ISSN 2007-9737

Digital Image Steganography Scheme for Speech Signals in the Discrete Wavelet Transform Domain

Ariel Rodríguez Méndez, Clara Cruz Ramos, Rogelio Reyes Reyes, Volodymyr Ponomaryov

Instituto Politécnico Nacional,
ESIME Culhuacán,
Mexico

arodriguezm1409@egresado.ipn.mx, {ccruzra, rreyesre, vponomar}@ipn.mx

Abstract. In recent years, the evolution and expansion of the multimedia world has allowed digital images to be widely used as a carrier medium for facilitating the sending and receiving of confidential digital content such as text, audio and images. However, there are few steganographic schemes that hide digital speech signals into digital images; also, the state-of-the-art methods have some disadvantages such as low embedding capacity, low quality in the modified image, and/or low audible quality in the recovered speech signal. In this paper, a new steganographic scheme that hides a digital speech signal into a color image is presented, where the embedding process is performed in the *HL* and *LH* channels of the first decomposition level of the discrete wavelet transform (DWT) over the YCbCr color space; additionally, a chaotic map method is implemented to encrypt the speech signal and to spread it throughout the whole image. Experimental results have demonstrated that the proposed scheme results in better performance against other state-of-the-art methods, due to their ability to embed and recover speech signals with duration up to 16.384 seconds using color images of 512×512 pixels, obtaining 32 *dB* and 0.92 in PSNR and SSIM, respectively for the stego-image, and 41 *dB* in SNR for the recovered speech signal.

Keywords. Steganography, digital image, speech signal, DWT, chaotic map.

1 Introduction

The main objective of steganography is to establish forms of secret communication between two points by hiding a message on a carrier medium, in order to prevent an intermediary from detecting the existence of such communication [1]. The steganographic techniques can be applied on the spatial or frequency domain however, the spatial domain techniques are fragile against to compression or noise contamination attacks, among others. On the other hand, frequency domain techniques, such as discrete cosine transform (DCT) or discrete wavelet transform (DWT) offer better robustness against various intentional and unintentional common attacks [2]. In addition, it is possible to perform a hybrid steganography method that uses more than one frequency domain as shown in [3].

However, the data hiding methods implemented in the frequency domain, even though they are robust to different manipulation attacks on the carrier medium, still present low embedding capacity, low quality in the modified carrier medium, and/or low quality in the recovered secret message.

Digital images are the most preferred medium to apply steganography, as they contain large amount of redundant data. Image steganography embeds the secret message into the image by modifying pixel intensities in a manner that alterations are hardly noticeable to human eye [4].

Nevertheless, the type of digital content (text, image, audio) to hide into the carrier image can produce different effects on the three main characteristics of steganography: imperceptibility, robustness and embedding capacity. This means that each steganographic method must be designed taking into account the three main characteristics mentioned above in order to obtain the best performance.

On the other hand, digital speech signals and digital images are not commonly employed together in steganography schemes due to the existing difficulties regarding their sizes.

For example, just a 5-second speech signal with a sampling rate of 8 *kHz* can hold up to 40,000 samples and, if it is necessary to represent the information as a binary sequence, there would be 640,000 bits for 16-bit resolution per sample. Thus, it is necessary to develop steganography techniques that allow embed a large amount of binary data, maintaining a high trade-off between the three characteristics mentioned above, and ensuring that the extracted speech signal from the stego-image (carrier image with embedded secret message) is intelligible.

In this paper, a digital image steganography scheme for speech signals is designed. This new steganographic framework appears to demonstrate high embedding capacity of speech signals and robustness against JPEG compression. The embedding process is implemented in the DWT domain of the luminance channel using the YCbCr color space, additionally, a chaotic map method is implemented to encrypt the speech signal and spread it throughout the whole image.

2 Related Work

Previously proposed steganographic schemes that embed speech signals into digital images are discussed below.

Talbi et al. [5], proposed a scheme that embeds and extracts a speech signal using digital color images as a carrier. The embedding process is performed in the frequency domain using the DCT. The speech signal is amplitude modulated before to be embedded into the DCT coefficients of the red color channel. For extraction process, the original carrier image is required to correctly extract the secret signal from the stego-image, however, the maximum SNR value for the recovered speech signal is 16 *dB*.

