

ADVERTIMENT. La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX (www.tesisenxarxa.net) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

ADVERTENCIA. La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR (www.tesisenred.net) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

WARNING. On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX (www.tesisenxarxa.net) service has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized neither its spreading and availability from a site foreign to the TDX service. Introducing its content in a window or frame foreign to the TDX service is not authorized (framing). This rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Departament d'Enginyeria Electrònica

METHODS OF COVERT COMMUNICATION OF SPEECH SIGNALS BASED ON A BIO-INSPIRED PRINCIPLE

Thesis submitted in partial fulfillment of the
requirement for the PhD Degree issued by the
Universitat Politècnica de Catalunya, in its
Electronic Engineering Program

Author: Dora Maria Ballesteros Larrotta

Advisor: PhD. Juan Manuel Moreno Aróstegui

May, 2013

To my loves: my son and my husband
for believing in me
for giving me support in difficult times
for helping me every day to keep going
for understanding me when I have been busy
... for these years together with me

Abstract

This work presents two speech hiding methods based on a bio-inspired concept known as the ability of adaptation of speech signals. A cryptographic model uses the adaptation to transform a secret message to a non-sensitive target speech signal, and then, the scrambled speech signal is an intelligible signal. The residual intelligibility is extremely low and it is appropriate to transmit secure speech signals. On the other hand, in a steganographic model, the adapted speech signal is hidden into a host signal by using indirect substitution or direct substitution. In the first case, the scheme is known as Efficient Wavelet Masking (EWM), and in the second case, it is known as improved-EWM (iEWM). While EWM demonstrated to be highly statistical transparent, the second one, iEWM, demonstrated to be highly robust against signal manipulations. Finally, with the purpose to transmit secure speech signals in real-time operation, a hardware-based scheme is proposed.

Key words: speech security, ability of adaptation, steganography, cryptography, similarity, transparency, hiding capacity, robustness, residual intelligibility.

Like the chameleon which adapts to the surrounding environment, changing its color to become "imperceptible" and not be detected by enemies, a good mechanism to hide a speech signal is to adapt it to a non-sensitive speech one.

Ballesteros and Moreno, 2012

Acknowledgments

To my advisor, prof. Juan M. Moreno, for his academic support and orientation in the development of the thesis.

To my sponsors of Colombia: Univerity Militar Nueva Granada (UMNG) and Colciencias, for the financial support.

To my family, for their advice, affection, time and confidence in me.

For all of you,

Thank you.

Index

List of Figures	xii
List of Tables	xv
Acronyms	xviii
1. Introduction	1
<i>1.1. Research topic</i>	2
<i>1.2. Research problem</i>	3
<i>1.3. Macro Hypotheses</i>	5
<i>1.4. Aim and objectives</i>	6
<i>1.5. Methodology</i>	7
<i>1.6. Chapter descriptions</i>	8
2. Speech security: background and survey	10
<i>2.1. Basic concepts of information security</i>	11
2.1.1. Steganography	11
2.1.2. Cryptography	14
2.1.3. Comparison between Steganography and Cryptography	15
<i>2.2. Security</i>	17
2.2.1. Secure Steganography	17
2.2.2. Secure Cryptography	18
<i>2.3. Steganalysis and Cryptanalysis</i>	19

2.3.1. Techniques for Steganalysis	19
2.3.2. Techniques for Cryptanalysis	21
2.4. <i>Methods of Speech Steganography: theory</i>	23
2.4.1. Least Significant Bit (LSB) substitution	23
2.4.2. Frequency Masking (FM)	24
2.4.3. Shift Spectrum Algorithm (SSA)	26
2.4.4. Spread Spectrum (SS)	27
2.4.5. Tone Insertion	28
2.5. <i>Survey of Speech Steganography</i>	29
2.6. <i>Permutation-based speech scrambling systems: theory</i>	32
2.6.1. Time-Segment Permutation	32
2.6.2. Frequency Domain Scrambling	33
2.6.3. Time-Frequency Scrambling	34
2.7. <i>Survey of permutation-based Speech Scrambling</i>	36
2.8. <i>Summary</i>	39
3. Ability of adaptation of speech signals	41
3.1. <i>Introduction</i>	42
3.2. <i>Histogram-based analysis of speech signals</i>	43
3.3. <i>Hypothesis formulation and statements</i>	46
3.4. <i>Experimental validation</i>	55
3.4.1. Between different kinds of sounds	55
3.4.2. Between different language and gender of the speaker	58
3.5. <i>Summary</i>	61

4. Speech scrambling and the ability of adaptation of speech signals	62
4.1. Motivation	63
4.2. The proposed scheme	67
4.3. Experimental validation	71
4.3.1. Relationship between Γ_{nd} and ρ^2	72
4.3.2. Relationship between Γ_{nd} and the ratio of the non-silent time	74
4.3.3. Relationship between HD and the ability of adaptation	75
4.4. Security Analysis	76
4.4.1. Exhaustive key search	76
4.4.2. Cipher-text only attack	77
4.4.3. Statistical attack and perfect secrecy	77
4.5. Summary	78
5. Speech steganography using Efficient Wavelet Masking	79
5.1. Introduction	80
5.2. Efficient Wavelet Masking	82
5.2.1. Embedding module	82
5.2.2. Extraction module	85
5.3. Performance of EWM	87
5.3.1. Statistical analysis	88
5.3.2. Hiding Capacity and other quality parameters	92
5.4. Improved Efficient Wavelet Masking	94
5.4.1. Embedding module	94
5.4.2. Extraction module	97
5.5. Relationship between robustness and transparency of the iEWM	100

5.5.1. Selecting SBH	100
5.5.2. Comparison of the proposed and classical schemes	105
5.6. Summary	110
6. Speech hiding on hardware devices	112
6.1. Motivation	113
6.2. Real-time, Speech-in-speech hiding scheme	115
6.2.1. Embedding module	116
6.2.2. Extraction module	119
6.3. Principle of Perfect Reconstruction (PR)	122
6.4. Hardware design of the speech-in-speech hiding scheme	126
6.4.1. Decomposition and reconstruction	126
6.4.2. Sorting and reverse	130
6.4.3. Delay	133
6.5. Hardware performance	134
6.5.1. Hardware resources	134
6.5.2. Reconstruction error	136
6.5.3. Validation of the entire design	138
6.5.4. Comparing to related works: dwt-idwt blocks	140
6.5.5. Comparing to related works: the entire design	143
6.6. Summary	146
7. Conclusions	147
7.1. General Conclusions	148
7.2. Future work	150

8. Thesis results dissemination	151
8.1. Journals: published papers	152
8.2. Journals: under review	153
References	154

List of Figures

Chapter 2

Figure 2.1. Global scheme of secret key steganography	12
Figure 2.2. The magic triangle of data hiding	14
Figure 2.3. Global scheme of symmetric cryptography	15
Figure 2.4. Illustration of the LSB substitution method	23
Figure 2.5. Shift Spectrum Principle	27
Figure 2.6. Spread Spectrum Principle	27
Figure 2.7. Example of TSP-based speech scrambling	33
Figure 2.8. Example of FDS-based speech scrambling	33
Figure 2.9. Example of TFS-based speech scrambling	34

Chapter 3

Figure 3.1. Speech signals in time domain and frequency domain	43
Figure 3.2. Wavelet coefficients of the speech signals	44
Figure 3.3. Histogram of the non-zero wavelet coefficients	44
Figure 3.4. Speech signal, entire time-scale and zoom of the signal	48
Figure 3.5. Target speech signal, entire time-scale and zoom of the signal	48
Figure 3.6. Sorted coefficients of target signal and speech signal	53
Figure 3.7. Wavelet coefficients of target and adapted-speech signal	53
Figure 3.8. Secret message, target signal and adapted-secret message	54

Chapter 4

Figure 4.1. Flowchart of the scrambling process	70
Figure 4.2. Example of adaptation	68
Figure 4.3. Flowchart of the descrambling process	69
Figure 4.4. Similarity and Normalized displacement	73
Figure 4.5. <i>Ratio</i> and Normalized displacement	75

Chapter 5

Figure 5.1. EWM: flowchart of the embedding module	82
Figure 5.2. EWM: flowchart of the extraction module	85
Figure 5.3. Difference in the temporal steganalysis test	89
Figure 5.4. Difference in the frequency domain steganalysis test	90
Figure 5.5. Difference in the wavelet steganalysis test	91
Figure 5.6. Block diagram of the improved-EWM embedding module	95
Figure 5.7. Block diagram of the improved-EWM extraction module	97
Figure 5.8. Lossy compression test: quality of the recovered secret message	102
Figure 5.9. Resampling test: quality of the recovered secret signal	104

Chapter 6

Figure 6.1. Block diagram of the Embedding module	117
Figure 6.2. Block diagram of the extraction module	119
Figure 6.3. Decomposition and reconstruction: non-polyphase scheme	122
Figure 6.4. General design of the <i>dwt-idwt</i> stages	124
Figure 6.5. Scheme of the <i>dwt</i> block	128
Figure 6.6. Scheme of the <i>idwt</i> block	129
Figure 6.7. Sorting process	131
Figure 6.8. Reverse process with the non-overlapped scheme	132

Figure 6.9. Scheme of the delay block	133
Figure 6.10. Block diagram of the decomposition-reconstruction system	136
Figure 6.11. Simulation of dwt and idwt blocks	137
Figure 6.12. Simulation of the embedding & extraction modules	138
Figure 6.13. Simulation of the speech-in-speech hiding scheme	140
Figure 6.14. Output at the transmitter and at the receiver	143

List of Tables

Chapter 2

Table 2.1. Steganography & encryption	16
Table 2.2. Performance of the speech hiding schemes	31
Table 2.3. Performance of the speech scrambling schemes	15

Chapter 3

Table 3.1. Squared Correlation Coefficient & Ratio: examples of adaptation	56
Table 3.2. Squared Correlation Coefficient & Ratio: summary of the tests	57
Table 3.3. Results by scenario	60

Chapter 4

Table 4.1. Speech signals in time domain and wavelet domain	70
---	----

Chapter 5

Table 5.1. Signals for HC=100%: Input signal & Difference signal	88
Table 5.2. Performance in other selected quality parameters	93
Table 5.3. Lossy compression test: statistical transparency	101
Table 5.4. Resampling test: statistical transparency	103
Table 5.5. Re-quantization test	104
Table 5.6. Performance results without signal manipulation	106
Table 5.7. Performance results: lossy compression attack	107

Table 5.8. Performance results: resampling attack	108
Table 5.9. Performance results: re-quantization attack	109

Chapter 6

Table 6.1. Nomenclature in the speech-in-speech hiding scheme	116
Table 6.2. Resource utilization and longest path delay	135
Table 6.3. Macro statistics of the sorting block	135
Table 6.4. Comparison of multiplierless-based schemes	142
Table 6.5. Quality of the stego signal and the recovered secret message	145

Acronyms

<i>DWT</i>	Discrete Wavelet Transform	<i>FFT</i>	Fast Fourier Transform
<i>IDWT</i>	Inverse Discrete Wavelet Transform	<i>HAS</i>	Human Auditory System
<i>EWM</i>	Efficient Wavelet Masking	<i>SNR</i>	Signal-to-noise ratio
<i>iEWM</i>	improved-EWM	<i>SPCC</i>	Squared Pearson Correlation Coefficient
<i>LSB</i>	Least Significant Bit	<i>TSP</i>	Time-segment permutation
<i>MSB</i>	Most Significant Bit	<i>FDS</i>	Frequency-domain scrambling
<i>FM</i>	Frequency masking	<i>TFS</i>	Time-frequency scrambling
<i>SS</i>	Spread Spectrum	<i>HD</i>	Hamming Distance
<i>SSA</i>	Shift Spectrum Algorithm		

1. Introduction

The aim of this chapter is to illustrate the motivation of the research and the objectives to overcome the problem. It gives the reader an overview of the research and how it is carried out in different phases.

1.1. Research topic

With the growth of internet, the quantity and kind of information which is transmitted increases day by day. Everybody wants to transmit data into secure channels; but despite the levels of security have improved the ways for stealing the information have improved, too. At this point, is it possible to transmit sensitive information through vulnerable channels -as internet- without compromising the secrecy of data? The answer is related to data hiding which involves cryptography, watermarking and steganography.

In cryptography, secret data is transformed according to a key so that they resemble unknown messages. If the encrypted message is intercepted by a non-authorized user, he/she knows that a secret message is being transmitted; however he/she cannot discover the secret message without the knowledge of the secret key. Therefore, the aim of cryptography is to save the secrecy of data.

On the other hand, in watermarking and steganography, the secret messages are hidden into host signals, e.g. images, audio or video. While watermarking is mainly focused on copyright protection, steganography is focused on covert communication. The transmitted signals, watermarked or stego, are legible signals with high similarity to the host signals and the purpose is to not generate suspicions about the existence of the secret message. It means that if the transmitted signal is intercepted by a non-authorized user, he/she does not suspect about the secrecy of the information.

Although the purpose of watermarking and steganography is not the same, they satisfy, with different order of priority, the following characteristics: transparency, hiding capacity (HC) and robustness. Transparency means a high similarity between the transmitted and the host signal, hiding capacity is the quantity of information that is hidden into the host signal, and robustness is the ability to resist signal manipulation

1.2. Research problem

Nowadays, concealment of speech signals is a great interest area for both users and researchers. Since a speech signal contains more information than a single plain-text (e.g. rhythm and gender of the speaker) and it can be viewed as a signature of their owner, the theme of secure speech signals is a topical issue. But, are the current techniques of data hiding able to transmit secure speech signals?

In the case of encryption, most techniques have been focused on encryption of plain-text, however, some methods to encrypt speech signals have been proposed. The classical approaches are based on permutation (in time, frequency or time-frequency domain) in which data are relocated according to a secret key. Some works have used Pseudo-Noise (PN) generators, and others, chaotic sequences. However, the problem to encrypt speech signals with long time-scale has not been overcome. Another group of techniques are based on amplitude scrambling in which the amplitude of the speech signal is distorted so that it resembles a noise signal. The main disadvantage of these schemes is that the secret message is not recovered if the amplitude of the encrypted signal is slightly modified (e.g. by filtering, re-sampling or re-quantization, among others).

Like cryptography, in the case of steganography, most techniques have been proposed to hide plain-text. One of the most known methods is the Least Significant Bit (LSB) substitution in which some bits of the host signal are replaced with the bits of the secret message. LSB substitution allows hiding speech signals into speech signals, but the behavior of the hiding capacity or/and the robustness is the opposite of the transparency, it means, if the transparency of the stego signal increases, then at least one between HC and robustness decreases. Spread Spectrum (SS) and Shift Spectrum Algorithm (SSA) give a higher transparency than LSB substitution, but the hiding

capacity is lower. Therefore, the time-scale of the secret message must be lower than the time-scale of the host speech signal. On the other hand, Frequency Masking (FM) is a method that directly takes advantage of the Human Auditory System (HAS) in which a weak sound is masked by a stronger sound. Although its hiding capacity is higher than in SS and SSA and its robustness is better than in LSB substitution, the masking process is not efficient enough.

1.3. Macro Hypotheses

In order to overcome the limits of the well-known methods of speech hiding, the following macro hypotheses have been used in the current research:

- (i) A permutation-based speech encryption scheme which uses an adaptive mechanism to relocate data is a good enough solution to transmit speech signals.
- (ii) A steganography model with an efficient application of the masking property gives a better trade-off among transparency, hiding capacity and robustness than its predecessors.
- (iii) Both schemes, encryption and steganography, can be based on the same principle of adaptation. In the first case, adaptation can help to scramble the secret message, while in the second case adaptation can help to mask the secret message.
- (iv) In real-time implementation, the adaptive secret key should be obtained from small frames. It allows having a secure output with small latency.

The above macro hypotheses are the basis of the research work. In the rest of the document, the hypotheses are validated.

1.4. Aim and objectives

Once the problem has been detected, the following step is to identify the aim and the objectives of the research work.

The aim of the research work consists in **proposing a novel scheme of speech-in-speech hiding that satisfies the features of security, transparency, robustness and hiding capacity.**

To achieve this aim, three objectives are identified that have a strong relationship with the macro hypotheses, as follows:

- (i) To propose and validate a novel cryptographic scheme of speech signals based on the principle of adaptation of speech signals *.
- (ii) To propose and validate a novel speech-in-speech hiding scheme based on the principle of adaptation which has a good enough trade-off among transparency, hiding capacity and robustness.
- (iii) To propose and validate a novel real-time speech-in-speech hiding scheme with adaptive-key generation.

* It is worth noting that the first specific objective is new in relation to the original proposal of the research work.

1.5. Methodology

According to the macro hypothesis, both cryptography and steganography schemes should be based on an adaptation criterion, and therefore the first step is to propose a hypothesis of adaptation of speech signals. Therefore, the hypothesis must give a response to the question: *is it feasible to adapt a speech signal to a target speech signal?* And if the answer is positive, *which are the requirements of adaptation?*

Once the hypothesis has been proposed, the following step is to validate it through exhaustive tests.

If the results demonstrate that adaptation is feasible, then, the third step is to apply speech adaptation into a scrambling scheme. It includes several tests to validate speech adaptation as a useful key-generator.

The fourth step is to apply adaptation into a steganography scheme. The idea is to use adaptation to generate an effective masking between the secret message and the host signal. The tests validate the transparency, the hiding capacity and the robustness of the stego signal.

Finally, the scheme of speech-in-speech hiding is modified so that it can be used in real-time operation. Nevertheless, the idea of an adaptive-key is preserved.

1.6. Chapter descriptions

The current document encompasses eight chapters. The following seven chapters are summarized as follows.

Chapter 2 shows a background of speech security in terms of steganography and cryptography. Firstly, some definitions of the above techniques are presented, secondly, the most important methods of each one are explained and finally, a survey of works in the area is shown.

Chapter 3 defines the hypothesis of adaptation of speech signals. In this chapter the idea behind the ability of adaptation, the formulation of the hypothesis, the requirements of adaptation, and an algorithm to adapt a speech signal to a target speech signal are presented. At the end of the chapter the ability of adaptation is tested in two ways: vowels to phrases and vice versa, and phrases to phrases in different language or/and gender of the speaker.

Chapter 4 validates the adaptation as an efficient key-generator into a speech scrambling system. Several tests were carried out in order to measure two parameters which are strongly related to the residual intelligibility: the number of displacements (I) and the number of elements which are not coincident in the same positions (HD).

Chapter 5 presents two schemes of speech-in-speech hiding. The first one is known as Efficient Wavelet Masking (EWM) and the second one as improved-EWM (iEMW). Both of them use the ability of adaptation of speech signals to take advantage of the masking property of the HAS. EWM is validated in terms of the statistical transparency while iEWM in terms of the robustness.

Chapter 6 presents a scheme of speech-in-speech hiding on hardware devices. Since the schemes presented in Chapter 4 and 5 are not useful for real-time operation, a

new scheme is proposed. However, it takes advantage of the strengths of its predecessors. The hardware performance and the quality of the recovered secret message are measured.

Chapter 7 presents the conclusions of the current works, in terms of the novelty, strengths and limits.

Finally, in Chapter 8 the publications derived from the research work are listed.

2. Speech security: background and survey

This chapter presents an overview of security techniques applied to speech signals. Firstly, the most important concepts of data hiding and encryption are shown; secondly, a review of the classical schemes is presented.

2.1. Basic concepts of information security

In the area of information security, there are three clearly distinguishable concepts: cryptography, steganography and watermarking. Although they can be used to transmit information in a secure form, the purpose and the techniques are different among them. In the following subsections the main concepts of steganography and cryptography focused on speech signals are explained.

2.1.1. Steganography

It encompasses pure steganography, secret key steganography and public key steganography. Since the current work uses secret key steganography to transmit the secret message, some definitions are selected in order to explain it, as follows:

“In secret key steganography the sender chooses a cover c and embeds the secret message into c using a secret *key* k . If the *key* used in the embedding process is known to the receiver, he can reverse the process and extract the secret message. Anyone who does not know the secret *key* should not be able to obtain evidence of the encoded information”. The cover c and the stego-object can be perceptually similar”. [1]

“Classical steganography concerns itself with ways of embedding a secret message (which might be a copyright mark, a covert communication, or a serial number) in a cover message (such as a video film, an audio recording, or computer code). The embedding is typically parameterized by a *key*; without knowledge of this *key* (or a related one) it is difficult for a third party to detect or remove the embedded material”. [2]

“The embedded data is the message that one wishes to send secretly. It is usually hidden in an innocuous message referred to as a cover-text, or

cover-image or cover-audio as appropriate, producing the stego-text or other stego-object. A stego-key is used to control the hiding process so as to restrict detection and/or recovery of the embedded data to parties who know it (or who know some derived key value)". [3]

"Steganography (from the greek "steganos" – covered) is a term denoting mechanisms for hiding information within a "cover" such that, generally, only an intended recipient will (i) have knowledge of its existence, and (ii) will be able to recover it from within its cover". [4]

According to the above definitions, four agents interact in a steganography system: the secret message, the cover signal, the stego signal and the secret *key*. The stego signal is the output of the system and the others are inputs to the system. The secret message is hidden into the cover signal according to the secret *key* and the result is the stego signal. To recover the secret message, the authorized user must know the stego signal and the secret *key*. Additionally, only the intended recipient should know about the existence of the secret message. It is illustrate in Figure 2.1.

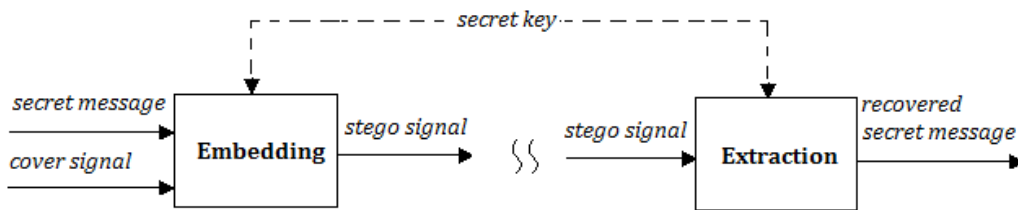


Figure 2.1. Global scheme of secret key steganography

In any steganographic system there are three inter-related characteristics that define its quality: the transparency, the hiding capacity and the robustness of the stego signal. Everyone is explained as follows.

a) *Transparency*: the stego signal is transparent if it does not generate suspicious about the existence of the secret message.

“The stego signal is transparent if an average human subject is unable to distinguish between the host signal and the stego signal”. [5]

b) *Hiding Capacity*: it is related to the amount of information (i.e. quantity of bits) hidden into the host signal. In the case of speech-in-speech hiding, it can be measured in terms of the total number of bits hidden by frame or in terms of the time-scale of the secret message hidden into a normalized time-scale of the host signal. For example, if a speech signal (with sampling frequency, f_s , of 8K Hz, and quantization, q , of 16 bits) hides 4 bits per sample, then $HC=32K$ [bits/s] or $HC=0.25*HC_{max}$ (for $HC_{max} = f_s * q$). On the other hand, if a speech signal of 1-second (with $f_s=8K$ Hz and $q=16$ -bits) hides a secret message of 1-second (with $f_s=8K$ Hz and $q=4$ -bits), then HC is 100% in terms of time-scale. Although in both cases the total number of replaced bits per frame is the same, in the first case it is not guaranteed that the time-scale of the secret message and the host signal is the same. In a similar way, if the quantization of the secret message is 6-bits, HC remains in 100% even if the total number of replaced bits has increased. For this reason, in the specific case of speech-in-speech hiding it is suggested to take into account both kinds of measurements.

c) *Robustness*: it is related to the ability of the stego signal to preserve the secret message even if signal manipulations are applied, such as filtering, lossy compression, re-quantization and re-sampling.

“A system is called robust if the embedded information cannot be altered without making drastic changes to the stego signal”. [1]

Since there is a compromise among the above features, they cannot be optimized at the same time, and therefore if one of them is optimized a reasonable deterioration is obtained in at least one of the others [5]. It is known as the magic triangle.

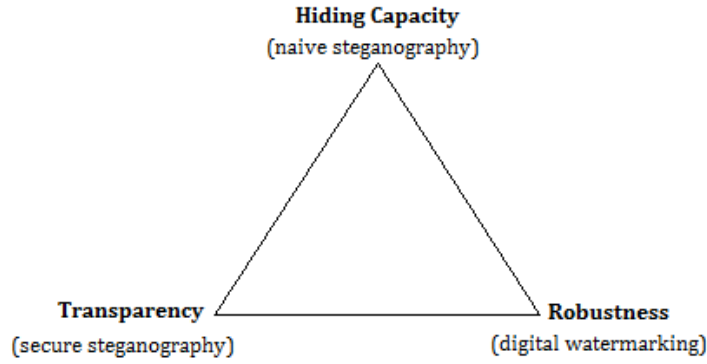


Figure 2.2. The magic triangle of data hiding. Based on [5].

Every feature is related to one method of data hiding. For example, while in digital watermarking the most important feature is the robustness, in the case of secure steganography is the transparency. Nevertheless, all the features should be satisfied in any data hiding system.

2.1.2. Cryptography

In a similar way as for steganography, some definitions of cryptography are presented, as follows:

“An encryption scheme or cryptosystem is a tuple $(P, C, K, \mathcal{E}, D)$ with the following properties: (i) P is a set. It is called the plaintext space. Its elements are called plaintexts. (ii) C is a set. It is called the ciphertext space. Its elements are called ciphertexts. (iii) K is a set. It is called the *key* space. Its elements are called keys. (iv) $\mathcal{E} = \{E_k : k \in K\}$ is a family of functions $E_k : P \rightarrow C$. Its elements are called *encryption functions*. (v) $D = \{D_k : k \in K\}$ is a

family of functions $D_k : C \rightarrow P$. Its elements are called decryption functions.

(vi) For each $e \in K$, there is $d \in K$ such that $D_d(E_e(p)) = p$ for all $p \in P$.

[6]

“Cryptography is the study of methods of sending messages in disguised form so that only the intended recipients can remove the disguise and read the message. The message we want to send is called the *plaintext* and the disguised message is called the *ciphertext*. The *plaintext* and *ciphertext* are written in some alphabet consisting of a certain number N of letters. The term "letter" (or "character") can refer not only to the familiar A-Z, but also to numerals, blanks, punctuation marks, or any other symbols that we allow ourselves to use when writing the messages. The process of converting a *plaintext* to a *ciphertext* is called enciphering or encryption, and the reverse process is called deciphering”. [7]

According to the above definitions, the encryption system has three agents: the plain-text, the cipher-text and the key. Unlike steganography, the plain-text is not hidden, instead of that it is “mapped” according to the key. It is illustrated in Figure 2.3.

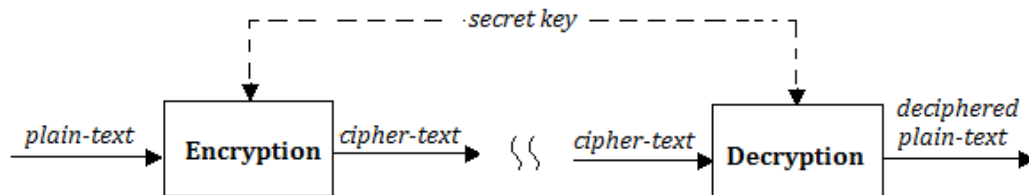


Figure 2.3. Global scheme of symmetric cryptography

2.1.3. Comparison between steganography and cryptography

Summarizing, while cryptography tries to conceal the plain-text of the secret message, the purpose of steganography is to try to conceal the existence of the secret

message. Both methods use a secret key in the embedding/encryption and extraction/decryption processes, however there are important differences between them which are illustrated in Table 2.1.

Table 2.1. Steganography & encryption

Method	Inputs	Output	Process
Steganography	Secret message; cover signal; key	Stego signal	Hiding
Cryptography	Plain-text; key	Cipher-text	Mapping

In the case of speech encryption, both the plain-text and the cipher-text are speech signals, while in the case of speech steganography, both the secret message, the cover (or host) signal and the stego signal are speech signals. It is worth noting that in the classical approach of speech encryption, a second signal (the cover signal) is not used in the process.

2.2. Security

This section presents the definition of security in both steganographic and cryptographic systems.

2.2.1. Secure Steganography

A steganographic system is secure if the following four requirements are satisfied:

“(i) Messages are hidden using a public algorithm and a secret key; the secret key must identify the sender uniquely; (ii) Only a holder of the correct key can detect, extract, and prove the existence of the hidden message. Nobody else should be able to find any statistical evidence of a message's existence; (iii) Even if the enemy knows (or is able to select) the contents of one hidden message, he should have no chance of detecting others; (iv) It is computationally infeasible to detect hidden messages”. [1]

In other words, a secure steganography system must have the following characteristics:

- a) The stego signal must be statistically transparent. It guarantees that the enemy does not detect the presence of the secret message.
- b) Key-generator must create a different key every time. It means that the key-space is long enough with the purpose to increase the effort to discover them.
- c) The secret message is recovered by a unique key. It guarantees that if the non-authorized user works with a wrong key, the secret message is not recovered.

2.2.2. Secure Cryptography

A cryptographic system is secure if the following conditions are satisfied:

“Let $|C| = |K|$ and $\Pr(p) > 0$ for any plaintext p . Our cryptosystem has perfect secrecy if and only if the probability distribution on the key space is the uniform distribution and if for any plaintext p and any ciphertext c there is exactly one key k with $E_k(p) = c$. Therefore, for each ciphertext c there is exactly one key k with $E_k(p) = c$ ”. [8]

In relation to the strength of the system, the authors of [9] present the following idea:

“The strength of crypto is based not on the secrecy of the algorithm, but on the secrecy of the key”. [9]

According to the above statements, it is clear that the most important aspect of a crypto-system is the *key*. In the first definition of security, it was presented that all *keys* must have the same probability and only one *key* must be used to map the plain-text to the cipher-text and vice versa. In other words, if there are N plain-texts and N cipher-texts, the total number of *keys* is exactly equal to N . Therefore, if a wrong *key* is used to decipher the encrypted message, a wrong plain-text must be obtained.

Summarizing, in both steganography and cryptography the *key*-space plays an important role in the security of the system. In both cases every pair of secret-message & stego signal or plain-text & cipher-text must have only one key and therefore if a wrong key is used (in the extraction or deciphering process), a wrong recovered secret message will be obtained.

2.3. Steganalysis and cryptanalysis

Steganalysis is the process of discovering the existence of secret messages while the purpose of cryptanalysis is to reveal the secret message.

Some definitions are presented, as follows:

“Steganalysis is the set of techniques that aim to distinguish between cover-signals and stego-signals. A passive warden simply examines the signal and tries to determine if it potentially contains a hidden message. If it appears that it does, then the signal is stopped; otherwise, it will go through. An active warden, on the other hand, can alter signals intentionally, even though there may not be any trace of a hidden message, in order to foil any secret communication that nevertheless can be occurring”. [10]

“Cryptanalysis: 1) the steps, operations, and processes performed to convert encrypted text into plain text without knowledge of the key employed in the encryption. 2) The study of encrypted texts. 3) An analysis of a cryptosystem to obtain sensitive information legally or clandestinely when applicable key is not available. Note: Cryptanalysis is usually performed with the aid of computer hardware and software”. [11]

Some of these techniques are presented in the following subsections.

2.3.1. Techniques for Steganalysis

The techniques of steganalysis are related to the nature of the stego-object; for example, techniques for stego-image detection are different from techniques for speech-stego detection. In the first case the characteristics of the Human Visual System

(HVS) are taken into account, while in the second one the Human Auditory System (HAS).

Although there is a wide variety of techniques to identify image-stego signals there are not too many techniques able to identify speech-stego signals. However, in the recent years some techniques have been proposed. Most of them are based on the statistical features of the speech signals -in time domain, frequency domain of time-frequency domain- and these features are the input of the classifier.

Steganalysis of speech signals in time domain: this kind of technique uses the statistics of the speech signal, in time domain, to identify the stego signals. In [12], the logarithm of the speech signal is applied before of calculating its statistics. In [13], the amplitude co-occurrence matrix is used as input for the classification system. The authors found that the detection rate is better in the logarithmic version of the speech signals instead of the original speech signal.

Steganalysis in frequency domain: the statistics of the spectrum of the speech signal are taken into account. The spectrum of the 2nd to 4th order derivate of the speech signal is affected when data have been embedded and this is more appreciable in higher frequencies [14], [15]. Then, it is expected that the statistics of stego signals are significantly different to the statistics of cover signals, and therefore the stego signal can be identified. Other authors use the cepstrum of the speech signal (instead of its spectrum) to identify stego signals [16].

Steganalysis in time-frequency domain: the statistics of the wavelet coefficients of the speech signal are used as features for the classifier [17].

Other type of technique uses some metrics of the speech signal (e.g. signal to noise ratio and Log-likelihood ratio [18] or fraction of false neighbors [19]) as features to detect the stego signal.

2.3.2. Techniques for Cryptanalysis

Since the purpose of cryptanalysis is to reveal the plain-text without the knowledge of the key, the techniques are classified according to the information that the attacker knows. The following cases are explained for the permutation-based speech scramblers.

Know plain-text attack: in this case, both a plain-text and a cipher-text are known by the attacker. The highest amplitude of the plain-text and the highest amplitude of the cipher-text are found and then the relative places among them gives one value of the *key*. The process continues with the rest of data (samples or spectral coefficients) and then all the relative positions give the entire *key* [20]. With the *key*, a new cipher-text is deciphered and then the corresponding plain-text is obtained.

Cipher-text only attack on a fixed permutation system: since in a real world the plain-text is not known by the attacker, the process consists on revealing the key according to a criterion of optimization based on the envelope of the spectrum. It works with a smooth spectrum as the reference spectrum and then the objective is to relocate the spectral components of the cipher signal to minimize the error between the spectra [21]. In this attack it is not necessary to relocate all data in right places, only a sufficient number to recover intelligible speech [20].

Cipher-text only attack on a varying permutation system: when the system uses different keys, the problem to decipher the plain-text without the knowledge of the key requires a higher effort. However, the complexity can be reduced if only the

bandwidth of 300 to 500 Hz is used in the attack. In this range the speech signal is still intelligible and the advantage is that the total number of coefficients to relocate is significantly lower than in the entire spectrum. Once the spectrum has been separated into blocks, the process is the same as in the previous attack [\[20\]](#), [\[21\]](#) .

When the key does not map the plain-text to the cipher-text in a relation of one to one, for example in the case of codeword permutations, the plain-text can be deciphered if the number of blocks is small enough. The relocation process is carried out by an optimization criterion (e.g. cepstral distance) and the use of neural networks and genetic algorithms [\[22\]](#).

2.4. Methods of Speech Steganography: Theory

This section explains some of the most important methods of steganography on speech signals. It encompasses LSB substitution, Frequency Masking, Shift Spectrum Algorithm, Spread Spectrum and Tone Insertion.

