# Highly Transparent and Secure Scheme for Concealing Text Within Audio

Diego Renza(✉), Camilo Lemus, and Dora M. Ballesteros L.

Universidad Militar Nueva Granada, Bogotá DC, Colombia
{diego.renza,camilo.lemus,dora.ballesteros}@unimilitar.edu.co

**Abstract.** This paper presents a highly transparent and secure scheme for concealing text within audio, based on Quantization Index Modulation and Orthogonal Variable Spreading Factor. The audio signal is decomposed through the Discrete Wavelet Transform and the approximation coefficients are selected to embed the text. Every character of the text is represented by a 256-bit orthogonal code, through mapping operations between the ASCII integer representation of the character and an external key. For improving the quality of recovered data, a repetition code is applied in the embedding process. Several tests were performed in order to measure the transparency of the output audio signal (i.e. stego signal) and the security of the recovered one. The main advantage of our proposal is the good trade-off among transparency, security and hiding capacity.

**Keywords:** Steganography · Quantization Index Modulation · Orthogonal Variable Spreading Factor · Discrete Wavelet Transform · Security

## 1 Introduction

Data hiding has been adopted as a way to securely embed and transport information within a digital media. There are three main parameters to measure the effectiveness of a data embedding system: transparency, hiding capacity and robustness. Transparency measures the level of imperceptibility of a hidden message within a media; hiding capacity refers to the amount of information that a scheme is able to successfully hide without adding perceptual/statistical distortion; and finally, robustness evaluates the resistance of the hidden data against attacks or modifications of the stego signal. Data hiding systems can be classified into watermarking and steganography; in watermarking the most important parameter is the robustness, while in steganography it is transparency [9].

Although there are a lot of works of image and video steganography, steganographic techniques have also been applied to audio signals [1]. Specifically, concealing text into audio can be used in applications of non-perceptual marking [7], authentication [10] and data transportation [6]. In general, current methods for audio steganography can be divided into three main categories depending


C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 27–35, 2017.
DOI: 10.1007/978-3-319-52277-7_4

on the domain that is used in each one: temporal domain techniques, frequency domain techniques and transform domain techniques [5]. Many temporal domain techniques are based on bit modification methods, like the least significant bit (LSB); they are characterized by an easy implementation offering a high embedding capacity and transparency, but a weak security against intentional attacks [3]. In the second category, frequency domain techniques, the embedding process is applied in the frequency components of the signal, using, in most cases, techniques such as frequency masking [3], spread spectrum, shifted spectrum or Discrete Cosine Transform [4]. In the third category, the Discrete Wavelet Transform (DWT) has been used to embed the secret message into time-frequency coefficients of the audio signal. The most important advantage of the time-frequency domain over the other ones is a better relationship between transparency and hiding capacity [1].

Regarding hiding capacity, different methods have been focused on enhancing hiding capacity without missing transparency, Quantization Index Modulation (QIM) is one of them. This method is widely used nowadays in information hiding systems, because of its good performance in terms of transparency, quality of the recovered data and computational cost [2]. Nevertheless, the security level of the secret message can be a weakness for this kind of schemes. The application of spreading by means of orthogonal codes in the text data before embedding can improve security of the secret message [7].

According to the above, in this paper we propose a scheme for concealing text within audio, by using the QIM method in the wavelet domain to obtain transparent stego signals and by using Orthogonal Variable Spreading Factor (OVSF) codes to increase security of the secret message.

## 2 Proposed Audio Steganography Model

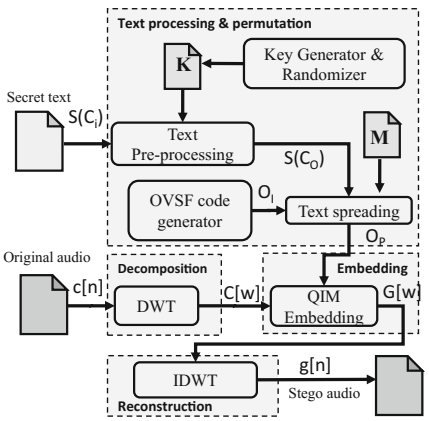
### 2.1 Embedding Module

The embedding module involves five main parts: *text processing and permutation, decomposition, repetition code, embedding,* and *reconstruction* (Fig. 1). The audio file and the secret text to be embedded are the inputs of this module.