In [6], a steganographic method based on the singular value decomposition (SVD) for speech signals is presented. Firstly, the speech signal is transformed into a gray-scale image, then, the singular value decomposition (SVD) is applied to the speech gray-scale image; also, the SVD is applied to the *HH1* subband of the first level of the lifting wavelet transform (LWT) of the color carrier image.

Data hiding is performed by replacing the singular values of *HH1* with the singular values of the speech gray-scale image; then, SVD is applied to obtain the modified *HH1* band. Finally, an 8-bit signature is generated using the *U* and *V* matrices of the speech gray-scale image to embed it into the *LL4* and *HH4* subbands of the fourth level decomposition of the LWT. For the extraction process, the embedded original 8-bit signature is compared with the extracted from the stego-image and, if the signatures match, the secret speech signal is reconstructed applying SVD to the modified subband, extracting the SVs and using the U_w and V_w orthogonal matrices.

Talbi [7] presents a steganographic method based on [6], which additionally uses an encryption technique to increase the security of the hidden speech signal. Experimental results obtained in [6, 7] have demonstrated low embedding capacity, and regular performance in the quality of the stego-image (PSNR ≈ 40 *dB*) and the quality of the extracted speech signal (SNR of 25 *dB*) for maximum 8-second recording speech signals using 512×512 carrier color images.

Punidha et al. [8], described a scheme based on the integer wavelet transform (IWT) and the YCbCr color space. For the embedding process, the secret message is generated using the approximation coefficients of the first level IWT of the speech signal and it is embedded into the *LL1* subband of the first level IWT decomposition of the *Cb* channel. During the extraction process, only the stego-image is required, however, only the *Cb* channel is used for hiding in order to maintain the visual quality of the stego-image, resulting in limited embedding capacity of this method.

In summary, steganographic systems that embed speech signals into digital images mainly have the disadvantage of low embedding capacity; thus, when a speech signal is embedded, too much pixels in the carrier image are modified, degrading the visual quality of the stego-image.

3 Methodology

The proposed scheme allows to embed digital speech signals in color images with a semi-blind steganographic approach.