2.4.1. Least Significant Bit (LSB) substitution

One of the most popular schemes in steganography is LSB substitution because it is a very simple and general method to hide data. It has been used in images, video and audio. The objective is to replace some of the LSBs of the hosts signal with the bits of the secret message.

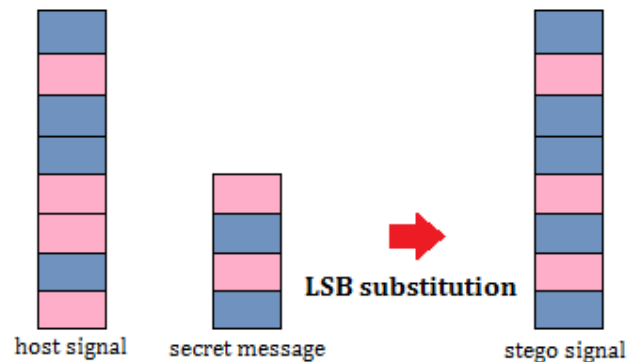


Figure 2.4. Illustration of the LSB substitution method.

Figure 2.4 shows an example of the LSB substitution method in which the host signal is 8-bits, secret message is 4-bits, and the 4-LSBs of the host signal are replaced with the secret message. Secret message can be e.g. the ASCII representation of characters, the binary representation of a speech signal or the representation of a binary image.

This method can be carried out in time domain (samples), frequency domain (e.g. spectral coefficients, cepstral coefficients) or time-frequency domain (e.g. wavelet coefficients).

The trade-off among transparency, hiding capacity and robustness is easily analyzed in this technique. If the total number of replaced bits by sample (or coefficient) increases, HC increases and transparency decreases. But if only 1-LSB is replaced, transparency increases and HC decreases. The robustness depends on the position of the replaced bit (or bits). For example if the host sample is 10110011_b and the fifth-LSB-place is modified with '0', the stego sample is 10100011_b ; but if the first-LSB-place is modified (with '0') the stego sample is 10110010_b . Although in both cases $HC=1\text{bit/sample}$, the robustness of the first case is better than in the second case, or in other words, the stego signal of the first case can tolerate signal manipulations (that slightly modify the value of the sample) while the second cannot. On the other hand, the transparency of the second case is better than in the first case. Since in steganography the most important feature is the transparency, the total number of replaced bits varies according to the desirable transparency. Some works have revealed that the transparency in wavelet domain is higher than in time domain for the same number of replaced LSBs [23], [24].

In the case of speech-in-speech hiding the host signal and secret message are speech signals. It is a common practice to attenuate the secret message in order to decrease its number of bits, and therefore, e.g. a secret message of 8-bits is hidden into a host signal of 16-bits. In this case HC is 100% in terms of the time scale (since both signals have the same time-scale) or HC is 8-bits/sample (or 50% of the total number of bits).

2.4.2. Frequency Masking (FM)

This method takes advantage of the masking property of the HAS in which one sound may be masked by another if one produces high levels while the other remains

faint [25]. A high enough threshold between the level of the high sound and the level of the faint sound produces a masking phenomenon and the faint sound would not be perceptible.

The dynamic range of the secret message must be four times lower than the dynamic range of the host signal (attenuation is applied if needed). Then, both signals are transformed to frequency domain (e.g. Fast Fourier Transform –FFT-). Once the secret's coefficients and the host's coefficients have been obtained, a search algorithm is used, as follows:

- i) The first secret's coefficient is compared to every one of the host's coefficients and it stops when the masking criteria is satisfied, i.e., when the amplitude of the host's coefficient is at least 4-times higher than the amplitude of the secret's coefficient.
- ii) The secret's coefficient is hidden into the LSBs of the "selected" host's coefficient. The output is the stego's coefficient.
- iii) Steps (i) and (ii) are repeated until the last secret's coefficient has been compared to at least one host's coefficient.

Finally, the stego's coefficients are transformed to time domain (e.g. IFFT) and the stego signal is obtained.

The search procedure is illustrated with an example. Suppose that the host's coefficients are = [5 12 8 16 10 12 17 12] and the secret's coefficients are = [2 1 2 3 4 3 3 4], then the first secret's coefficient (2) is compared to the first host's coefficient (5) and the searching process continues because the masking criterion is not satisfied (since $5 < 2 * 4$). Then, 2 is compared to 12 (the second host's coefficient) and this is selected because the masking criterion is satisfied (since $12 \geq 2 * 4$). The process is repeated until the eighth secret's coefficient (4) is compared to at least one of the host's

coefficients. In some cases, for the last places of the secret's coefficients it is not possible to find a host's coefficient that satisfies the masking criterion and then this technique does not guarantee that all of the secret's coefficients will be hidden (unlike the LSB technique). Therefore, HC in FM is equal or lower than in the LSB scheme.

2.4.3. Shift Spectrum Algorithm (SSA)

This technique is used in the frequency domain (or time-frequency domain) of the host signal and the secret message. The spectrum of the secret message is shifted to the highest subband of the spectrum of the host signal [26]. The perceptual transparency is based on the fact that the HAS is less sensible to the highest frequencies; therefore, the secret message can be hidden without suspicion of its existence. For example, if the secret message has a bandwidth of 4K Hz and the host signal has a bandwidth of 20K Hz, the range of 16-20K Hz of the host signal can be replaced with the secret message. It is possible to take two options: the spectrums of the signals are overlapped or the LSBs of the host's spectrum are replaced with the bits of the secret's spectrum.

Although the HC of SSA is lower than in the cases of LSB and FM, the transparency is better because only one portion of the host's signal has been modified and it represents the less sensible range of the HAS.

Figure 2.5 illustrates the SSA method in which the spectrum of the secret message is lower than the spectrum of the host signal and therefore the high frequencies of the host signal are replaced with the spectrum of the secret message.

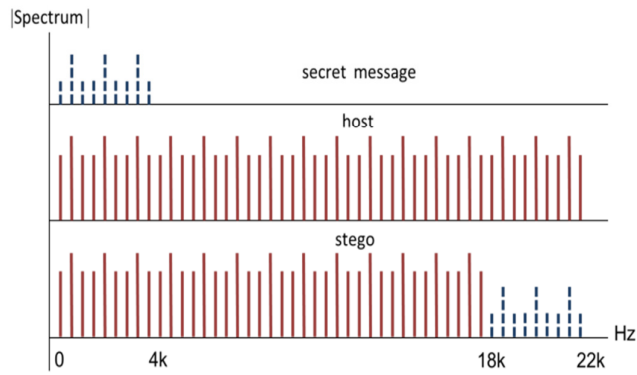


Figure 2.5. Shift Spectrum Principle. [56].

2.4.4. Spread Spectrum (SS)

In the classical approach, the secret message is spread out by a constant called the chip rate and then modulated with a pseudorandom signal [27]. The disadvantage is that the computational cost for implementing the scheme is high. A solution consists on spreading the secret's spectrum along host's spectrum [26]. Because the bandwidth of the host signal is higher, the number of spectral coefficients is higher too; therefore, the secret's coefficients are relocated in interspersed positions of the host's spectrum.

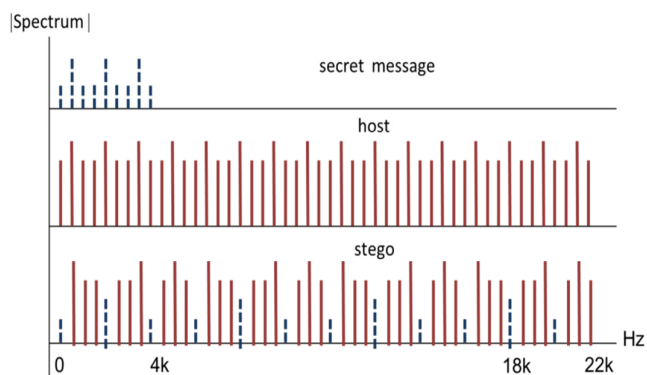


Figure 2.6. Spread Spectrum Principle. [56].

Figure 2.5 illustrates the principle proposed by the authors of [26]. In the current example, every four coefficients of the host's spectrum are replaced with one coefficient of the secret's spectrum.

Like SSA, the HC of SS is lower than the obtained from LSB or FM. Since the low frequencies of the host signal are modified, it is expected that the transparency is lower than in SSA.

2.4.5. Tone Insertion

It takes advantage in an indirect form of the masking property of the HAS. Two tones at frequencies f_0 and f_1 are used for embedding one bit. The value of the bit ('0' or '1') depends of the ratio of the power between the tones [28], [29]. For example, if the power of f_1 is 0.1% of the power of f_0 , then a bit with value of '0' will be embedded. Similarly, if the power of f_0 is 0.1% of the power of f_1 , then a bit with value of '1' will be embedded. This process is repeated in small non-overlapped frames. The advantage of this method is that the stego signal is robust against signal manipulations, but the disadvantage is the very low hiding capacity.

2.5. Survey of Speech Steganography

This section presents the state of the art of speech steganography and the comparison among the schemes found in literature.

Tone Insertion: the scheme proposed by Gopalan uses two frequencies to insert one bit according to the ratio between the powers of the frequencies. It is carried out in the frequency domain [30] or in the cepstral domain [31], [32]. In both cases, the HC is lower than 256 bits into a signal of 256.000 bits.

LSB substitution: Cvejic and Sepanem used the wavelet domain to embed bits. They found that the highest number of replaced bits without significant degradation of the quality of the signal is 8-LSBs. In this case, the HC is up to 352.800 bits into a signal of 705.600 bits [33]. The transparency and robustness depend on the number of replaced bits, the higher the number the lower the transparency but the higher the robustness. However, since 8-bits only represent the ~0.4% of the amplitude of the signal, the stego signal is not robust enough against signal manipulations. Shirali and Manzuri proposed a scheme in which the number of LSBs depends on the amplitude of the wavelet coefficient. The higher the amplitude, the higher is the number of replaced LSBs. With the purpose of increasing the transparency of the signal, the silent regions are not used to embed data. In average, the highest HC is 3 bits per coefficient [34].

Frequency masking (FM): in the proposal of Djebbar et al, the speech signal is divided in frames of 4ms and its spectrum is calculated [35]. The secret message is hidden in the first 28 coefficients (of the 64 by frame). Since FM depends on the masking criterion, the HC is not fixed and depends on the host signal. The HC is up to 14300 bits into a signal of 256000 bits. With the purpose of increasing the robustness of the stego signal, they modified the scheme and the bits are not replaced from the 1-

LSB, the first position is selected. If the first replaced bit begins in the 4-LSB position, the hiding capacity significantly decreases (HC~3kbps) [36].

Shift Spectrum Algorithm (SSA): Djebbar et al proposed a scheme in which the secret message is embedded in the 8-LSBs of the coefficients of the host signal in the range of 7K Hz to 8K Hz [37]. Since the HAS is low sensible in the selected range, the perceptual transparency of the stego signal is high, nevertheless, the secret message is lost if the stego signals is filtered with a high-pass filter. The HC is up to 8000 bits into a signal of 256000 bits. Rabie and Guerchi proposed a speech-in-speech hiding scheme based on SSA and Code-Excited Linear Prediction (CELP) [38]. The 32-CELP parameters of the secret message are hidden into the high frequency of the host signal (the last 32 coefficients of the 80 coefficients by frame). Like the scheme of [37], the weakness is that the information related to the secret message is lost if the stego signal is filtered with a high pass filter. Finally, the scheme proposed by Dimitry et al shifts the spectrum of the voice signal to the range of 18-22 K Hz of the host signal. The stego signal has the spectrum of the host signal up to 18 KHz and the spectrum of the secret message from 18 KHz to 22 KHz. The hiding capacity is 4 of every 22 spectral coefficients [39].

Spread Spectrum (SS): in [39] is proposed a scheme of SS in which an attenuated speech signal is hidden into an audio signal. The spectral coefficients of the attenuated secret message are interlaced with the spectral coefficients of the host signal every n places. According to their results, the stego signal resists MP3 compression with a bit rate of 320 kbps but does not with bit rate of 256 kbps. If 8-LSBs are replaced every four spectral coefficients, the ratio of the capacity is 8 of every 64 bits.

The comparison among transparency, hiding capacity and robustness of the above techniques is shown in Table 2.2.

Table 2.2. Performance of the speech hiding schemes.

Method	Scheme	Transparency	Hiding Capacity	Robustness
Tone Insertion	[30-32]	High	HC < 0.1%	High
LSB	[33]	Middle to high	HC ≤ 50%	Middle to low
	[34]	High	HC < 20%	Middle
FM	[35]	High	HC < 6%	Middle to high
	[36]	High	HC < 1.5%	High
SSA	[37]	High	HC < 4%	Low
	[38]	High	HC ≤ 40%	Low
	[39]	High	HC ≤ 20%	Low
SS	[39]	High	HC ≤ 12.5%	Middle to low

According to the results shown in Table 2.2, it is worth noting that when the HC increases, the transparency or/and the robustness decreases. The best scheme in terms of HC is the worst scheme in terms of transparency and the best scheme in terms of robustness is the worst scheme in terms of HC. Until now, none of the known schemes has a good enough trade-off among transparency, hiding capacity and robustness. Since the speech signal has a higher number of bits than a plain-text, HC plays an important role in the design of a speech-in-speech hiding scheme.

2.6. Permutation-based speech scrambling systems: theory

This section explains some of the most important methods of speech scrambling. Although the methods are divided in permutation-based and amplitude scrambling, only the permutation-based schemes are explained because the proposed design (presented in Chapter 4) is a special case of permutation-based speech scrambling. It encompasses: Time-Segment Permutation (TSP), Frequency-Domain Scrambling (FDS) and Time-Frequency Scrambling (TFS). Everyone is briefly explained in the following subsections.

2.6.1. Time-Segment Permutation

TSP is one of the oldest and simplest techniques of speech scrambling. The speech signal is divided in small blocks (typically 16 to 32 ms) and then the samples are relocated according to a *key*. Without loss of generality, there are M blocks each one with L samples. The samples are permuted into the block and each block can have (or not) a different *key*.

The weaknesses of this method are listed as follows:

- a) The residual intelligibility is not low enough: it depends on the size of the key.
- b) The key space is not long enough: e.g. a block with L samples has up to $L!$ combinations, but only a small percentage ($\sim 0.1\%$ [40]) is usable. For example, a *key* related to single delay or inversion is not usable.
- c) The bandwidth of the scrambled speech signal can be higher than of the original speech signal.
- d) A trained listener can discover the original speech signal.

Figure 2.7 illustrates an example of TSP scheme in which the place within the block is permuted.

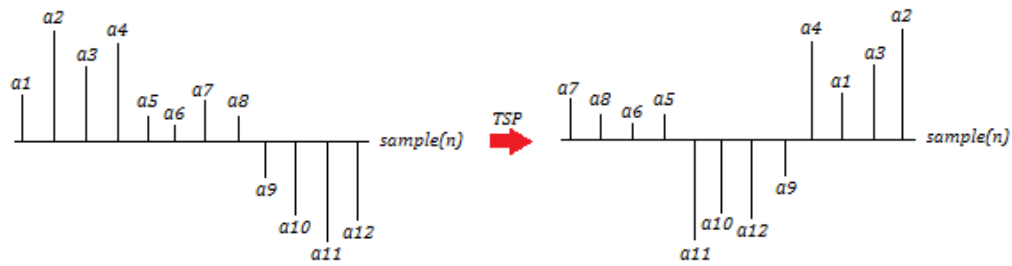


Figure 2.7. Example of TSP-based speech scrambling

In the above example every block has twelve samples which are relocated according to a *key* and then the number of possible combinations by block is 12!

2.6.2. Frequency-Domain Scrambling

Unlike the TSP scheme, the permutation process is carried out in the frequency domain. The speech signal is separated in subbands and then the sub-bands are permuted (it is known as band-splitting). The higher the number of subbands, the higher is the number of possible combinations. If there are P subbands, the total number of possible combinations is $P!$ Figure 2.8 illustrates an example of the scheme of FDS.

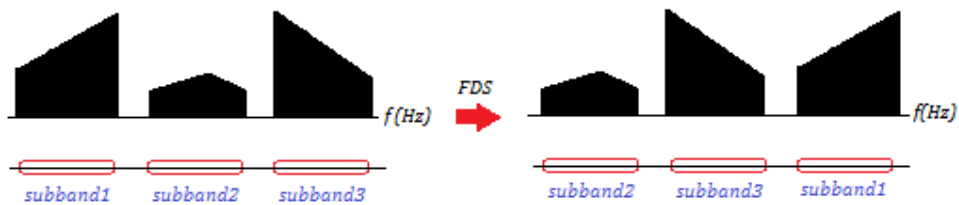


Figure 2.8. Example of FDS-based speech scrambling

The advantages of FSD are:

- a) The bandwidth of the speech signal is preserved.

- b) If band-splitting is combined with inversion, the total number of possible combinations increases up to $P!2^P$. It is worth noting that unlike TSP, in FDS the inversion is a usable permutation and then the percentage of effective keys can increase (e.g. up to 5% [40]).

Despite the residual intelligibility is better than in TSP it is not low enough.

2.6.3. Time-Frequency Scrambling

This technique combines TSP and FDS. The speech signal is split in P subbands and every subband is divided in M blocks of length L . The permutation process is made inter blocks and inter coefficients. The total number of possible combinations by subband is $L! * M!$ and therefore the higher the values of M and L , the lower is the residual intelligibility. However, since M and L are small numbers, the key-space is not long enough. Despite of this, TFS overcomes the problem of residual intelligibility of its predecessors, and like FDS, it preserves the bandwidth of the speech signal [41].

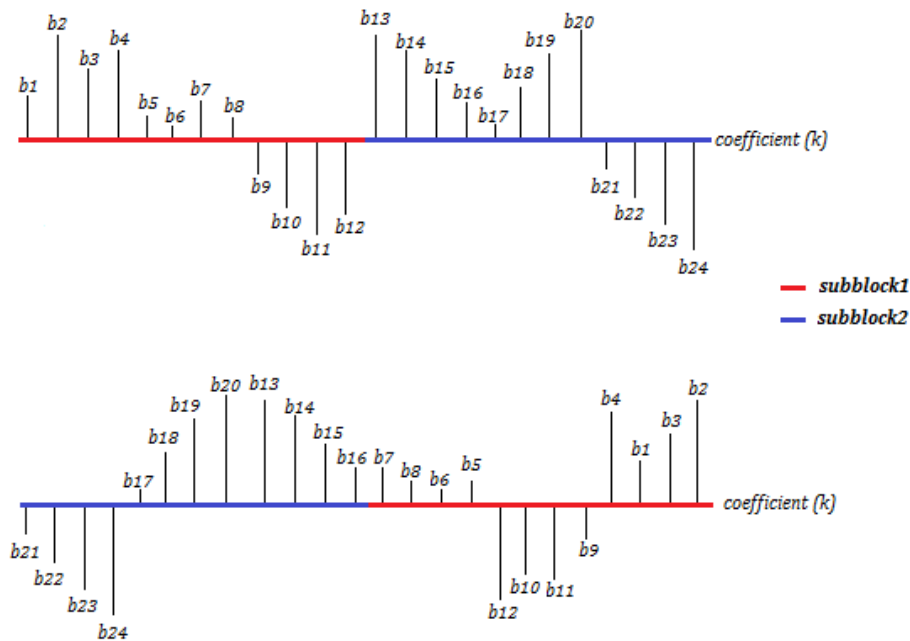


Figure 2.9. Example of TFS-based speech scrambling

Figure 2.9 illustrates an example of TFS in wavelet domain. The permutation process is carried out inter blocks of the same sub-band according to a key. It is worth noting that the amplitude of the coefficients is not “destroyed”, but in fact the plaintext of the speech signal is. In the current example, the total number of subbands is 2 and the number of coefficients per subband is 12. Therefore, the total number of possible combinations is $2! \cdot 12!$

2.7. Survey of permutation-based speech scrambling

The purpose of this section is to give a survey about the schemes of speech scrambling. At the end, they are compared in order to illustrate the strengths and weaknesses.

Time Segment Permutation, TSP: one of the oldest works of speech scrambling is the one developed by Philips, Lee and Thomas [42]. They analyzed the relationship between the highest level of displacement of the samples and the residual intelligibility of the signal. They found that if the samples are placed in reverse order, the residual intelligibility is high; it means that the secret message can be revealed. However, if the samples are placed in positions close to the original ones, the residual intelligibility is high, too, and therefore the best case is an intermediate value. If the size of the frame is 15 ($N=15$), the proposed level of normalized displacement is 0.5.

Frequency Domain Scrambling, FDS: Matsunaga et al proposed a scheme in which the speech signal is compressed and then it is transformed to frequency domain by using the FFT [43]. Once the spectral coefficients have been obtained, they are permuted and a dummy spectrum is added; finally the IFFT is calculated. This process is carried out in frames of 32 ms and the number of coefficients to be permuted is 83. The descrambled speech signal is similar to the original speech one. Woo and Leung used the 256 spectral coefficients to scramble a speech signal with all the coefficients placed in different position than their original ones (derangement) [44]. They found that the residual intelligibility is very low.

TFS and FDS: the scheme proposed by Mosa, Messiha and Zahran uses chaotic maps for permuting the speech signal in time domain [45]. The chaotic keys have a size up to several thousands and therefore the effort to discover them is very high. Together with that, once the samples have been permuted, the output is divided in

small blocks and their spectrums are relocated. It guarantees that the residual intelligibility is very low. According to their results, the secret message is recovered even if additive noise is mixed to the scrambled speech signal.

Time-Frequency Scrambling, TFS: the scheme proposed by Fulong, Jun and Yumin uses the multi-level Wavelet Transform to scramble the speech signal [46]. The speech signal is transformed to time-frequency domain by the DWT and then the wavelet subbands are permuted. The reconstruction of the permuted sub-bands is the scrambled speech signal. In a similar way, Sadkhan, Abdulmuhsen, Al-Tahan proposed a scheme in which the speech signal is scrambled in wavelet domain [47]. In this case, the speech signal is divided in blocks of 16ms and its wavelet coefficients are permuted (128 coefficients by frame). Then, the blocks of the permuted wavelet coefficients are concatenated and the inverse wavelet transform is applied. According to their results, the recovered speech signal is legible even if the scrambled speech signal has been manipulated with additive noise.

The above proposals are compared in terms of residual intelligibility, quality of the descrambled speech signal (recovered secret message) and robustness against signal manipulations. The comparison is shown in Table 2.3.

Table 2.3. Performance of the speech scrambling schemes.

Method	Scheme	Security (key size)	Residual intelligibility	Resistance against attacks
TSP	[42]	Key-space = 15!	Middle	Low
FDS	[43]	Key-space = 83!	Middle	Low
	[44]	Key-space = 256!	Low	Middle
TSP + FDS	[45]	Key-space ₁ ≥ 1000! Key-space ₂ is NP*	Very low	High
TFS	[47]	Key-space = 128!	Low	Middle

*NP = not provided

According to the results, the best scenario is the proposed by the authors of [45] because they used a long key. It is worth noting that the chaotic key generator can be a better solution than the classical pseudo-noise generator of the rest of the approaches. However, it is expected that the residual intelligibility of TFS is better than the obtained by TSP+FDS if the length of the key space is high enough.

2.8. Summary

According to the state of the art of speech steganography and speech scrambling schemes, the following ideas summarize the chapter:

- a) The main difference between a stego signal and a scrambled signal is that the first one seems to contain non-sensitive information, and the second one looks like a manipulated signal. Therefore, an attacker employs his effort in trying to reveal the secret message of the scrambled signal but does not in the stego signal. By using traditional approaches, if the transparency of the stego signal is high enough, it can be a more secure way to transmit speech signals.
- b) The most important feature in a steganography system is the transparency followed by the hiding capacity and the robustness. In the specific case of speech-in-speech hiding, the HC is significantly higher than in the case of text-in-speech hiding. Although the robustness is not the most important feature in the design, it is desirable that the stego signal can resist signal manipulations like lossy compression, filtering, re-quantization and additive noise.
- c) In the proposals found in literature for speech hiding, most of them satisfy one or two of the features, but none of them has a good enough trade-off among transparency, hiding capacity and robustness.
- d) In the case of speech scrambling, the most important aspects to take into account are the residual intelligibility of the speech signal and the size of the key. The lower the residual intelligibility and the higher the size of the key, the better is the scrambling scheme.

- e) The schemes based on TFS have a lower residual intelligibility than that obtained in TSP or FDS because both the places and frequency of the sounds are modified.
- f) Although most of the classical approaches use a PN sequence to relocate the samples or coefficients, it is not an efficient key generator because the length of the key is not long enough. Alternative solutions can be based on chaotic sequences.

3. Ability of adaptation of Speech Signals

The hypothesis of adaptation of speech signals is presented in this chapter. This mechanism is the core of the two proposals: a permutation-based speech scrambling scheme and a speech-in-speech hiding scheme.

3.1. Introduction

In the subject of digital speech processing, many techniques have been proposed with the aim to enhance the quality of the signal (e.g. [48]-[50]), to detect and stand out characteristics (e.g. [51]-[53]) and to classify sounds (e.g. [54]-[55]), among others. These techniques can be used in time domain, frequency domain or time-frequency domain. A speech signal can be manipulated so that it sounds with different tone and for example, a voice signal from a female-adult speaker can be transformed so that it sounds like a voice signal from a female-child speaker. But until now, it has never been proposed a technique able to modify the plain-text (and gender, rhythm and language) of the speech signal so that it resembles a target speech signal. This concept is the core of the current thesis and it is known as the ability of adaptation of speech signals.

This chapter is divided in three parts. In the first one, the idea behind the ability of adaptation of speech signals is explained. The relationship between different speech signals is analyzed in terms of their histogram and kurtosis. In the second one, the hypothesis of adaptation of speech signals is introduced and the conditions under which the hypothesis is true. A deterministic method to adapt an original speech signal to a target speech signal is also explained in the second part. Finally, the hypothesis of adaptation is validated in two ways: firstly in terms of the type of sounds (vowels and words) and secondly in terms of the language and the gender of the speaker.

3.2. Histogram-based analysis of speech signals

Speech signals can be considered as a signature of its owner because both the rhythm and tone are special characteristics that vary among people. For example, if the same plain-text is pronounced by two people, the time representation of their voices can be similar but their frequency representations are not. It is true even if the gender (and age) of the speaker is the same. Additionally, if the plain-text is modified, both the time and frequency representations of the speech signals will be completely different.

It can be easily illustrated with an example. Suppose there are two speech signals with different plain-text, for example speech_1 with the plain-text “good morning everybody” and speech_2 with the plain-text “see you the next week”. Both signals are from a female speaker, with $f_s=8\text{KHz}$ and time-scale=2s. Figure 3.1 shows the signals.

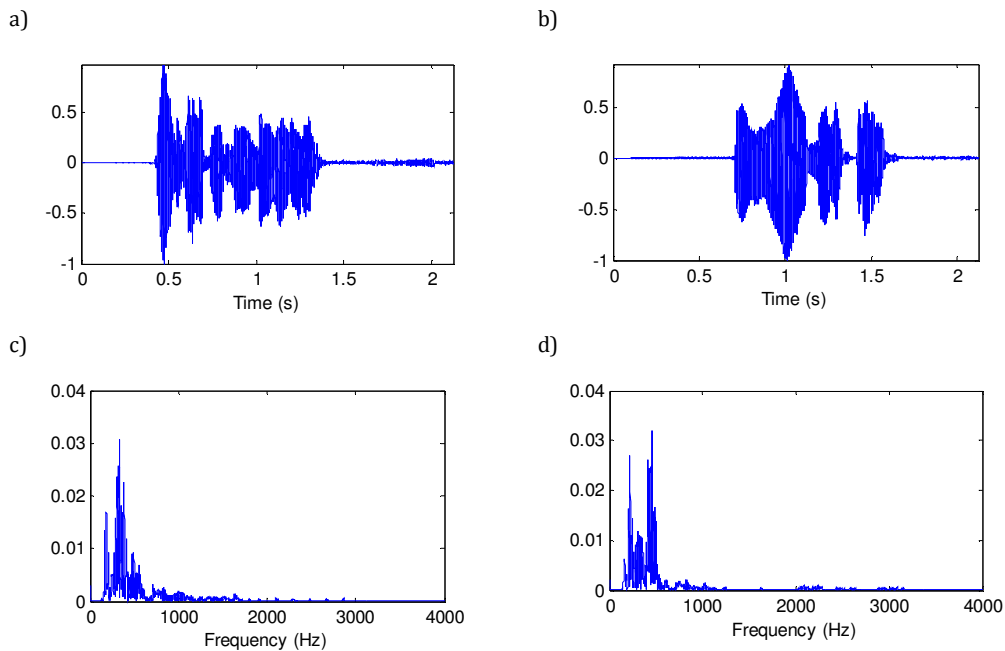


Figure 3.1. Time domain: a) speech_1 ; b) speech_2 . Frequency domain: c) speech_1 ; d) speech_2

As it is expected, both time and frequency representations are different in each case. If the signals are represented in time-frequency domain by using the Discrete

Wavelet Transform, their wavelet coefficients are different, too. Figure 3.2 shows the 1D-arrays of the wavelet coefficients of the speech signals.

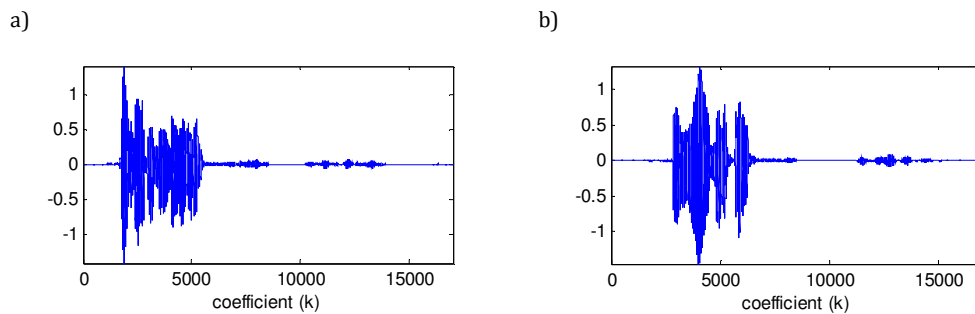


Figure 3.2. Wavelet coefficients: a) speech₁; b) speech₂

Now, two non-zero arrays are made from the non-zero wavelet coefficients of speech₁ and speech₂. Since there are a lot coefficients with magnitude close to zero, a threshold is set, th , which classifies the zero or the non-zero wavelet coefficients. If the magnitude of the wavelet coefficient is lower than th , then the thresholded coefficient is set to zero, but if this is higher than (or equal to) th , the amplitude of the coefficient is preserved. Once the two non-zero arrays have been obtained, their histograms are calculated. Figure 3.3 shows the histograms of the non-zero wavelet coefficients of the two speech signals.

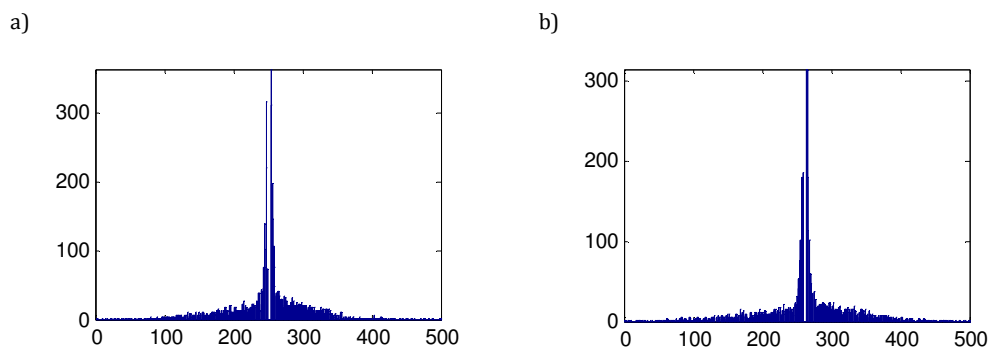


Figure 3.3. Histogram of the non-zero wavelet coefficients: a) from speech₁ signal;
b) from speech₂ signal.

According to Figure 3.3, it is noticed that the histograms have similar shape. Since the *kurtosis* reflects the shape of a distribution, it is expected that the kurtosis of the above histogram is similar. The *kurtosis* is obtained as follows:

$$k = \frac{\sum_{i=1}^N (w_i - \mu)^4}{(N-1)\sigma^4} \quad (3.1)$$

Where μ is the mean, σ^2 is the variance, k is the kurtosis, N is the total number of elements and w is the 1D-array of the non-zero wavelet coefficients of the speech signal.

In the current example, the *kurtosis* from the signals is 5.5 for speech₁ and 4.6 for speech₂. Since a similar shape of the histograms is related to a similar density of data, similar value of *kurtosis* means that the density distribution of the wavelet coefficients is similar, too. In other words, although speech₁ and speech₂ sound different, the density distributions of their non-zero wavelet coefficients are similar. Therefore if the wavelet coefficients of speech₂ are relocated so that they resemble the wavelet coefficients of speech₁, the adapted speech signal may sound similar to speech₁. This is the idea behind the ability of adaptation of speech signals.

In this context, speech₂ may sound similar to speech₁ (and vice versa) because their kurtosis (of the non-zero wavelet coefficients) is similar. Then, the adaptation is feasible if and only if the kurtosis and the number of the non-zero wavelet coefficients are similar between the speech signals. The hypothesis of adaptation is presented in the next section.

3.3. Hypothesis formulation and statements

This section proposes and explains the concept of adaptation based on a relocation process in wavelet domain. The hypothesis of adaptation is formulated as follows:

any speech signal may seem similar to a target speech signal if its wavelet coefficients are sorted [56],[57]

In the above hypothesis:

- (i) The term “any” speech signal corresponds to legible voice signals. It discards silence signals and highly noisy speech signals (SNR_{\min} should be 20 dB).
- (ii) The term “may” implies that the adaptation is feasible.
- (iii) The term “similar” means that the Squared Pearson Correlation Coefficient (SPCC) of the target speech signal and the adapted secret signal is close to 1. In other words, the sound of the adapted speech signal is perceptually identical to the sound of the target speech signal.

The Squared Pearson Correlation Coefficient, ρ^2 , is obtained as follows:

$$\rho^2(S, \hat{S}) = \left[\frac{\sum_{i=1}^m (S(i) - S_{mean})(\hat{S}(i) - \hat{S}_{mean})}{\sqrt{\sum_{i=1}^m (S(i) - S_{mean})^2} \cdot \sqrt{\sum_{i=1}^m (\hat{S}(i) - \hat{S}_{mean})^2}} \right]^2 \quad (3.2)$$

Where S , \hat{S} , S_{mean} and \hat{S}_{mean} are the original, extracted secret signal, mean of original and mean of extracted secret signal, respectively.

- (iv) The term “sort” is related to a relocation process.

The hypothesis is true if the speech signal and the target speech signal satisfy the following conditions:

- a) The same sampling frequency.

