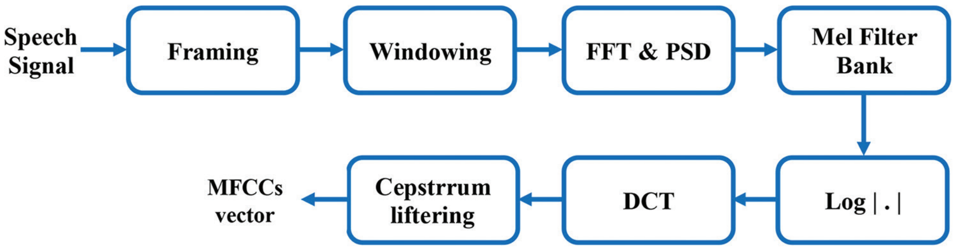


Fig. 3. Main successive steps of extracting the Mel-frequency cepstral coefficients.

B. Features Extraction

The frame width (*fw*) and the overlap length (*ol*) between the adjacent frames determine the number of frames, in which the speech signal could be segmented into. The *fw* is typically taken between 20 and 40 ms (30 ms is standard) which means 1440 samples per frame, and the default *ol* is taken $\frac{2}{3}fw$ (20 ms). The number of the segmented frames (*F*) can be calculated as following in equation (7):

$$F = T\,/\,(fw - ol) \tag{7}$$

For tradition values of fw = 0.03 s then ol = 0.02 s, the number of created frames will be as in equation (8):

$$F = 100T \tag{8}$$

Where *T* is the total time of the speech clip in seconds. Here, we can notice that the sampling rate *fs* does not affect the number of the obtained segmented frames.

For the selected example above, there are 1000 frames. The extraction steps of the MFCCs are applied to each frame. Then, 26 MFCCs will be returned per frame as explained in Section II. This leads that the total extracted features will be as 1000 × 26 matrix (26,000 features). Before the encoding step, the MFCCs features are rounded into two digits after the decimal point to simplify the quantization step. Equation (9) shows how to calculate the number of the extracted MFCCs features.

$$MFCCs = 100T\,M_F \tag{9}$$

Where MFCCs is total number of the Mel frequency coefficients features and M_F is the number of the coefficients per frame. A single pitch frequency (fundamental frequency represented by a series of pulses) for each time segment (frame) of the speech is also computed by the Short-Term Fourier Transform. It is included as an extra component to the features vector.

C. Encoding of the Pitch and MFCCs Features

For normal human speech of about 55 dB loudness, it is found that the dynamic range of the extracted MFCCs is real values usually no more than 10. This range is found by testing many speech utterances for many persons and with different time lengths. A specific lookup table (LUT) is created to encode the all-possible values of the features using 10-binary digits analog to digital conversion process with 1024 quantization levels and a resolution equal to 0.01. This means the minimum value takes the code (0000000000) and the maximum value has a code of full ones (1111111111). Table I shows a part of the LUT. By this LUT, all MFCCs features could be converted into digital form during the embedding step inside the digital cover media.

A single pitch frequency is inferred to each time frame of the speech signal. Because of the pitch frequency may take values up to about 4 kHz (human voice frequency band), then using 12 bits word in binary system to represent the pitch values.

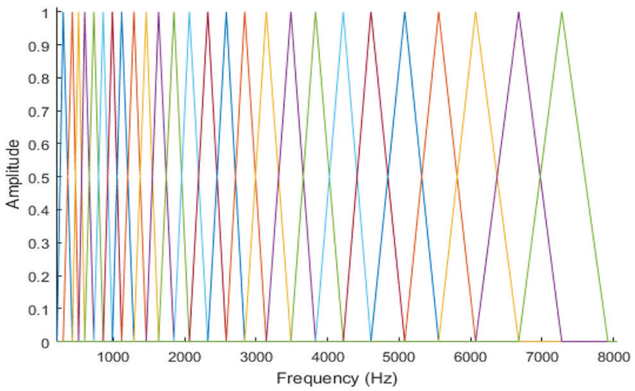


Fig. 4. Mel frequency filter-bank (Oliveira, Cerqueira and Filho, 2018).

D. Cover Color Image Preparations

A digital color image is selected to be the cover media. First, the color image is decomposed into its three components (Red, Green, and Blue), each pixel is represented by three decimal values ranged (0–255) and then the decimal numbers are converted into 8-bit binary representation. Traditional image steganography methods attempt to hide their messages directly in areas of the LSBs of the spatial domain where human visual system does not perceive. On the other hand, various image stego-analysis schemes have been developed to detect the presence of any secret messages in the LSBs area of the cover image, and then it might be easy to discover the embedded data by estimation different scenarios.

Therefore, the current research will hide the secret speech message represented by its features (MFCCs and pitches) within the higher significant bits (HSBs) and using the LSBs only to indicate the necessary information required to recover where the embedded messages. Actually, hiding