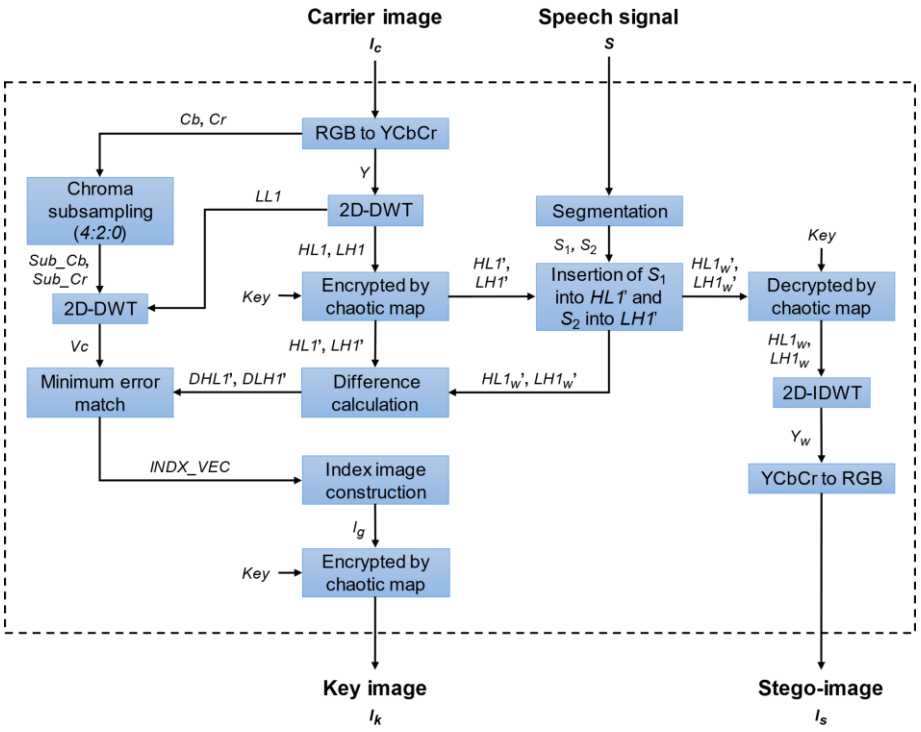


Fig. 1. Block diagram for the embedding process of the proposed scheme

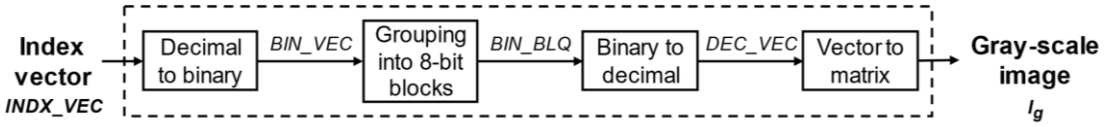


Fig. 2. Indexes normalization process for grayscale image construction

The embedding and extraction processes are developed in the DWT domain (Haar wavelet) and a sine chaotic map is used to increase the security of the hidden message and spread it throughout the whole image avoiding its direct accessibility. The designed scheme is mainly divided in two process. The first one consists of the embedding stage, which is presented in figure 1, and the second process forms the extraction stage, which can be observed in figure 3.

3.1 Speech Signal Embedding

Figure 1 shows the detailed block diagram for the speech signal embedding process.

Firstly, the carrier image is transformed from the RGB color space to the YCbCr color space [9]:

$$Y = 0.257R + 0.504G + 0.098B + 16, \quad (1)$$

$$Cb = -0.148R - 0.291G + 0.439B + 128, \quad (2)$$

$$Cr = 0.439R - 0.368G - 0.071B + 128. \quad (3)$$

Then, the *Y* channel is transformed to the first-level DWT decomposition in order to obtain the *LL1*, *HL1*, *LH1* and *HH1* subbands, where *HL1* and *LH1* are encrypted by the sine chaotic map [10] using the equation (4), obtaining the chaotic subbands *HL1*' and *LH1*'.

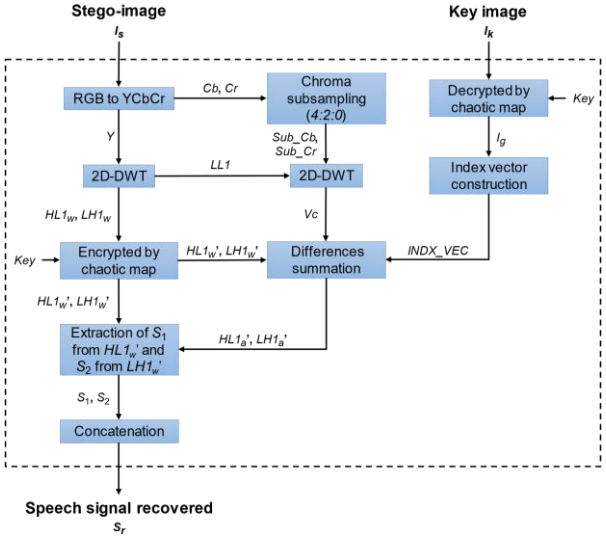


Fig. 3. Block diagram for the extraction process of the proposed scheme

$$x_{n+1} = f(x_n, r) = r\sin(\pi x_n), \tag{4}$$

where x_{n+1} is the resulting chaotic sequence and *r* is a control parameter that determines the bifurcation and entropy property of the chaotic performance.

The digital speech signal is divided in two segments of the same size (S_1 and S_2), which are embedded into the *HL1*' and *LH1*' coefficients according to following equation:

$$A_w = A + \propto B, \tag{5}$$

where *A* is the carrier coefficient, $\propto$ is a parameter that defines the embedding strength, *B* is a value of the secret signal, and A_w is the new carrier coefficient modified.

In this way, the chaotic subbands with the secret signal $HL1_w$' and $LH1_w$' are obtained and decrypted using the sine chaotic maps system. Then, the modified luminance channel Y_w is calculated by applying the inverse DWT (IDWT) to the *LL1*, $HL1_w$, $LH1_w$ and *HH1* subbands. The stego-image I_s is obtained by the YCbCr to RGB conversion as follows:

$$R = 1.164(Y - 16) + 1.596(Cr - 128), \tag{6}$$

$$G = 1.164(Y - 16) - 0.813(Cr - 128) - 0.392(Cb - 128), \tag{7}$$

$$B = 1.164(Y - 16) + 2.017(Cb - 128). \tag{8}$$

After the stego-image is obtained, the differences between the subbands *HL1*', *LH1*' and $HL1_w$', $LH1_w$' respectively, are calculated as follows:

$$DHL1' = HL1' - HL1_w', \tag{9}$$

$$DLH1' = LH1' - LH1_w'. \tag{10}$$

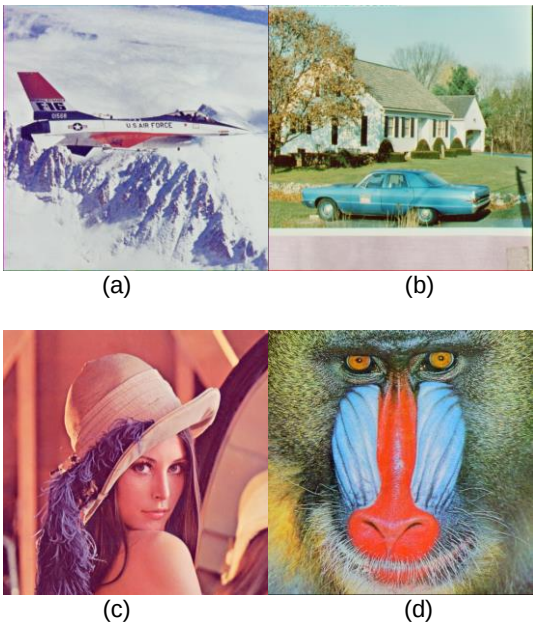
Additionally, the *Cb* and *Cr* channels are resampled using the 4:2:0 chroma subsampling format, obtaining *Sub_Cb* and *Sub_Cr*. Then, a vector *Vc* is performed from the coefficients in the *HL2*, *LH2*, and *HH2* subbands of the second-level DWT decomposition of *LL1*, and the coefficients of the *Sub_Cb* and *Sub_Cr* transformed to the first-level DWT decomposition.

In this way, each value in *DHL1*' and *DLH1*' is used in (11) to calculate the absolute difference with each element in *Vc*. Subsequently, is necessary to locate and store the index of the coefficient in *Vc* where the absolute difference has the smallest value (minimum error) obtaining *INDX_VEC*:

$$DA = |a - b|, \tag{11}$$

where $a \in \{DHL1', DLH1'\}$ and $b \in Vc$.

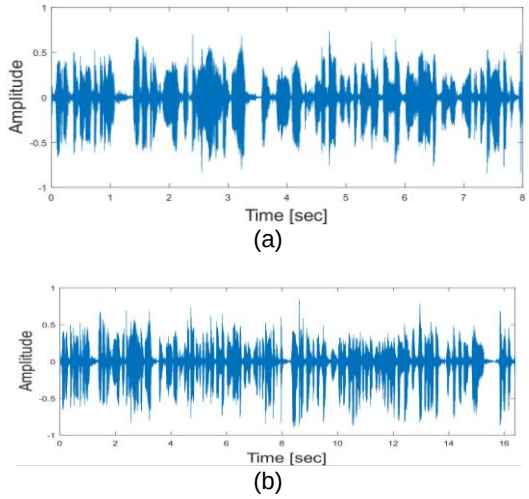
The obtained indexes *INDX_VEC* are normalized to a range of values between 0 and 255 to construct a grayscale image, which is encrypted using the sine chaotic map. Figure 2 shows the normalization process of the indexes. As can be seen, a matrix is performed from the normalized indexes vector. The dimensions of this matrix are established from the maximum information capacity that can be embedded and the necessary number of bits to represent the largest index of *Vc*. For example, for a 512×512 image, it is possible to hide up to 131,072 audio samples and 18 bits are required to represent the largest index of *Vc*; therefore, there are 294,912 normalized values in the range of 0 to 255. Thus, the matrix should be dimensioned using the following equation:



(a) (b)

(c) (d)

Fig. 4. Carrier images: (a) Airplane, (b) House, (c) Lena, (d) Baboon



(a)

(b)

Fig. 5. Speech signals used as secret messages: (a) 8 sec, (b) 16.384 sec

$$M = \lfloor \sqrt{INs} \rfloor, N = \lceil \sqrt{INs} \rceil, \quad (12)$$

where *M* is the vertical dimension (height), *N* is the horizontal dimension (width) and *INs* is the normalized value vector size. So, for *INs* = 294,912, the dimensions of the matrix should be 543×544, obtaining the gray-scale image I_g. Finally, the key image I_k is obtained by applying the sine chaotic map to I_g.