- b) The same time-frame.
- c) The same wavelet base in the decomposition and reconstruction stages.
- d) The ratio of the (number of the) non-zero wavelet coefficient between the signals is close to 1.

It is measured according to:

$$ratio = \frac{non_zero(w_1)}{non_zero(w_2)} \quad (3.3)$$

Where $non_zero(w)$ is the total number of the non_zero wavelet coefficients of the speech signal, $\{w_1, w_2\}$ are the wavelet coefficients of $speech_1$ and $speech_2$, respectively.

A frame is considered as a segment of the speech signal with a quasi-constant dynamic range and SNR. For example, if a speech signal contains whisper sounds and screaming voice signals, the signal must be separated in non-overlapped segments of whisper sounds and screaming sounds. The same process is applied if the SNR abruptly changes into the signal. On the other hand, a coefficient is classified as non-zero if its magnitude is higher than a threshold (e.g. 1% of the highest amplitude).

The hypothesis is theoretically supported by the following development. Suppose there are two speech signals: one (discrete) speech signal, $s[n]$, and one (discrete) target speech signal, $tg[n]$. Figure 3.4 illustrates the speech signal and Figure 3.5 shows the target speech signal.

Every signal has m samples located in an integer place of the discrete time and therefore they can be modeled as follows:

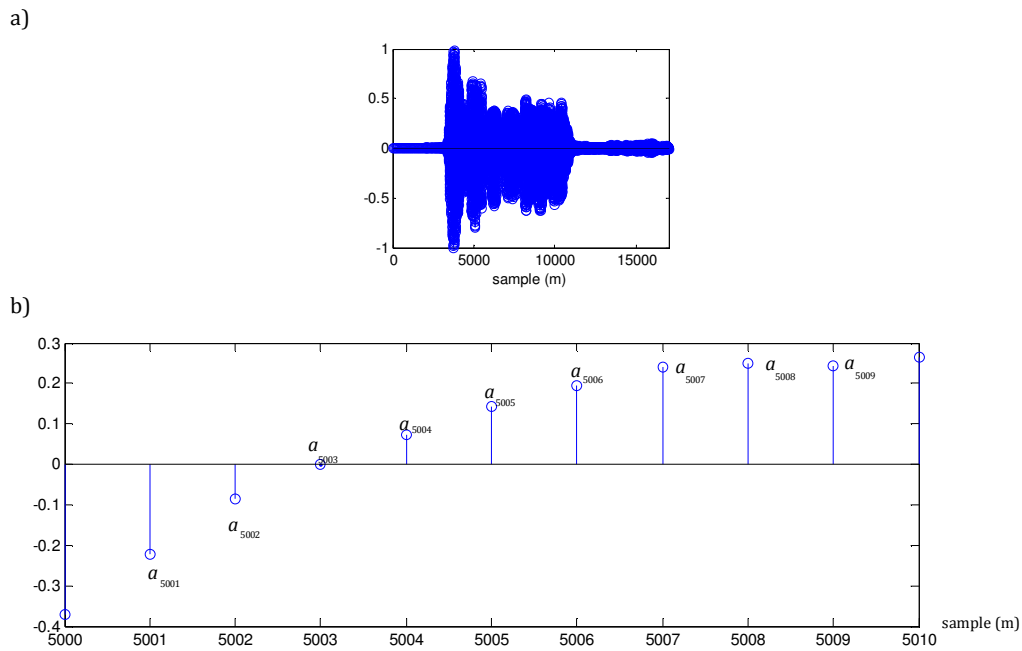


Figure 3.4. Speech signal, $s[n]$: a) entire time-scale; b) zoom of the speech signal

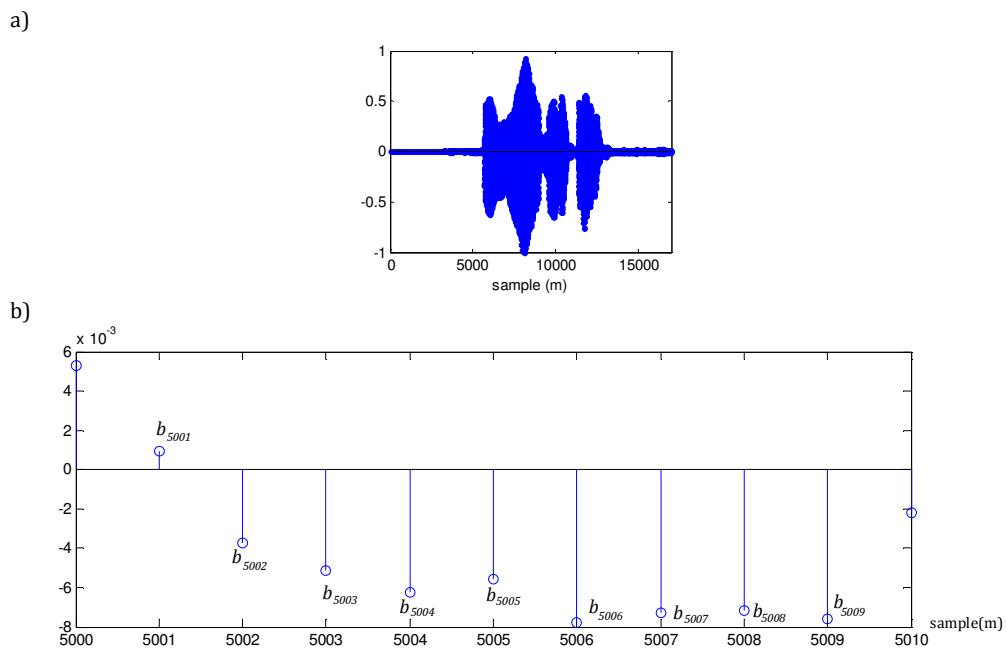


Figure 3.5. Target speech signal, $tg[n]$: a) entire time-scale; b) zoom of the speech signal

$$\begin{aligned}
s[n] &= a_0 \delta[n] + a_1 \delta[n-1] + \dots + a_{(m-1)} \delta[n-(m-1)] \\
s[n] &= \sum_{n_0=0}^{m-1} a_k \delta[n-n_0] \quad \{n, n_0\} \in \mathbb{Z} \quad a_i \in \mathfrak{R}
\end{aligned} \tag{3.4}$$

$$\begin{aligned}
tg[n] &= b_0 \delta[n] + b_1 \delta[n-1] + \dots + b_{(m-1)} \delta[n-(m-1)] \\
tg[n] &= \sum_{n_0=0}^{m-1} b_k \delta[n-n_0] \quad \{n, n_0\} \in \mathbb{Z} \quad b_i \in \mathfrak{R}
\end{aligned} \tag{3.5}$$

with

$$s[n] \neq tg[n] \tag{3.6}$$

Where n is the discrete time, $\delta[n-n_0]$ is the delayed impulse at $n=n_0$ and $\{a_{n_0}, b_{n_0}\}$ are the amplitudes of the impulses of the speech signal and target speech signal, respectively. These signals are perceptually different if a_i and b_i are not correlated.

The signals are represented on time-frequency domain by the DWT, according to e.q. (3.5) and (3.6):

$$s[n] \xrightarrow{DWT} S(k) \quad k \in \mathbb{Z} \tag{3.7}$$

$$tg[n] \xrightarrow{DWT} Tg(k) \tag{3.8}$$

Where $\{S(k), Tg(k)\}$ are the wavelet representations of the speech signal and the target speech signal, respectively, and k is the time-frequency axis. The wavelet coefficients includes coarse and detail coefficients, as follows:

$$S(k) \equiv \{c_s(k), d_s(k)\} \tag{3.9}$$

$$Tg(k) \equiv \{c_{tg}(k), d_{tg}(k)\} \tag{3.10}$$

In the above equations $\{c_s(k), d_s(k)\}$ and $\{c_{tg}(k), d_{tg}(k)\}$ are coarse and detail coefficients of the speech signal and the target speech signal, respectively.

In a similar way to equations (3.4) and (3.5), the wavelet coefficients can be modeled as the sum of delayed impulses in the time-frequency domain, according to:

$$c_s(k) = \sum_{k_0=0}^{M-1} g_{1k_0} \delta[k - k_0] \quad \{k, k_0, M\} \in Z, \quad g_{1k_0} \in \mathfrak{R} \quad (3.11)$$

$$d_s(k) = \sum_{k_0=0}^{M-1} g_{2k_0} \delta[k - k_0] \quad g_{2k_0} \in \mathfrak{R} \quad (3.12)$$

$$c_{tg}(k) = \sum_{k_0=0}^{M-1} p_{1k_0} \delta[k - k_0] \quad p_{1k_0} \in \mathfrak{R} \quad (3.13)$$

$$d_{tg}(k) = \sum_{k_0=0}^{M-1} p_{2k_0} \delta[k - k_0] \quad p_{2k_0} \in \mathfrak{R} \quad (3.14)$$

Where $\{g_{1k_0}, g_{2k_0}\}$ are coarse-weights and detail-weights of the wavelet coefficients of the speech signal, and $\{p_{1k_0}, p_{2k_0}\}$ are coarse-weights and detail-weights of the wavelet coefficients of the target speech signal. The value of M corresponds to the total number of detail (or coarse) coefficients of every signal and it is related to the number of samples m and the order of the filters of the DWT.

If the speech signal and the target speech signal are perceptually different, their coarse-weights and detail-weights will be different, too. In other words:

$$\text{if } s[n] \neq tg[n] \Rightarrow S(k) \neq Tg(k) \quad \therefore \quad \{g_{1k_0} \neq p_{1k_0} \vee g_{2k_0} \neq p_{2k_0}\} \quad (3.15)$$

Then, only if a sorting (relocation) process is applied, the speech signal would be perceptually identical to the target speech signal. If the coarse-weights and detail-weights of the speech signal are relocated so that they resemble the coarse-weights and detail-weights of the target speech signal, the adapted speech signal would sound like the target speech signal. Then, the hypothesis is true if the wavelet representation of the adapted speech signal, $S_a(k)$, is highly correlated to the wavelet representation of the target speech signal, $Tg(k)$, according to:

if

$$\rho^2(S_a(k), Tg(k)) \approx 1 \Rightarrow s_a[n] \approx tg[n] \quad (3.16)$$

with

$$S_a(k) \xrightarrow{IDWT} s_a[n] \quad (3.17)$$

In the above equations, ρ^2 is the index of similarity, $IDWT$ is the Inverse Discrete Wavelet Transform and $s_a[n]$ is the adapted speech signal, in time domain. At this point, the purpose is to find a wavelet representation of the adapted speech signal which satisfies eq. (3.16). There are at least two ways to find the adapted speech signal: by a deterministic and by a heuristic search. In this thesis a deterministic search is proposed.

To find an adapted speech signals that it resembles the target speech signal, the following steps should be carried out:

- (i) The speech signal and the target speech signal are decomposed using the DWT with the same wavelet base, according to eq. (3.7) and eq. (3.8).
- (ii) The coarse-weights and detail-weights of the speech signal are grouped in a 1D-array, as follows:

$$G = [g_1 \quad g_2] \quad (3.18)$$

- (iii) The coarse-weights and detail-weights of the target speech signal are grouped in a 1D-array, as follows:

$$P = [p_1 \quad p_2] \quad (3.19)$$

- (iv) The 1D arrays G and P are sorted in descending order. The initial positions of the weights are kept in the arrays u_g and u_p , respectively.
- (v) Every weight of G is relocated, according to:

$$G_a(u_p) = G(u_g) \quad (3.20)$$

Where G_a is the 1D-array that looks similar to P. It means:

$$\rho^2(P, G_a) \approx 1 \quad (3.21)$$

- (vi) With the 1D-array G_a the wavelet coefficients of the adapted speech signals are found, as follows:

$$c_{s_a}(k) = \sum_{k_0=0}^{M-1} G_{a_{k_0}} \delta[k - k_0] \quad (3.22)$$

$$d_{s_a}(k) = \sum_{k_0=M}^{2M-1} G_{a_{k_0}} \delta[k - k_0] \quad (3.23)$$

The first-half of the normalized array G_a corresponds to the coarse-weights and the second-half corresponds to the detail-weights; $\{c_{s_a}, d_{s_a}\}$ contains the coarse and detail coefficients of the adapted speech signal.

- (vii) The Inverse Wavelet Transform, of $\{c_{s_a}, d_{s_a}\}$ is calculated, according to:

$$\{c_{s_a}, d_{s_a}\} \xrightarrow{IDWT} s_a[n] \quad (3.24)$$

Where $s_a[n]$ is the adapted speech signal, in time domain.

- (viii) Finally, the dynamic range of the adapted speech signal is set to the same dynamic range of the target speech signals, as follows:

$$S_a[n] = S_a[n] * \left(\frac{\max |t_g[n]|}{\max |S_a[n]|} \right) \quad (3.25)$$

The output signal has the same plain-text, rhythm and gender of the speaker of the target speech signal if and only if the conditions described at the beginning of the section were previously satisfied.

The following example illustrates the above steps. The speech signal has the plain-text “good morning everybody” and the target speech signal has the plain-text “see you the next week”. Both signals have been sampled with $f_s=8\text{KHz}$ and are from a

female speaker. The above signals were used in the first part of the current section as speech_1 and speech_2 (Figure 3.1a and 3.1b).

Their wavelet coefficients are calculated by using the 5/3 base and they are grouped in 1D arrays (Figure 3.2). Then, these arrays are sorted in descending order (Figure 3.6) and their original positions are kept in two 1D arrays, one array per signal.

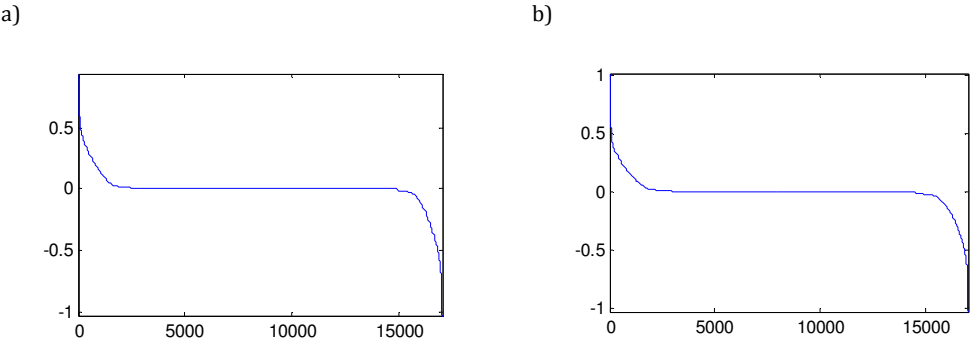


Figure 3.6. Sorted coefficients: a) target signal, b) speech signal

According to Figure 3.6, the sorted target's coefficients and the sorted speech's coefficients have a similar behavior. The difference lies on the positive amplitude which is higher in the second signal. If the speech's coefficients are relocated according to the information contained into the arrays of their original positions, the adapted-speech's coefficients resemble the target's coefficients (Figure 3.7).

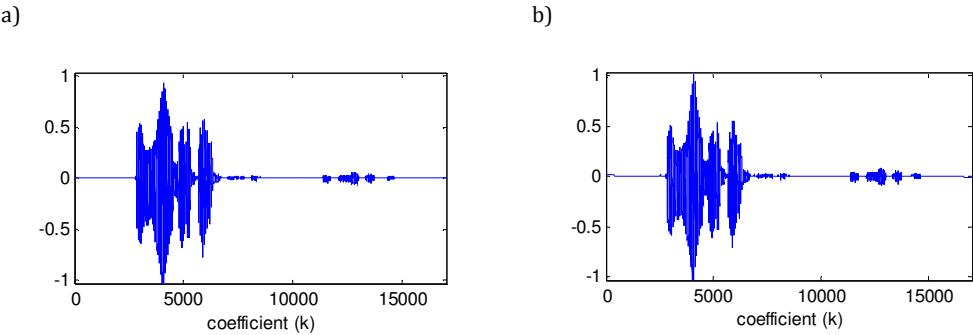


Figure 3.7. Wavelet coefficients: a) target's coefficients, b) adapted-speech's coefficients

It is worth noting that the adapted-speech's coefficients look similar to the target's coefficients. Finally, the IDWT is applied to the adapted-speech's coefficients and then the adapted speech signal is obtained (Figure 3.8).

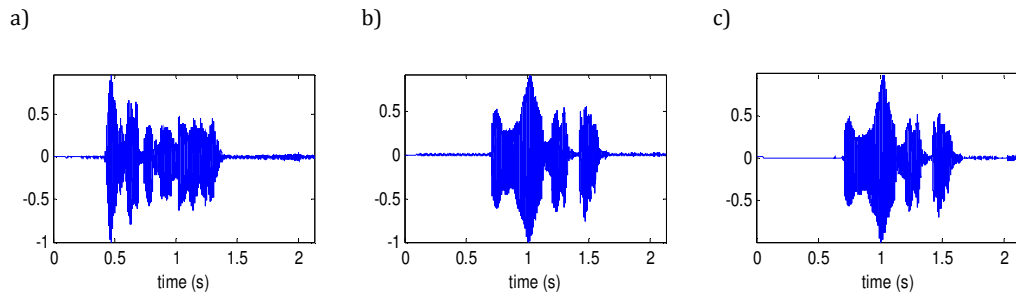


Figure 3.8. Time signals: a) secret message, b) target speech signal, c) adapted-secret message

The target speech signal and the adapted-speech signal have the same plain-text with the same rhythm as gender of the speaker.

In the current example, the *ratio* of the non-zero coefficients is 0.846, and the level of similarity is 0.995. Additionally, as the *ratio* is close to 1, the perceptual similarity between the target speech signal and the adapted-speech signal is high.

3.4. Experimental validation

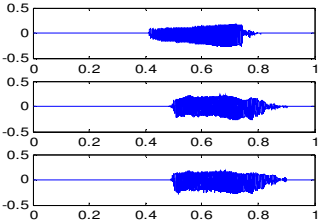
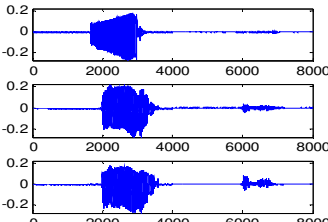
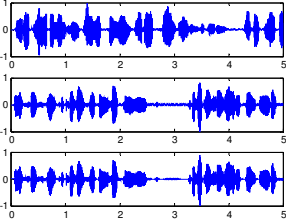
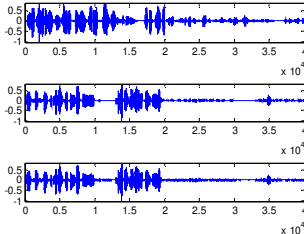
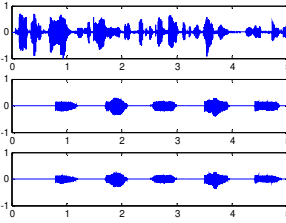
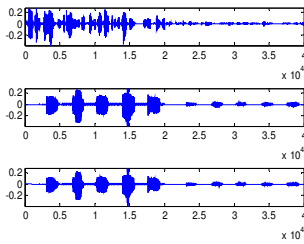
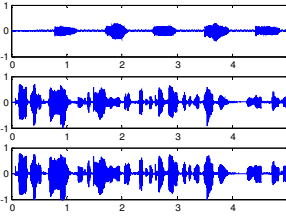
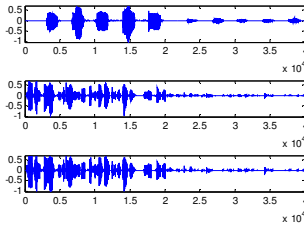
With the purpose to validate the hypothesis of adaptation, several tests have been performed. In the first part, the objective is to verify if adaptation is feasible in different cases as vowel to vowel, words to words, words to vowel and vowel to words. In the second part, the purpose is to verify if adaption depends on the language and gender of the speaker.

3.4.1. Different kinds of sounds

In this set of tests, the speech signals are divided in two groups: vowel signals and voice signals. In the first group there are 5 vowel sounds in English language while in the second group there are six voice signals belonging to female and male speakers in English, French and German. The Sound Quality Assessment Material (SQAM) was selected as the database of the second group [58]. All the speech signals are sampled to 8K Hz and quantized with 16-bits with a time-scale of five seconds.

The following cases are taken into account: vowel to vowel, words to words, words to vowel, and vowel to words. An example of each case is illustrated in Table 3.1. The first column enunciates the case of adaptation. The second column shows the level of similarity and the ratio of non-zero coefficients between the speech signal and the target speech signal. The third column plots (top-down) the speech signal, the target speech signal and the adapted speech signal. In the last column the speech's coefficients, the target-speech's coefficients and the adapted-speech's coefficients are shown. The coefficients are 1D-arrays which include the coarse and detail of the signal, according to equations (3.18) and (3.19).

Table 3.1. Squared Correlation Coefficient & Ratio: examples of adaptation. [57]

Case	ρ^2 & ratio	Time Domain	Wavelet Coefficients
Case 1: vowel to vowel	$\rho^2 = 0.95$ ratio = 1.13		
Case 2: words to words	$\rho^2 = 0.99$ ratio = 0.94		
Case 3: words to vowels	$\rho^2 = 0.98$ ratio = 0.86		
Case 4: vowel to words	$\rho^2 = 0.98$ ratio = 1.17		

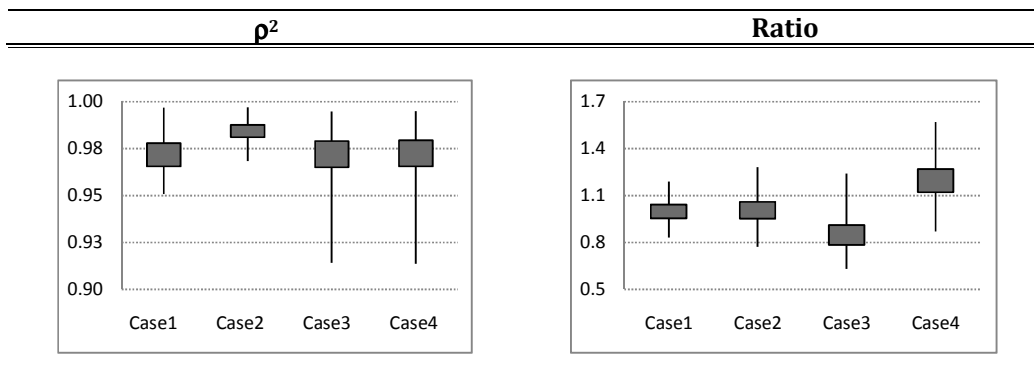
According to the results shown on Table 3.1, it is noticed that:

- (i) It is not necessary a time synchronization between the speech signal and the target speech signal. Due to the ability of adaptation, the adapted speech signal is in fact synchronized with the target speech signal.

- (ii) The ratio of the non-zero coefficients can be slightly higher or lower than 1, nevertheless the speech distortion index is close to 1 (and always $\rho^2 > 0.95$).
- (iii) The proposed method works both with single sounds (vowels, syllables) and words. A group of vowels can imitate words and the contrary is also possible.

The summary of the tests is illustrated in Table 3.2. In every case the lowest and highest value and the confidence interval of the 95% are plotted. The first case, vowel to vowel, consists in 20 tests. The five vowels work both as speech signal and as target speech signal and everyone is adapted to the rest of them. The second case, words to words, includes 30 tests. Six records of female and male speakers in three languages are adapted between them. The third case, six word messages are adapted to five vowel signals in 30 tests. In the fourth case, five vowel signals are adapted to six words messages in 30 tests, too.

Table 3.2. Squared Correlation Coefficient & Ratio: summary of the tests. [57]



According to Table 3.2, all of the adapted speech signals have a speech distortion index, ρ^2 , higher than 0.9, although the ratio of the non-zero coefficients is not exactly equal to 1. Nevertheless, it is noticed that the two best cases (first and second one)

have a ratio of non-zero coefficients closer to 1 than rest of them (third and fourth cases). It is expected that if the ratio is in [0.8 1.20], the adapted speech signal will have ρ^2 higher than 0.95. The parameter ρ^2 can be interpreted as the percentage of the coefficients of the adapted secret signal that are linearly correlated to the coefficients of the target speech signal.

On the other hand, the p-value was taken into account in these experiments. It defines if the linear correlation is due to a coincidence or not; if the p-value is lower than 0.05 the idea about a coincidence is rejected. Since the p-values were always smaller than 0.05, then $\rho^2 > 0.9$ is significant and the hypothesis of the adaptation of the speech signals is listed as true.

3.4.2. Different language and gender of the speaker

In this second group of tests, the hypothesis of adaptation is tested in relation to the gender and the language of the speech signals. Four scenarios are analyzed, as follows:

- (i) The language of the messages and the gender of the speakers of both the speech and the target speech signals are the same.
- (ii) The language of the messages is the same, but the gender of the speakers is different.
- (iii) The gender of the speakers is the same, but the language of the messages is different.
- (iv) Both the gender of the speakers and the language of the messages are different.

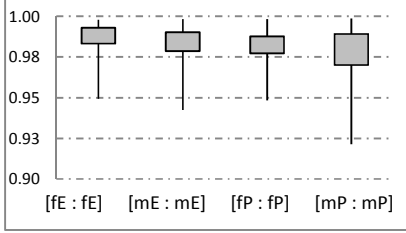
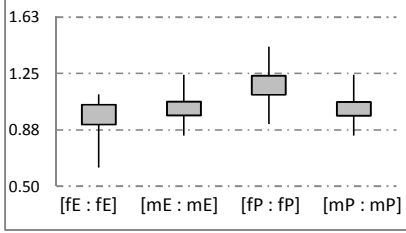
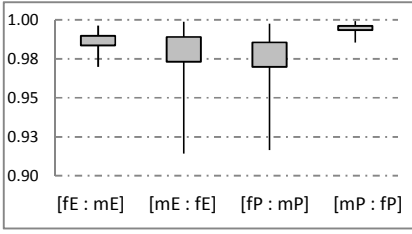
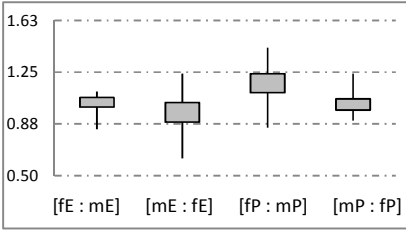
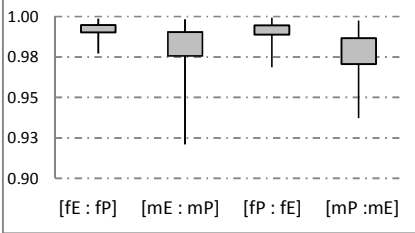
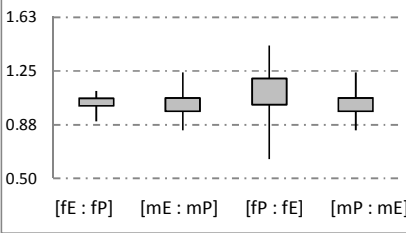
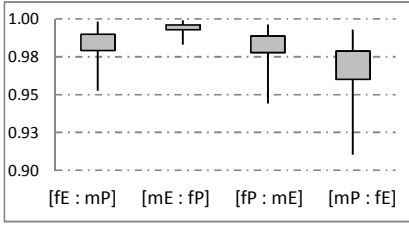
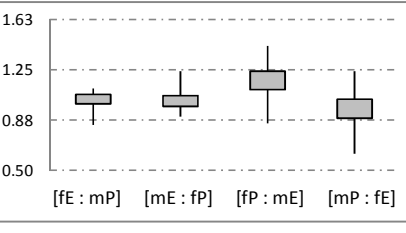
The speech signals used in these tests correspond to 10 speech signals from female speakers in English language, 10 speech signals from male speakers in English

language, 10 speech signals from female speakers in Polish language and 10 speech signals from male speakers in Polish language. Therefore, the total number of messages is 40 which correspond to 40 speakers. English and Polish languages were selected because they have strong dissimilarities in terms of phonetic sounds.

The results by scenario are shown in Table 3.3. It includes the level of similarity (second column) and the *ratio* of the non-zero wavelet coefficients (third column). There are 4 cases by scenario, each one with 25 tests, for a total by scenario of 100 tests. Every case is represented by its highest and lowest value, and the confidence interval of 95%. It uses the following notation to name the case: the first and the second letter are related to the gender and the language of the original speech signal, respectively and the third and the fourth letter are related to the gender and the language of the target speech signal, respectively. For example, the case [fP: mE] means that the speech signal is from a female (f) speaker in Polish (P) language and the target speech signal is from a male (m) speaker in English (E) language.

According to Table 3.3, all of the results of the four scenarios have level of similarity higher than 0.9 (and their confidence ranges are higher than 0.98). It is worth noting that the first and second cases of the fourth scenario have *ratio* into the interval [0.88 1.25] and their index of similarity higher than 0.95; while the third and fourth cases have *ratio* outside the above interval and their index of similarity fell to 0.90. However, the index of similarity of the third case is better than of the fourth case. Therefore, it can be concluded that the quality of the adapted speech signal has a strong relationship with *ratio*, and if it is the range [0.88 1.25] it is expected that the similarity between the adapted speech signal and the target speech signal will be high. Together with that, if *ratio* is outside of the above range, it is more desirable a value slightly higher than 1.25 instead of a value slightly lower than 0.88.

Table 3.3. Results by scenario. [59]

Scenario	ρ^2	Ratio
<i>Same gender and language</i>		
<i>Different gender but same language</i>		
<i>Same gender but different language</i>		
<i>Different gender and different language</i>		

Summarizing, a speech signal can be adapted so that it resembles another speech signal if the theoretical requirements are satisfied even so if the gender of the speaker and/or the language of the message is not the same.

3.5. Summary

The main ideas of this chapter are summarized as follows:

- (i) A powerful hypothesis of speech processing is presented. This is known as the ability of adaptation of speech signals, formulated as:

*any speech signal may seem similar to a target speech
signal if its wavelet coefficients are sorted.*

- (ii) The idea behind the ability of adaptation is related to the histograms of the non-zero wavelet coefficients of speech signals. Although two speech signals have different time-behavior and frequency-behavior, their histograms (of the non-zero wavelet coefficients) can be similar. Therefore, if the time-frequency elements of a speech signal are relocated it is feasible that the output signal looks and sounds similar to a target speech signal.
- (iii) Several test of adaptation between vowel sounds and words sounds demonstrate that the adaptation is feasible between different kinds of sounds.
- (iv) Together with that, it is demonstrated that adaptation is feasible even if the language or/and the gender of the speaker are changed between the speech signal and the target speech signal.

4. Speech scrambling and the ability of adaptation of speech signals

A novel scheme of speech scrambling is presented in this chapter. Unlike the traditional approach in which the scrambled speech signal is a non-intelligible signal, the current proposal supplies a scrambled signal which is a perfectly legible signal, but with a plain-text different from the original speech signal. The idea is based on the ability of adaptation of the speech signals.

4.1. Motivation

In order to give protection to speech signals, many techniques of analog speech scrambling and digital encryption have been proposed. Among others, there are three aspects to take into account in any scrambling system: to produce a residual intelligibility as low as possible, to supply a high quality of the recovered signal even if the scrambled speech signal is manipulated, and to generate a long effective number of keys (key-space) for resisting cryptanalysis. Usually, the techniques are classified in permutation-based and amplitude scrambling (AS) [40], [41].

Time-Segment Permutation, *TSP*, Frequency-Domain Scrambling, *FDS*, and Time-Frequency Scrambling, *TFS*, are techniques of permutation-based speech scrambling. In the first case, *TSP*, the speech signal is divided in small blocks (typically 16 to 32 ms) and the permutations are made into the blocks according to a scrambling key, generated usually by a Pseudo-Noise (PN) generator [42], [60], [61]. Although it is a simple technique, it has some disadvantages as a small key-space, not low enough residual intelligibility, and low resistance to cryptanalysis. In the second case, *FDS*, the permutation process is carried out in the frequency domain [43], [44], [62]-[66]. The residual intelligibility may be lower than in *TSP*, but the key can be discovered using known cipher-text attacks [20], [21]. In the third case, *TFS*, the speech signal is split in subbands and every subband is divided in segments [46], [47]. It overcomes the disadvantage of its predecessors in terms of the residual intelligibility, but until now the problem of the small key-space has not been overcome.

On the other hand, in the amplitude scrambling (AS) systems, the scrambled speech signal is not obtained by a permutation process, instead of that, the amplitude of the speech signal is modified so that it resembles a white noise signal [67]-[71]. These systems have mainly two disadvantages: firstly, it has been demonstrated that the scrambled speech signals do not overcome cryptanalysis attacks [72] and secondly, the

robustness against signal manipulations like MP3 compression, additive noise, filtering, among others, is not guaranteed.

Since two of the most important requirements of speech scrambling are a priori not satisfied in AS systems, the effort should be focused on improvement the key-generation in the permutation-based speech scrambling schemes. It encompasses to create a long key-space and adequate sequences that guarantee the very low residual intelligibility of the speech signal. In terms of long key-space, if the speech signal is not divided in sub-blocks and the entire speech signal is permuted, then the total number of possible combinations is significantly higher and consequently the effort to reveal the key by brute effort attack is not feasible. In terms of adequate sequences, the key-generator must avoid the reverse and delay sequences because these do not destroy the intelligibility of the speech signal. In order to objectively classify if a sequence is appropriate or not, there are two parameters that quantify the level of displacement of data and the total number of data displaced out of their original places. They are known as the normalized displacement, Γ_{nd} , and the Hamming Distance, HD , respectively. If Γ_{nd} increases the residual intelligibility decreases, however there is a turning point in which the residual intelligibility increases again [42]. Since every limit value ($\Gamma_{nd} \sim 0$ and $\Gamma_{nd} \sim 1$) is related to the delay sequence and the reverse sequence, an appropriate value must be significantly higher than 0 and simultaneously very distant from 1. In the case of HD , if all elements are displaced out of their original places ($HD=100\%$), the scrambled speech signal would have little residual intelligibility [44]. But if $HD > 90\%$ and if the unpermuted elements are distributed randomly, the residual intelligibility is sufficiently low [43]. Summarizing, adequate sequences must have simultaneously HD higher than 90% and Γ_{nd} around the turning point.

In the traditional approaches, the key-generator is based on PN sequences with a low number of elements (e.g. ~ 90). However, in the last years alternative solutions have

been proposed. The authors of [73] use high dimension matrix transformation to relocate the samples of the speech signal. Although their scrambled speech signals are robust against MP3 attack, neither the low residual intelligibility nor the resistance against cryptanalysis is guaranteed. On the other hand, the authors of [74] use a cellular automaton (CE) to generate the permutation sequence. The advantage of the proposal is that the length of the key is up to the total number of samples of the speech signal and then the effort to discover the key is high. However, the residual intelligibility depends strongly on initial control conditions like the number of generations (NOG) and the neighborhood rule. If these parameters are not selected appropriately, the very low residual intelligibility is not reached. Consequently, the issue of an efficient key-generator for permutation-based speech scrambling systems has not been overcome yet.