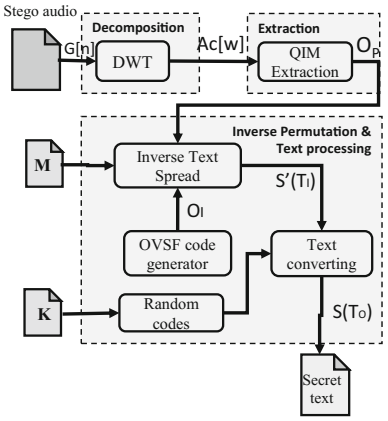
***Text Processing and Permutation:*** The input of this block is the secret text ($S$) and the outputs are the random vector ($K_1$) and the selected orthogonal codes ($O_p$). Firstly, a random vector $K_1$ is generated with the numbers of 1 to 256 in disorderly places. Secondly, the ASCII integer value of every character of $S$ is obtained and kept in the vector $I$. For example, if $S = HOLA$, then ASCII($H$) = 72, ASCII($O$) = 79, ASCII($L$) = 76, ASCII($A$) = 65, and the result is $I$ = [72 79 76 65]. In the third place, with the values (or elements) of $I$, locations of $K_1$ are sought with the following mapping function 1.

$$T_0 = K_1(I) \tag{1}$$

Suppose that $K_1(72) = 15$, $K_1(79) = 5$, $K_1(76) = 246$, $K_1(65) = 129$. Then, the result of this step is $T_o$ = [15 5 246 129]. Fourthly, a 256-bit OVSF code is




(a) Embedding Module.




(b) Recovering Module.

**Fig. 1.** Proposed audio steganography model.

constructed. The result is the matrix $O_i$ which has 256 rows each one of 256 bits. Fifthly, with the numbers contained in $T_o$, the rows of the matrix $O_i$ are selected. The output is the $O_p$ matrix, whose number of rows equals to the number of characters in the secret text. Each row has 256 bits, according to Eq. 2.

$$O_p = \begin{bmatrix} O_i(T_o(1)) \\ O_i(T_o(2)) \\ \dots \\ O_i(T_o(N)) \end{bmatrix} = \begin{bmatrix} O_i(15) \\ O_i(5) \\ O_i(246) \\ O_i(129)) \end{bmatrix} \tag{2}$$

Where $N$ is the total number of characters of the secret text. The right side matrix shows the example where $T_o = [15\ 5\ 246\ 129]$. In Summary, every character of the text is transformed to 256 bits, according to Table 1.

**Table 1.** Example of the result of the text processing and permutation block.

| Text | I | $T_o = K_1(I)$ | $O_p = O_i(T_o)$ |
|---|---|---|---|
| H | 72 | 15 | $O_i(15)$ |
| O | 79 | 5 | $O_i(5)$ |
| L | 76 | 246 | $O_i(246)$ |
| A | 65 | 129 | $O_i(129)$ |

***D*ecomposition:** The aim of this step is to obtain the wavelet coefficients of the audio signal, according to Eq. 3.

$$[A_c, D_c] = DWT(c) \tag{3}$$

The $DWT$ function performs the wavelet decomposition of the signal $c$. The outputs are the approximation coefficients $A_c$ and the detail coefficients $D_c$. Since the approximation coefficients have higher energy than the detail coefficients, they are selected to embed the binary word, $O_p$, repeated $M$ times (see repetition code block).

***Repetition Code:*** To increase the accuracy of the recovered text, the binary word $O_p$ is repeated $M$ times for adding a repetition code to itself.

$$M = \left\lfloor \frac{L}{256 \times N} \right\rfloor \tag{4}$$

Where $L$ is the total number of approximation coefficients, $N$ is the total number of characters of the secret text and $\lfloor . \rfloor$ is the integer part of data. With the current example, suppose that $L = 10240, N = 4$, then $M = 10$. It means, the secret text is repeated ten times within the audio signal. The result of this step is a binary data string of length L, named $O_w$, and organized as follows:

Finally, the *secret key* is constructed with the values of $M$ and $K_1$.