3.2 Speech Signal Extraction Process

The speech signal recovery process requires the stego-image and the key image. Figure 3 explains the procedures required to extract and reconstruct the speech signal. Firstly, the stego-image is transformed to the YCbCr color space using (1)-(3); then, the *Y* channel is divided into the *LL1*, $HL1_w$, $LH1_w$ and *HH1* subbands derived from the first-level DWT decomposition. $HL1_w$ and $LH1_w$ contain the secret message, however, it is necessary to encrypt this subbands by the sine chaotic map before starting extraction.

As shown in figure 3, the vector *Vc* is constructed in the same way as in the embedding process but with the stego-image. On the other hand, the key image is decrypted using the sine chaotic map in order to convert each pixel to a binary sequence and, knowing the number of bits necessary to represent the largest index of *Vc*, binary blocks are generated, which are converted to decimal values. The new normalized values are used as indexes to locate the coefficients in *Vc* that are closer to the differences *DHL1*' and *DLH1*' calculated at the embedding stage.

The "*Differences summation*" block shown in figure 3 consists of computing the approximations of the original chaotic subbands *HL1*' and *LH1*' using following equations:

$$HL1_a' = HL1_w' + DHL1_a', \quad (13)$$

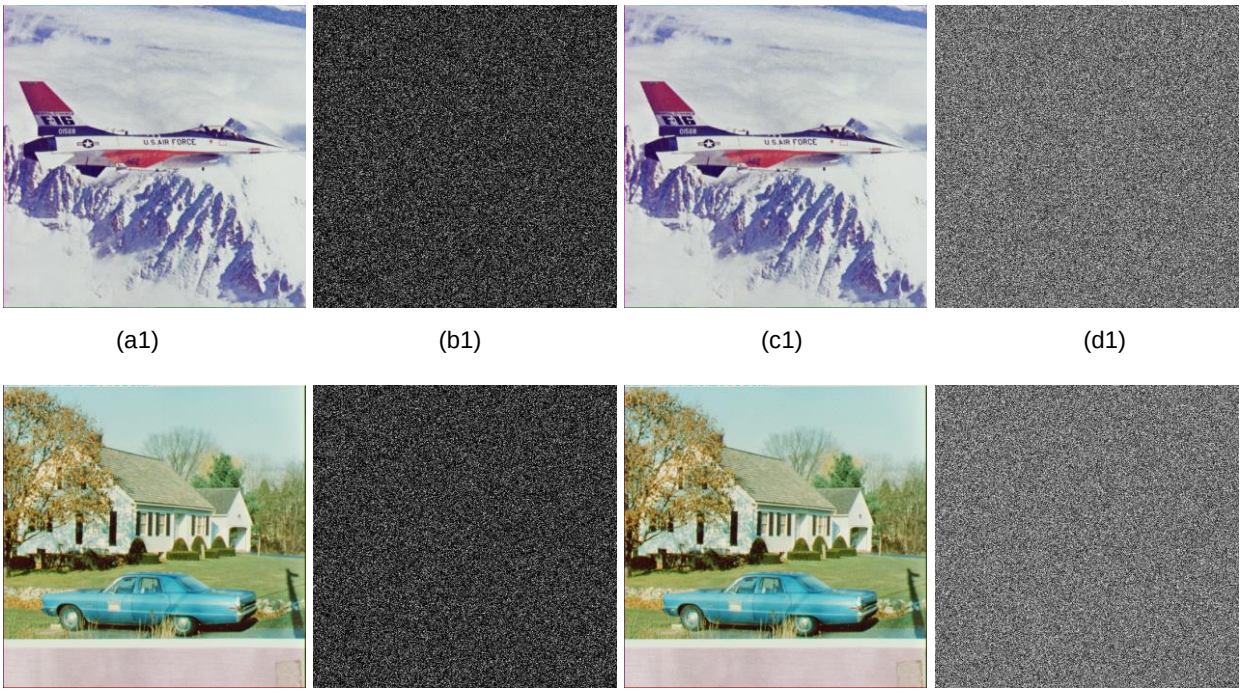
$$LH1_a' = LH1_w' + DLH1_a', \quad (14)$$

where $DHL1_a$' and $DLH1_a$' are the approximations of the differences *DHL1*' and *DLH1*', respectively.