The aim is to generate a scrambled speech signal with the following characteristics:

- (i) Perfect Secrecy: it is satisfied if the key-space is equal to the secret-space, and the mapping process between the secret message and the scrambled speech signal is one-to-one [75]. If the above conditions are satisfied, the system resists the brute force and the known-cipher attacks.
- (ii) Very low residual intelligibility: if the permutations satisfy the condition of normalized displacement and the condition of Hamming Distance, little residual intelligibility is obtained.
- (iii) Robustness against signal manipulations attacks: the secret message is recovered even if the scrambled speech signal has been manipulated (e.g. MP3 compression, and additive noise, filtering).

Therefore, it is proposed a speech scrambling system that simultaneously satisfies the above desirable conditions. The core is the ability of adaptation of speech signals

presented in Chapter 3. The secret message is adapted to the target speech signal which has non-sensitive information. Unlike traditional approaches in which the scrambled speech signal sounds like a white noise signal, in the proposal, the scrambled speech signal sounds like the target speech signal. The secret key is the mapping between the secret message and the target speech signal. With the secret key, the adaptation process is reversed and then the secret message is recovered. Consequently, the proposal is focused on protecting public data with private key. Suppose that you want to publish (e.g. on a web site) a speech signal which has sensitive information (secret message), but you want to protect the secrecy. The idea is to manipulate the secret message so that it resembles a target speech signal (which has non-sensitive information), and then the scrambled speech signal is published instead of the secret message. Therefore, although anyone can access the scrambled speech signal, the secret message is protected. Only the authorized user can reveal the secret message through the secret key (previously obtained through another channel).

The rest of the chapter is organized as follows. The idea behind the ability of adaptation of speech signals as a key-generator in a scrambling scheme is explained in Section 4.2. Remarkable results of the performance of the scrambling system are shown in Section 4.3. Cryptanalysis is shown in Section 4.4. The chapter is summarized in Section 4.5.

4.2. The proposed scheme

The scrambling system is based on the ability of adaptation of speech signals already presented in Chapter 3. It works with two speech signals: the secret message and the target speech signal. Once the conditions of adaptation have been verified (see Section 3.2), the scrambled speech signal is obtained as shown in Figure 4.1.

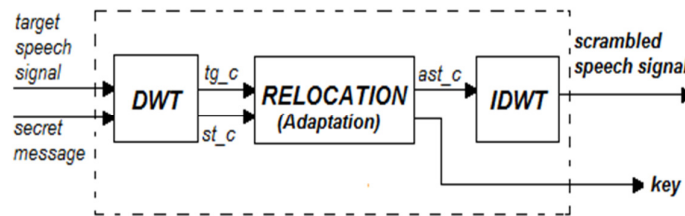


Figure 4.1. Flowchart of the scrambling process

The steps are explained as follows:

- a) The speech signals are decomposed by using the Discrete Wavelet Transform with one level of decomposition. The wavelet base must be the same in both cases. Since the wavelet decomposition of a signal gives coarse and detail coefficients, they are grouped in a one-dimensional array. Therefore, there is one 1D array per signal. At the output of this step the secret's coefficients, st_c , and the target's coefficients, tg_c , are obtained.
- b) The secret's coefficients are relocated with the purpose to resemble the target's coefficients. For example, if the target's coefficients are $tg_c=[10, 9, 6, 4, 8, 12, 14, 16]$ and the secret's coefficients are $st_c=[2, 5, 6, 9, 8, 4.5, 4, 3]$, then the adapted secret's coefficients, ast_c , are $ast_c=[5, 4.5, 3, 2, 4, 6, 8, 9]$. Figure 4.2 shows the three groups of coefficients. It is remarkable the similarity of ast_c to tg_c although st_c has a completely different behavior.

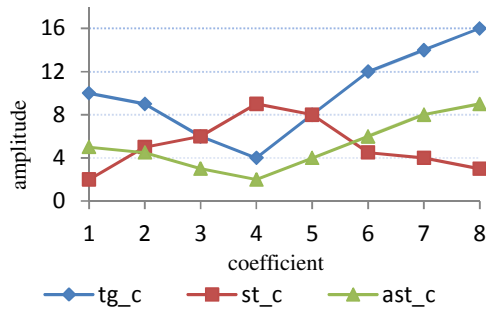


Figure 4.2. Example of adaptation. [76]

The *key* is formed from the positions of *ast_c* in relation to the original positions in *st_c*. The first value of the key contains the original position of the first value of *ast_c*, the second value of the *key* contains the original position of the second value of *ast_c*, and so on. In the current example, $key=[2, 6, 8, 1, 7, 3, 5, 4]$.

- c) In the last step, the adapted secret's coefficients are reconstructed by using the Inverse Discrete Wavelet Transform. The output is the scrambled speech signal.

Then, the scrambled speech signal is transmitted together with the *key*. At the receiver, the secret message can be recovered with a reverse process of adaptation. The descrambling process is explained as follows (Figure 4.3):

- a) The scrambled speech signal is decomposed by the DWT, mono-level. The wavelet base is the same used in the scrambling process. The coarse and detail coefficients are put together into an 1D array, *ast_c*.
- b) The adapted secret coefficients, *ast_c*, are relocated according to the *key*. For example, if $ast_c=[5, 4.5, 3, 2, 4, 6, 8, 9]$ and $key=[2, 6, 8, 1, 7, 3, 5, 4]$, the recovered secret's coefficients, *rst_c*, are obtained as $rst_c=[2, 5, 6, 9, 8, 4.5, 4, 3]$. It is worth noting that *rst_c* is equal to *st_c*.

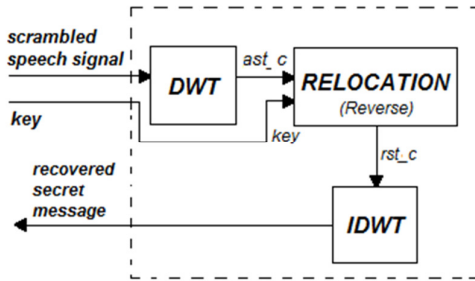


Figure 4.3. Flowchart of the descrambling process

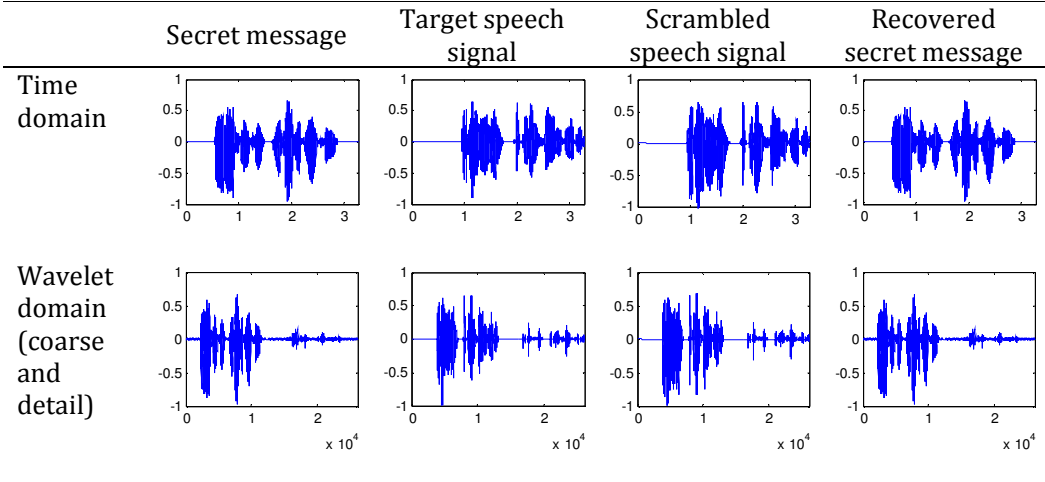
- c) Once the adapted process has been reverted in the wavelet domain, the IDWT is applied to the recovered secret's coefficients. The output is the recovered secret message. If the scrambled speech signal has not been manipulated, it is expected that the recovered secret message will be exactly equal to the original secret message. Nevertheless, although the scrambled speech signal suffers small amplitude changes, the recovered secret message will be very similar to the original secret message.

The proposed scheme is illustrated with an example. Suppose we have two speech signals with different language and gender of the speaker. The secret message is in English language from a male speaker with the plain-text *in the last lecture, we dealt with unit four* and the target speech signal is in Catalan language from a female speaker with the plain-text *tornem un moment al que vam fer a l'última classe*. Then, the secret message is adapted so that it resembles the target speech signal. The scrambled (or adapted) speech signal sounds highly similar to the target speech signal –with the same rhythm, gender and plain-text-. Therefore, the scrambled signal is transmitted together with the *key*. At the receiver, the adaptation process is reversed and the recovered secret message is obtained.

Table 4.1 shows the secret message, the target speech signal, the scrambled speech signal and the recovered secret message. In the current example, the value of *ratio* is equal to 0.9516, the level of similarity between the target speech signal and the

scrambled speech signal is 0.9782 and the level of similarity between secret message and recovered secret message is ~ 1 . Since the plain-text of the secret message is completely different from the plain-text of the target speech signal, it is expected that the residual intelligibility will be close to zero. On the other hand, since the adaptation can be completely reversed, the recovered secret message is equal to the original secret message.

Table 4.1. Speech signals in time domain and wavelet domain



Finally, in terms of key-generation, the proposal has the following characteristics [76]:

- (a) The key length is the same as the secret’s coefficients length. It is expected that the key length is at least 8K per second (for a speech signal sampled at 8K Hz). The higher the time-scale of the speech signal, the higher is the key length.
- (b) If the key length is m , it has m non-repetitive numbers in the range $[1 m]$.
- (c) Computational cost to obtain the *key* in the scrambling procedure is very low. Since the kernel of adaptation is the sorting process and two arrays are sorted, the computational cost to create the *key* is the double of $O(m \log m)$.
- (d) Computational cost to discover the key is very high. An eavesdropper needs $m!$ attempts to obtain the right key.

4.3. Experimental validation

In order to validate the ability of adaptation as a key-generator into a speech scrambling system, several tests were conducted to measure the level of permutation. Two sets of speech signals have been used; the first one corresponds to English messages of a male speaker and the second one to Catalan messages of a female speaker. These records have been taken from the database of the SLT at the Universitat Politècnica de Catalunya [77]. Firstly, ten English messages are adapted to ten Catalan messages and vice versa. Secondly, the English messages are adapted between them, and the same process is carried out with the Catalan messages. At the end, there are 100 tests of adaptation of English messages to Catalan messages, 100 test of Catalan messages to English messages, 90 tests of adaptation between English messages and 90 tests of adaptation between Catalan messages.

In every case three parameters are measured: the normalized displacement (Γ_{nd}), the level of derangement (HD) and the *ratio* of the non-silent time of every pair of speech signals, (*ratio*). Together with that, the *level of similarity* between the scrambled speech signal and the target speech signal is taken into account.

The normalized displacement, Γ_{nd} , is measured as follows [60]:

$$\Gamma_{nd} = \frac{\sum_{q=1}^N |\gamma_q|}{\Gamma_{\max}} ; \quad \Gamma_{\max} = \begin{cases} N^2/2 & \text{if } N \text{ is even} \\ (N^2 - 1)/2 & \text{if } N \text{ is odd} \end{cases} \quad (4.1)$$

Where N is the number of elements in the array, γ is the value of the extent of the shifting and $|\cdot|$ is the magnitude symbol. The value of Γ_{nd} is in the range $[0 \ 1]$. If $\Gamma_{nd}=0$, it means that the elements have not been permuted; but if $\Gamma_{nd}=1$, the elements were placed in the most distant position possible.

The level of derangement is measured through the Hamming Distance (HD) according to:

$$HD = \frac{\sum_{k=1}^N d(k)}{N} ; \quad d(k) = \begin{cases} 1 & \text{if } i(k) \neq p(k) \\ 0 & \text{if } i(k) = p(k) \end{cases} \quad (4.2)$$

Where $i(k)$ is the original array, $p(k)$ is the permuted array and $d(k)$ is the difference array. If all elements are permuted, HD is equal to 1 (or 100%) and it is known as derangement.

The *ratio* and the *similarity* were defined in equations (3.3) and (3.2), respectively. Similarity is measured through Squared Pearson Correlation Coefficient because it has been demonstrated that this parameter can be viewed as a speech distortion index; and it gives an indication on the strength of the linear relationship between two speech signals [78].

In [57], the relationship between the *ratio* of the non-silent time and the *similarity* between the original speech signal and the adapted (or scrambled) speech signals was presented. Now, the aim is to establish the value of the normalized displacement and the level of derangement to guarantee a right performance of the ability of adaptation into a scrambling system.

4.3.1. Relationship between Γ_{nd} and ρ^2

The aim is to analyze the relationship between the level of similarity and the normalized displacement. Since the objective of destroying the intelligibility of the secret message is satisfied if the scrambled speech signal is highly similar to the target speech signal, the current purpose is to guarantee that the value of ρ^2 is the highest possible. Consequently, at the end of the test, a suggested value of Γ_{nd} is obtained.

The results are grouped in four scenarios: adaptation from English to Catalan messages, adaptation from Catalan to English messages, adaptation between Catalan messages and adaptation between English messages. Figure 4.4 shows the results, where tg represents the target speech signal and st the secret message. The red dotted line is the threshold for the lowest level of similarity (0.9). The desirable behavior is found above this line. Since the *ratio* values of the selected English messages have the lowest dispersion, their values of Γ_{nd} and ρ^2 have the lowest dispersion, too (Fig 4.4.d). In the opposite case, since the *ratio* values between English to Catalan message (and vice versa) have the highest dispersion, their values of Γ_{nd} and ρ^2 have the highest dispersion, too (Fig 4.4.a, 4.4.b).

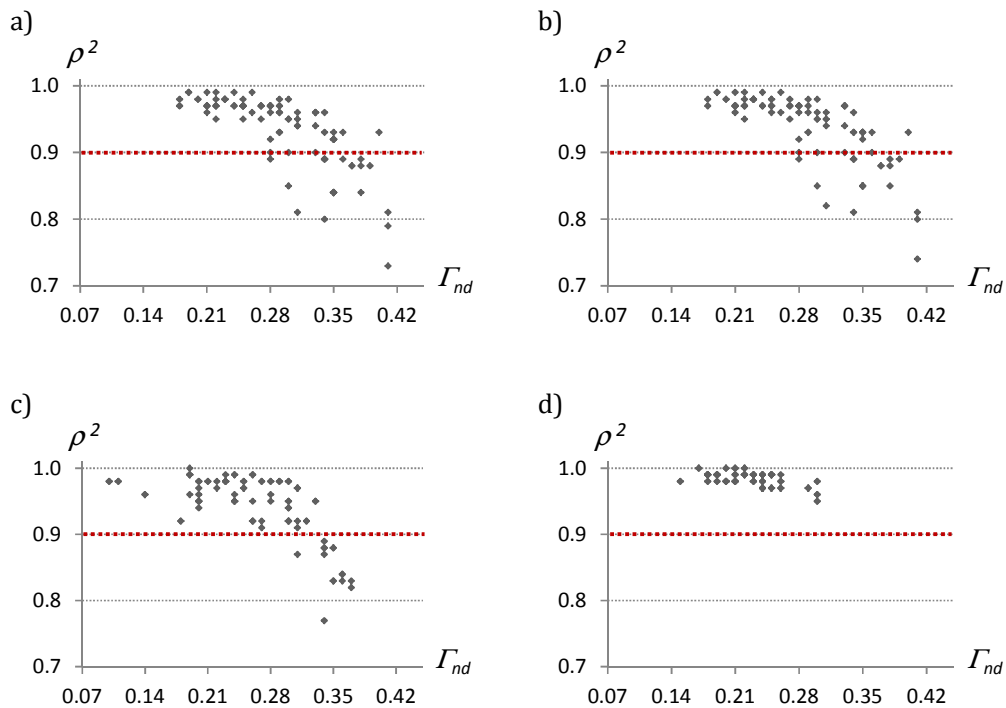


Figure 4.4. Similarity and Normalized displacement: a) tg =Catalan messages, st =English messages; b) tg =English messages, st =Catalan messages; c) $[tg\ st]$ =Catalan messages, d) $[tg\ st]$ =English messages. [76]

According to Figure 4.4.a and 4.4.b, if Γ_{nd} is lower than ~ 0.25 , all values of ρ^2 are higher than (or equal to) 0.9. In the case of Figure 4.4.c and 4.4.d if Γ_{nd} is lower than ~ 0.3 , all values of ρ^2 are higher than (or equal to) 0.9. Because the threshold of Γ_{nd} should satisfy all the scenarios, the lower value among them is selected as the suggested value. Then, if the normalized displacement, Γ_{nd} , is lower than 0.25, it is expected that the adaptation process is successful and the scrambled speech signal is highly similar to the target speech signal; therefore, the residual intelligibility is very low.

4.3.2. Relationship between Γ_{nd} and the ratio of the non-silent time

Once the suggested value of Γ_{nd} has been selected, the following aim is to identify the range of the *ratio* values that guarantees a successful adaptation of the secret message. Figure 4.5 plots the results of this test. In this case, a red dotted line represents the threshold for the highest level of normalized displacement. The desirable behavior is found left of this line.

First of all, it is important to remark that the slope of *ratio* in Fig 4.5.a is negative because the non-mute time of the selected Catalan messages is lower than the non-mute time of the selected English messages, and consequently in Fig 4.5.b the slope of *ratio* is positive. On the other hand, because the *ratio* values of the selected Catalan messages are more dispersed, data in Fig 4.5.c is more dispersed than in Fig 4.5.d.

According to Figure 4.5, if *ratio* is in the range [0.8 1.3], most of the points of Γ_{nd} are lower than the threshold fixed in 0.25. Therefore, if *ratio* is in the range [0.8 1.3], the displacement of the secret's coefficients is such that the adaptation is successful (high similarity between the scrambled speech signal and the target speech signal) and then the scrambled speech signal does not keep trace of the secret message.

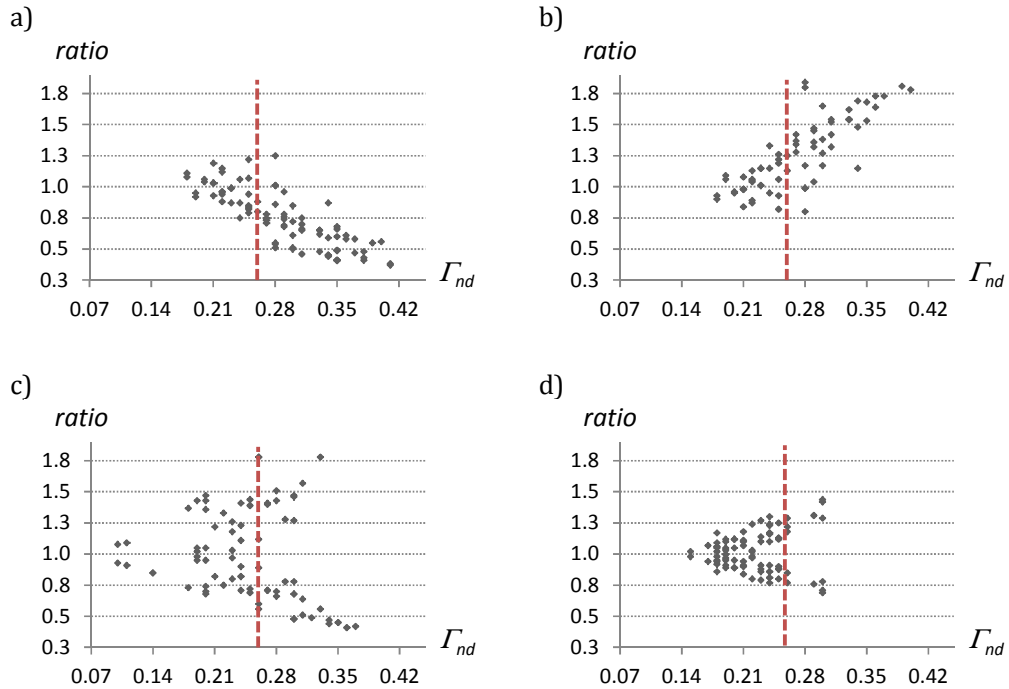


Figure 4.5. *Ratio* and Normalized displacement: a) tg=Catalan messages, st=English messages; b) tg=English messages, st=Catalan messages; c) [tg st]=Catalan messages, d) [tg st]=English messages. [76]

4.3.3. Relationship between *HD* and the ability of adaptation

In the 380 tests, it has been found that the number of permuted elements was always higher than 97.5% and in most cases higher than 99.9%. For example, if a speech signal has 40,000 wavelet coefficients, at least 39,000 of them are placed out of their original position.

Since $HD \sim 100\%$, it is expected that a scrambling system based on the ability of adaptation of the speech signals would have a high enough level of derangement to destroy the intelligibility of the original secret message.

4.4. Security Analysis

Once the speech scrambled system has been tested in terms of the level of derangement and the normalized displacement, the following step to validate the current proposal is in terms of security analysis. Among the tests to overcome are: exhaustive key search (brute force attack), cipher-text only attack and statistical attack.

4.4.1. Exhaustive key search

The first attack that a good cryptosystem must overcome is the brute force attack. A long enough key-space guarantees that the key will not be discovered by an exhaustive key search. If the key size is m , the total number of possible combinations is $m!$ Nevertheless, unlike plain-text in which every letter is represented by only one character, in the case of speech signals every sound (vowel or syllable) is represented by several samples and therefore several wavelet coefficients. Consequently, there are several keys (close to the right key) that produce a descrambled speech signal with the same plain-text of the secret message. Without loss of generality, suppose that an average person speak n sounds (vowel or syllable) per second and the total number of different sounds is p (with p significantly higher than the number of symbols in a language). Therefore, the total number of different plain-texts per second is $(p!)^n$. If the speech signal has t seconds, the above value increases up to $(p!)^{n \cdot t}$. Consequently, an eavesdropper needs to test between $(p!)^{n \cdot t}$ to $m!$ attempts. For example, suppose that the secret message has 5-seconds with 40K wavelet coefficients, then the total number of possible combinations is $(40K)!$ If the secret message encompasses only vowels (the most simple case), $p=5$ and suppose that $n=8$. Then, in the best scenario the lowest number of attempts is $(5!)^{8 \cdot 5} = (5!)^{40} = 1.46 \cdot 10^{83}$ which is long enough to be discovered. In the current example, an eavesdropper needs to test between $1.46 \cdot 10^{83}$ and $(40K)!$ attempts. [76]

4.4.2. Cipher-text only attack

It is a well-known method of cryptanalysis in which the aim is to discover the key based on the envelope of the spectrum of the scrambled speech signal [20], [21]. In classical approaches, the spectrum of the scrambled speech signal has several discontinuities, and therefore, the *key* can be revealed if the blocks into the spectrum are relocated to form a smooth envelope. This technique is useful in scrambling schemes of *FDS*. Since in the proposal the scrambled speech signal looks like an intelligible speech signal with a smooth envelope of the spectrum, the *key* is not revealed with this type of attack. [76]

4.4.3. Statistical attack and perfect secrecy

According to Shannon's theory, a cryptosystem has perfect secrecy if the number of secret messages is equal to the number of enciphered messages and the relationship between them is one-to-one [75]. In our case, because the length of the secret's coefficients is equal to the length of the *key* and each *key* produces a different scrambled speech signal, our proposal has perfect secrecy. It is worth noting that although there are several speech signals that can sound with the same plain-text, the mapping process between the secret message and the scrambled speech signal is one-to-one. In other words, the message-space length is exactly equal to the key-space length and the scrambled-space length. In terms of confusion and diffusion, if an eavesdropper intercepts the scrambled speech signal and he/she would have enough time to try all possibilities, he/she does not have certainty of which of them is the right secret message, because the distribution probability of the message-space is uniform. Through adaptation, there is not a prior relationship between the secret message and the *key*, or between the secret message and the scrambled speech signal. [76]

4.5. Summary

The current chapter is summarized as follows:

- (i) The ability of adaptation of speech signals has been used to scramble speech signals in wavelet domain. The system can be viewed as a special case of Time-Frequency Scrambling. Unlike traditional approaches, the *key* is not an input of the system; it is generated in the adaptation from the secret message to the target speech signal.
- (ii) The derangement level and the displacement value of the adaptation process give a residual intelligibility very low. Therefore, the most important feature in a scrambling system is satisfied.
- (iii) The effort required to obtain the secret key in the scrambling module is low, with a complexity of $O(m \log m)$, where m is the secret's coefficients length. Nevertheless, the effort to find the key by an eavesdropper is hard because it is up to $m!$ Consequently, it is concluded that the system overcomes the brute force attack.
- (iv) The system works with perfect secrecy because the key-space length is equal to the secret-space length, there are as many secret messages as scrambled speech signals, and the mapping between inputs and outputs is one-to-one.

5. Speech steganography using Efficient Wavelet Masking

This chapter shows two schemes of speech steganography which take advantage of the ability of adaptation of speech signals and the masking property of the Human Auditory System (HAS). The schemes are known as Efficient Wavelet Masking (EWM) and improved-EWM (iEWM). The first one is optimized in terms of statistical transparency and the second one in terms of robustness. Both schemes can hide a speech signal into another one of the same time-scale.

5.1. Introduction

The second analyzed method of speech hiding is steganography. The main difference related to scrambling is that the secret message is hidden into the host signal instead of modifying the secret message so it resembles a target speech signal (which can be a legible or non-legible speech signal).

The output signal, known as the stego signal, must be perceptually equal to the host signal with the purpose of not generating suspicion about the existence of the secret message (transparency). The higher the number of bits of the secret message that are hidden, the higher is the hiding capacity. Additionally, a robust stego signal overcomes signal manipulations like lossy compression, re-quantization or resampling, among others. All of the features (transparency, hiding capacity and robustness) are known as the “magic triangle” and there is a strong relationship among them; when one increases at least one of the others decreases [5]. Therefore, it is not possible to simultaneously optimize the three above features.

In the literature the following techniques of speech hiding are well known: Least Significant Bit (LSB) substitution, Frequency Masking (FM), Spread Spectrum (SS) and Shift Spectrum Algorithm (SSA) [3], [27], [79], [80]. In the case of LSB, the least significant bits of the host signal are replaced with the bits of the secret message [23], [24], [81]-[84]. If the number of replaced bits per sample increases, then HC and robustness increase too, but the transparency decreases. In the second scheme, FM, every coefficient of the secret message is hidden into one coefficient of the host signal if the masking criterion has been previously satisfied [35], [36], [85]. Since the hiding process follows masking criteria, the transparency and robustness are satisfied, however the hiding capacity can become lower than in LSB. In the third case, SS, the bandwidth of the secret message is spread into the bandwidth of the host signal [26],

[86]. Finally, in the case of SSA, the bandwidth of the secret message is delayed to the highest subband of the bandwidth of the host signal [26], [87], [88]. Both SS and SSA have the smallest hiding capacities, but in some cases they have the highest value of transparency.

Since none of the classical schemes of speech steganography have a good enough trade-off among transparency, hiding capacity and robustness, a novel scheme of speech-in-speech hiding based on the ability of adaptation of speech signals is proposed. This scheme is known as Efficient Wavelet Masking (EWM). An improved version of EWM in terms of robustness is known as iEWM. Both EWM and iEWM have the same hiding capacity in terms of time-scale.

The rest of the chapter is organized as follows. Section 5.2 explains the proposed scheme known as Efficient Wavelet Masking and Section 5.3 shows its results in terms of statistical transparency. Section 5.4 explains the scheme iEWM and Section 5.5 illustrates the results in terms of robustness. The chapter is summarized in Section 5.6.

5.2. Efficient Wavelet Masking (EWM)

In nature, one of the best examples of adaptation is the chameleon which adapts to the surrounding environment, changing its color, to become "imperceptible" and not be detected by enemies. In a similar way, the best form to hide data is by adapting them to the host signal. Because the main purpose of any steganographic model is that the secret message seems "imperceptible" into the host signal, the ability of adaptation of speech signals is used as the core of the proposed speech-in-speech hiding scheme known as Efficient Wavelet Masking [56].

Like FM, EWM is based on the masking property of the HAS but the main difference lies on the "efficient" masking of the secret message due to the principle of adaptation.

The scheme encompasses two modules: the embedding module at the transmitter and the extraction module at the receiver. They are explained as follows.

5.2.1. Embedding module

It is carried out by the following steps: decomposition and scaling, efficient sorting, indirect LSB replacement, reconstruction and post-scaling (Figure 5.1).

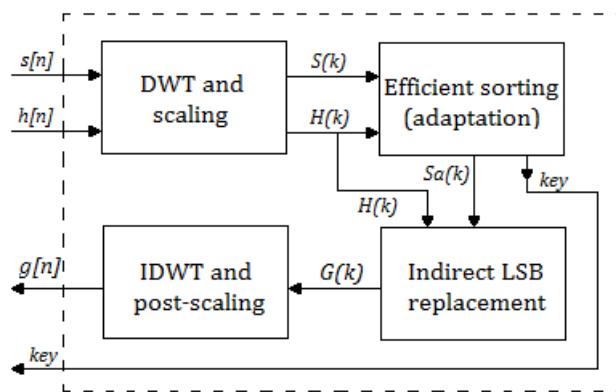


Figure 5.1. EWM: flowchart of the embedding module

The steps are explained as follows:

- (i) *Decomposition and scaling*: both the secret message and the host signals are decomposed by using the DWT.

$$s[n] \xrightarrow{DWT} S(k) \quad k \in Z \quad (5.1)$$

$$h[n] \xrightarrow{DWT} H(k) \quad (5.2)$$

Where $s[n]$ is the secret message, $h[n]$ is the host signal, $S(k)$ is the group of secret's coefficients and $H(k)$ is the group of host's coefficients.

Then, the secret's coefficients are attenuated -12dB under the dynamic range of the host's coefficients. If the host's coefficients are quantized to 16 bits in the range $[-2^{15}+1$ to $2^{15}-1]$, then, the secret's coefficients are quantized to 14 bits in the range $[-2^{13}+1$ to $2^{13}-1]$.

- (ii) *Efficient sorting (adaptation)*: the secret's coefficients are relocated so that they resemble the host's coefficients. The original and final places of the secret's coefficients are used to build the *key* (in a way similar to that explained in Chapter 4).
- (iii) *Indirect LSB replacement*: once the adapted-secret's coefficients have been obtained, the division between every pair of coefficients (from host's coefficients and adapted-secret's coefficients) is calculated, according to:

$$div(k) = \frac{S_a(k)}{H(k)} \quad (5.3)$$

Where $S_a(k)$ is the group of adapted-secret's coefficients.

Since the dynamic range of the secret's coefficients is a quarter of the dynamic range of the host's coefficients, it is expected that the division is close to 0.25; nevertheless, it could be higher. Then, the value of *div* is normalized so that it can be represented by 5-bits, as follows:

$$Pd(k) = 31 * \frac{div(k)}{\max(div(k))} \quad (5.4)$$

Where Pd is called as the Percentage data and $\max(\cdot)$ is the maximum function. Once Pd has been obtained, this is hidden into the 5-LSBs of the host's coefficient, according to:

$$St(k) = \left\{ \left\lfloor \frac{H(k)}{2^5} \right\rfloor * 2^5 \right\} + Pd(k) \quad (5.5)$$

Where $St(k)$ is the group of stego's coefficients and $\lfloor \cdot \rfloor$ is the floor function. For example if $H(1) = 3455 = 0000110101111111_b$ and $Pd(1) = 20 = 10100_b$, then $St(1) = \{ \lfloor 3455/32 \rfloor * 32 \} + 20 = 3444 = 0000110101110100_b$. It is equal to replace the 5-LSBs of $H(1)$ with $Pd(1)$ and it is obtained $St(1) = 0000110101110100_b$.

The advantage of using an indirect substitution form is that the number of bits replaced is less than in a direct form. Although only five bits are replaced in every wavelet coefficient, a speech signal of 14 bits is indirectly hidden into a speech signal of 16 bits. For this reason, it is expected that the transparency in the stego signal is higher than in other schemes.

- (iv) *Reconstruction and post-scaling*: in the last step, the stego's coefficients are reconstructed by using the IDWT with the same wavelet base of the first step.

$$G(k) \xrightarrow{IDWT} g[n] \quad (5.6)$$

Where $st[n]$ is the stego signal in time domain. Finally, the signal is set in the dynamic range of $[-1 \ 1]$.

5.2.2. Extraction module

The secret *key* and the stego signal are the inputs of the module, while the output is the recovered secret message. This module is constituted by the following subsystems: decomposition and scaling, recovering, reconstruction and post-scaling (Figure 5.2).

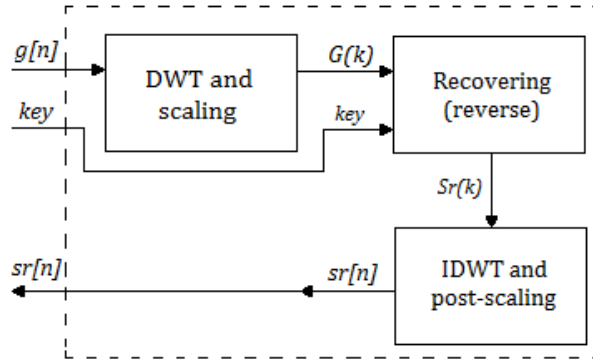


Figure 5.2. EWM: flowchart of the extraction module

The steps are explained, as follows

- (i) *Decomposition and scaling*: the stego signal is decomposed by using the DWT.

$$g[n] \xrightarrow{DWT} G(k) \quad (5.7)$$

- (ii) *Recovering*: the 5-LSBs of the stego's coefficients are extracted, according to:

$$Pd(k) = G(k) - \left\{ \left\lfloor \frac{H(k)}{2^5} \right\rfloor * 2^5 \right\} \quad (5.8)$$

Then, the adapted-secret's coefficients are obtained by the multiplication between the stego's coefficients and the *percentage data*, *Pd*, as follows:

$$S_a(k) = Pd(k) * S(k) \quad (5.9)$$

Finally, the adapted-secret's coefficients, *Sa(k)*, are relocated according to the *key* and then the recovered-secret's coefficients, *Sr(k)*, are obtained. At

this point, the adaptation process is reversed and the output, $Sr(k)$, has the same behavior of the secret's coefficients.

- (iii) *Reconstruction and post-scaling*: the recovered-secret's coefficients are reconstructed by using the IDWT.

$$Sr(k) \xrightarrow{IDWT} sr[n] \quad (5.10)$$

Where $sr[n]$ is the recovered-secret message.

A post-scaling is applied so that their dynamic range is [-1 1].