***Embedding:*** Once the text has been converted to $256 \times M$ bits per character, the bit string $O_w$ is embedded into the approximation coefficients by the $QIM$ technique. The process is to quantize every approximation coefficient, according to the bit to be hidden and the quantization step ($\Delta$). With the current example, if there are 10240 approximation coefficients, and $O_w$ has 10240 bits, every coefficient is quantized according to the bit value and $\Delta$, as follows:

$$A'_c(i) = \begin{cases} \Delta \left\lfloor \frac{A_c(i)}{\Delta} \right\rfloor & if \quad O_w = 0 \\ \Delta \left\lfloor \frac{A_c(i)}{\Delta} \right\rfloor + \frac{\Delta}{2} & if \quad O_w = 1 \end{cases} \tag{5}$$

Where $A_c(i)$, $A'_c(i)$ are the approximation coefficient before and after the quantization process, respectively; $O_w(i)$ is the bit to be hidden; and $i$ is the index position.

***Reconstruction:*** The last stage of this module consists of reconstructing the audio signal from $A'_c$ and the original $D_c$.

$$G = DWT^{-1}\left[A'_c, D_c\right] \tag{6}$$

After performing all the above steps, the stego audio $G$ signal is generated and transmitted.

### 2.2 Recovering Module

In the extraction process, $G$ and $K_1$ are the inputs of the recovering module, and the output is the recovered text. This process consists of applying three main steps: *decomposition*, *extraction* and *inversing permutation and text processing* (Fig. 1(b)).

***Decomposition:*** The DWT is applied to the input signal $G$ to obtain the approximation and the detail coefficients of the stego signal ($A_{cs}$ and $D_{cs}$). It uses the same conditions of the wavelet base and number of decomposition levels of the embedding module.

***Extraction:*** Inverse QIM is used for extracting the binary string from the approximation coefficients of the stego signal, through Eq. 7, where $O_{wr}(i)$ is the recovered bit.

$$O_{wr}(i) = \begin{cases} bit = 1 & if \quad \frac{\Delta}{4} < A_{cs}(i) - \Delta \left\lfloor \frac{A_{cs}(i)}{\Delta} \right\rfloor \leq \frac{3\Delta}{4} \\ bit = 0 & if \quad Otherwise \end{cases} \tag{7}$$

***Inverse Permutation and Text Processing:*** Once the inverse QIM has been performed and the system knows all the recovered bits (one per approximation coefficient), this string is divided in 256-bit frames. The total number of characters of the text is obtained through the value of $M$ contained into the *secret key*, as follows: $N = \left\lfloor \frac{L}{M \times 256} \right\rfloor$.

Where $L$ is the total number of approximation coefficients, $M$ is the number of redundancy, N is the number of characters, and $\lfloor \rfloor$ is the integer part of data. For example, if $L = 10240$ and $M = 10$, then $N = 4$.

The string $O_{wr}$ is organized in a similar way of the string $O_w$ (Fig. 2). Then, every $N \times 256$ bits, the recovered bits should be the same due to the repetition code. For example, with the current data, every 1024 bits (i.e. $4 \times 256$) should be the same.

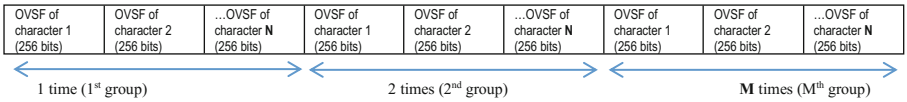



**Fig. 2.** Distribution of the $O_w$ vector.

The following step consists of comparing the first 256-bit string with every row of the matrix $O_i$, until a match is found. The number of the row is kept in the vector $T'_o(1)$. If none of the rows of $O_i$ is equal to the bit string, the bits corresponding to the first character in the $2^{nd}$ group are selected, and a comparison to the rows of $O_i$ is made again. The above procedure is performed until a match between the OVSF code of the first character and a row of the matrix $O_i$ is found. Then, the second 256-bit string is selected and the comparison process begins again. The length of $T'_o$ is equal to the value of N. Subsequently, the values in $T'_o$ vector are sought in the $K_1$ vector (obtained from the *secret key*) and the index of every match is kept in the vector $I'$. In other words, a vector $I'$ that satisfies Eq. 8 is obtained:

$$K_1(I') = T'_0 \tag{8}$$

For example if $T'_0 = [15\ 5\ 246\ 129]$ and $K_1(72) = 15$, $K_1(79) = 5$, $K_1(76) = 246$, $K_1(65) = 129$, then, the vector $I'$ that satisfies Eq. 8 is $I' = [72\ 79\ 76\ 65]$. The last step consists of looking for the corresponding ASCII value of $I'$, the result is $S'$. For the current example, the recovered text is $S' = HOLA$.