Therefore, the difference between $HL1_w$' and $HL1_a$' allows the first S_1 segment extraction of the secret message, and the difference between $LH1_w$' and $LH1_a$' permits the second S_2 segment extraction using equation (15). The secret message extracted from both segments are concatenated to obtain the recovered speech signal S_r.:

$$B = (A_w - A_x)/\propto, \quad (15)$$

ISSN 2007-9737



(a1) (b1) (c1) (d1)

(a2) (b2) (c2) (d2)

Fig. 6. Stego-images and key images obtained from the embedded process for two different speech signals: (a1, a2) Stego-images and (b1, b2) Key images with 8 sec embedded speech signal; (c1, c2) Stego-images and (d1, d2) Key images with 16.384 sec embedded speech signal

Table 1. Objective evaluation comparison of the stego-images and recovered speech signals for the pro-posed scheme and the state-of-the-art methods. (NS – Not supported)

Carrier image	Scheme	Stego-image				Recovered speech signal	
		PSNR (*dB*)		SSIM		SNR (*dB*)	
		8 sec	**16.38 sec**	**8 sec**	**16.38 sec**	**8 sec**	**16.38 sec**
Airplane	Proposed	34.42	**31.85**	0.87	**0.82**	**42.72**	**40.58**
	[5]	31.70	30.69	0.89	0.81	13.82	12.74
	[6]	41.16	NS	**0.99**	NS	27.0	NS
	[7]	**43.48**	NS	**0.99**	NS	14.24	NS
	[8]	NS	NS	NS	NS	NS	NS
House	Proposed	34.43	**31.87**	0.95	**0.92**	**43.29**	**40.62**
	[5]	30.55	29.51	0.94	0.91	13.82	12.90
	[6]	37.84	NS	**0.99**	NS	11.31	NS
	[7]	**41.01**	NS	**0.99**	NS	18.23	NS
	[8]	NS	NS	NS	NS	NS	NS

ISSN 2007-9737

where A_w are the values obtained from $HL1_w$' or $LH1_w$', and A_x are the approximate values obtained from $HL1_a$' or $LH1_a$'.

4 Experimental Results

To evaluate the performance of the proposed system, different tests have been carried out on

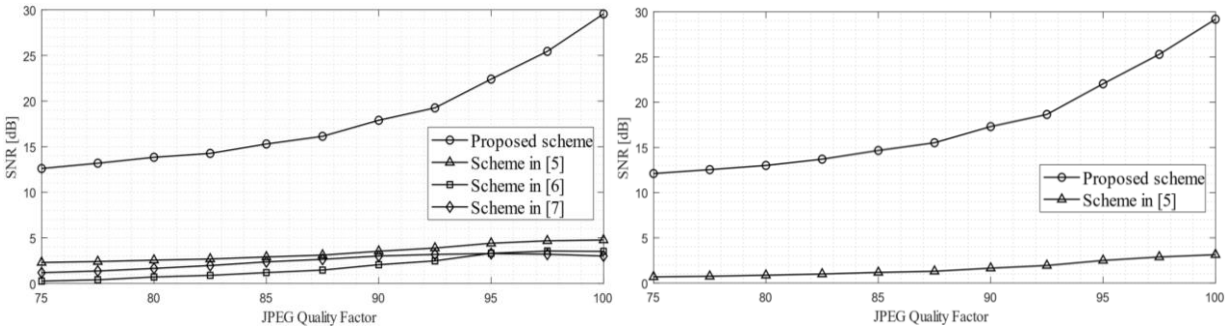


Fig. 7. Robustness evaluation comparison of recovered speech signals against JPEG compression at different quality factors for the stego-image *Airplane* using: (a) 8 sec speech signal, (b) 16.384 sec speech signal