5.3. Performance of EWM

The statistical transparency is considered in this thesis to establish the quality of the stego signal and its robustness against some steganalysis techniques. The objective in any steganalysis test is to find signs about the existence of a secret message into the speech signal. Most steganalysis methods use an intelligent system which is trained with statistics of host and stego signals. Then, the speech signal is analyzed and it is classified as a host or stego signal. In this work, the stego signals are tested by three steganalysis methods to measure the difference between the statistics of the host signal and the stego signal and determine if this is smaller than a threshold. If the criterion is satisfied, the stego signal does not create suspicion about the existence of the secret message and it can be transmitted in a secure channel. If this difference is large for any of the statistics, the stego signal could be identified by an expert system and therefore the message will be vulnerable.

Three domains have been used in this work to assess the statistical transparency: time domain, frequency domain, and wavelet domain. In the time domain, the test is based on the log function of the speech signal proposed by [12]. In the frequency domain, the test is based on the second-order derivative of the audio signal proposed by [15]. Finally, in the wavelet domain, the test is based on the statistical analysis of wavelet subbands proposed in [17]. The statistical analysis is carried out by obtaining the fourth first moments of these functions: average (μ), variance (σ^2), skewness (sk) and kurtosis (k). The difference in the statistics is estimated in the five methods considered in this chapter: LSB, FM, SS, SSA, and the proposed one, EWM. Additionally, five hiding capacities have been taken into account in order to evaluate the performance of every method against the size of the secret message.

5.3.1. Statistical transparency

The performance of the five schemes against three steganalysis tests is tested. In every experiment, the differences between the statistics of the host and the stego signals are calculated. To organize the experiments, five values of hiding capacity (in terms of the time-scale: 25%, 33%, 50%, 75% and 100%) are analyzed. The capacity corresponds to the percentage of the time-scale of the secret message in relation to the time-scale to the host signal; if both signals have the same time-scale, the hiding capacity is 100% even if the number of replaced LSBs is not the same. The SSA scheme is only tested with its maximum capacity, 25%, SS scheme is tested with 25% and 33%; EWM, LSB and FM are tested with the five capacities. One host signal and five secret messages are used, one for every hiding capacity. The length of the host signal is 2-seconds and the frequency sampling of each one is 8 KHz.

Table 5.1. Signals for HC=100%: Input signal & Difference signal. [56]

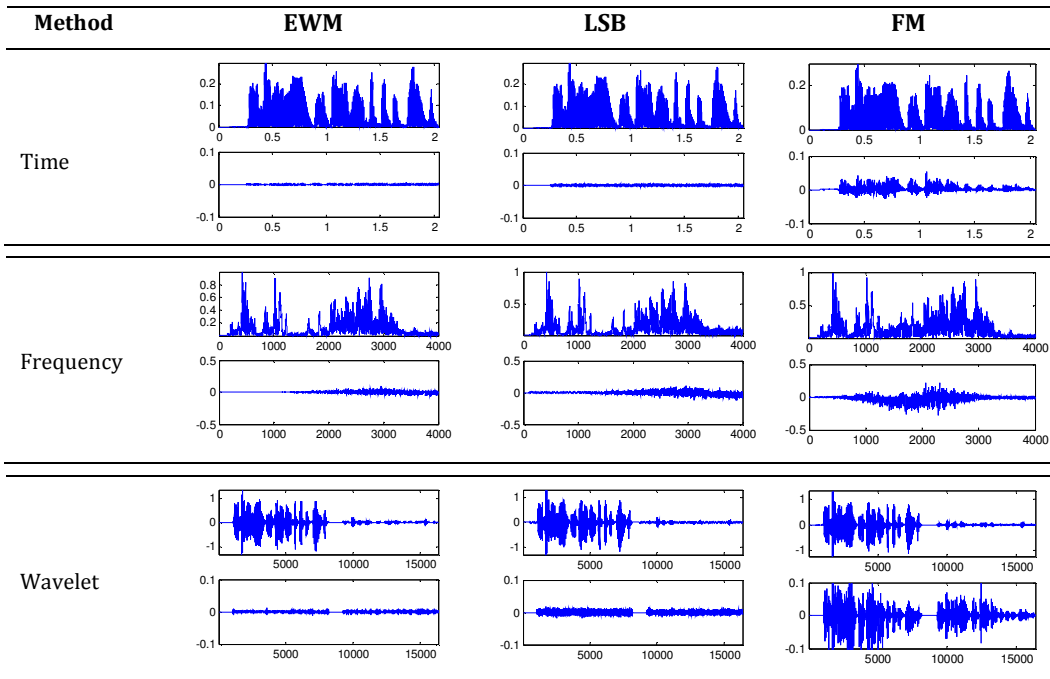


Table 5.1 illustrates some differences among the three steganalysis methods with HC=100%. In every row two signals are plotted, the host signal (in logarithmic form, spectrum or wavelet decomposition) and the difference signal (between the host signal and the stego signal in the selected domain). It is confirmed that with the EWM scheme the transparency is better than with other methods, such as LSB and FM. This can be easily seen in the steganalysis test in wavelet domain.

Now, Figures 5.3 to 5.5 show the difference between the statistics of the host signal and the statistics of the stego signal, for every steganalysis test and steganography method.

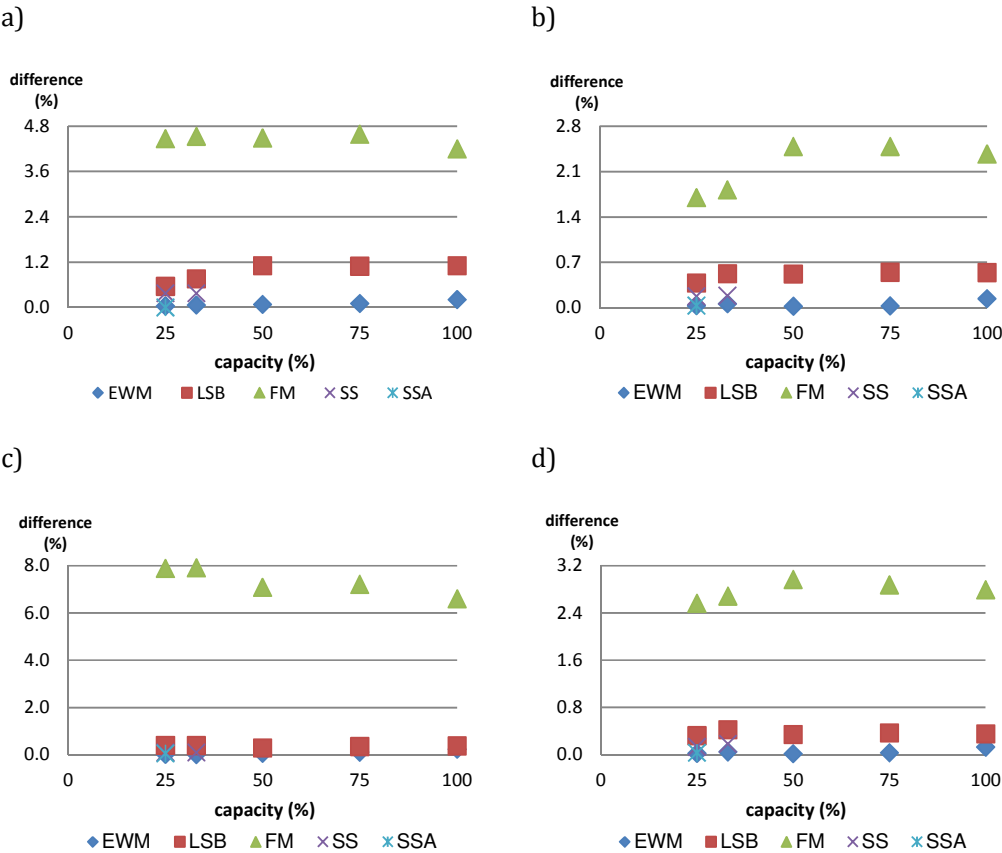


Figure 5.3. Difference (%) in the temporal steganalysis test: a)average, b)variance, c)skewness, d)kurtosis. [56]

In the steganalysis test in time domain, the maximum difference (7.8%) belongs to the skewness in the FM scheme. EWM scheme has a result similar to SSA in low capacities, while in high capacities, the best performance corresponds to the EWM scheme. The difference in EWM is ever lower than 0.5%, while in the LSB scheme is lower than 1%. In this test, the schemes didn't give any sign about the existence of the secret message, because the statistics of the stego signals were very similar to the statistics of the host signals. In other words, it is difficult for a classifier to identify the stego signals (from the current schemes) based on the statistics of the logarithm of the speech signal.

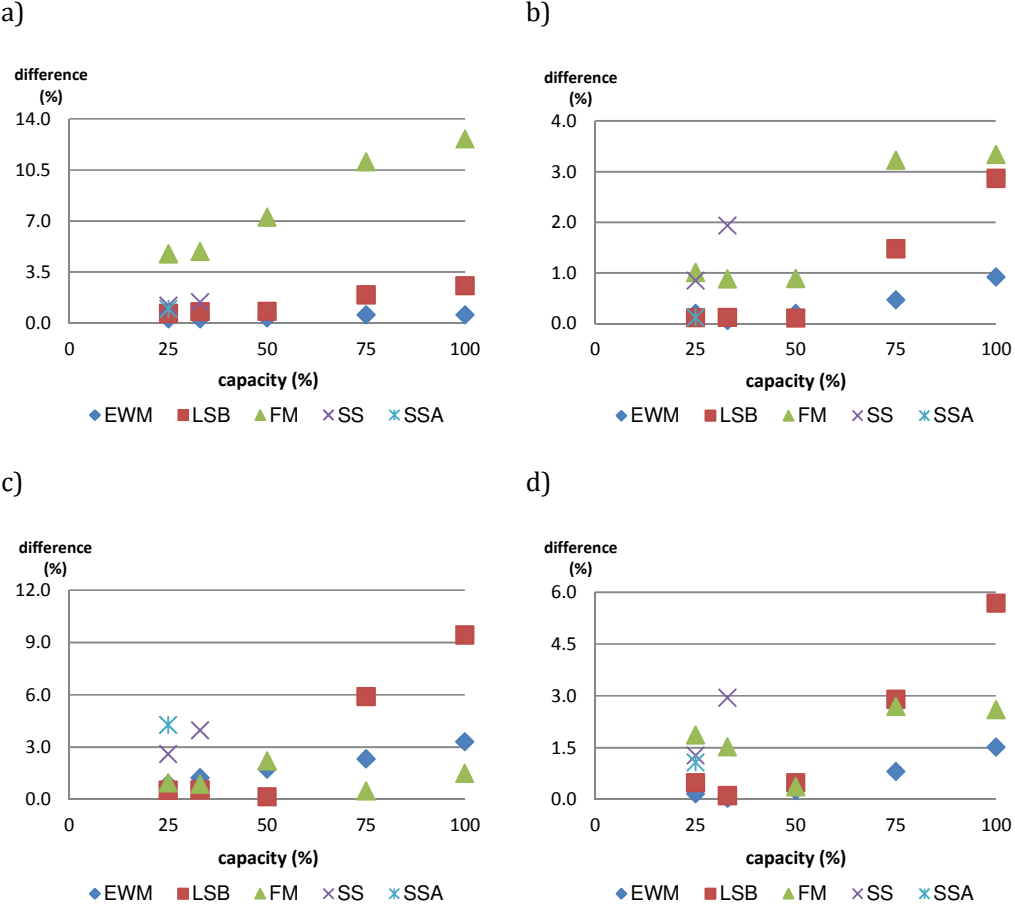


Figure 5.4. Difference (%) in the frequency domain steganalysis test: a) average, b) variance, c) skewness, d) kurtosis. [56]

According to Figure 5.4, for low capacities (25% and 33%) EWM has the smallest differences in 75% of the statistics, while LSB has the smallest in 25% of the statistics. For high capacities, EWM's differences are lower than 3.5%; LSB's differences are lower than 10%, while FM's differences are lower than 13%. It means that EWM is the best scheme both in low and high capacities, in terms of the transparency in frequency domain.

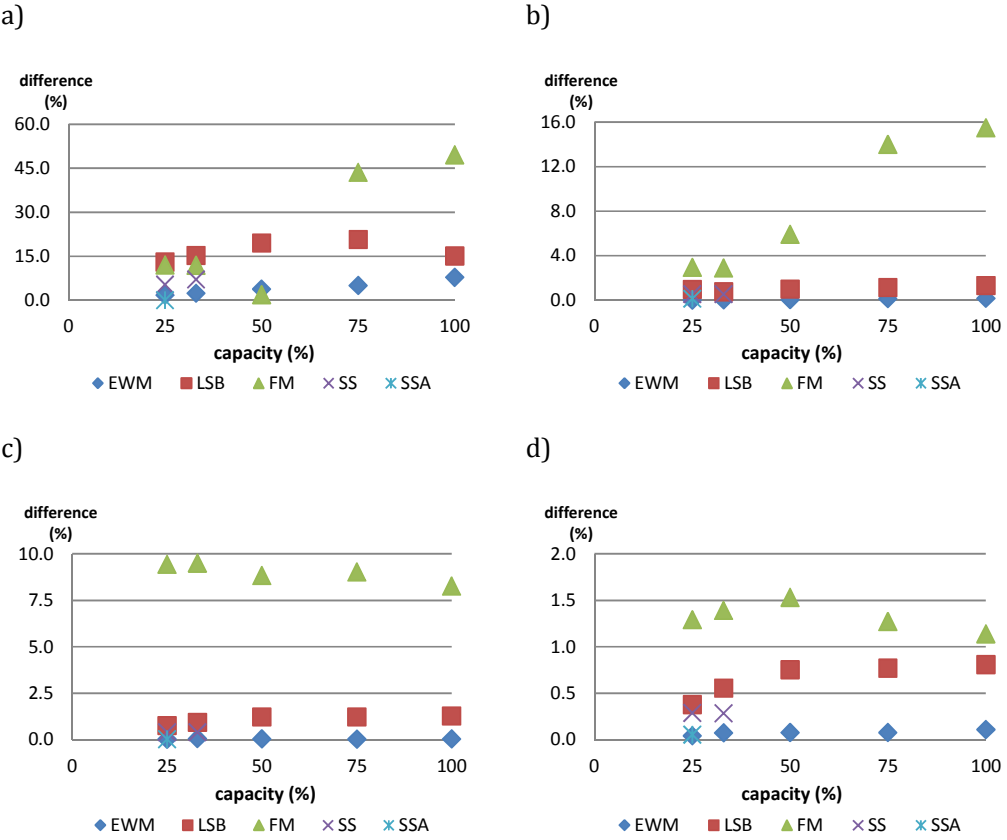


Figure 5.5. Difference (%) in the wavelet steganalysis test: a) average, b) variance, c) skewness, d) kurtosis. [56]

Finally, the steganalysis test in wavelet domain is presented in Figure 5.5. This test gives the highest difference in the statistics between the host and the stego signal. It implies that if we guarantee a small difference between the wavelet representation of

the host and the stego signals, then the stego signals should pass any steganalysis method based on time, frequency or time-frequency domain. According to the results, the worst case corresponds to the FM scheme, because its difference became 50%, and the most stable scheme is EWM, since the difference increases very little from a capacity to another and 95% of its statistics are lower than in the other schemes.

Summarizing, it has been found that the most stable model is EWM, because even if the size of the secret message increases with the hiding capacity, the maximum difference between the statistics of the host and the stego signals remains below 10%. The remaining methods either increase its error with the capacity or between tests.

5.3.2. Hiding Capacity and other quality parameters

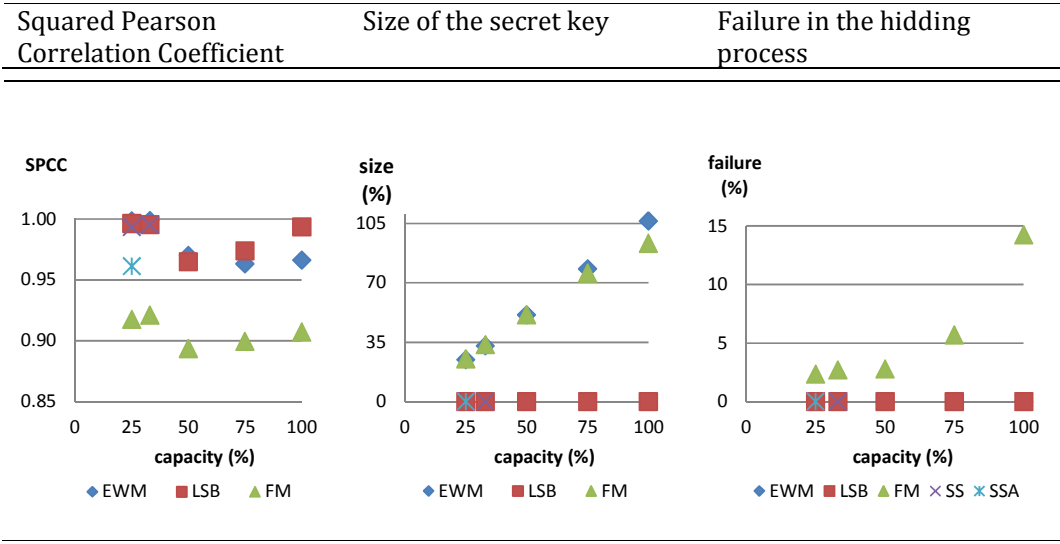
In addition to the statistical transparency, the quality of the recovered secret message plays an important role in any steganalysis scheme. In Table 5.2, (first column), the correlation coefficient in every scheme is illustrated. Every value of hiding capacity has one pair of host signal and secret message.

In low capacities (HC=25%, 33%) the performance of the EWM, LSB, and SS schemes is similar and it is better than the performance of the SSA and FM schemes. In high capacities (HC \geq 50%), LSB is significantly better than FM and slightly better than EWM. Summarizing, in terms of quality in the recovered secret message, the best scheme is the LSB and the worst is the FM scheme; while EWM scheme has the second position.

Second, it is analyzed the relation between the hiding capacity and the size of the secret key. The LSB, SS and SSA schemes do not need a secret key for recovering the secret message, but EWM and FM use a secret key to keep the positions of the secret coefficients. In the second column of Table 5.2, the percentage of the size of the secret

key in relation to the size of the host signal is shown. The size of the EWM scheme is slightly higher than the FM, and this value is proportional to the hiding capacity. In an ideal way, these sizes should be equal, but the difference is due to the failure to hide information in FM. In the third column of Table 5.2, the percentage of failure in data hiding in every scheme demonstrates that FM does not guarantee that all of the secret coefficients can be hidden into the host signal.

Table 5.2. Performance in other selected quality parameters, for HC=25%, 33%, 50%, 75% and 100%. [56]



5.4. Improved Efficient Wavelet Masking

The EWM is an “optimized” scheme in terms of the statistical transparency of the stego signal but it is not good enough in terms of the robustness against signal manipulations. Since only 5-LSBs of the host’s coefficients are modified, the stego signal does not tolerate small changes (e.g. ~0.1%) in its amplitude and then the bits related to Pd can be lost. Therefore, an “optimized” scheme in terms of robustness based on the EWM is presented in this section. Nevertheless, there is a trade-off among the transparency, the hiding capacity and the robustness and it is expected that the new scheme, the improved-EWM (or iEWM), is less transparent than its predecessor.

The core of iEWM is the ability of adaptation of speech signals and the selective Significant-Bit-to-Hold (SBH). With the purpose of increasing the robustness, the number of replaced LSBs depends on the amplitude of the host’s coefficient instead of a fixed number of LSBs of its predecessor, the EWM scheme. The larger the host’s coefficient, the higher the number of replaced LSBs. Since the larger coefficients would hide a higher number of bits, it is expected that the robustness of the stego signal will improve.

The embedding and extraction modules are described as follows.

5.4.1. Embedding module

The purpose of this module is to hide a secret signal into the host signal. The procedure is illustrated in Figure 5.6. It contains the following steps: decomposition, efficient sorting and scaling, selective Significant-Bit-to-Hold, reconstruction and post-scaling. Every step is detailed as follows:

- (i) *Decomposition and scaling*: both signals, the host and the secret one, are decomposed by using the Discrete Wavelet Transform. To obtain the same

number of wavelet coefficients, the number of the samples of the signals and the wavelet base used in the decomposition must be equal in both cases. The relation between the input and the output is defined by:

$$s[n] \xrightarrow{DWT} S(w) \quad (5.11)$$

$$h[n] \xrightarrow{DWT} H(w) \quad (5.12)$$

Where $s[n]$, $h[n]$, $S(w)$, $H(w)$ are the secret signal, the host signal, the secret's coefficients and the host's coefficients, respectively. Unlike the EWM scheme, the secret signal is not attenuated by -12dB in relation to the host signal.

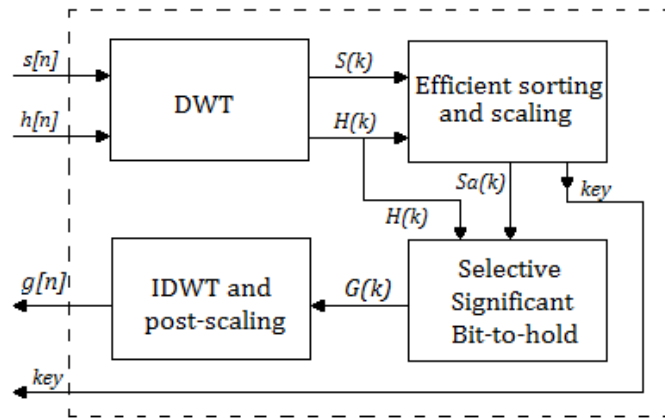


Figure 5.6. Block diagram of the improved-EWM embedding module

- (ii) Efficient sorting and scaling: the secret's coefficients are relocated so that they resemble the host's coefficients. It uses the ability of adaptation of the speech signals proposed in Chapter 3. Once the coefficients have been relocated, the dynamic range of the adapted-secret's coefficients and the host's coefficients are modified to work with integer values in the next step, in the dynamic range $[-2^{15}+1 \ 2^{15}-1]$. The design works with a resolution of 16-bits, but the scheme can be easily extrapolated.

Like the EWM scheme, the original positions of the secret's coefficients are kept in a 1D-array. With the original and new positions after the sorting process, the secret *key* is created.

- (iii) *Selective Significant Bit-to-hold*: the aim of this block is to hold some of the most significant bits (MSBs) of the host's coefficients and replace the rest of them. Since every host's coefficient has specific amplitude, the number of replaced bits depends on its amplitude and the selective Significant-Bit-to-Hold (SBH). The larger the host's coefficient, the higher the number of replaced LSBs. Without loss of generality, assume that the minimum number of bits to represent the host's coefficient is n , and then the number of replaced bits is $n-SBH$. If SBH is high, only a few bits are replaced and the transparency of the stego signal would be high, otherwise the number of replaced bits is large and the transparency would be low. Nonetheless, the lower the number of replaced bits, the lower the robustness against signal manipulations. The trade-off between the robustness and the transparency related to the value of SBH will be discussed in section 5.5.

The stego's coefficient is calculated from the host's coefficient, the adapted-secret's coefficient, n and SBH, according to:

$$G(w) = \left\{ \left\lfloor \frac{H_N(w)}{2^{n-SBH}} \right\rfloor * 2^{n-SBH} \right\} + \left\lfloor \frac{S_a(w)}{2^{SBH+1}} \right\rfloor \quad (5.13)$$

Where $G(w)$ is the stego's coefficient. It is noticed that the adapted-secret's coefficient, $S_a(w)$, is attenuated by the factor $1/2^{SBH+1}$. For example, if $H(1)=14102$, $S_a(1)=12800$, $n=14$ and $SBH=4$, the stego's coefficient is calculated as $G(1)=\left\{ \left\lfloor 14102/210 \right\rfloor * 210 \right\} + \left\lfloor 12800/25 \right\rfloor = 13312 + 400 = 13712$. In binary format, $H(1)_b = \mathbf{11011100010110}$, $\lfloor S_{S_N}(1)/25 \rfloor_b = 0110010000$ and

$G(1)_b = \mathbf{11010110010000}$. In the current example, the 10-LSBs of the host's coefficient have been replaced.

- (iv) *Reconstruction*: the stego signal, $g[n]$, is obtained from the stego's coefficients by using the Inverse Discrete Wavelet Transform (IDWT), as follows:

$$G(w) \xrightarrow{IDWT} g[n] \quad (5.14)$$

Finally, the dynamic range of the stego signal is set in the interval $[-1 \ 1]$.

5.4.2. Extraction module

The aim of the extraction module is to obtain an estimate of the secret message from the stego signal. The steps are plotted in Figure 5.7.

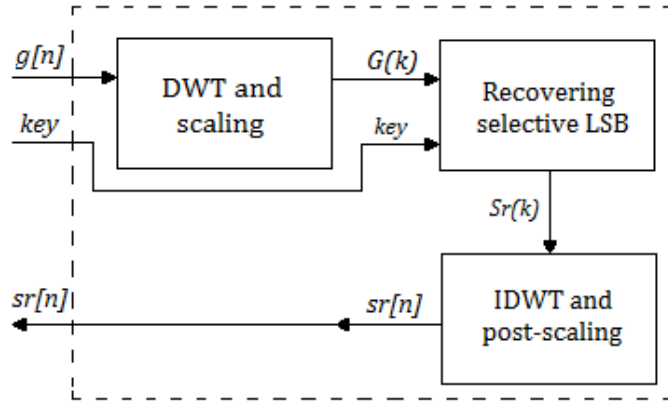


Figure 5.7. Block diagram of the improved-EWM extraction module

It includes decomposition, recovering selective LSB, and reconstruction. They are detailed as follows.

- (i) *Decomposition and scaling*: the stego signal, $g[n]$, is decomposed by using the Discrete Wavelet Transform, as follows.

$$g[n] \xrightarrow{DWT} G(w) \quad (5.15)$$

Where $G(w)$ is the stego's coefficients. The wavelet base is the same used for the embedding module. The stego's coefficients are scaled to obtain integer numbers in the range $[-2^{-15}+1 \ 2^{15}-1]$.

- (ii) *Recovering selective LSB*: the purpose of this step is to recover the bits related to the secret's coefficients. Like in the embedding process, the value of SBH is used to calculate the output, according to:

$$S_a(w) = \left[G(w) - \left\{ \left\lfloor \frac{G(w)}{2^{n-SBH}} \right\rfloor * 2^{n-SBH} \right\} \right] * 2^{SBH+1} \quad (5.16)$$

Where $S_a(w)$ is the group of adapted-secret's coefficients. In this case, the value of n is the minimum number of bits used to represent the stego's coefficient. For example, if $G(1)=13712$, $SBH=4$ and $n=14$, it is obtained that $S_a(1)=[13712-\lfloor 13712/2^{10} \rfloor * 2^{10}] * 2^5 = 12800$. In the current example, the result is equal to extract the 10-LSBs of the stego's coefficient and then scale it according to 2^{SBH+1} . It is noticed that the result of S_a is the same as in the hiding process presented in the current section.

Finally, the adaptation process is reversed according to the *key*. At the output, the recovered-secret's coefficients, $Sr(w)$, are obtained.

- (iii) *Reconstruction and post-scaling*: the recovered-secret's coefficients are reconstructed by the IDWT and the recovered-secret's message is obtained, as follows:

$$Sr(k) \xrightarrow{IDWT} sr[n] \quad (5.17)$$

Where $sr[n]$ is the recovered-secret message. A post-scaling is applied to set the dynamic range of the signal in the interval $[-1 \ 1]$.

It is worth noting that the main difference between EWM and iEWM lies on the LSB substitution step, in the first it uses an indirect fixed substitution and in the later it uses a direct non-fixed substitution based on the SBH criteria.

5.5. Relationship between robustness and transparency of the iEWM

In this section, several tests are conducted to demonstrate the robustness of the improved Efficient Wavelet Masking, iEWM. The speech signals from the Sound Quality Assessment Material (SQAM) are used in the tests. They belong to female and male speakers in English language [58]. Before the signal manipulations, the sounds are re-sampled to 8 KHz, the resolution is preserved in 16 bits, and the Bit Rate (BR) is 128 kbps. The following signal manipulations (attacks) are selected to test the robustness of the proposed scheme: lossy compression, resampling and re-quantization. Firstly, the stego signal is lossy compressed with four Bit Rates ($BR_1=64\text{kbps}$, $BR_2=48\text{kbps}$, $BR_3=32\text{kbps}$ and $BR_4=24\text{kbps}$). Secondly, the stego signal is decimated/interpolated by the factor Q ($Q_1=5/4$, $Q_2=4/3$, $Q_3=5/3$, $Q_4=2$). Thirdly, the stego signal is quantized at 8-bits. The performance of the speech hiding schemes in terms of the statistical transparency and the quality of the recovered secret signal are measured. In the first one, the statistics of the host signal and the manipulated stego signal are taken into account. In the second one, the speech distortion index between the secret signal and the recovered secret signal from the manipulated stego signal is calculated. Thereafter, the iEWM and EWM schemes will be compared through different values of the SBH. Finally, once the SBH has been selected, iEWM is compared to some of the speech-in-speech hiding schemes in order to illustrate the high robustness of the proposed method.

5.5.1. Selecting SBH

The iEWM and its predecessor are compared in order to select an adequate SBH which satisfies the trade-off between transparency and robustness. The tests values are

SBH₁=1, SBH₂=2, SBH₃=4 and SBH₄=6. The objective is to preserve the high transparency while increasing the robustness against standard benchmark attacks.

The selected signal manipulations are analyzed as follows.

Lossy compression: table 5.3 plots the histograms of the logarithm of the host signal and the logarithm of the compressed stego signals (from EWM and iEWM). It is shown for the following Bit Rates: BR₁=64, BR₂=48, BR₃=32 and BR₄=24. Since the uncompressed host signal has BR=128, the Compression Ratio (CR) of each case is CR₁=2, CR₂=2.6, CR₃=4 and CR₄=5.3.

Table 5.3. Lossy compression test: statistical transparency

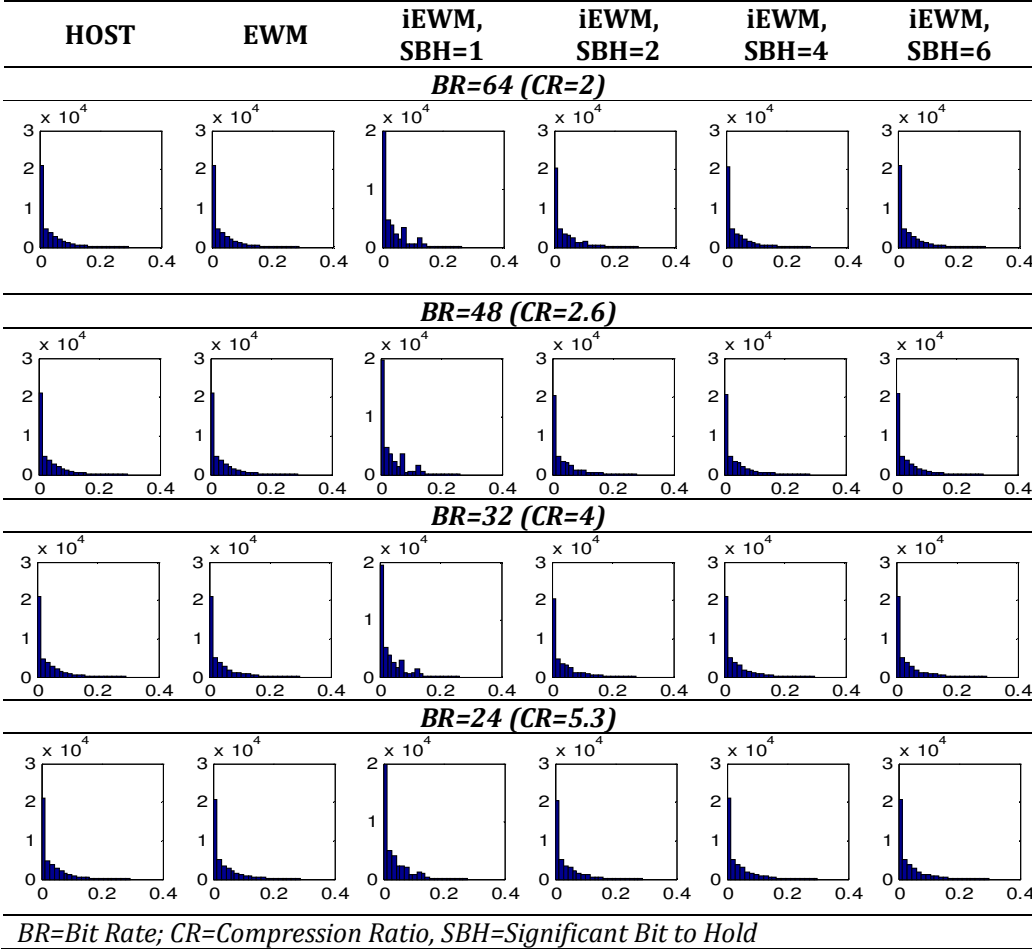


Figure 5.8 shows the speech distortion index of the recovered secret message. The x-axis corresponds to the Compression Rate while the y-axis to the Squared Pearson Correlation Coefficient.