## 3 Experimental Results and Discussion

Experimental validation is focused on the measurement of transparency and security analysis of the proposed scheme. Since the proposed method can work with different audio signals regardless sampling frequency, it was tested with 20 music files downloaded from www.freesound.org database and two speech files recorded by the authors. Also the length of the text ranged from 1 to 12 characters. According to preliminary tests, the *Haar* base is selected to decompose the audio signal because of the requirement of perfect reconstruction. This parameter is fixed in both the embedding and recovering modules.

### 3.1 Transparency

Transparency is measured through the following parameters: PSNR (Peak Signal to Noise Ratio), and HER (Histogram Error Ratio). PSNR calculates the correlation between the maximum value of original signal and the noise added by embedding process; HER measures the normalized difference between two histograms. They are calculated as follows:

$$PSNR = 10Log_{10}(\frac{MAX^2}{MSE}) \tag{9}$$

Where MAX is the maximum value of the host signal (C), and MSE is the mean squared error between the stego (G) and the host signal. A higher value of PSNR means less distortion between two signals [2].

$$HER = \frac{\sum_{i=2}^{N}(Ch_i - Gh_i)^2}{\sum_{i=2}^{N}(Ch_i)^2} \tag{10}$$

Where, $Ch$ and $Gh$ are the histogram of the host and stego signals, respectively. The ideal value for HER is 0 [8]. Figure 3 shows the PSNR results for different $\Delta$ values. All approximation coefficients of the first decomposition level were modified. Figure 4 shows the comparison between a zoom of the original audio signal and the stego audio signal in time domain. According to the results, the PSNR is higher than 100 dB, which represents a very low distortion in the host signal and then high transparency. In HER case, values are in the order of $10^{-16}$, which means that data distribution of the stego signal is very similar to the one of the host signal, and again, it works with high transparency.

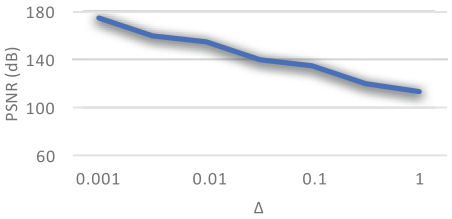
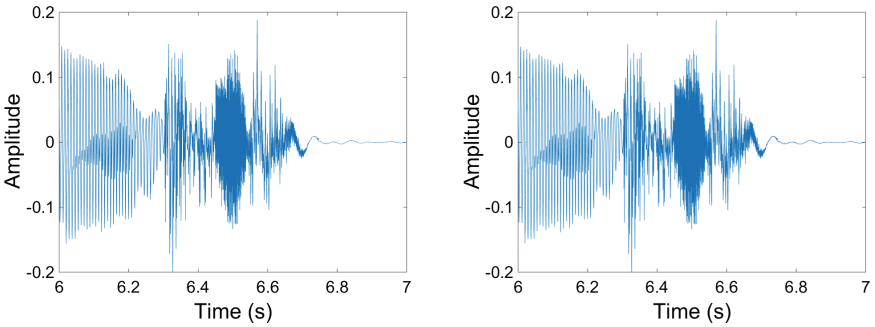



**Fig. 3.** PSNR average of the stego signal for all the tested signals when $\Delta = [0\ 1]$.




(a) Original audio signal. (b) Stego signal.

**Fig. 4.** Comparison of a zoom of original audio signal and stego-signal.

### 3.2 Security

The most important contribution of this proposal is the security of the text through the use of OVSF codes and the vector $K_1$. It is only, if the user has both data, that she/he can reveal the secret content. Since the total number of rows of the matrix $O_i$ (which has the orthogonal codes) is 256, the total number of mapping choices between the input character and its orthogonal code is 256. If the secret text has N characters, each one can be mapped with 1 of the 256 orthogonal codes and then, the total number of available choices is: $\#keys = 256^N$. Where $\#keys$ is the total number of available keys that an intruder must test to discover the secret text. For example, if $N = 4$, an intruder must test around $4.2 \times 10^9$ choices. The higher the value of $N$, the longer is the total number of available keys.