TABLE I
The Lookup Table of MFCCs Coding

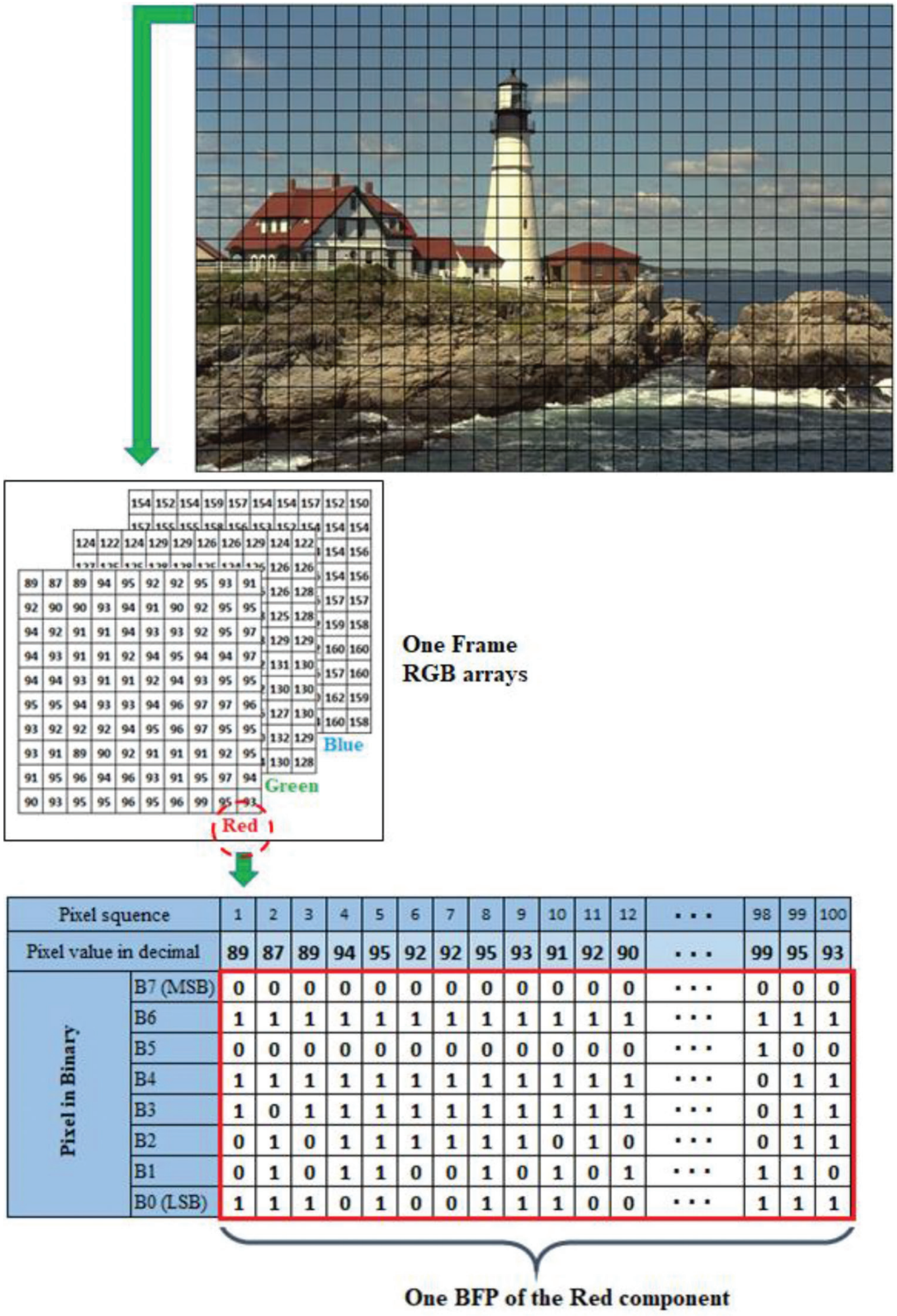
Sequence	MFCCs	Code (using 10 Binary Digits)
1	0	0100101100
2	0	0100101100
3	0	0100101100
4	0	0100101100
5	0	0100101100
6	0	0100101100
7	0.19	0100111111
8	0.58	0101100110
9	−0.05	0100100111
10	−0.12	0100100000
11	0.56	0101100100
12	2.71	1000111011
↓	↓	↓
25999	0.12	0100111000
26000	0.19	0100111111

MFCCs: Mel-frequency cepstral coefficients the secrete data within the HSBs will not increase the capacity of the cover image, but it improves the security of the steganography process. The information hidden in the LSBs is meaningless and useless if it falls into the hands of unauthorized persons. Only the target person has the key of how to use this information inserted in the LSBs to cover up the secret data from the HSBs.

Before the hiding process, a virtual spatial framing of the cover image should be done. Each of the three components of the cover digital image is divided into equally square frames (10 × 10 pixels' sub images). After that, each frame is reshaped to be 1 × 100 vector. Each decimal pixel value in the frame is converted to its 8-bit binary representation. Then, a new digital array form is created with dimensions 8 × 100 that can be named binary frame panel (BFP), as shown in Fig. 5.

E. Secrete Data Hiding

The hiding process used in this research made modifications to the method presented by Abdulraman,



89	87	89	94	95	92	92	95	93	91
92	90	90	93	94	91	90	92	95	95
94	92	91	91	94	93	93	92	95	97
94	93	91	91	92	94	95	94	94	97
94	94	93	91	91	92	94	93	95	95
95	95	94	93	93	94	96	97	97	96
93	92	92	92	94	95	96	97	95	95
93	91	89	90	92	91	91	91	92	95
91	95	96	94	96	93	91	95	97	94
90	93	95	95	96	95	96	99	95	93

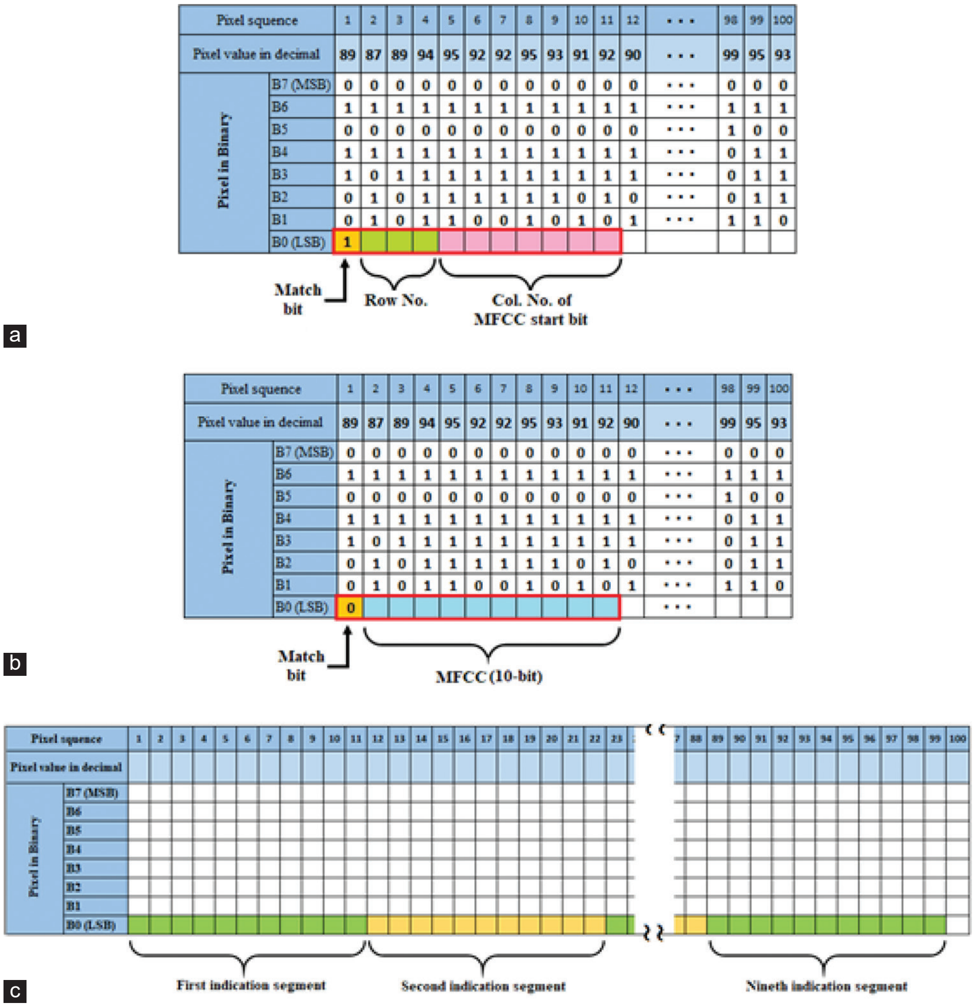
Pixel squence		1	2	3	4	5	6	7	8	9	10	11	12	• • •	98	99	100
Pixel value in decimal		89	87	89	94	95	92	92	95	93	91	92	90	• • •	99	95	93
Pixel in Binary	B7 (MSB)	0	0	0	0	0	0	0	0	0	0	0	0	• • •	0	0	0
	B6	1	1	1	1	1	1	1	1	1	1	1	1	• • •	1	1	1
	B5	0	0	0	0	0	0	0	0	0	0	0	0	• • •	1	0	0
	B4	1	1	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B3	1	0	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B2	0	1	0	1	1	1	1	1	1	0	1	0	• • •	0	1	1
	B1	0	1	0	1	1	0	0	1	0	1	0	1	• • •	1	1	0
	B0 (LSB)	1	1	1	0	1	0	0	1	1	1	0	0	• • •	1	1	1