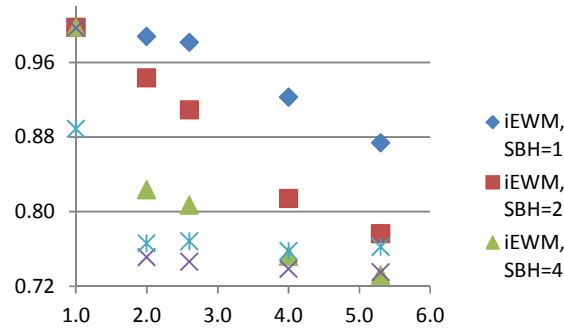


Figure 5.8. Lossy compression test: quality of the recovered secret message

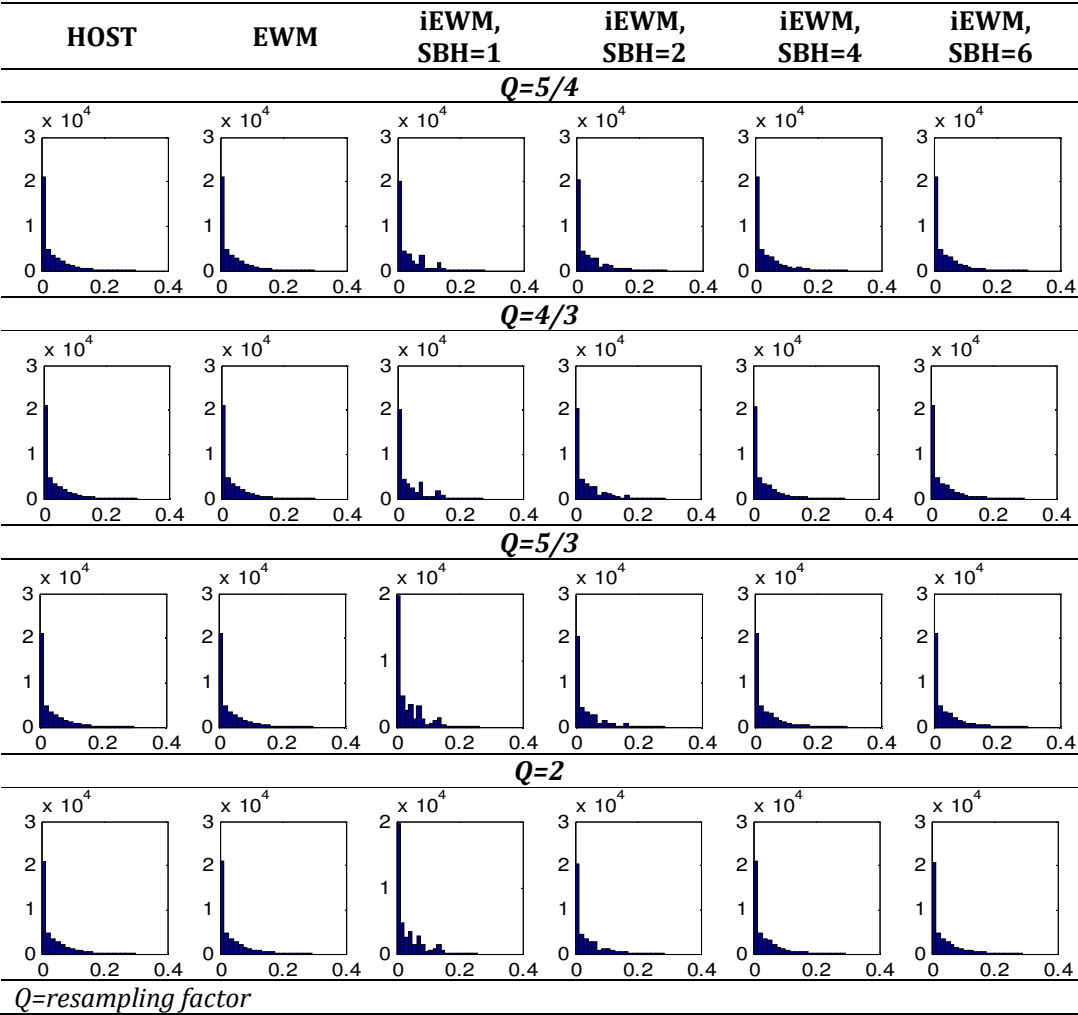
According to the results of Table 5.3 and Figure 5.8, it is noticed that the statistical transparency of the iEWM scheme is close to the EWM when SBH=4 and SBH=6. In the case of SBH=1, the histogram of the compressed stego signal is highly different to the host signal. On the other hand, the higher the CR, the more degraded are the recovered secret signals. Nevertheless, iEWM with SBH<4 demonstrated better performance than the EWM scheme, while the results of SBH=4 are better in three of the five cases. Analyzing the trade-off between statistical transparency and quality of the recovered secret signal, it is found in the current case that SBH=4 has the best relationship.

Resampling: in this attack the sampling frequency, f_c , of the speech signal is modified. Firstly, the signal is decimated and secondly the signal is interpolated, by a factor of Q . The higher the value of Q , the lower is the number of samples after the decimation process. In this test, the stego signals are decimated/interpolated by the

factors: $Q_1=5/4$, $Q_2=4/3$, $Q_3=5/3$ and $Q_4=2$. The histograms of the logarithm of the host signal and the logarithm of the resampled stego signals are illustrated in Table 5.4.

According to Table 5.4, iEWM is close to the host's histogram (and the EWM's histogram) when $SBH=4$ and $SBH=6$. The graphs of $SBH=1$ and $SBH=2$ can give suspicion about the existence of a secret signal.

Table 5.4. Resampling test: statistical transparency



The quality of the recovered secret signal from the manipulated stego signals are shown in Figure 5.9. The x-axis corresponds to the Q factor and the y-axis to the Squared Pearson Correlation Coefficient.

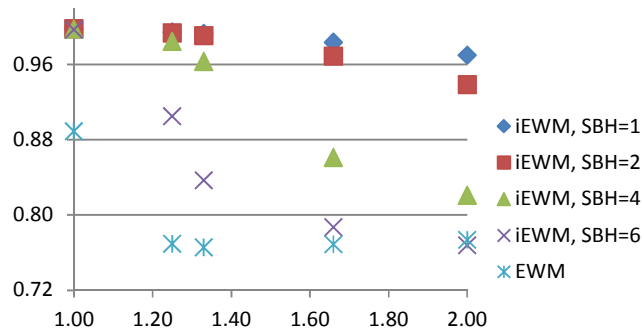
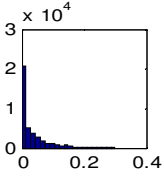
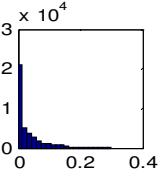
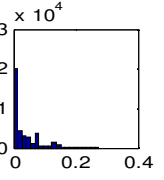
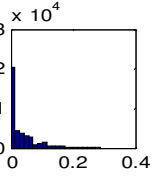
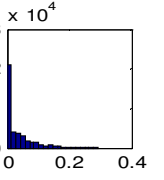
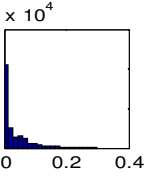


Figure 5.9. Resampling test: quality of the recovered secret signal

Unlike the compression attack, the quality of the recovered secret signal for SBH<6 is always better than in the EWM scheme; only for Q=2, the EWM scheme is better than one of the iEWM cases.

Re-quantization: the quantization of the speech signals is transformed from 16 to 8 bits. The statistical transparency (histogram) and the quality of the recovered secret signal (speech distortion index) are shown in Table 5.5.

Table 5.5. Re-quantization test: statistical transparency & quality of the recovered secret signal

HOST	EWM	iEWM, SBH=1	iEWM, SBH=2	iEWM, SBH=4	iEWM, SBH=6
Histogram					
					
Speech distortion index					
	0.779	0.996	0.993	0.959	0.795

According to Table 5.5, only the histogram from SBH=4 is similar to the host's histogram. On the other hand, the quality of the recovered secret signal is significantly higher in SBH=4 than from the EWM scheme.

Summarizing, the iEWM scheme has the best trade-off between statistical transparency and quality of the recovered secret signal in the three analyzed attacks (lossy compression, resampling and re-quantization) when SBH=4. For this reason, it is suggested to use the iEWM scheme with the above value.

5.5.2. Comparison of the proposed and classical schemes

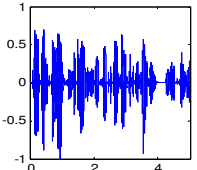
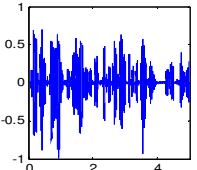
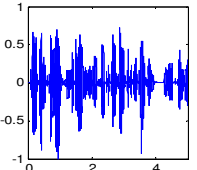
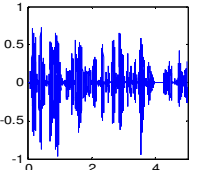
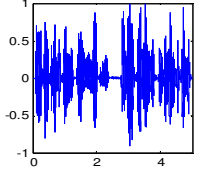
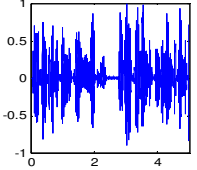
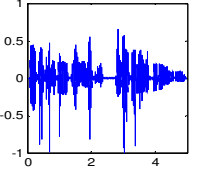
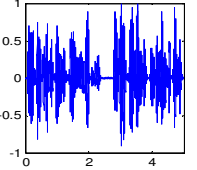
Once the SBH has been selected, the next step is to compare the performance of the proposed scheme with other speech-in-speech hiding methods. The LSB and FM schemes have been selected because they permit to hide a speech signal into another speech signal of the same time-scale, like in the EWM and iEWM ones.

The current test is divided in two parts. Firstly, the statistical transparency and the quality of the recovered secret message are measured before the signal manipulations. Secondly, the same features are taken in account for lossy compression, resampling and re-quantization attacks. A predefined value is used in each case, CR=5.3 (BR=24) in lossy compression, Q=2 in resampling and resolution of 8-bits in re-quantization. The test signals are from the Sound Quality Assessment Material (SQAM). The host signal is a female English record while the secret signal is a male English one. Both of them have a time-scale of five seconds.

Table 5.6 shows the performance of the three analyzed schemes in relation to the quality of the stego signal and the recovered secret signal. The signals in the time domain and the objective measurement parameters are plotted for each scheme. The parameters considered are the statistical transparency as given by the difference between the statistical moments of the logarithm of the host signal and the logarithm of

the stego signal, (esk : difference percentage in the skewness and ek : difference percentage in the kurtosis) and the quality of the recovered secret signal by the speech distortion index, ρ^2 .

Table 5.6. Performance results without signal manipulation

Performance	Original (Host or Secret)	LSB	FM	iEWM
Host signal & Stego signal				
[$esk\%$; $ek\%$]		[0.72%; 0.75%]	[1.99%; 2.59%]	[1.90%; 2.29%]
Secret signal & Recovered secret signal				
ρ^2		0.9962	0.7389	0.9978

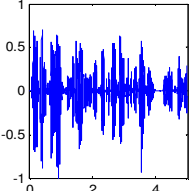
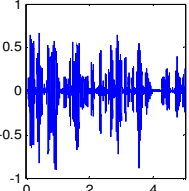
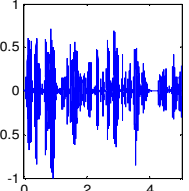
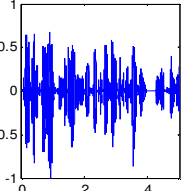
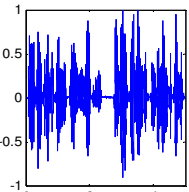
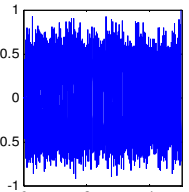
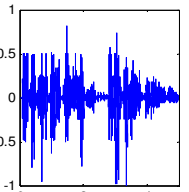
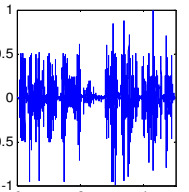
According to Table 5.6, the statistical transparency of the stego signal is better in the LSB scheme, but the quality of the recovered secret signal is slightly better in the iEWM scheme. Nevertheless, in all cases, both esk and ek are lower than 3%. The worst performance corresponds to the FM scheme.

The second part of the current test consists on applying signal manipulations on the stego signals. Firstly, the stego signals are transformed to MP3 format and then are forwarded into its original format. The statistical transparency is measured with the new stego signals and the recovered secret signals extracted from them (Table 5.7).

According to Table 5.7, the statistical transparency is not highly affected by the compression attack and in fact it can improve. On the other hand, the quality of the recovered secret signal from the attacked stego signal is strongly degraded; the worst

case is in the LSB scheme (the recovered signal is not legible). Although the FM scheme gives a moderate quality performance it is noticed that the best result is from the iEWM scheme.

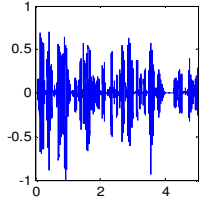
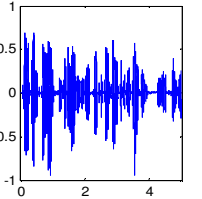
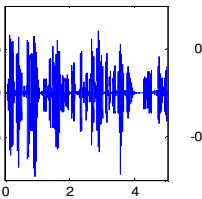
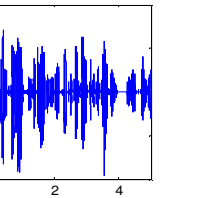
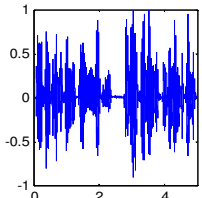
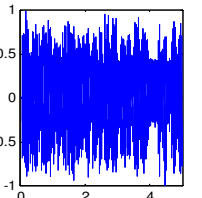
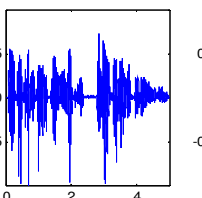
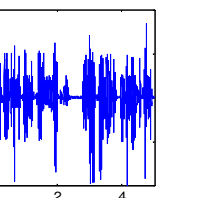
Table 5.7. Performance results: lossy compression attack (BR=24)

Performance	Original (host or secret)	LSB	FM	iEWM
Host signal & Stego signal				
[esk%; ek%]		[2.04%; 1.92%]	[1.51%; 1.56%]	[0.34%; 1.50%]
Secret signal & Recovered secret signal				
ρ^2		0.0009	0.5916	0.7342

Secondly, the stego signals are decimated by half of f_c and then are interpolated by the double of the last f_c . At the end, the attacked stego signal and the host signal have the same number of samples by second. The results are shown in Table 5.8.

The robustness against the resampling attack is higher in the iEWM scheme since the quality of the recovered secret signal is closer to the secret signal and the measurement parameters of the statistical transparency remain below 3%. The FM scheme is the second scheme in terms of quality and statistical transparency. Although the statistical transparency of the LSB scheme is higher than for the others methods, the secret message cannot be recovered if the stego signal has been manipulated. On the other hand, it is remarkable that both FM and iEWM schemes give better results than those obtained in the compression attack.

Table 5.8. Performance results: resampling attack (Q=2)

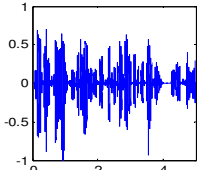
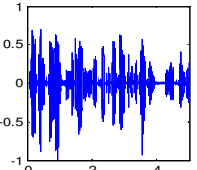
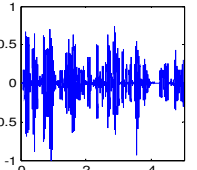
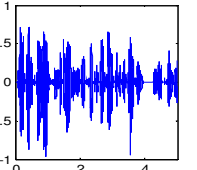
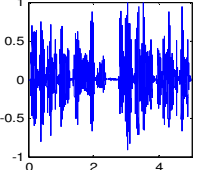
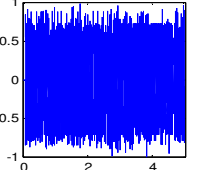
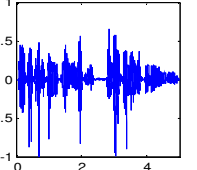
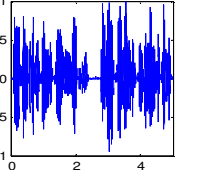
Performance	Original (Host or Secret)	LSB	FM	iEWM
Host signal & Stego signal				
[esk%; ek%]		[0.16 %; 0.12 %]	[2.52%; 3.58%]	[1.99%; 2.66%]
Secret signal & Recovered secret signal				
ρ^2		0.0059	0.7365	0.8049

Finally, we tested the stego signals with the re-quantization attack. The resolution of the stego signals is transformed from 16 to 8 bits. Once the stego signal has been re-quantized, the secret signal is extracted. Again, the statistical transparency of the attacked stego signal and the quality of the recovered secret signal is measured. Table 5.9 shows the results.

According to Table 5.9, the iEWM scheme has a good robustness against the re-quantization attack; its recovered secret signal is very closer to the original one. Together with that, the statistical transparency is high since the *esk* and *ek* parameters are under 3%. Unlike the LSB scheme, the FM scheme permits to recover the secret signal with a moderate quality index.

Summarizing, in the three studied attacks the iEWM scheme has a better robustness than the LSB and FM schemes. The FM scheme permits to recover the secret signal with a low to moderate quality index while the LSB scheme does not allow it.

Table 5.9. Performance results: re-quantization attack (resolution=8-bits)

Performance	Original (Host or Secret)	LSB	FM	iEWM
Host signal & Stego signal				
[esk%; ek%]		[0.51 %; 0.62 %]	[2.35%; 2.93%]	[1.92%; 2.25%]
Secret signal & Recovered secret signal				
ρ^2		0.0052	0.7364	0.9537

5.6. Summary

Two schemes of speech-in-speech hiding have been proposed with the following characteristics:

- (i) Both schemes, EWM and iEWM, are based on the ability of adaptation of speech signals taking in advantage the masking property of the HAS. The embedding and extraction processes are in wavelet domain.
- (ii) The first one, EWM, uses an indirect LSB substitution based on a parameter, Pd , which relates the amplitude of the host signal with the amplitude of the adapted-speech signal. The 5-LSBs of the host's coefficients are replaced with the parameter Pd .
- (iii) Since only 5 bits of the host's coefficients are changed in the embedding process, the transparency of the EWM is higher than in other schemes such as LSB and FM (with 8-bits of substitution).
- (iv) The maximum hiding capacity of EWM is significantly higher than in SS and SSA and equal to LSB and FM.
- (v) The weakness of EWM is the low robustness against signal manipulations. However, if the stego signal is not manipulated, the recovered secret message is highly similar to the original secret message. It has the same plain-text, intonation, rhythm and gender of the speaker.
- (vi) Unlike EWM, iEWM uses direct LSB substitution. In this case, the adapted secret message is directly hidden into the host signal, in wavelet domain. The number of bits varies according to the amplitude of the host's coefficients and therefore, the higher the amplitude, the higher is the number of replaced bits. Nevertheless, the MSB of the host's coefficients are kept and it is controlled with the parameter Significant-Bit-to-Hold (SBH).

- (vii) According to the results of the tests performed, the most appropriate value of SBH is 4. In this case, the transparency is slightly lower than in EWM but the robustness is significantly better.
- (viii) In terms of robustness iEWM is an advisable scheme because it allows recovering the secret message with better quality than in other schemes such as LSB, FM and EWM.

6. Speech hiding on hardware devices

This chapter shows the design and simulation of a real-time speech hiding scheme on hardware devices. The scheme encompasses wavelet decomposition, sorting unit and reconstruction. The secret's coarse-coefficients are relocated based on a descending order criterion and then they are hidden into the host's coarse-coefficients. The key keeps the original places of the secret's coefficients and this is hidden into the host's detail-coefficients. The advantage of the proposed architecture is that not side information is required to recover the secret message because the key is hidden into the transmitted signal.

6.1. Motivation

The schemes of speech hiding developed in this research work have several advantages in relation to the schemes found in literature, but, they are not suitable for real-time implementation. Since the three schemes (EMM, iEWM, speech scrambling) are based on the ability of adaptation of speech signals, they need to know the entire host signal (or target speech signal) to carry out the adaptation process. Therefore, there is long latency between the original speech signal and the stego (or scrambled) speech signal. Consequently, it is necessary to review the characteristics of the hardware covert communications schemes.

In literature there are some schemes of speech hiding on hardware devices. For example, the authors of [89] use a secret *key* of a Pseudo-Noise (PN) sequence (by performing a XOR operation with the clock signal) to generate an encrypted speech signal (which is like a noise signal and is clearly dissimilar to the original speech one). In [45], the authors use a *key* based on the Euler's numerical solution of chaotic equations to generate the encrypted speech signal. On the other hand, hardware-based speech steganographic schemes use LSB substitution, SS or SSA techniques. In [90] a speech hiding scheme is proposed which uses a SS scheme and a PN sequence. The approach presented in [91] uses a chirp signal to embed the secret message instead of a PN sequence. Summarizing, most of the hardware covert communications schemes use a PN or chaotic sequence to hide the secret message and the *key* it is related to control parameters. These kinds of approaches have two disadvantages: firstly, the secret *key* must be transmitted as side information; secondly, if the control parameters are discovered, the secret message is discovered too.

Taking into account the strengths and weaknesses of the known schemes, a desirable hardware speech hiding scheme should have the following characteristics:

- (i) The *key* should be generated by an adaptive process. A fixed-*key* can be more vulnerable than an adaptive-key because the former depends on initial control parameters and the latter does not. On the other hand, the *key* should be hidden into the transmitted speech signal and therefore side information is not necessary to recover the secret message.
- (ii) The hiding capacity of the SS and SSA schemes can be enhanced if the full wavelet coefficients are used to hide the secret information. Nevertheless, the average number of bits hidden into every wavelet coefficient should not be higher than 8 (to obtain a highly transparent stego signal).
- (iii) Finally, the masking property of the HAS should be applied by frames. It will decrease the latency of the system, so that the system can work in real-time.

The purpose of the current proposal is to supply an embedded speech hiding scheme with higher hiding capacity than the related works and with a self-adjusted and self-contained secret key. The scheme works into a steganographic model.

The rest of the chapter is organized as follows. A brief state of the art of the embedded covert communications systems is presented in Section 6.2. The embedding and extraction modules are explained in Section 6.3 and their hardware design is described in Section 6.4. The main results of the proposed scheme are shown in Section 6.5. Some concluding remarks are provided in Section 6.6 and the references are listed in Section 6.7.

6.2. Real-time, Speech-in-speech hiding scheme

Like the EWM and iEWM schemes the proposed embedded speech-in-speech hiding scheme works with two modules: the embedding module and the extraction module. In the first one, the secret speech signal is hidden into the host speech signal by using an adaptive key. The stego signal transmits both the secret message and the key. In the second one, the secret speech signal is recovered from the stego speech signal. Unlike EWM and iEWM the secret key is not side information and the system works in real-time operation.

The current scheme is LSB-based with the following characteristics:

- (i) The hiding process is carried out in the wavelet domain. Both signals, the speech signal and the host speech signal, are decomposed by using the DWT.
- (ii) Only half of the secret's coefficients are hidden into the host signal. Since the coarse coefficients keep the most relevant energy of the signal, only the coarse-secret's coefficients are hidden and the detail-secret's coefficients are discarded. Therefore, the compression ratio (CR) into the system is two.
- (iii) An adaptive *key* relocates the coarse-secret's coefficients before the hiding process
- (iv) The coarse-host's coefficients hide the relocated coarse-secret's coefficients.
- (v) The *key* is hidden into the detail-host's coefficients. The system does not require side information to recover the secret message.
- (vi) The recovered secret message has the same plain-text of the original secret message but with slightly lower quality.

Table 6.1 shows the nomenclature used in the modules. Each module is described as follows.

Table 6.1. Nomenclature in the speech-in-speech hiding scheme. [92]

Embedding module and extraction module			
Symbol	Definition	Symbol	Definition
h	host signal, time domain	$rkey$	recovered key
hc	host's coarse-coefficient	$rssc$	recovered ssc
hd	host's detail-coefficient	rsc	recovered sc
hcd	hc with delay	rs	recovered secret message
hdd	hd with delay	g	stego signal, time domain
s	secret message, time domain	gc	stego's coarse-coefficient
sc	secret's coarse-coefficient	gd	stego's detail-coefficient
ssc	sorted sc		
key	Key		

6.2.1. Embedding module

In this module a speech signal of 8-bits is hidden into a speech signal of 16-bits of the same time-scale and with m samples. The host message is transformed in wavelet domain and its coarse- and detail-coefficients are obtained, while the secret message is decomposed and its coarse-coefficients are obtained. With the purpose to increase the difficulty to discover the secret message, the secret's coarse-coefficients are relocated by a sorting process. Therefore, the sorted secret's coarse-coefficients are hidden into the host's coarse-coefficients and their original positions (key) are hidden into the host's detail-coefficients. The key is self-adjusted and self-contained in the transmitted signal (stego signal). Since the module is suitable for real-time operation, the sorting block works with N coefficients, with $N \ll m$. Once the secret's coarse-coefficients and the key have been hidden into the host's coefficients, the stego signal is obtained by the wavelet reconstruction of the modified host's coefficients. The procedure is illustrated in Figure 6.1. It contains the following blocks: $dwthost$, $dwtsecret$, $sorting$, $delay$, $idwtstego$. Every block is described as follows.

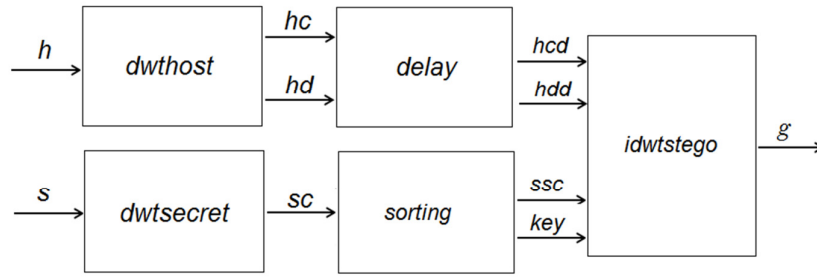


Figure 6.1. Block diagram of the embedding module. [92]

(a) *dwtthost*: the host signal is decomposed by using the Discrete Wavelet Transform (DWT). At the output of this block, the host's coarse-coefficients and the host's detail-coefficients are obtained. It is represented by the following equation:

$$h \xrightarrow{DWT} \begin{cases} hc \\ hd \end{cases} \quad (6.1)$$

Where h corresponds to the host signal in time domain, hc corresponds to the host's coarse-coefficients and hd to the host's detail-coefficients. One host's coarse-coefficient and one host's detail-coefficient are calculated every two clock cycles.

(b) *dwtsecret*: in this block, the secret's coarse-coefficients are obtained according to:

$$s \xrightarrow{DWT} sc \quad (6.2)$$

Where s is the secret signal in time domain and sc are the secret's coarse-coefficients. In the current block, the secret's detail-coefficients are not calculated. In a similar way to the previous block, one secret's coarse-coefficient is calculated every two clock cycles.

The *dwtthost* and the *dwtsecret* blocks use the same wavelet base (filters of decomposition).

(c) *Sorting*: this block sorts sc in descending order within a frame equal to N . There are two outputs: the sorted secret's coarse-coefficients, ssc , and their original positions, key . For example, if $sc=[20, 22, 15, 30, 12, 28, 19, 24]$, then $ssc=[30, 28, 24, 22, 20, 19, 15]$ and $key=[4, 6, 8, 2, 1, 7, 3]$. Since one secret's coarse-coefficient is generated every two clock cycles, to sort N secret's coarse-coefficients $2N$ clock cycles are required, and the sorted data is available in the following two clock cycles. The choice of N is driven by the trade-off among the robustness of the key, the hardware complexity and the delay of the system. If N increases, the robustness of the key is better, but the latency and the hardware complexity increase, too. It is explained in detail in Section 6.3.

(d) *Delay*: the purpose of this block is to synchronize the delays into the embedding module. The host's coarse-coefficients and the host's detail-coefficients are delayed $2N+2$ clock cycles. The outputs of this block provide a delayed version of hc (hcd) and a delayed version of hd (hdd).

(e) *idwstego*: this block reconstructs the stego signal from the stego's coarse-coefficients and the stego's detail-coefficients. The stego's coarse-coefficients, gc , are obtained from hcd and ssc . Without loss of generality, if the length of hcd is 17-bits and the length of ssc is 9-bits, gc is calculated according to:

$$gc=hcd(16:9)\&ssc(8:0) \quad (6.3)$$

where $\&$ is the concatenation operator. In this notation, the Most Significant Bit (MSB) is 16 and the LSB is 0.

In a similar way, the stego's detail-coefficients, gd , are obtained from hdd and the key . If the length of hdd is 16 bits and the length of key is 7 bits, gd is defined as:

$$gd=hdd(16:8)\&key(6:0)\&'0' \quad (6.4)$$

Since the length of the *key* is 7-bits, N is up to 127. The least significant bit of the stego's detail-coefficient, gd , is forced to be an even value in order to minimize the reconstruction error. This is fully explained in Section 6.4.

Once gc and gd have been calculated, the stego signal in time domain is their Inverse Wavelet Transform (IDWT) following the equation:

$$\left. \begin{array}{l} gc \\ gd \end{array} \right\} \xrightarrow{IDWT} g \quad (6.5)$$

Where g is the stego signal, in time domain. This speech signal has embedded both the secret message and the adaptive-*key*. It is expected that the stego signal will be similar to the host signal since the most significant bits of the host's coarse-coefficients and host's detail-coefficients are preserved.

6.2.2. Extraction module

In this module the secret message is recovered from the stego signal. Firstly, the stego signal is decomposed using the DWT; secondly the *key* is obtained from the stego's detail-coefficients while the secret's coarse-coefficients are obtained from the stego's coarse-coefficients. Once the secret's coarse-coefficients have been relocated according to the *key*, the IDWT is applied. The output is the recovered secret message. It is expected that the recovered secret message will be similar but not equal to the original secret message since the secret's detail-coefficients were not hidden. A small difference between them exists.

This module contains the following blocks: *dwtstego*, *reverse* and *idwtsecret*. It is illustrated in Figure 6.2.

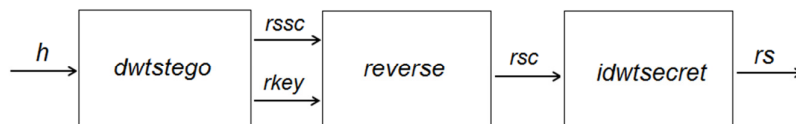


Figure 6.2. Block diagram of the extraction module. [92]

(a) *dwtstego*: the stego signal is decomposed by using the DWT. At the output of this block, the stego's coarse-coefficients and the stego's detail-coefficients are obtained, according to:

$$g \xrightarrow{DWT} \begin{cases} gc \\ gd \end{cases} \quad (6.6)$$

The *gc* contains the secret's coarse-coefficients while *gd* contains the *key*. It is necessary to extract the least significant bits of the above coefficients, according to:

$$rsc = gc(8:0) \quad (6.7)$$

$$rkey = gd(7:1) \quad (6.8)$$

where *rsc* and *rkey* are the recovered *sc* and the recovered *key*. The length of *rsc* is 9-bits, while the length of *rkey* is 7-bits (these lengths corresponds to the case that the *host* signal is 16-bits and the *secret* message is 8-bits).

(b) *reverse*: in this block, the relocation process done in the sorting block is reversed. The purpose is to relocate *rsc* with the information contained into *rkey* to obtain the recovered secret's coarse-coefficients, *rsc*. For example, if *rsc*=[30, 28, 24, 22, 20, 19, 15] and *rkey*=[4, 6, 8, 2, 1, 7, 3] then *rsc*=[20, 22, 15, 30, 12, 28, 19, 24]. It is worth noting that *rsc* is the same *sc* used in the example of the embedding module. In an ideal case in which the principle of perfect reconstruction of the wavelet transform is satisfied, *rsc* must be equal to *sc*, and in a similar way *rkey* must be equal to *key*. In Section 6.3 we will explain further this concept.

(c) *idwsecret*: in this block the recovered secret message, *rs*, is obtained from the recovered secret's coarse-coefficients, *rsc*, according to:

$$rsc \xrightarrow{IDWT} rs \quad (6.9)$$

Since the secret's detail-coefficients were not hidden in the embedding module, the recovered secret message is similar but not equal to the original secret message. Nevertheless, the recovered secret message has a good quality and it is legible.

6.3. Principle of Perfect Reconstruction (PR) and general design of the Discrete Wavelet Transform

One of the most important aspects to take into account in the design of the *dwt-idwt* blocks in hardware is to guarantee the principle of perfect reconstruction. If a signal is decomposed by using the DWT and then it is reconstructed by using the IDWT -with the same wavelet base-, it is expected that the reconstructed signal is exactly equal to the original one. This is the principle of perfect reconstruction.

In hardware, the weights of the decomposition and reconstruction filters can be slightly different from the theoretical ones because of the quantization process. The higher the quantization error, the higher the reconstruction error, and therefore the reconstructed signal will be not similar to the original one. Because of that, the quantization of the weights is an important aspect to take into account in the design of the topology of the *dwt-idwt* blocks.

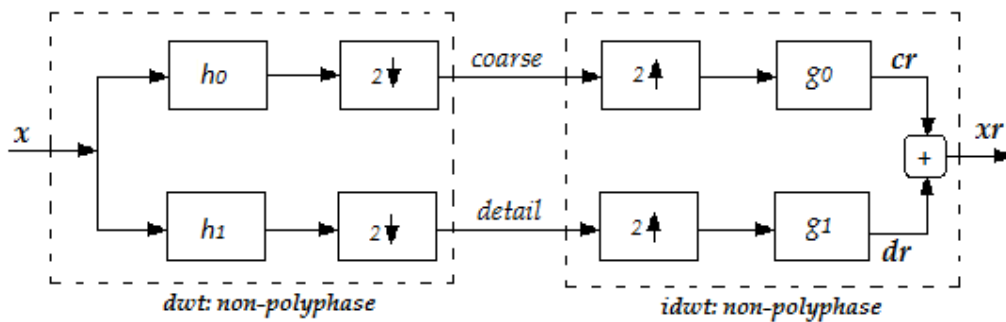


Figure 6.3. Decomposition and reconstruction: non-polyphase scheme. [93]

The non-polyphase scheme of the decomposition and reconstruction stages is illustrated in Figure 6.3. The low-pass decomposition filter is represented by h_0 , the high-pass decomposition filter by h_1 , the low-pass reconstruction filter by g_0 , the high-pass reconstruction filter by g_1 , while $2\downarrow$ represents a down-sampling process and $2\uparrow$ an up-sampling process, by factor of two.

To guarantee aliasing cancellation and perfect reconstruction, the following conditions must be satisfied [94], [95]:

$$1/2 \{ [H_0(z).G_0(-z)] + [H_1(z).G_1(-z)] \} = 0 \quad (6.10)$$

And

$$1/2 \{ [H_0(z).G_0(z)] + [H_1(z).G_1(z)] \} = z^{-m} \quad (6.11)$$

If a constant value, kte , is factorized in all the four filters, equations (6.10) and (6.11) are written as follows:

$$\begin{aligned} & 1/2 \left\{ kte^2 \left[\frac{H_0(z)}{kte} \cdot \frac{G_0(-z)}{kte} \right] + kte^2 \left[\frac{H_1(z)}{kte} \cdot \frac{G_1(-z)}{kte} \right] \right\} = \\ & 1/2 kte^2 \left\{ \left[\frac{H_0(z)}{kte} \cdot \frac{G_0(-z)}{kte} \right] + \left[\frac{H_1(z)}{kte} \cdot \frac{G_1(-z)}{kte} \right] \right\} = 0 \end{aligned} \quad (6.12)$$

And

$$\begin{aligned} & 1/2 \left\{ kte^2 \left[\frac{H_0(z)}{kte} \cdot \frac{G_0(-z)}{kte} \right] + kte^2 \left[\frac{H_1(z)}{kte} \cdot \frac{G_1(-z)}{kte} \right] \right\} = \\ & 1/2 kte^2 \left\{ \left[\frac{H_0(z)}{kte} \cdot \frac{G_0(-z)}{kte} \right] + \left[\frac{H_1(z)}{kte} \cdot \frac{G_1(-z)}{kte} \right] \right\} = z^{-m} \end{aligned} \quad (6.13)$$

According to (6.12) and (6.13) a topology that factorizes the term kte in all the four filters and includes a post-amplifier block with gain of kte^2 satisfies the principle of PR as its original topology.