With the purpose to illustrate the security of the proposed scheme, Table 2 is presented. The original vector $K_1$ is slightly modified and used it to recover the text. The result is completely different to the original secret text.

Although a text is always recovered, there is not a way to know if the recovered text matches the original text, since an alphanumeric string will always be obtained and the original text could not have a logic or linguistic sequence.

**Table 2.** Test on extracting secret text with slightly modification of ($K_1$).

| Audio type | Secret text | Text length (characters) | Recovered text |
|---|---|---|---|
| Music | UMNG2016 | 8 | td Í “ ˆ |
| Music | UMNG2016 | 8 | SÆ‘ÊRWi |
| Music | 12Stego34 | 9 | •(2Mõ%o!Q |
| Music | 12Stego34 | 9 | f>&\fi,{Ó |
| Music | 12Stego34 | 9 | 5 − ‘0¶>$D\pm$ |
| Music | M@CB%ok4? | 10 | )Áó×QJp 3/4 Ü |
| Music | <T”EST]> | 10 | B5><éK5 |
| Music | r9e803bd? | 10 | Ÿ“–$^2$Ê0nL,, |
| Music | M@CB%ok4? | 10 | ç ž T“ 1_ |

## 4 Conclusion

A scheme of concealing text into audio (voice or music) has been presented in this paper. The strengths of the proposal are: high transparency of the stego signal and high security of the secret content. The first characteristic obeys the very low distortion introduced by the quantization process (i.e. by the QIM method) in the wavelet domain. A good selection of ($\Delta$) gives stego signals with PSNR higher than 100 dB. The second characteristic is related to the mapping process between every character of the secret text and a 256-bit orthogonal code by the use of OVSF codes. Only the correct key (composed by $K_1$ and $M$) allows obtaining the correct text. On the other hand, a repetition code is used for providing a fortress against possible errors in recovering bits. Then, some errors are tolerated in the detection of the OVSF codes, because the secret text is embedded many times into the approximation coefficients of the host signal. Results of several simulations show perfect recovering of the secret text.

**Acknowledgment.** This work is supported by the “Universidad Militar Nueva Granada-Vicerrectoría de Investigaciones” (grant IMP-ING-2136 of 2016).

## References

1. Ballesteros L, D.M., Moreno A, J.M.: Highly transparent steganography model of speech signals using efficient wavelet masking. Expert Syst. Appl. **39**(10), 9141–9149 (2012)
2. Basu, P.N., Bhowmik, T.: On embedding of text in audio. In: International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 203–206. IEEE (2010)
3. Djebbar, F., Ayad, B., Hamam, H., Abed-Meraim, K.: A view on latest audio steganography techniques. In: International Conference on Innovations in Information Technology, pp. 409–414. IEEE (2011)

4. Famili, Z., Faez, K., Fadavi, A.: A new steganography based on $\chi^2$ technic. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 1062–1069. Springer, Heidelberg (2009). doi:10.1007/978-3-642-10268-4_124
5. Lin, Y., Abdulla, W.H.: Audio Watermark. Springer, Heidelberg (2015)
6. Qadir, M.A., Ahmad, I.: Digital text watermarking: secure content delivery and data hiding in digital documents. IEEE Aerosp. Electron. Syst. Mag. **21**(11), 18–21 (2006)
7. Renza, D., Ballesteros, D.M., Ortiz, H.D.: Text hiding in images based on QIM and OVSF. IEEE Lat. Am. Trans. **14**(3), 1206–1212 (2016)
8. Shahadi, H.I., Jidin, R., Way, W.H.: Lossless audio steganography based on lifting wavelet transform and dynamic stego key. Indian J. Sci. Technol. **7**(3), 323–334 (2014)
9. Zamani, M., Manaf, A., Ahmad, R.B., Jaryani, F., Taherdoost, H., Zeki, A.M.: A secure audio steganography approach. In: International Conference for Internet Technology and Secured Transactions, pp. 1–6. IEEE (2009)
10. Zmudzinski, S., Steinebach, M.: Perception-based audio authentication watermarking in the time-frequency domain. In: Katzenbeisser, S., Sadeghi, A.-R. (eds.) IH 2009. LNCS, vol. 5806, pp. 146–160. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04431-1_11