Fig. 5. Cover color image preparations and binary frame panel creation.

et al. (2019). These modifications help to increase the hiding capacity and reducing the number of unused bits. The hiding steps begin in the red array component of the color image by taking the first MFCC feature (10-bit word) and make searching inside the first BFP rows. The searching starts from the 1st row of the BFP up to the 7th row and leaving the 8th row (LSBs row) for indication. The searching tries to find any match between the ten-bit of the first MFCC feature and any 10-bit combination in this row. If the searching process does not find the similar word in first 10 bits, it will continue the searching by shifting one bit to the right to compare with a next 10-bit combination and so on until the end of the whole BFP. For the result of searching, there are two possibilities, either can find an exact similar 10-bit combination or cannot find any match in that BFP.

The LSB row (last row) of each BFP is virtually divided into nine indication segments. Each one consists of 11 bits that means 99 bits are used for the indication and leaving the last bit unused. Each segment is dedicated to indicate information about one MFCC feature where the exact match is found. This leads to ability of hiding at most nine successive MFCCs features per BFP. In case an exact matching is found, then logic 1 is assigned to first bit of the first segment, this bit is named the matching bit. The next three bits represent the row number (in binary), where the exact similar 10-bit combination is located. The last seven bits are used to record the column number (in binary) in the BFP where the exact 10-bit combination starts from. Fig. 6a shows the matching case.

If no exact word is found, then logic 0 is assigned to the matching bit of the first indication segment. Fig. 6c shows all the nine indication segments per one BFP. Now, the 10-bit MFCC themselves are directly copied into the rest ten locations of the segment. These steps are repeated for all other MFCCs feature and so on. Fig. 6b shows the no matching case. This procedure continues to hide at most nine successive MFCCs features per BFP. At the end of hiding 26 MFCCs (features of 1-time frame of the speech signal) using three BFPs, nine features in both 1st and 2nd BFPs, respectively, and the rest eight features in the 3rd BFP. The last 12 bits of the third BFP remains and it is dedicated to insert the pitch frequency binary value of that time frame represented by 12-bits binary number. The same steps are carried out to hide



Pixel squence		1	2	3	4	5	6	7	8	9	10	11	12	• • •	98	99	100
Pixel value in decimal		89	87	89	94	95	92	92	95	93	91	92	90	• • •	99	95	93
Pixel in Binary	B7 (MSB)	0	0	0	0	0	0	0	0	0	0	0	0	• • •	0	0	0
	B6	1	1	1	1	1	1	1	1	1	1	1	1	• • •	1	1	1
	B5	0	0	0	0	0	0	0	0	0	0	0	0	• • •	1	0	0
	B4	1	1	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B3	1	0	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B2	0	1	0	1	1	1	1	1	1	0	1	0	• • •	0	1	1
	B1	0	1	0	1	1	0	0	1	0	1	0	1	• • •	1	1	0
	B0 (LSB)	1															

Pixel squence		1	2	3	4	5	6	7	8	9	10	11	12	• • •	98	99	100
Pixel value in decimal		89	87	89	94	95	92	92	95	93	91	92	90	• • •	99	95	93
Pixel in Binary	B7 (MSB)	0	0	0	0	0	0	0	0	0	0	0	0	• • •	0	0	0
	B6	1	1	1	1	1	1	1	1	1	1	1	1	• • •	1	1	1
	B5	0	0	0	0	0	0	0	0	0	0	0	0	• • •	1	0	0
	B4	1	1	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B3	1	0	1	1	1	1	1	1	1	1	1	1	• • •	0	1	1
	B2	0	1	0	1	1	1	1	1	1	0	1	0	• • •	0	1	1
	B1	0	1	0	1	1	0	0	1	0	1	0	1	• • •	1	1	0
	B0 (LSB)	0												• • •			

Fig. 6. The indication segment details (a) in case of full matching is found (b) in case of no matching is found (c) indication segments per binary frame panel.

all the extracted features using the three image arrays. Last information has to be embedded which is the speech signal time period in seconds (*T*) that is represented by 12-bit that helps in the recovery side to know the number of embedded MFCCs. The proposed method suggests inserting this data in the last 12 bits of LSB row in the last BFP of the blue array. Fig. 6 shows the indication segment details. Fig. 7 shows the over view of the proposed methodology.

It is worth noting the following points:

- In one row, it can hide more than one feature word, and on the other hand, it may not be possible to find any match within it. In one row, there are 91 possibilities of 10-bit combinations and then 637 possibilities per BFP.