With the purpose to have a more efficient architecture of the decomposition stage, the non-polyphase scheme is replaced with a polyphase scheme and the input signal is down-sampled (split) before the filtering process. Unlike the non-polyphase scheme half of the results are not wasted. Figure 6.4 illustrates the general design in which h_{0even} and h_{1even} filters the even part of the input signal, x_{even} , while h_{0odd} and h_{1odd} filters the odd part of the input signal, x_{odd} .

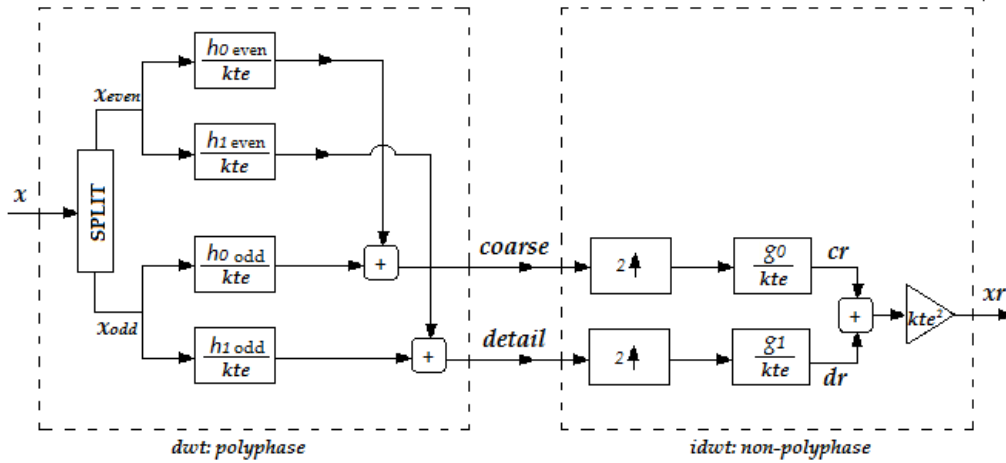


Figure 6.4. General design of the $dwt-idwt$ stages. [93]

Then, the coarse (c) and detail (d) coefficients are obtained as follows:

$$\begin{aligned}
 c(z) &= \left\{ X_{\text{even}}(z) \cdot \frac{H_{0\text{even}}(z)}{kte} \right\} + \left\{ X_{\text{odd}}(z) \cdot \frac{H_{0\text{odd}}(z)}{kte} \right\} \\
 c(z) &= \left\{ X(z) \cdot (2 \downarrow) \frac{\langle h_0[0] + h_0[2] \cdot z^{-1} + h_0[4] \cdot z^{-2} + \dots \rangle}{kte} \right\} + \\
 &\quad \left\{ z^{-1} X(z) \cdot (2 \downarrow) \frac{\langle h_0[1] + h_0[3] \cdot z^{-1} + h_0[5] \cdot z^{-2} + \dots \rangle}{kte} \right\}
 \end{aligned} \tag{6.14}$$

$$\begin{aligned}
 d(z) &= \left\{ X_{\text{even}}(z) \cdot \frac{H_{1\text{even}}(z)}{kte} \right\} + \left\{ X_{\text{odd}}(z) \cdot \frac{H_{1\text{odd}}(z)}{kte} \right\} \\
 d(z) &= \left\{ X(z) \cdot (2 \downarrow) \frac{\langle h_1[0] + h_1[2] \cdot z^{-1} + h_1[4] \cdot z^{-2} + \dots \rangle}{kte} \right\} + \\
 &\quad \left\{ z^{-1} X(z) \cdot (2 \downarrow) \frac{\langle h_1[1] + h_1[3] \cdot z^{-1} + h_1[5] \cdot z^{-2} + \dots \rangle}{kte} \right\}
 \end{aligned} \tag{6.15}$$

And the reconstructed signal, x_r , is calculated as follows:

$$\begin{aligned}
x_r &= kte^2 \cdot [c_r(z) + d_r(z)] \\
x_r &= kte^2 \cdot \left[\left\{ (2 \uparrow c(z)) \frac{G_0(z)}{kte} \right\} + \left\{ (2 \uparrow d(z)) \frac{G_1(z)}{kte} \right\} \right] \\
x_r &= kte^2 \cdot \left[\left\{ (2 \uparrow c(z)) \frac{\langle g_0[0] + g_0[1] \cdot z^{-1} + g_0[2] \cdot z^{-2} + \dots \rangle}{kte} \right\} + \right. \\
&\quad \left. \left\{ (2 \uparrow d(z)) \frac{\langle g_1[0] + g_1[1] \cdot z^{-1} + g_1[2] \cdot z^{-2} + \dots \rangle}{kte} \right\} \right]
\end{aligned} \tag{6.16}$$

Where c_r contains the reconstructed-coarse coefficients and d_r contains the reconstructed-detail coefficients. The above equations not only guarantee the PR; they provide an efficient scheme for the decomposition and reconstruction of the signal.

6.4. Hardware design of the speech-in-speech hiding scheme

The proposed scheme is LSB-based with adaptive *key*, in wavelet domain. The generic blocks of the embedding and extraction modules are: *dwt*, *idwt*, *sorting*, *reverse* and *delay*. The blocks *dwthost*, *dwtsecret* and *dwtstego* are based on the generic block *dwt* while the blocks *idwtstego* and *idwtsecret* are based on the generic block *idwt*. Since the Discrete Wavelet Transform plays an important role in the proposed scheme, in the first part of this section we will discuss the design of the blocks *dwt-idwt* which satisfies the principle of Perfect Reconstruction. Thereafter, we will present the design of the blocks *sorting*, *reverse* and *delay*.

6.4.1. Decomposition and reconstruction: dwt and idwt blocks

With the purpose to have efficient hardware architecture of the wavelet transform, the following characteristics have been selected:

- (a) Biorthogonal base. The symmetry of this kind of wavelets reduces the quantity of operations. Specifically, it is selected the 5/3 wavelet base.
- (b) Multiplierless scheme. In this topology the multiplications are replaced with shifts and therefore the hardware resources decrease.
- (c) Quantization of the weights of the filters based on rational integers. The quantization error is significantly lower than in fixed point format.

Taking into account the above characteristics and the design presented in Section 6.3, the weights of the 5/3 wavelet base are represented as follows:

$$\frac{h_0(k)}{\sqrt{2}} = h_{0a} \cdot h_{0b} \quad h_{0a}(k) = \{-1 \quad 2 \quad 6 \quad 2 \quad -1\} \quad h_{0b}(k) = 1/8 \quad (6.17)$$

$$\frac{h_1(k)}{\sqrt{2}} = h_{1a} \cdot h_{1b} \quad h_{1a}(k) = \{1 \quad -2 \quad 1\} \quad h_{1b}(k) = 1/4 \quad (6.18)$$

$$\frac{g_0(k)}{\sqrt{2}} = g_{0a} \cdot g_{0b} \quad g_{0a}(k) = \{1 \ 2 \ 1\} \quad g_{0b}(k) = 1/4 \quad (6.19)$$

$$\frac{g_1(k)}{\sqrt{2}} = g_{1a} \cdot g_{1b} \quad g_{1a}(k) = \{1 \ 2 \ -6 \ 2 \ 1\} \quad g_{1b}(k) = 1/8 \quad (6.20)$$

Where $h_0(k)$, $h_1(k)$, $g_0(k)$, $g_1(k)$ are the lowpass-decomposition, highpass-decomposition, lowpass-reconstruction and highpass-reconstruction filters. It is worth noting that the weights of the filters have been divided by the term $\sqrt{2}$. This is equal to *kte* presented in Section 6.3.

Since all the terms ($h_{0a}(k)$, $h_{0b}(k)$, $h_{1a}(k)$, $h_{1b}(k)$, $g_{0a}(k)$, $g_{0b}(k)$, $g_{1a}(k)$, $g_{1b}(k)$) can be represented as a sum of power of two (SPT), they can be computed by right-shifts and left-shifts, in binary representation. A left-shift is a multiplication by power of two (i.e. $2^0 \cdot \text{data}$, $2^1 \cdot \text{data}$, $2^2 \cdot \text{data}$,...) while a right-shift is equal to the ceiling operator of the division by a power of two (i.e. $\lfloor \text{data}/2^0 \rfloor$, $\lfloor \text{data}/2^1 \rfloor$, $\lfloor \text{data}/2^2 \rfloor$,...). Therefore, the truncation error appears only in the division process if data is an odd number. For example, if $\text{data} = 101101_b$, a division by 2 with one-right-shift is 10110_b and the error is $\frac{1}{2}$ LSB, but if data is an even number, i.e. $\text{data} = 101100_b$, the division by 2 with one-right-shift is 10110_b and the error is 0-LSBs. The grouping of the weights of the filters by integer constants and the post-amplifier stage by power of two permit to carry out all of the hardware operations by right-shifts and left-shifts.

Additionally, the proposed design satisfies the desirable condition of the Quadrature Mirror Filters (QMF) in which $G_0(z) = H_1(-z)$ and $G_1(z) = -H_0(-z)$. It is a sufficient condition to guarantee anti-aliasing and it is also an efficient condition to have a low hardware cost.

The structure of the *dwt* block is illustrated in Figure 6.5. The input signal is split and the even (x_{even}) and odd parts (x_{odd}) are shifted and added to obtain *coarse* and

detail coefficients. The scheme does not use multiplier units and instead of that all operations (multiplications and divisions) are carried out with five right-shifts and five left-shifts. It takes advantage of the symmetry property of the biorthogonal filters (e.g. one left-shift is used to compute $2\{x_{odd}(n)+x_{odd}(n-1)\}$). Additionally, a small number of adders are used in the topology.

Since the input signal is split, the *coarse* and *detail* coefficients are updated every two cycles. In the current design, the input signal is 16-bits, the coarse coefficients are 17-bits and the detail coefficients are 16-bits.

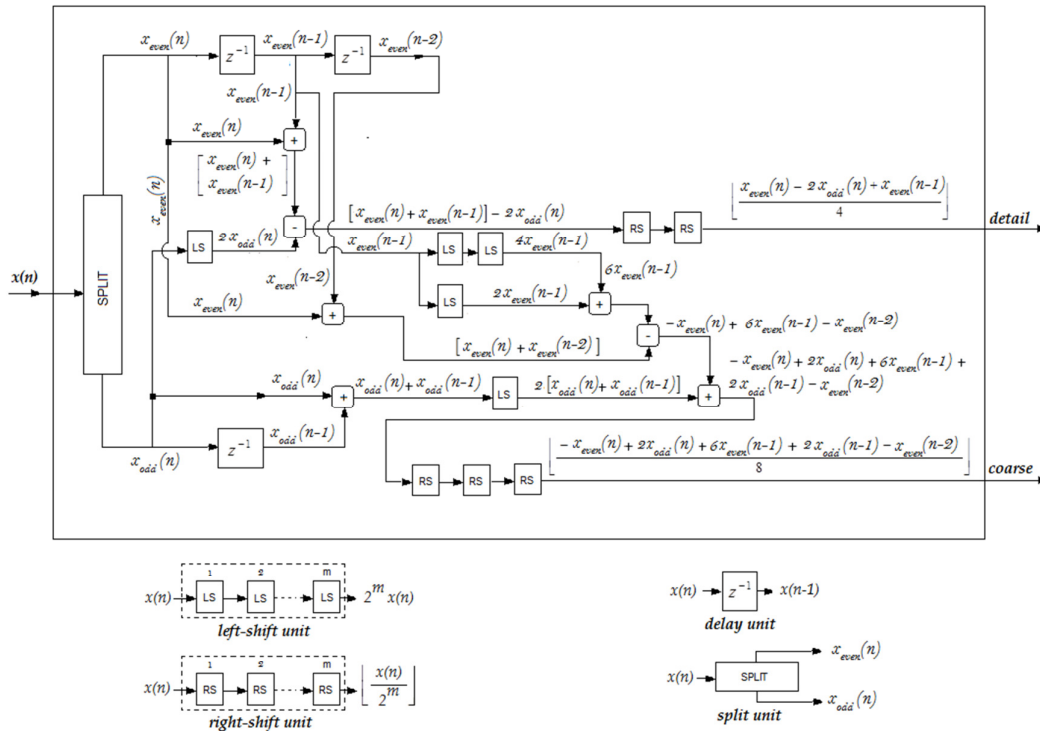


Figure 6.5. Scheme of the *dwt* block. [93]

To reconstruct the signal, the *idwt* block is designed. Like the *dwt* block, it uses a multiplierless topology. Figure 6.6 illustrates the design: *detail* coefficients are represented as $d(n)$, *coarse* coefficients as $c(n)$, oversampled *detail* coefficients as

$d_{over}(n)$, oversampled *coarse* coefficients as $c_{over}(n)$, reconstructed *detail* coefficients as $dr(n)$, reconstructed *coarse* coefficients as $cr(n)$ and the denoising signal as x_{den} . Firstly, the *coarse* and *detail* coefficients are oversampled, secondly, $d_{over}(n)$ and $c_{over}(n)$ are shifted and added, thirdly, $dr(n)$ and the $cr(n)$ are added and finally $[cr(n)+dr(n)]$ is multiplied by the term kte^2 (by using one left-shift). Since the wavelet coefficients are oversampled, the reconstructed signal is obtained in every clock cycle. In the current design, the coarse coefficients are 17-bits, the detail coefficients are 16-bits and the reconstructed (denoising) signal is 16-bits.

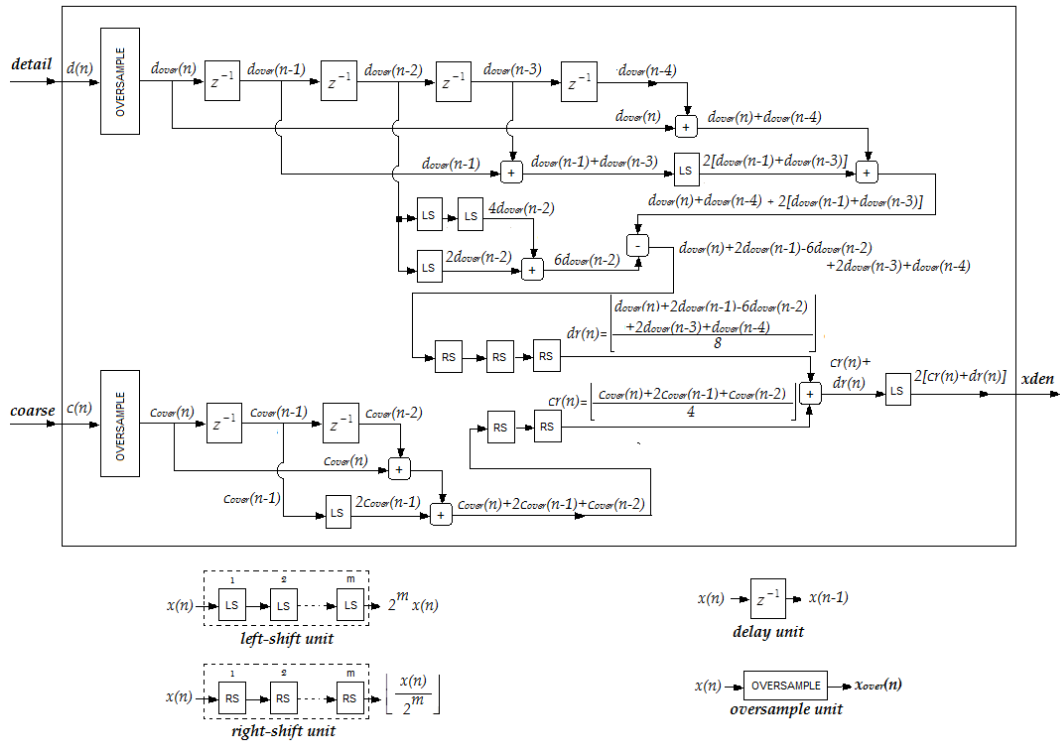


Figure 6.6. Scheme of the $idwt$ block. [93]

The highest reconstruction error of the dwt and $idwt$ blocks is 2-LSBs and it means that if the input signal is 16 bits (in signed format), the highest error is up to 3 of 32767 or in other words it is 0.0092%.

6.4.2. Sorting and reverse

There are several architectures proposed in the literature for sorting data. The schemes can be classified as sorting networks and linear sorters. The former approach sorts parallel data while the latter sorts serial data. Since in the proposed design a new data appears every clock cycle, the scheme based on sorting networks is discarded. In the case of linear sorters, the traditional approaches use bidirectional data-shifts in a continuous stream. Since the length of the array is fixed, one number is deleted from the array to give a place to the new number in every clock cycle, and the criterion is based on, for example, a *first input- first output* (FIFO) scheme [96], [97]. Like the traditional approaches, the length of the current arrays is fixed too, but the *sorting* block works with non-overlapped frames. For example, the first N numbers are sorted and the result is given in the following N cycles of the clock signal at the same time that the second frame is sorted. In other words, while the frame k is sorted, the results of the frame $k-1$ are supplied in every clock cycle. Unlike the schemes of linear sorter based on a FIFO scheme which works with two arrays (one for the sorted-array and one for the rank-array), the proposed scheme works with four arrays, two sorted-arrays and two rank-arrays. One sorted-array (and rank-array) is the original and the other is the copy.

Figure 6.7 shows an example of descending sorting using a FIFO-based scheme and a non-overlapped scheme, for $N=8$. In the first N clock cycles, both schemes sort the data in the same form, therefore the sorted-array and rank-array have the same results. The difference between them is that in the non-overlapped scheme a copy of the rank-array and sorted-array is made in the clock cycle N . In the clock cycle equal to 9 (or $N+1$), the results of the sorted-array and rank-array are completely different between the schemes, i.e. in the FIFO-based scheme the new data (7) is added to the sorted array while in the non-overlapped scheme the process of sorting begins again. When the

clock cycle is equal to 18 (or $2N$), the sorted-arrays of the schemes are equal, but the rank-arrays are different. It is noticed that the highest value of the rank-array in the non-overlapped scheme is N , while in the FIFO-based scheme is the current clock cycle. Again, a copy of the sorted-array and rank-array is made in the clock cycle equal to $2N$. This process is repeated until the total of data is reached.

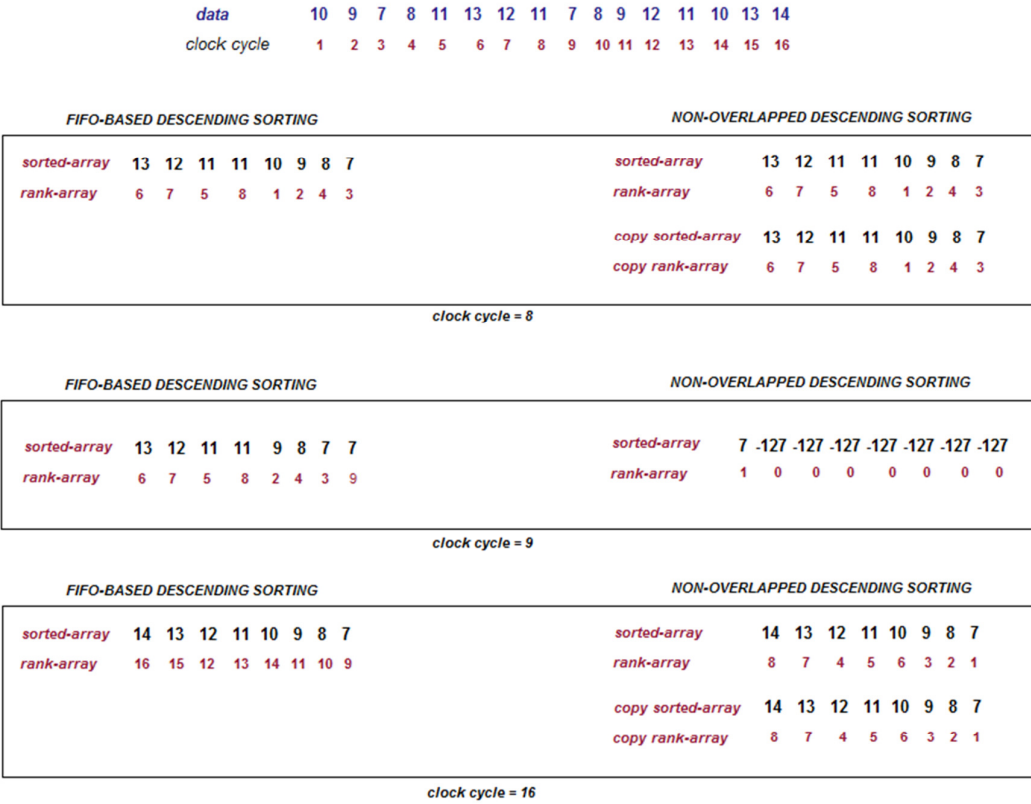


Figure 6.7. Sorting process with the FIFO-based and non-overlapped schemes.

[92]

In hardware, the non-overlapped *sorting* block includes comparators, multiplexers, D-type flip-flops (FF) and a counter. At the beginning, the FFs of the sorted-arrays are set to the lowest number into the range, i.e. -127 if the data is encoded with 8-bits, while the FFs of the rank-arrays are set to zero. The sorting

process is done between cycles 1 and N of the counter; when the counter reaches N , the sorted-array and rank-array are copied and the counter is set to 1, again. In the following cycles, while the sorting process is done again, the sorted data and their ranks of the above group are provided one by one in every clock cycle.

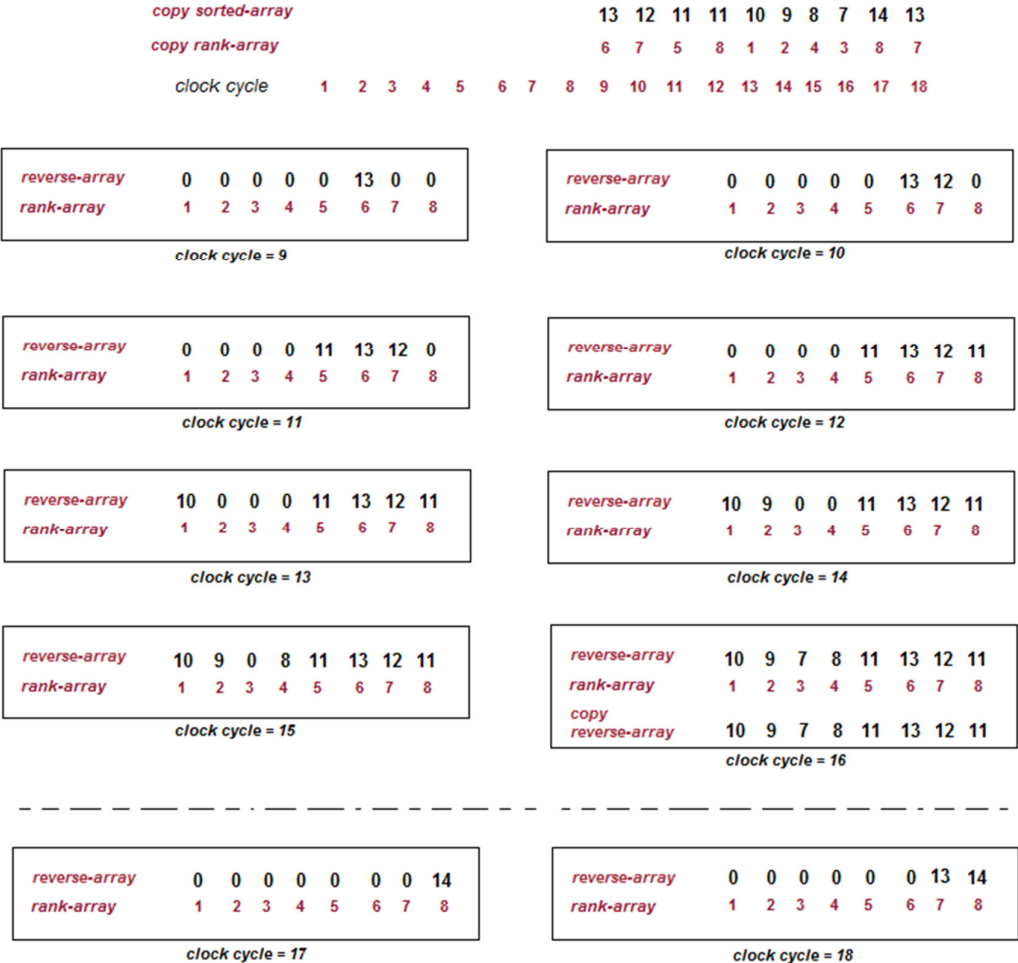


Figure 6.8. Reverse process with the non-overlapped scheme. [92]

On the other hand, the aim of the *reverse* block is to put the data in the original places. It uses the sorted-array and the rank-array. Since one *sorted* and one *rank* number enters every clock cycle, the reverse process uses N cycles to relocate the

places. Figure 6.8 shows an example of the reverse process. The first group of N elements is reversed between the clock cycles $N+1$ and $2N$. At the beginning, the reverse-array contains null-data and in every clock cycle one position is replaced to the current data of copy sorted-array. Once the N elements have been reversed, a copy of the reverse-array is made and the reverse process begins again.

In a similar way to the *sorting* block, the hardware resources of the *reverse* block include comparators, FFs and a counter. When the counter reaches the value of N , a copy of the reverse-array is made and the reverse process begins again in the following clock cycle. One value of the copy of the reverse-array is provided every clock cycle. Since the secret's coarse-coefficients are updated every two clock cycles, the real latency between the first secret's coarse-coefficient and the first sorted secret's coarse-coefficient is $2(N+1)$ instead of $(N+1)$ of the previous example.

6.4.3. Delay

The aim of this block is to synchronize the data into the embedding module. It uses z^{-1} units interconnected in two synchronous arrays and the total delay by signal is $2N+2$. Since the host's coarse-coefficients and the host's detail-coefficients must be delayed, the *delay* block works with two inputs and two outputs (Figure 6.9). The length of the host's coarse-coefficients is 17-bits while the length of the host's detail-coefficients is 16-bits (for a *host* signal encoded with 16-bits).

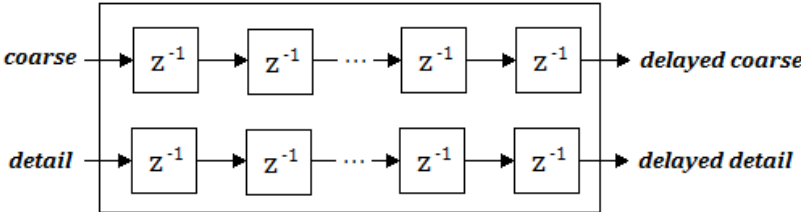


Figure 6.9. Scheme of the *delay* block.

6.5. Hardware performance

In this section we validate the hardware architecture of the speech-in-speech hiding scheme. The embedding and extraction modules are modeled using VHDL and they are compiled and simulated using ISE Foundation 12.4 and ModelSim SE 6.4a, respectively. The *host* signal is encoded with 16-bits, the *secret* message with 8-bits, the *stego* signal with 16-bits, and the *recovered secret* message with 8-bits.

6.5.1. Hardware Resources

With the purpose of measuring the hardware resources of the proposed design, we select the Spartan-6 xc6slx45 device for the implementation. In Table 6.2 the total amount of resources by block (compiled separately), the percentage of the used resources, and the latency, are given. The total of resources of the selected FPGA is supplied between brackets. As expected, the used resources in *main* (entire design) are not the sum of the used resources of the eight blocks. If N is 8, the latency of the embedding (or extraction) module will be 23 *clock cycles* and the total latency of the system will be 47 *clock cycles*. If $f_s=8$ KHz, the latency of each module will be 2,87 *ms* and the total latency will be 5,87 *ms*; but if N is 127, the latency of every module will be 65 *ms*.

According to Table 6.2, the maximum delay per block is extremely low in relation to the time between consecutive samples, Δt , in speech signals (typically $\Delta t=125000$ ns). Therefore, the delay between the *host* signal and the *stego* signal would not be perceptible by the HAS. Since for a real-time speech communication system, the highest mouth-to-ear delay should be up to 200 *ms*, the low latency of the embedding module allows that the speech signal can be hidden into a high quality transmission scheme. On the other hand, the speech signal can be recovered in real-time, too.

Table 6.2. Resource utilization and longest path delay. [92]

Block	Slice Registers (54576)	Slice LUTs (27288)	LUT-FF pairs (1219)	Bounded IOBs (218)	Max. delay (ns)	Latency (clock cycles)
<i>dwthost</i>	99(<1%)	130 (<1%)	87 (7%)	51 (23%)	5.32	$L_1=2$
<i>dwtsecret</i>	43 (<1%)	68 (<1%)	20 (2%)	19 (9%)	6.04	$L_2=2$
<i>delay</i>	318(<1%)	299 (~1%)	299 (25%)	68 (31%)	2.64	$L_3=2(N+1)$
<i>sorting (N=8)</i>	240 (<1%)	408(~1%)	139 (11%)	25(11%)	6.14	$L_4=2(N+1)$
<i>idwtstego</i>	160 (<1%)	138(<1%)	68(6%)	50(22%)	5.11	$L_5=3$
<i>dwstego</i>	79(<1%)	130(<1%)	66(5%)	32(15%)	5.17	$L_6=2$
<i>reverse</i>	156(<1%)	185(<1%)	150(12%)	26(12%)	3.49	$L_7=2(N+1)$
<i>idwtsecret</i>	40(<1%)	27(<1%)	22(2%)	19(9%)	2.94	$L_8=3$
<i>Main (total)</i>	797(1%)	849(3%)	427(35%)	34(15%)	7.28	* $L_T=$ $L_{T1}+L_{T2}+1$

* L_T : total latency; L_{T1} : latency embedding module; L_{T2} : latency extraction module;
 $L_{T1}=L_2+L_4+L_5$; $L_{T2}=L_6+L_7+L_8$

In relation to the hardware resources, the design is extremely simple and uses only small percentage of the available resources of the selected FPGA. It is remarkable that the *dwt-idwt* blocks use lower resources than the *sorting-reverse* blocks. It means that the selected scheme (polyphase), the representation of the weights of the FIR filters by integer data, and the multiplierless topology (using left-shifts and right-shifts) are adequate options to obtain a low cost hardware and low reconstruction error.

The generation of the adaptive-key is carried out by the *sorting* block. Since N is up to 127 and the current design uses $N=8$, it is important to estimate the hardware resources when N is higher. Firstly, the macro statistics of the *sorting* block for $N=8$ are shown in Table 6.3.

Table 6.3. Macro statistics of the sorting block. [92]

Adders/Subtractor	Register			Comparators	Multiplexers	
5-bit adder	1-bit	5-bit	9-bit	9-bit	5-bit 2-to-1	9-bit 2-to-1
2	1	18	16	7	41	42

Most of the hardware resources of the *sorting* block are comparators and multiplexers. The number of comparators is $N-1$ while the number of multiplexers is

up to $2 \cdot (N-1) \cdot (N-2)$. If $N=100$ it is expected that the total number of comparators will be 99 and the total of multiplexers will be 19400, approximately. Although the hardware resources increases with respect to the current resources ($N=8$), these resources would be less than the available resources of the FPGA. In other words, the speech-in-speech hiding architecture may work with a higher value of N .

6.5.2. Reconstruction error

Since one of the most important characteristics in the proposed model is to have an extremely low reconstruction error according to the principle of perfect reconstruction, the validation of the *dwt-idwt* blocks in terms of the reconstruction error is taken into account. It is measured as the difference between the input data and the output data expressed in the total number of LSBs. To test this error, the architecture is connected as in Figure 6.10.

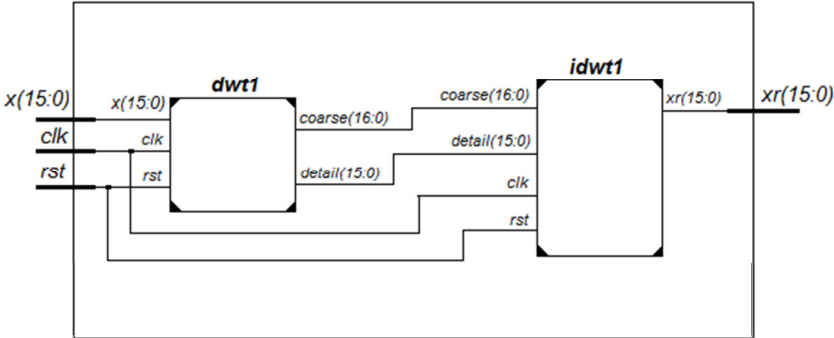


Figure 6.10. Block diagram of the decomposition-reconstruction system. [93]

Then, if the reconstructed signal, xr , is very close to the input signal, x , the reconstruction error is low; otherwise the reconstruction error is high. Figure 6.10 illustrates a simulation of the process. The plot shows the clock signal (*clk*), the reset signal (*rst*), the input signal (*x*), the reconstructed signal (*xr*), the coarse coefficient (*xcoarse*), the detail coefficient (*xdetail*) and the enable signal for even cycles (*div2*).

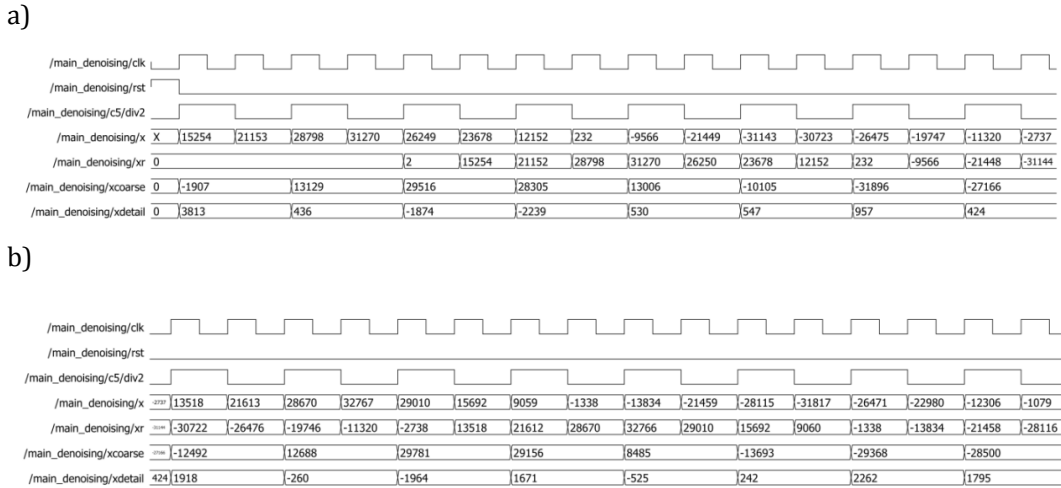


Figure 6.11. Simulation of dwt and idwt blocks: (a) 1st to 16th cycles, (b) 17th to 32th cycles. [93]

In a frame-by-frame design, a border extension is applied to the input signal with the purpose of smoothing the first and last coefficients, but in a real-time design it is not suitable. For this reason, the first coarse and first detail coefficients are not a proper representation of the input signal (and they should be ignored). The proper output is in the following even cycle and therefore the latency of the *dwt* block is two. For example, the 1st input of Figure 11a (15254) has its coarse (13129) and detail coefficients (426) in the 3rd clock cycle. These coefficients are updated every two cycles according to the theory. On the other hand, the 1st output (15254) is in the 6th cycle; therefore the latency of the *idwt* block is three. The 1st - 5th outputs should be zero; a small value is due to the quantization process. Finally, it is worth noting that if the input signal is an even number, the reconstructed one is an even number too and the reconstruction error is zero; but if the input signal is an odd number, the reconstructed signal is an even number and the error is +/-1. Therefore, the highest reconstruction error is equal to 2-LSBs.