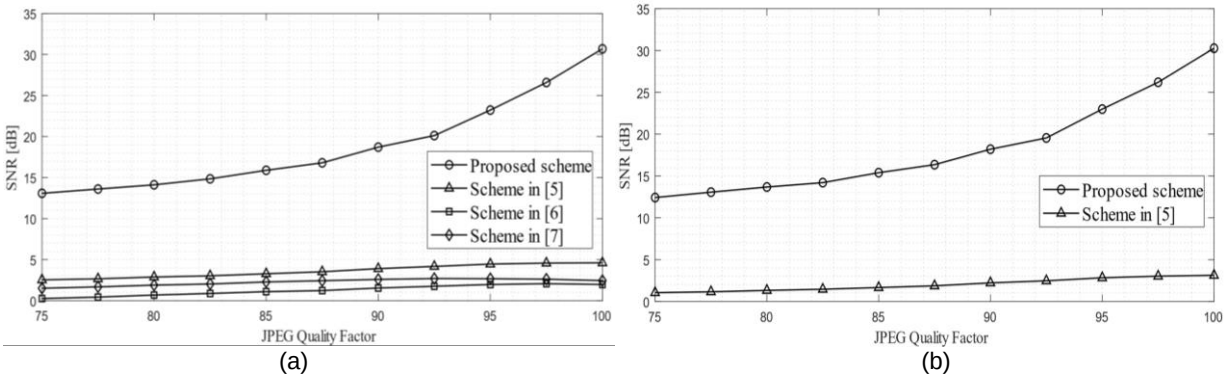


Fig. 8. Robustness evaluation comparison of recovered speech signals against JPEG compression at different quality factors for the stego-image *House* using: (a) 8 sec speech signal, (b) 16.384 sec speech signal

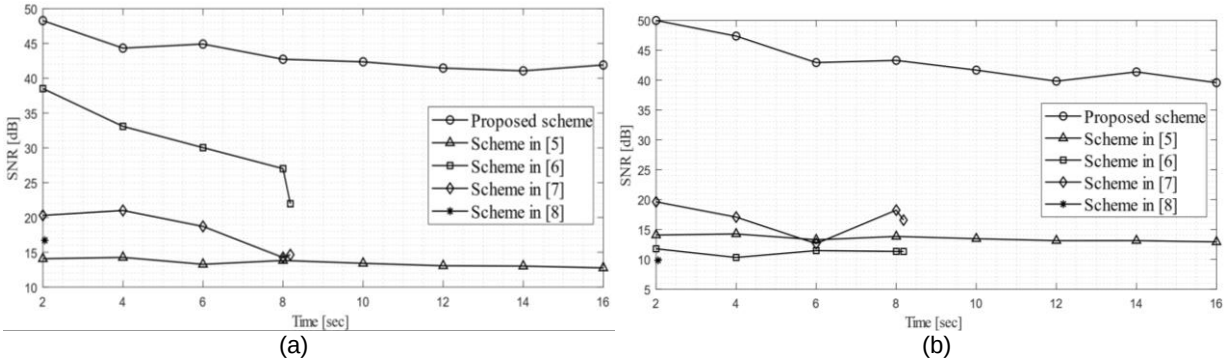


Fig. 9. Recovered speech signals audio quality comparison between the proposed scheme and state-of-the-art methods for different recording times, (a) Stego-image *Airplane*, (b) Stego-image *House*

RGB images with dimensions of 512×512 pixels. The main characteristics of the digital speech signals used as the secret message to perform the tests are: sampling frequency *Fs* = 8000 *Hz*, bits of quantization *bQ* = 16, monoaural in .wav format, from 2 to 16.384 seconds of recording time values, and the embedding strength has been adjusted to $\propto = 50$.

Figure 4 shows some color images used as carriers in the proposed scheme.

Figure 5 shows two speech signals with different recording times used as secret messages. It is important to mention that the speech signal shown in figure 5(b) allows to embed the maximum possible information capacity into the carrier images.

Figure 6 shows the stego-images and the key images obtained from the embedded process for two different speech signals.

Table 1 shows the experimental results comparison between the proposed scheme and state-of-the-art schemes for the imperceptibility of the secret message and the final visual quality of the stego-image, when the objective evaluation criteria were the PSNR and SSIM values between the original carrier images and the stego-images. Additionally, the audio quality of the recovered speech signal has been evaluated using the signal-to-noise ratio (SNR) values between the original speech signal and the reconstructed one.

Figures 7 and 8 show the objective audio measure SNR for the recovered speech signals against JPEG compression, applied to the stego-images *Airplane* and *House* respectively, with different quality factors, in order to evaluate the robustness of the proposed scheme and the state-of-the-art methods. As we can be seen in figure 7 and figure 8, the method presented in [8] cannot embed speech signals larger than 2.048 seconds of recording time into color images of 512×512 pixels. The same conclusion can be achieved for schemes presented in [6] and [7] (see figures 7(b) and 8(b)), because they can only embed speech signals with a maximum recording time of 8.192 seconds.