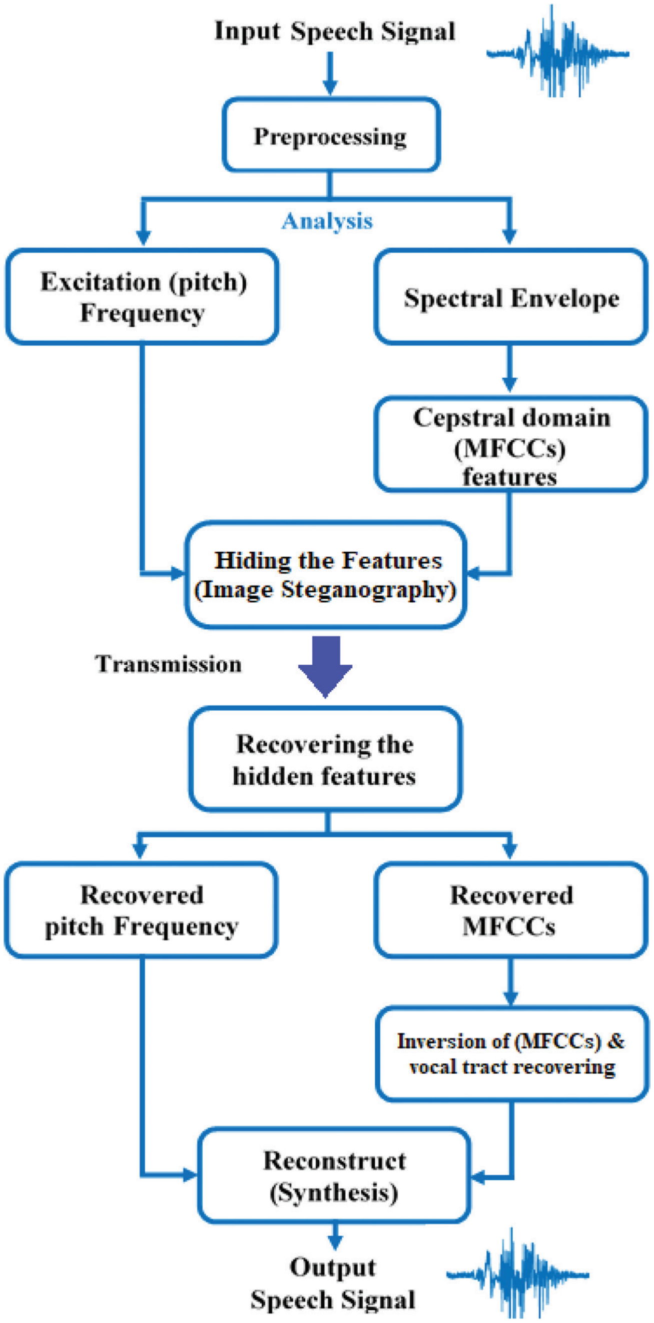

Input Speech Signal
Preprocessing
Analysis
Excitation (pitch) Frequency
Spectral Envelope
Cepstral domain (MFCCs) features
Hiding the Features (Image Steganography)
Transmission
Recovering the hidden features
Recovered pitch Frequency
Recovered MFCCs
Inversion of (MFCCs) & vocal tract recovering
Reconstruct (Synthesis)
Output Speech Signal

Fig. 7. Over view of the proposed methodology.

- The same 10-bit combination in a BFP can be reused for hiding more than one feature word.
- Preserve the sequence of the MFCCs features during the embedding process is essential.
- The selected cover image should be able to accommodate all MFCCs features and the pitches. Each three BFPs can provide indication for 1-time frame features (26 MFCCs + 1 Pitch). For the suggested example above, there are 1000 segmented time frames that need 3000 virtual BFPs in the cover image. Because of using a color image (three arrays), then 1000 BFPs are required per a component array. The square image frame has 100-pixel resolution N_p (10 × 10), then it needs a cover color image with a pixel's resolution no <100,000. In general, the color image resolution can be determined by the suggested formula in equation (10):

$$RES_{req} \geq \frac{3T\,N_P}{3\,(fw-ol)} \tag{10}$$

Where, RES_{req} is the minimum required resolution of the cover image (number of pixels in an image = height pixels X width pixels), *T* is the duration of the speech signal (s), N_p is the number of pixels per square image frame, *fw* is the time frame width (s), and *ol* is the overlap length (s) between adjacent time frames.

The users usually deal with the utterance period (*T*) in seconds and the cover color image resolution, then, using the tradition values such as N_p = 100, *fw* = 0.03 s, and *ol* = 0.02 s, then the minimum required resolution is TRES_{req} ≥10,000 *T*and the maximum speech duration $T_{\max}$ can be hidden in a specific color image with resolution (RES) can be calculated as:

$$T_{max} \leq \frac{RES}{10000} \tag{11}$$

Increasing the image resolution with about 10% could be suggested and encouraged to enhance and ensure the accommodation capacity. Here in MFCCs_based proposed technique, both the minimum required resolution and T_{max} do not depend on the sampling frequency.

For direct speech samples embedding technique, (direct_based) the minimum required resolution (RES_{req}) to hide a specific speech signal with duration time t can be calculated as in equation (12).

$$RES_{req} \geq \frac{10 \times T \times F_s}{3} \tag{12}$$

Or the maximum duration (T_{max}) of a speech clip can be embedded in a specific color image as in equation (13).

$$T_{max} \leq \frac{3\,RES}{10\,F_s} \tag{13}$$

In the direct_based technique, both the minimum required resolution and T_{max} depend on the sampling frequency value.

Concerning to the *accommodation bit capacity*, in the tradition direct_based steganography, when all the samples of the speech signal are required to be embedded directly behind the cover media. At least, if each sample is digitized by 10-bit and using sampling frequency *fs*, this leads the total number of bits (*Nbit*) in the digitized speech (*Nbit*) is as in equation (14):

$$Nbit = 10T\ fs \tag{14}$$

Where 10 means ten bits per sample. For example, If *fs* is taken as 48 kHz, then Nbit = 480000 *T*.

For the introduced method, *Nbit* can be calculated through equation (15):

$$Nbit = \left[100T \times MCF \times 11\right] + [100T \times 12] \tag{15}$$

Where, *MCF* is the number of the Mel Coefficients per time Frame, 100*T* is the total number of the segmented time frames as in equation (8), 11 means the *Nbits* required per MFCC, and 12 is the *Nbits* to represent the pitch value of each frame.

The suggested value of *MCF* is 26, and then equation (15) will be as in equation (16):

$$Nbit \cong 30000T \tag{16}$$

By a rational comparison between the *Nbit* in these two cases above, the required *Nbits* to represent the speech clip is reduced to about 6–7%. In other words, the accommodation capacity of the cover image could be increased by about 16 times.

For the assumed example of 10 s utterance duration which is sampled by *fs* (48 kHz). This means 4,800,000 bits have to be embedded inside the cover image if a tradition steganography is achieved. By the proposed method, 300,000 bits are required to represent this 10 s speech. By simple rational comparison, it is clear that the reduction ratio of the required bits is about 6.25%. In other words, if a color cover image can hide 10 s speech signal (as maximum capacity), then by the proposed method, 160 s can be embedded in the same image size.

F. Recovering of the Secrete Data

In the receiver side, the authorized person should have the technique how to deal with information inserted in the LSBs rows in the BFPs of the received stego-image. In the recovering side, the technique implements, the same steps of image decomposition into its three components (RGB), virtual division into (10 × 10) frames, and creating the (8 × 100) BFPs. The recovering stage starts by checking all the indication segments in the rows of the BFPs sequentially one by one. Each segment provides the enough information to recover a specified feature.

If the matching bit of the segment is logic 1 (exact matching existing), then the next three bits gives the number of the row in the HSBs of the BFP that under process. The rest seven bits means the column number in BFP that the required 10-bit combination starts from. Now, the 10-bit combination location is discovered and its contents copied into the MFCCs recovery matrix. If the matching bit of the segment is logic 0 (no exact matching was found), then the next 10 bits represent the feature word itself directly. They are also directly copied into the recovery matrix. These steps are sequentially implemented for the successive BFPs to keep the order of the MFCCs. Last 12 bits of the third BFP represent the pitch frequency value associated to the recovered MFCCs features. The previous steps in the recovering stage are repeated for the whole BFPs sequentially to recover all the embedded features (MFCCs and their associated pitch frequencies). At the end of the recovering process, all recovered MFCCs features are stored in a specific matrix called (MFCCs recovery matrix). This matrix is a set of MFCCs column vectors; each column vector represents the features of 1-time frame. All recovered pitch frequencies are kept in another vector called (recovery pitch vector). The MFCCs and their related pitches are converted back to the decimal values. The conversion of the MFCCs is done using the same specific LUT in Table I that has to be provided to the recovering side.

G. Magnitude Spectrum Recovering

The recovered MFCCs features are used to re-build the vocal tract filter. This is achieved by reversing the MFCCs back to a smoothed magnitude spectrum using an IDCT and anti-Log (exponential) operation. The required excitation signal is generated from the series of the recovered pitch pulses. This enables the location of the spectral peaks (main formants) in the speech model to be determined. The amplitudes of the peaks were determined from the smoothed spectral.

As shown in Fig. 3, the process by which the MFCCs features are extracted from a speech signal has a number of invertible steps. It is possible to make certain approximations to the information that has been discarded to allow an inverse to be calculated (as the phase c/cs is lost).

The first stage of inverting the MFCCs vector into a magnitude spectral representation needs a logarithmic filter-bank vector. An inverse DCT to be computed as in equation (17):

$$\log C_v(k) = \sum_{n=0}^{K-1} c_v(n) cos\left[\frac{(2k+1)n\pi}{2K}\right] \tag{17}$$

Where *K* is the number of the Mel filter-bank channels (in this work, is 26).

Equation (17) gives a smoothed of the logarithmic filter-bank vector. The log operation can be reversed, by the using of the exponential operation, and gets the Mel-filter-bank vector. Again, by a short notation, the recovering of the vocal tract frequency response can be in the reverse direction of equation (6) and as in equation (18) below:

$$\widehat{PSD} = Mel^{-1}[10^{IDCT(c_v)}] \tag{18}$$

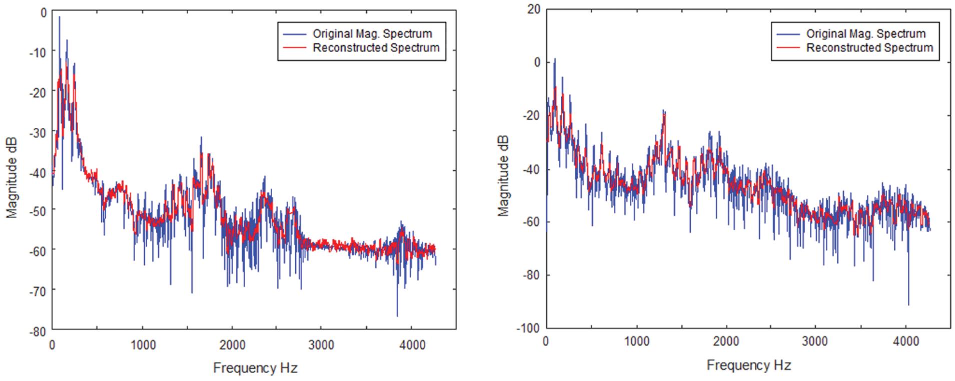


Fig. 8. Two randomly selected recovered magnitude spectrum versus original one.

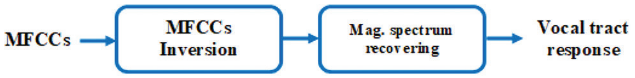


Fig. 9. The steps of vocal tract recovering.

Where $\widehat{PSD}$ referese to the recovered PSD (estimated magnitude spectrum).

Fig. 8 shows two random recovered magnitude spectrums in red line versus original one in blue line.

The area under the triangular filters is used in the Mel-filter-bank analysis increases at the upper frequencies. The effect of this is to impose a high frequency deviation on the resulting Mel-filter-bank channels which make distortion in the generated magnitude spectrum. This deviation can be canceled in the frequency domain by scaling the Mel-filter-bank outputs, by the area of the corresponding triangular Mel-filter.

H. Vocal Tract Recovering

The vocal tract filter coefficients can be computed using Wiener-Khintchine theorem (Kleijn and Paliwal, 1995) that relates the autocorrelation coefficients to the PSD using IFF. The excitation signal is easily reconstructed using a series of the recovered pitch pulses. A suitable value for gain can be added from the log energy element of the feature vector. The steps of vocal tract recovering are shown in Fig. 9.

I. Speech Reconstruction

The recovered vocal tract and the pitch-based excitation signal are merged to reconstruct the speech signal using equation (1). The over view of speech reconstruction is shown in Fig. 10.