Figure 9 shows a performance comparison between the proposed scheme and state-of-the-art methods, regarding to the recovered speech signals audio quality for different recording times from 2 to 16 seconds. As it can be noticed in experimental results, the proposed scheme appears to demonstrate an outstanding performance compared with state-of-the-art schemes in terms of imperceptibility for the secret message, excellent visual quality of the stego-image, improved embedding capacity, higher robustness against JPEG compression and superior audio quality of the recovered speech signal.

5 Conclusions

In this paper, a new steganographic scheme that hides a digital speech signal into a color image is presented. The proposed scheme has been evaluated with several speech signals with different recording times, demonstrating higher embedding capacity, better robustness and imperceptibility of the secret message, excellent visual quality of the stego-image and superior audio quality of the recovered speech signal in comparison with state-of-the-art schemes.

The designed scheme shows a non-visual perceptible difference between the original carrier image and stego-image, even if the maximum embedding capacity of the proposed scheme is used, obtaining an average values of 32 *dB* and 0.92 in PSNR and SSIM, respectively.

The main advantage of the proposed steganographic scheme is the robustness against JPEG compression at different quality factors (from 75 to 100), allowing to embed a speech signal up to 16.384 seconds of recording time into a carrier color image with 512×512 pixels, without any degradation of the audio quality of the recovered speech signal (up to 41 *dB* of SNR), even if the stego-image is JPEG compressed (up to 30 *dB* of SNR).

Acknowledgements

Authors would like to thank to Instituto Politécnico Nacional, Consejo Nacional de Ciencia y Tecnología (CONACYT) and Comisión de Operación y Fomento de Actividades Académicas (COFAA) del Instituto Politécnico Nacional, for their support during the development of this work.

References

1. **Boquera, M.C.E. (2003).** Servicios avanzados de telecomunicación. *Díaz de Santos*, Madrid.
2. **Lai, C. & Tsai, C. (2010).** Digital image watermarking using discrete wavelet transform and singular value decomposition. *IEEE Transactions on Instrumentation and Measurement*, Vol. 59, No. 11, pp. 3060–3063. DOI: 10.1109/TIM.2010.2066770.
3. **Saini, M. (2018).** LWT based hybrid digital watermarking scheme in YCbCr colour space. *International Conference on Intelligent Circuits and Systems IEEE*, pp. 206–211. DOI: 10.1109/ICICS.2018.000.
4. **Watni, D. & Chawla, S. (2019).** A comparative evaluation of jpeg steganography. *5th International Conference on Signal Processing, Computing and Control IEEE*, pp. 36–40. DOI: 10.1109/ISPCC48220.2019.8988383.
5. **Talbi, M., Ftima, S.B., & Cherif, A. (2017).** Speech modulation for image watermarking. *International Conference on Control, Automation and Diagnosis IEEE*, pp. 522–527. DOI: 10.1109/CADIAG.2017.8075713.
6. **Talbi, M. & Bouhlel, M.S. (2019).** Singular values decomposition and lifting wavelet transform for speech signal embedding into digital image. *Recent Advances in Electrical & Electronic Engineering*, Vol. 12, No. 2, pp. 138–151. DOI: 10.2174/2352096511666180511151646.
7. **Talbi, M. (2018).** Speech signal embedding into digital images using encryption and watermarking techniques. (*SETIT´18) Smart Innovation, Systems and Technologies*, Springer, Vol 147, pp. 3–13. DOI: 10.1007/978-3-030-21009-0_1.
8. **Punidha, R. & Sivaram, M. (2017).** Integer wavelet transform based approach for high robustness of audio signal transmission. *International Journal of Pure and Applied Mathematics*, Vol. 116, No. 23, pp. 295–304.
9. **Jack, K. (1997).** *YCbCr to RGB considerations.* https://www.renesas.com/us/en/www/doc/application-note/an9717.pdf
10. **Pak, C. & Huang, L. (2017).** A new color image encryption using combination of the 1D chaotic map. *Signal Processing*, Vol. 138, pp. 129–137. DOI: 10.1016/j.sigpro.2017.03.011.

Article received on 13/06/2020; accepted on 22/07/2020.
Corresponding author is Ariel Rodriguez Mendez.