IV. Results and Evaluation

The evaluation of the proposed method has two lines. The first one is to evaluate the process of hiding the secret information represented by human speech clip and its impact on the host image quality. This includes the extent of awareness of the unauthorized person of the possibility of data hidden or not. The results show the relationship of the image size with the length of the hidden speech clip and the amount of the hidden data. The second line deals with the reconstructed voice quality using the recovered features detected in the received stego-image.

As mentioned before, the direct embedding of the speech samples is a tradition method of steganography (direct_based). In this case, each sample should be converted to digital form so the size of the digitized speech clip depends on the time length T and on the digitization characteristics (sampling rate fs and $Nbits$ per sample). The minimum required resolution of the host image should follow the threshold value mentioned in equation (12) to hide a specific length of speech. In the same time, the maximum length of the speech has to satisfy equation (13) in case of using a specific cover image that has a limited hiding capacity.

Testing the image quality after the steganography is done by both eyes vision tests and by measuring the value of the peak signal to noise ratio (PSNR). The signal in this case is the original host image, and the noise is the bits introduced through the embedding steps. PSNR is an approximation to human perception of image quality. First, the PSNR is measured for the direct_based method. A color host image with 0.62 Mb resolution (720 × 900) is chosen to hide a speech clip. According to equation (13), the maximum loading of this image is about 16 s speech signal duration if the sampling frequency equal 12 kHz.

To test other standard sampling frequency values (24 kHz and 48 kHz), another larger image is required. Therefore, a color image with 2.3 MB resolution (1080 × 2240) is selected. Fig. 11a shows the PSNR for different partial and full loadings of speech duration time (25%, 50%, 75%, and 100%). As clear from the figure, the sampling frequency in direct embedding method has a great effect on the selected cover image size and on the time length of the speech that to be inserted in the host image. Doubling the sampling rate leads necessarily to reduce the speech time to the half or duplicating the image size. However, the most noticeable thing in Fig. 11a is for one ratio of speech loading, the PSNR values are almost equal regardless to the sampling frequency.

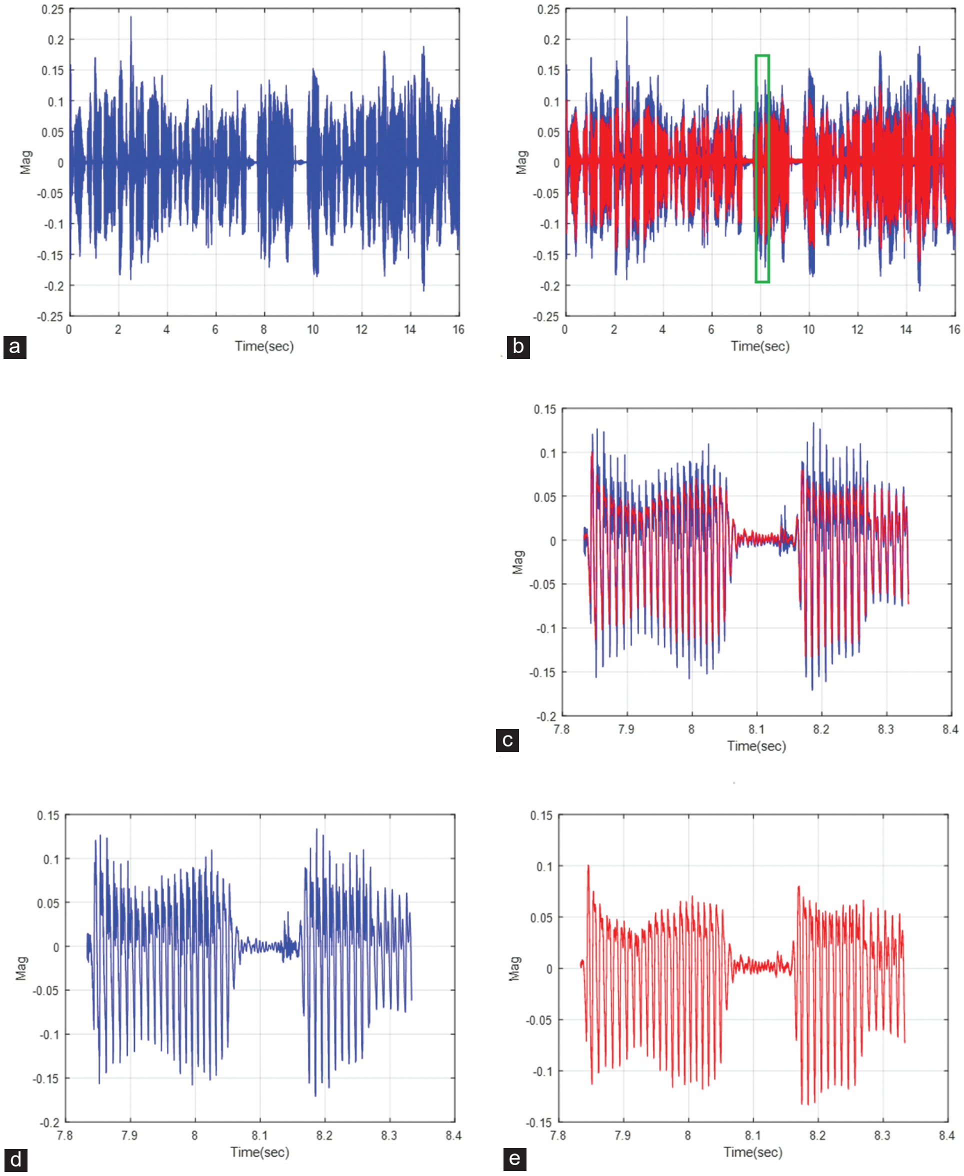


Fig. 10. Over view of speech reconstruction. (a) The original speech signal, (b) both original speech signal (blue line) and the recovered speech signal (red line), (c) zoom in for both original signal (blue line) and the recovered signal (red line), (d) zoom in for original speech signal, (e) zoom in for recovered signal.

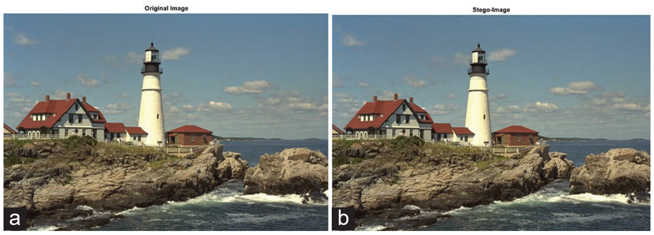


Fig. 11. (a) Peak signal to noise ratio (PSNR) for the direct based method of hiding. The legend includes the sampling frequency, max speech duration, and host color image resolution in Mb, respectively. (b) PSNR for the Mel-frequency cepstral coefficients based on proposed method of hiding for three standard values of sampling frequency.

Concerning to the evaluation for the proposed method of hiding the secret data using the MFCCs features (MFCCs_based) rather than all clip samples. The same color host image is chosen again that has size 2.3 Mb (720 × 900). For this case and according to equation (11), a speech of about 64 s can be full loading this image, that is, 4 times more than if the direct way to hide the samples is used. Fig. 11b shows the PSNR for different partial and full loading (25%, 50%, 75%, and 100%) using three normally used sampling rates (12 kHz, 24 kHz, and 48 kHz). The same observation appears in which the PSNR values for one speech-loading ratio are almost equal regardless to the used sampling rate.

Fig. 12 shows the original host image and the stego-image of full capacity speech loading using the proposed technique (MFCCs_based). It is clear that no major vision effects are appearing in the stego-image compared with original one.

Relating to the *Nbits* represents the all secret data that are used in the hiding process. *Nbits* has effects on the stego-image quality. Table II and Fig. 13 show the difference in the *Nbits* for both techniques of the data hiding, the direct_based hiding of all samples, and the MFCCs_based hiding. The results shown in Table II are got using equations (14) and (16). In the direct hiding technique, the *Nbits* very related to both of the value of the sampling frequency and on the speech duration time, whereas the proposed MFCCs_based hiding depends only on the duration and the sampling rate has no effect on the *Nbits* used in the hiding process. The *Nbits* also affect on the time required of the hiding and recovering processes.

TABLE II
Nbits Used in Both Direct and MFCCs Techniques

Speech time duration (s)	Nbits (Mb) (Direct_Based)			Nbits (Mb) (MFCCs_Based)
	fs 12 kHz	fs 24 kHz	fs 48 kHz	Any fs
10 s	1.2	2.4	4.8	0.3
20 s	2.4	4.8	9.6	0.6
30 s	3.6	7.2	14.4	0.9
40 s	4.8	9.6	19.2	1.2
50 s	6	12	24	1.5
60 s	7.2	14.4	28.8	1.8
70 s	8.4	16.8	33.6	2.1

MFCCs: Mel-frequency cepstral coefficients, Nbits: Number of bits

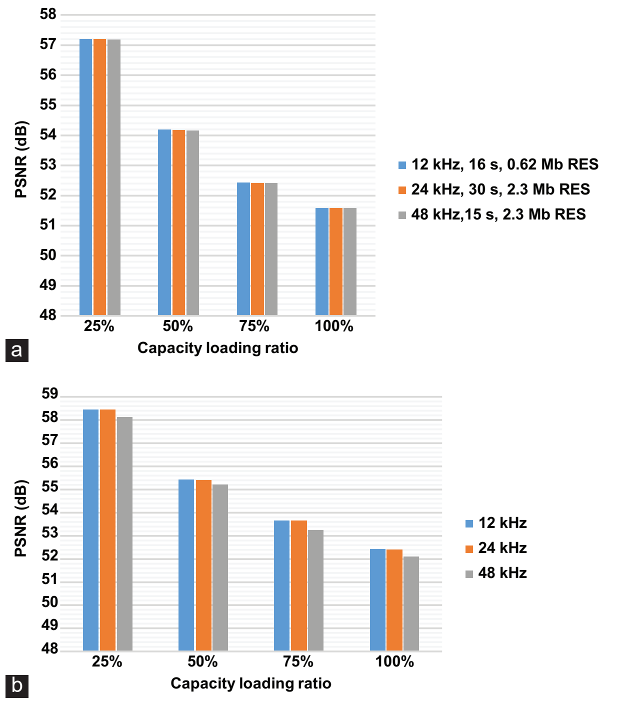


Fig. 12. (a) Original host color image. (b) Stego-image after full capacity speech signal loading.

The reduction ratio in *Nbits* using the proposed MFCCs_based is within (6.25%–25%).

The second line of the evaluation for the proposed technique is the rating of the quality for the reconstructed speech compared with the original one. Listening tests show that clear and very acceptable speech signals were produced. Fig. 8 shows both the recovered magnitude spectrum and that computed from the original speech signal for two random selected time frames. The magnitude spectrum estimated directly from the original frames is shown as the blue line whereas the magnitude spectrum got from the recovered MFCCs and pitches vectors is shown as the red line. The figures also show that the envelope of the magnitude spectrum has been reasonably well conserved. Some problems may appear where the magnitude spectrum gives formants merging into a single peak.

The comparison is also done by measuring the correlation coefficient between the original speech signal and the reconstructed one. Fig. 14 shows the correlation coefficients versus different number of extracted MFCCs from the speech signal to be embedded inside the cover image. The high number of the MFCCs leads to higher *Nbits* and then needing bigger image size or less speech time duration besides increasing the time required for the process. The selected number of MFCCs equation (26) is the best choice that corresponds to the requirements of the process in terms of the amount of data, the accuracy of the results (gives a correlation coefficient 94.24%), and the time required for the execution of the steps.

Table III presents a comparison between the proposed MFCCs based technique with some other existing models.

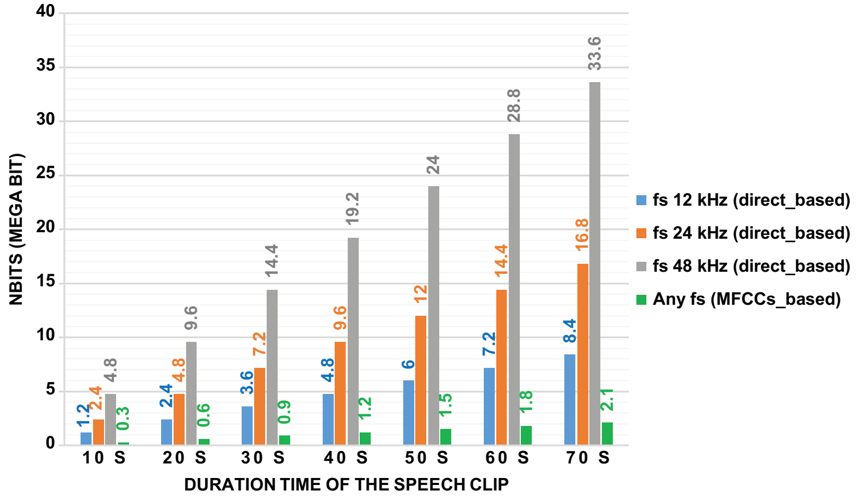


Fig. 13. Number of bits with respect to the speech duration time for different sampling rates.

TABLE III
Comparison between the Proposed MFCCs Based Technique with Some Other Existing Models

Method	Used domain	Host media	Secrete data	PSNR (dB)
(Nipanikar, Deepthi and Kulkarni, 2017)	DWT	Digital Gray Image	Speech signal	47.6
(Saroj and Dewangan, 2018)	Spatial and DCT	Digital Color Image	Audio signal	-
(Sharma, 2015) 28	Spatial	Digital Color Image	Speech signal	-
(Abdulraman, et al., 2019)	Spatial	Digital Gray Image	Text	51.9
(Jamel, 2019)	DWT	Digital Gray Image	Digital Gray Image	36.84
(Navas, Thampy and Sasikumar, 2008)	Integer WT	Digital Gray Image	Patient's records	44
(Al-Qershi and Khoo, 2011)	DWT	Digital Gray Image	Patient's records	41.25
Proposed MFCCs based technique	Spatial and Cepstral	Digital Color Image	Speech signal	52.41

MFCCs: Mel-frequency cepstral coefficients, PSNR: Peak signal to noise ratio

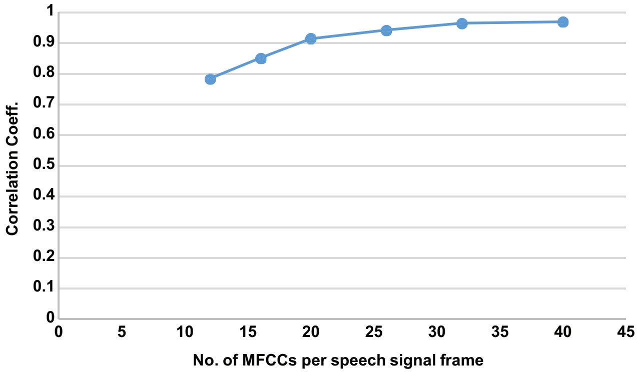


Fig. 14. The correlation coefficient with respect to the number of extracted Mel-frequency cepstral coefficients per speech frame.

V. Conclusion

This work presented a proposed technique of human speech signal hiding in a color cover digital image. The proposed technique introduces solutions for some challenges facing the steganography process, such as increasing the security of sending the secrete data, reducing the amount of the hidden data, and increasing the capacity of the host image. The presented method depends on sending the MFCCs features plus additional information relating the excitation signal and speech duration that should be inserted with the features. The speech reconstruction is possible from a stream of the recovered features using a model of speech production. Sending the speech features rather than the whole samples of the speech signal is the main reason to reduce the amount of the embedded data and can increase the security as well. The method used to hide the confidential data depends on including it in the host image at the higher significant levels and not as in the traditional methods do at the least significant levels. The most noticeable conclusion here that in the MFCCs_based steganography, the value of the sampling rate *fs* used in the speech digitization has no effect on the loading percentage inside the host image. Another conclusion, that for a specific speech-loading ratio, the PSNR is invariant with the change of the sampling frequency used. The possible MFCCs can be extracted from a speech signal which can be between 12 and 40. Therefore, it is concluded that 26 MFCCs is an optimal choice that considers both the low data amount and acceptable speech reconstruction degree with about 94.24% correlated with the original speech. If high number of MFCCs is used, better-quality speech reconstruction is possible from the MFCCs despite the missing of phase information in MFCCs_based steganography but with more hidden data, lower PSNR, and longer time of processing. The amount of the hidden data depends only on the duration time of the speech signal not on the sampling rate as in the direct method.

References

Abdulraman, L.S., Hma-Salah, S.R., Maghidid, H.S. and Sabir, A.T., 2019. A robust way of steganography by using blocks of an image in spatial domain. *Innovaciencia*, 7(1), pp.1-7.

Al-Qershi, O.M. and Khoo, B.E., 2011. High capacity data hiding schemes for medical images based on difference expansion. *Journal of Systems and Software*, 84(1), pp.105-112.

Chakroborty, S., Roy, A. and Saha, G., 2007. Improved closed set text independent speaker identification by combining mfcc with evidence from flipped filter banks. *International Journal of Signal Processing*, 4(2), pp.114-122.

Cox, I.J., Miller, M.L., Bloom, J.A., Fridrich, J. and Kalker, T., 2008. *Digital Watermarking and Steganography*. 2nd ed. Morgan Kaufmann Publishers, USA.

Davis, S. and Mermelstein, P., 1980. Comparison of Parametric Representations for Monosyllable Word Recognition in Continuously Spoken Sentences, 28(4), pp. 357-366.

Huang, X., Acero, A. and Hon, H.W., 2001. *Spoken Language Processing*. Prentice Hall, Inc., USA.

Jamel, E.M., 2019. Secure image steganography using biorthogonal wavelet transform. *Journal of Engineering and Applied Sciences*, 14, pp.9396-9404.

Kleijn, W.B. and Paliwal, K.K., 1995. *Speech Coding and Synthesis*. Elsevier Science Inc., New York, United States.

Navas, K., Thampy, S.A., Sasikumar, M. 2008. EPR hiding in medical images for telemedicine. *International Journal of Biomedical Science*, 2(1), pp.292-295.

Nipanikar, S.I., Deepthi, V.H. and Kulkarni, N., 2017. A sparse representation based image steganography using particle swarm optimization and wavelet transform. *Alexandria Engineering Journal*, 57, 2343-2356.

Oliveira, M.L.L., Cerqueira, J.J.F. and Filho, E.F.S., 2018. *Simulation of an Artificial Hearing Module for an Assistive Robot*. Intelligent Systems Conference, London, UK.

Saroj, N. and Dewangan, S.K., 2018. An implementation of hiding audio secure data in images using steganography. *Journal of Emerging Technologies and Innovative Research*, 5(9), 20-25.

Sharma, D., 2015. Steganography of Speech Signal into an Image. *Second International Conference on Recent Advances in Engineering and Computational Sciences*, USA.