6.5.3. Validation of the entire design

In this section, the speech-in-speech hiding scheme is simulated on ModelSim 6.4.a. The embedding and extraction modules have been interconnected in the same architecture. At the input, the host signal (*host*) and the secret message (8-bits) from two real speech signals with time-scale of 1s are supplied. The output provides the recovered secret message (8-bits). The internal signals are: the stego signal (*stego*), the secret's coarse-coefficients (*scoarse*), the recovered secret's coarse-coefficients (*srcoarse*), the key (*ks*) and the recovered key (*ksr*). In order to illustrate the adaptive-key generation with a small number of *clock cycles*, it is selected $N=8$. In Figure 6.12 the simulation results between clock cycles 2200 and 2260 are provided. The following notation to highlight the inter-block latency is used: stars for host signal and stego signal; triangles for secret message and recovered secret message; squares for secret's coarse-coefficient and recovered secret's coarse-coefficient; and circles for *key* and recovered *key*.

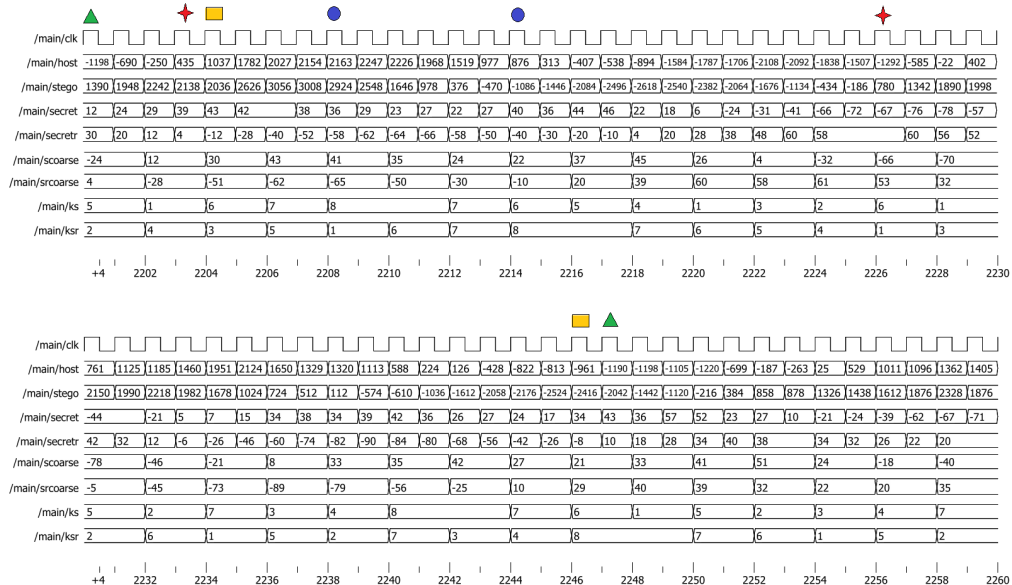


Figure 6.12. Simulation of the embedding & extraction modules, clock cycles [2200

2260]. [92]

To demonstrate the latency between the host signal and the stego signal (latency of the embedding module), the host signal equal to 435 is selected (*clock cycle* number 2203). The corresponding value is the stego signal equal to 780 (*clock cycle* number 2226). Then, the latency is 23 *clock cycles*. Secondly, we select the *key* equal to 8 (*clock cycle* number 2208) and its corresponding recovered *key* equal to 8 (*clock cycle* number 2214), in this case the latency is equal to 6. Thirdly, the secret's coarse-coefficient equal to 30 (*clock cycle* number 2204) and its corresponding recovered secret's coarse-coefficient equal to 29 (*clock cycle* number 2246) illustrate the latency between the above signals, which is equal to 42. Finally, the total latency of the system (embedding module + extraction module) is obtained from the secret message and the recovered secret message. The secret message equal to 12 (*clock cycle* number 2200) and the recovered secret message equal to 10 (*clock cycle* number 2247), then the total latency of the system is 47.

It is worth noting that the recovered *key* is exactly equal to the original *key* because in the embedding module the stego's detail-coefficient was forced to be an even value; then the reconstruction error is zero. In the case of the recovered secret's coarse-coefficient there is a small error (2-LSBs) in relation to the secret's coarse-coefficient, because the stego's coarse-coefficient is an even or odd number and the term $coefficient * sample / 8$ is calculated as $\lceil coefficient * sample / 8 \rceil$; where $\lceil . \rceil$ is the ceiling operator.

In order to illustrate the similarity between the host signal and the stego signal, and between the secret message and the recovered secret message, the result of the entire simulation is provided in Figure 6.13.

The simulation works with two speech signals (*host* and *secret*) with time-scale of 1-second and sampling frequency, fs , of 8 KHz. The *host* and the *stego* signals are in the range [-32768 32768], while the *secret* and the recovered secret message (*secret_r*) are

in the range $[-128, 128]$. According to Figure 6.11, it is remarkable that the *secret* message is hidden into the region of silence as the region of non-silence of the *host* signal; however, the *stego* signal is very similar to the *host* one. Since the latency of the embedding module (23 clock cycles) is very low in relation to f_s , the delay between the *host* and the *stego* signal is quasi-imperceptible. It allows to transmit the *stego* signal in real-time.

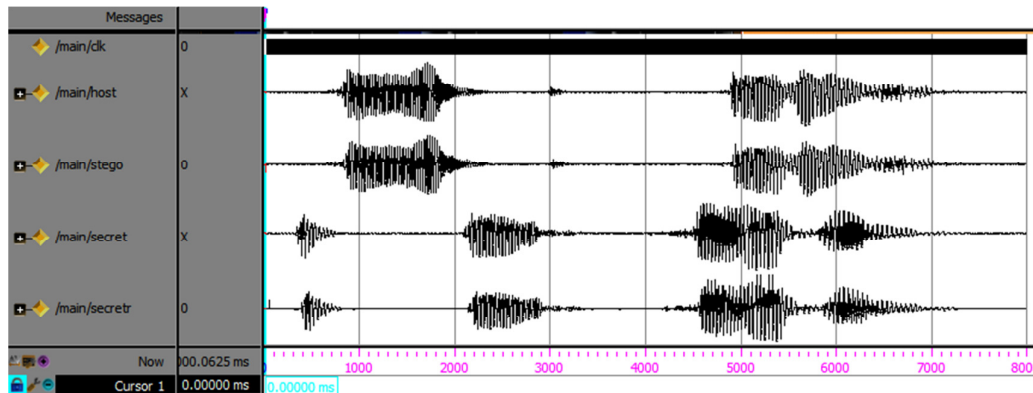


Figure 6.13. Simulation of the speech-in-speech hiding scheme. [92]

6.5.4. Comparing to related works: dwt-idwt blocks

In wavelet-based data hiding PR is an outstanding requirement of the system and therefore the quantization process plays an important role in the design. However, other parameters like latency and hardware resources are important, too. In this context, the selection of the “best” design is based on the good trade-off among reconstruction error, latency and hardware cost. In order to illustrate the strengths of our design, in this section some remarkable designs of multiplier-based and multiplierless-based schemes are analyzed.

Multiplier-based schemes: the schemes based on this topology use multiplier units to multiply the input signal by the weights of the FIR filters. Since the multiplier unit sums the size of its inputs, the product can exceed the minimum number of bits to

represent the data (e.g. $101 \cdot 10 = 01010$), and then the hardware resources are not as low as possible. This is the main weakness of the multiplier-based schemes. In [98] a multiplier topology of the lifting scheme is presented. Its main characteristic is that the size of the quantized weights can be selected according to a desired data precision. If the size increases, the precision increases too and the quantization error decreases, but the hardware cost increases. Since all the weights of the filters require long word-bits and it uses a multiplier topology, this design requires a higher number of resources. Unlike [98], the design presented in [99] works with fixed size of the quantization of the weights. The main disadvantage is that its quantization error is high (~15%) and therefore it is not appropriate for denoising systems (but it is for other kind of applications like detection). Both designs are complex in terms of hardware cost.

Multiplierless-based schemes: unlike the multiplier-based schemes, the current ones use shifts and sum operations to carry out the multiplication process. The main point is the representation of the weights of the filters with the minimum number of bits. The lower the number of nonzero bits, the lower is the number of shifts. Typically, the formats are fixed-point, Canonical Signed Digit (CSD) and ratio of integer numbers. The CSD format is a special case of fixed-point in which the bit 1 represents a positive power of two and $\bar{1}$ a negative power of two (e.g. $0.10\bar{1}_2$ is equal to $0.5 - 0.125 = 0.375$). The designs in [100]-[102] use the CSD format to compute the 9/7 wavelet base. According to their results, the best design in terms of quantization error is not the best in terms of latency. The number of SPT terms is at least 21.

With the purpose to reduce the quantization error, the weights of the filters can be represented as ratio of integer numbers. In [103], [104] is designed a 5/3 wavelet base for the lifting scheme. Although most of the weights have a finite representation as rational terms, the gain ($\sqrt{2}$) is approximated to $44/32$ and it gives a high quantization

error (~2.5%). The advantage is that all denominators are power of two and they can be easily made by right-shifts. In [105] is shown a wavelet-denoising system by using rational 9/7 wavelet base. The denominator of the rational terms is 64 and the numerators are in the range [1 46]. Therefore, the size of the internal signals is higher than the size of the input speech signal (more hardware resources) and they need at least 35 shifts. In our design, the highest denominator is 8 and the numerators of the rational terms are in the range [1 6]. It gives a low number of shifts operations (10) and therefore a low hardware complexity. The latency of our *dwt* block is significantly lower than in the above designs. Additionally, unlike other designs [106], the gain of the decomposition filter is the same as the reconstruction filter ($|H(0)|=|G(0)|$) and this satisfies the requirement of the QMFs. Some of the remarkable works are shown in Table 6.4.

Table 6.4. Comparison of multiplierless-based schemes. [93].

Design	Scheme	Quantization	Quantization error	Advantage	Disadvantage
[100]	Non-polyphase scheme	CSD	Up to 3.2% i.e. unquantized=0.037828455 quantized=0.0390625	Optimized to PR requirement	Long SPT terms (32) Long latency (23)
[101]	Polyphase scheme	CSD	Up to 7% i.e. unquantized=0.037828455 quantized=0.03515625	Optimized to PR requirement	Long SPT terms (32) Long latency (19)
[102]	Lifting scheme	CSD	Up to 0.024 % i.e. unquantized=0.8 quantized=0.7998046875	Optimized to PR requirement	Long SPT terms (21) Long latency (49)
[103], [104]	Lifting scheme	Integer	From 0% to 2.5% unquantized= $\sqrt{2}$ quantized=44/32	Optimized to PR requirement	Long SPT terms
[105]	Polyphase	Integer	Up to 0.0031%	Optimized to PR requirement	Long SPT terms Higher size of internal signals
Proposed scheme	Polyphase	Integer	Up to 0.0031% reconstruction error $E_T = [0.3 \cdot 2^{-15}]$ or up to 0.0092%	High tradeoff between PR, latency and hardware cost	Fixed to 5/3 wavelet base

6.5.5. Comparing to related works: the entire design

In this section, it is analyzed the quality of the *stego* signal and the recovered secret message of the proposed architecture. Firstly, the differences between *host* signal and *stego* signal, and between *secret* message and *recovered secret* message, are measured. Secondly, the Signal-to-Noise-Ratio (SNR) and the Squared Pearson Correlation Coefficient, ρ^2 , are calculated for every pair of signals. Finally, the results are compared to those obtained from other schemes. With the purpose to obtain the difference between *host* signal and *stego* signal, and between *secret* message and recovered secret message, we include the *diff* block which calculates the error between the above signals. Since the latency between the *host* signal and the *stego* signal is 23 (if $N=8$) and the latency between the *secret* message and the *recovered secret* message is 47, the *diff* block keeps 23 samples of the *host* signal and 47 samples of the *secret* message.

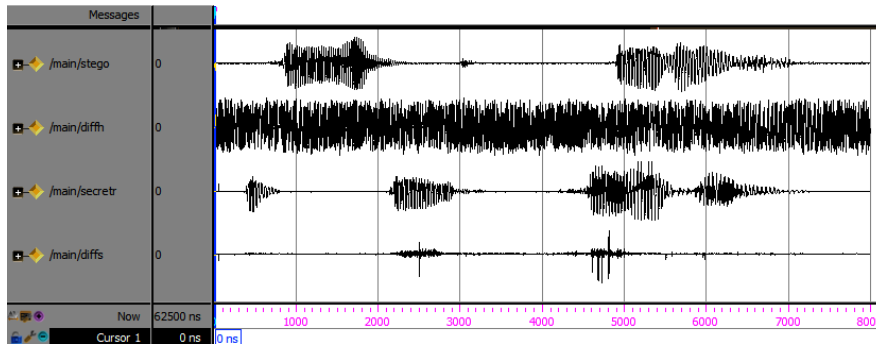


Figure 6.14. Output at the transmitter and at the receiver, and their error signals.

[92]

Figure 6.14 shows in descending order: the *stego* signal (*stego*), the difference between the *stego* and the *host* signal (*diffh*), the recovered secret message (*secret*) and the difference between the secret message and the recovered secret message (*diffs*). The *stego* signal is in the range $[-32768 \ 32768]$ while *diffh* is in $[-600 \ 600]$,

secretr is in [-128 128] and *diffs* is in [-128 128]. It can be noticed that *diffh* is similar to white noise while *diffs* is not. In terms of percentage, the difference between *host* and *stego* is up to 1.8%, per sample; while in the case of the *diffs*, the difference between *secret* and *secretr*, most of the samples are under 2% but there are a few up to 80%. Since the *recovered secret* message only contains the information from the secret's coarse-coefficients, the percentage in *diffs* is higher than the percentage in *diffh*; in other words, it is expected that the quality of the *stego* signal will be slightly higher than the quality of the *recovered secret* message.

In order to objectively assess the quality of the output signals (*stego* and *secretr*), the proposed scheme is simulated in Matlab together with LSB, FM and EWM, and the SNR and SPCC are measured in every scheme. Unlike LSB, FM and EWM schemes are not suitable for real time processing because they need to know in advance the host signal and the secret message. The detail of the algorithms LSB, FM and EWM are presented in [56]. The SNR is calculated as follows:

$$SNR = 10 * \log_{10} \left(\frac{\sum (x)^2}{\sum (x - y)^2} \right) \quad (6.21)$$

Where x , y , are the input signal and the output signal, respectively. At the embedding module, the input signal is the *host* and the output signal is the *stego*; while in the extraction module, the input signal is the *secret* and the output signal is the *recovered secret*. SNR measures the level of noise of the output signal in relation to the input signal, while ρ^2 measures the level of similarity between the input and output signals (according to eq. 3.2.). The higher SNR and ρ^2 , the better is the quality of the output signal. The results of the simulations are illustrated in Table 6.5.

Table 6.5. Quality of the stego signal and the recovered secret message. [92]

Method	Host & Stego		Secret & Recovered Secret	
	SNR	ρ^2	SNR	ρ^2
FM	22.66	0.993	13.99	0.974
iEWM	30.23	0.998	30.64	0.999
LSB	33.88	0.999	18.14	0.985
Proposed	33.88	0.999	16.65	0.978

In relation to the stego signal, the current proposal provides the same results as the LSB scheme and better than the FM scheme. In relation to the recovered secret message, the quality decreases in relation to the LSB scheme but it is better than the FM scheme, again. Although the best global results correspond to the iEWM scheme, that is not suitable for real-time implementation. On the other hand, the current proposal has the advantage over LSB than it uses an adaptive key and therefore the security of the system increases.

6.6. Summary

The wavelet-based speech-in-speech hiding scheme encompasses decomposition, sorting, substitution and reconstruction.

The *dwt-idwt* blocks use multiplierless topology with the following characteristics:

- (i) The 5/3 wavelet base is factorized so that the weights of the filters are represented by rational numbers of small integers. A post-scaling stage is added to obtain a reconstructed signal with the same dynamic range of the input signal.
- (ii) The symmetry property of the biorthogonal base (5/3) is taken into account.
- (iii) The reconstruction error is zero if the input signal is an even number and it is lower than 0.01% if the input signal is an odd number.

Finally, the entire design has the following characteristics:

- (i) At the transmitter, the coarse-secret's coefficients are relocated before the hiding process according to an adaptive key. The adaptive key is hidden into the detail-host's coefficients. The detail-secret's coefficients are discarded.
- (ii) At the receiver, the relocation process is completely reversed because the detail-host's coefficients were forced (at the transmitter) to be even numbers and therefore the recovered secret *key* is exactly equal to the original secret *key*. The recovered secret message is highly similar to the original secret message (it is not equal because the detail-secret's coefficients were not hidden).
- (iii) The latency and hardware resources of the entire design are extremely low.

7. Conclusions

Although in every chapter a summary section has been included, the purpose of this chapter is to present the general conclusions of the research work. The relationship among the chapters and the future work is also taken into account.

7.1. General conclusions

The general conclusions of the research work are:

- (i) The ability of adaptation of speech signals is a useful tool to transmit secure speech signals. It can be used in steganographic systems as well as cryptographic systems.
- (ii) It was demonstrated that the ability of adaptation is a feasible operation if some requirements are satisfied. The adaptation is carried out between sounds of different characteristics like their nature (vowels, words), the gender of the speaker (female, male) and the language of the plain-text.
- (iii) In the case of steganography, two schemes were proposed. The first one is known as Efficient Wavelet Masking (EWM) and the second one as improved Efficient Wavelet Masking (iEWM). They take advantage of the masking property of the HAS by using an *efficient* process of masking based on the adaptation of the secret message to the host signal.
- (iv) EWM demonstrates that the statistical transparency is significantly better than the obtained in LSB, FM, SS and SSA. The error between the statistics of the host signal and the stego signal was always lower than 15%. Additionally, the maximum hiding capacity is higher than in SS and SSA and similar to LSB and FM.
- (v) iEWM has better robustness against signal manipulation than in EWM, LSB and FM. Its transparency is slightly lower than in EWM and the hiding capacity is the same. In terms of trade-off among transparency, robustness and hiding capacity, iEWM is the best scheme in comparison with the reviewed schemes in the literature. However, in terms of statistical transparency, the best is EWM.

- (vi) In the case of cryptography, a novel scheme of speech scrambling was proposed. Unlike traditional approaches, the scrambled speech signal is a legible speech signal and the permutation process is based on the adaptation between the secret message and a target speech signal. The scheme can be viewed as a special case of Time-Frequency Scrambling, TFS.
- (vii) The main advantage of the proposed speech scrambling scheme over the known permutation-based speech scrambling schemes is that the perfect secrecy is guaranteed because the key-space is equal to the secret-space and the scrambled-space. The mapping between the input and the output is one-to-one. In addition, the low residual intelligibility is satisfied as the high quality of the recovered secret message.
- (viii) Since both the proposed steganography scheme and the scrambling scheme require knowing in advance the speech signals, they are not suitable for real-time operation. Therefore, in the proposal on hardware devices the adaptation is carried out in small frames. The stego signal is obtained quasi-immediately at the time that the speech signal and the host signal are pronounced. Additionally, the transparency is similar to the obtained in LSB scheme but the security is higher.

7.2. Future work

Although the aim of the research and its specific objectives has been covered in the current PhD project, some topics for a future work can be identified:

- (i) In the proposal, the ability of adaptation of speech signals is carried out by a deterministic search but there is at least one alternative to provide it. A heuristic search is an alternative solution and this can decrease (or not) the execution time. A research that compares the response time and the effectiveness of the algorithm is a future work.
- (ii) On the other hand, the adaptation per time-frames can be considered, too. For example, if the secret message is too long (several minutes), the adaptation can be carried out by time-frames of seconds and then, the execution time can decrease in comparison to the case when the entire speech signal is adapted. In this case, the key encompasses several sub-keys. The research should analyze if additional requirements are needed as well as the quality of the adapted speech signal by time-frames. In a similar way of the above point, the execution time should be compared, too.

8. Thesis results dissemination

The purpose of this chapter is to collect the results of the PhD research work. It encompasses published papers as well as accepted papers in international journals.

8.1. Journals: published papers

The following papers have been published as result of the research work:

D.M. Ballesteros L, J.M. Moreno A, On the ability of adaptation of speech signals and data hiding, Expert Systems with Applications, vol. 39, 2012, pp. 12574-12579.

<http://dx.doi.org/10.1016/j.eswa.2012.05.027>

D.M. Ballesteros L, J.M. Moreno A, Highly transparent steganography model of speech signals using Efficient Wavelet Masking, Expert Systems with Applications, vol. 39, 2012, pp. 9141-9149. <http://dx.doi.org/10.1016/j.eswa.2012.02.066>

D.M. Ballesteros L, J.M. Moreno A, (In Press) Real-time, speech-in-speech hiding scheme based on least significant bit substitution and adaptive key, Comput Electr Eng, 2013, <http://dx.doi.org/10.1016/j.compeleceng.2013.02.006>

D.M. Ballesteros L, J.M. Moreno A, (In Press) Wavelet-denoising on hardware devices with Perfect Reconstruction, low latency and adaptive thresholding, Comput Electr Eng, 2013, <http://dx.doi.org/10.1016/j.compeleceng.2013.03.005>

Dora M. Ballesteros L, Juan M. Moreno A, A bit more on the ability of adaptation of speech signals. Rev. Fac. Ing. Univ. Antioquia, vol. 66, Issue 1, 2013, pp. 82-90. <http://aprendeenlinea.udea.edu.co/revistas/index.php/ingenieria/article/view/15042/13127> (last checked 18.04.13).

8.2. Journals: under review

The following paper has been submitted and it is under review:

D.M. Ballesteros L, J.M. Moreno A, (Under Review) Speech scrambling based on the ability of adaptation of speech signals, Digital Signal Processing. (Submitted: 27.10.12).

References

- [1] S. Katzenbeisser, F. Petitcolas, Information Hiding: techniques for steganography and digital watermarking. Artech House, 2000, pp. 237.
- [2] R.J. Anderson, F. Petitcolas, On the limits of steganography, IEEE Journal on Selected Areas in Communications, 16 (1998) 474-481.
- [3] F. Petitcolas, R.J. Anderson, M.G. Kuhn, Information hiding-a survey, Proceedings of the IEEE 87 (1999) 1062-1078.
- [4] L. Liu, M. Tamer, Encyclopedia of Database Systems. Steganography. Springer Media, LLC, 2009, doi: 10.1007/978-0-387-39940-9_1487
- [5] J. Fridrich, Applications of data hiding in digital images, Fifth International Symposium on Signal Processing and Its Applications, ISSPA (1999), vol.1, pp.9.
- [6] J.A. Buchmann, Encryption (Chapter 3) in: Introduction to Cryptography. Ed. Springer, 2001, pp. 69-101.
- [7] N. Koblitz, Cryptography (Chapter 3) in: A Course in Number Theory and Cryptography. Springer-Verlag, 1987, pp. 53-80.
- [8] J.A. Buchmann, Probability and Perfect Secrecy (Chapter 4) in: Introduction to Cryptography, Ed. Springer, 2001, pp. 103-113.
- [9] E. Cole. Cryptography Explained (Chapter 2) in: Hiding in Plain Sight: Steganography and the Art of Covert Communication, Wiley Publishing, Inc, 2003, pp. 13-50.
- [10] I. Avcibas, Audio steganalysis with content-independent distortion measures, IEEE Signal Processing Letters, 13 (2006) 92-95.
- [11] Computer Science and Communications Dictionary, Kluwer Academic Publishers 2000, doi:10.1007/1-4020-0613-6_3995.

- [12] Q. Yin-Cheng, Y. Liang, L. Chong, Wavelet domain audio steganalysis for multiplicative embedding model, International Conference on Wavelet Analysis and Pattern Recognition, 2009, pp. 429-432.
- [13] Y. Liang, Q. Xiu-Juan, L. Lin-Jie, Z. Yi, Time domain speech steganalysis method based on multiplicative embedding model, International Conference on Wavelet Analysis and Pattern Recognition, 2012, pp. 148-151.
- [14] Q. Liu , A.H. Sung, M. Qiao, Spectrum Steganalysis of WAV Audio Streams, 6th International Conference on Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, 2009, Ed. Petra Perner, pp. 582-593.
- [15] L. Qingzhong, A.H. Sung, Q. Mengyu, Temporal Derivative-Based Spectrum and Mel-Cepstrum Audio Steganalysis, IEEE Transactions on Information Forensics and Security, 4 (2009) 359-368.
- [16] Q. Yin-Cheng, Z. Jing-Na, W. Wei-Liang, Time-Spread Echo Hiding Steganalysis Algorithm Based on Ensemble Learning, International Conference on Network Computing and Information Security, 2011, pp. 426-429.
- [17] L. Cairong, Z. Wei, A. Haojun, H. Ruimin, Steganalysis of Spread Spectrum Hiding Based on DWT and GMM, International Conference on Networks Security, Wireless Communications and Trusted Computing, 2009, pp. 240-243.
- [18] I. Avcibas, Audio steganalysis with content-independent distortion measures, IEEE Signal Processing Letters, 13 (2006) 92-95.
- [19] O.H. Kocal, E. Yuruklu, I. Avcibas, Chaotic-Type Features for Speech Steganalysis, IEEE Transactions on Information Forensics and Security, 3 (2008) 651-661.
- [20] B. Goldberg, S. Sridharan, E. Dawson, Design and cryptanalysis of transform-based analog speech scramblers, IEEE Journal on Selected Areas in Communications, 11 (1993) 735-744.

- [21] B. Goldberg, S. Sridharan, E. Dawson, Cryptanalysis of frequency domain analogue speech scramblers, IEE Proceedings I: Communications, Speech and Vision, 140 (1993) 235-239.
- [22] J.A. Apolinario, Jr., P.R.S. Mendonca, R.O. Chaves, L.P. Caloba, Cryptanalysis of speech signals ciphered by TSP using annealed Hopfield neural network and genetic algorithms, IEEE 39th Midwest symposium on Circuits and Systems, 1996, pp. 821-824.
- [23] N. Cvejic, T. Seppanen, A wavelet domain LSB insertion algorithm for high capacity audio steganography, IEEE 10th Digital Signal Processing Workshop, 2002, pp. 53-55.
- [24] N. Cvejic, T. Seppanen, Channel capacity of high bit rate audio data hiding algorithms in diverse transform domains, IEEE International Symposium on Communications and Information Technology, 2004, vol. 81, pp. 84-88.
- [25] H. Fastl, E. Zwicker, Psychoacoustics: Facts and Models, Springer Series in Information Sciences, Third Edition, 2007.
- [26] D.E. Skopin, I.M.M. El-Emary, R.J. Rasras, R.S. Diab, Advanced algorithms in audio steganography for hiding human speech signal, 2nd International Conference on Advanced Computer Control (ICACC), 2010, pp. 29-32.
- [27] P. Dutta, D. Bhattacharyya, T-H Kim, Data Hiding in Audio Signal: A Review, International Journal of Database Theory and Application, vol.2, No. 2, June 2009, pp. 1-8.
- [28] K. Gopalan, S.J. Wennedt, D. Haddad, Steganographic method for covert audio communications, U.S. patent No. 7,231,271, June 2007.
- [29] K. Gopalan, Audio steganography method and apparatus using cepstrum modification, U.S. patent No. 7,555,432, June 2009.

- [30] K. Gopalan, S. Wenndt, Audio Steganography for cover data transmission by imperceptible tone insertion, IASTED International Conference on Communication Systems and Applications, 2004.
- [31] K. Gopalan, Audio Steganography by Cepstrum modification. ICASSP 2005, pp. 481-484.
- [32] K. Gopalan, Audio steganography by modification of cepstrum at a pair of frequencies, 9th International Conference on Signal Processing, ICSP, 2008, pp. 2178-2181.
- [33] N. Cvejic, T. Seppanen, A wavelet domain LSB insertion algorithm for high capacity audio steganography, IEEE 10th Digital Signal Processing Workshop, 2002, pp. 53-55.
- [34] S. Shirali-Shahreza, M.T. Manzuri-Shalmani, High Capacity error free wavelet domain speech steganography, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 1729-1732.
- [35] F. Djebbar, K. Abed-Meraim, D. Guerchi, H. Hamam, Dynamic energy based text-in-speech spectrum hiding using speech masking properties, 2010 2nd International Conference on Industrial Mechatronics and Automation (ICIMA), 2010, pp. 422-426.
- [36] F. Djebbar, H. Hamam, K. Abed-Meraim, D. Guerchi, Controlled distortion for high capacity data-in-speech spectrum steganography, 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 212-215.
- [37] F. Djebbar, D. Guerchi, K. Abed-Meraim, H. Hamam, Text Hiding in High Frequency Components of Speech Spectrum, 10th International Conference on Information Science, Signal Processing and their Applications, ISSPA, 2010, pp. 666-669.

- [38] T. Rabie, D. Guerchi, Magnitude Spectrum Speech Hiding, IEEE International Conference on Signal Processing and Communications, ICSPC 2007, pp. 1147-1150.
- [39] D.E. Skopin, I.M.M. El-Emary, R.J. Rasras, R.S. Diab, Advanced algorithms in audio steganography for hiding human speech signal, 2nd International Conference on Advanced Computer Control (ICACC), 2010, pp. 29-32.
- [40] N.S. Jayant, Analog scramblers for speech privacy, Computers & Security, 1 (1982) 275-289.
- [41] J.F. de Andrade, M.L.R. de Campos, J.A. Apolinario, Speech privacy for modern mobile communication systems, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP (2008), pp. 1777-1780.
- [42] V.J. Philips, M.H. Lee, J.E. Thomas, Speech Scrambling by the Re-ordering of Amplitude Samples, Radio and Electronic Engineer, Vol. 41, No 3, March 1971, pp. 99-112.
- [43] A. Matsunaga, K. Koga, M. Ohkawa, An analog speech scrambling system using the FFT technique with high-level security, IEEE Journal on Selected Areas in Communications, 7 (1989) 540-547.
- [44] R.W. Woo, C. Leung, A new key generation method for frequency-domain speech scramblers, IEEE Transactions on Communications, 45 (1997) 749-752.
- [45] E. Mosa, N.W. Messiha, O. Zahran, Chaotic encryption of speech signals in transform domains, International Conference on Computer Engineering & Systems, ICCES 2009, pp. 300-305.
- [46] M. Fulong, C. Jun, W. Yumin, Wavelet transform-based analogue speech scrambling scheme, Electronics Letters 32 (1996) 719-721.
- [47] S.B. Sadkhan, N. Abdulmuhsen, N.F. Al-Tahan, A Proposed Analog Speech Scrambler Based on Parallel Structure of Wavelet Transforms, Radio Science Conference, 2007. NRSC 2007. National, 2007, pp. 1-12.

- [48] P. Krishnamoorthy, S.R.M. Prasanna, Enhancement of noisy speech by temporal and spectral processing, *Speech Communication*, 53 (2011) 154-174.
- [49] T.S. Gunawan, E. Ambikairajah, J. Epps, Perceptual speech enhancement exploiting temporal masking properties of human auditory system, *Speech Communication*, 52 (2010) 381-393.
- [50] M. Bahoura, J. Rouat, Wavelet speech enhancement based on time-scale adaptation, *Speech Communication* 48 (2006) 1620-1637.
- [51] B. Linkai, T.D. Church, Perceptual speech processing and phonetic feature mapping for robust vowel recognition, *IEEE Transactions on Speech and Audio Processing*, 8 (2000) 105-114.
- [52] A.K. Vuppala, J. Yadav, S. Chakrabarti, K.S. Rao, Vowel Onset Point Detection for Low Bit Rate Coded Speech, *IEEE Transactions on Audio, Speech, and Language Processing*, 20 (2012) 1894-1903.
- [53] Z. Chi, J.H.L. Hansen, Whisper-Island Detection Based on Unsupervised Segmentation With Entropy-Based Speech Feature Processing, *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (2011) 883-894.
- [54] T. Zheng-Hua, B. Lindberg, Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection, *IEEE Journal of Selected Topics in Signal Processing*, 4 (2010) 798-807.
- [55] M. Huijbregts, F. de Jong, Robust speech/non-speech classification in heterogeneous multimedia content, *Speech Communication*, 53 (2011) 143-153.
- [56] D.M. Ballesteros L, J.M. Moreno A, Highly transparent steganography model of speech signals using Efficient Wavelet Masking, *Expert Systems with Applications*, 39 (2012) 9141-9149.
- [57] D.M. Ballesteros L, J.M. Moreno A, On the ability of adaptation of speech signals and data hiding, *Expert Systems with Applications*, 39 (2012) 12574-12579.

- [58] EBU Technical, Sound Quality Assessment Material recordings for subjective tests, 2008. (<http://tech.ebu.ch/publications/sqamcd>) (last checked 28.02.12).
- [59] Dora M. Ballesteros L , Juan M. Moreno A, A bit more on the ability of adaptation of speech signals. Rev. Fac. Ing. Univ. Antioquia N.º 66 pp. 82-90. Marzo, 2013. (<http://aprendeenlinea.udea.edu.co/revistas/index.php/ingenieria/article/view/15042/13127>) (last checked 18.04.13).
- [60] E. Mosa, N.W. Messiha, O. Zahran, Random encryption of speech signal, International Conference on Computer Engineering & Systems, ICCES (2009), pp. 306-311.
- [61] S.C. Kak, B.E. Encryption of signals using data transpositions. Proc. IEE, Vol. 125, No. 12, (1978) 1327-1328.
- [62] E. Del Re, R. Fantacci, D. Maffucci, A new speech signal scrambling method for secure communications: theory, implementation, and security evaluation, IEEE Journal on Selected Areas in Communications, 7 (1989) 474-480.
- [63] S. Sridharan, E. Dawson, B. Goldberg, Fast Fourier transform based speech encryption system, IEE Proceedings Communications, Speech and Vision, 138 (1991) 215-223.
- [64] Y. C. Lim, J. W. Lee, S. W. Foo, Quality Analog Scramblers Using Frequency-Response Masking Filter Banks, Circuits, Systems, and Signal Processing, 29 (2010) 135-154.
- [65] M.S. Ehsani, S.E. Borujeni, Fast Fourier transform speech scrambler, Intelligent Systems, First International IEEE Symposium, (2002), pp. 248-251.
- [66] D.C. Tseng, J.H. Chiu, An OFDM speech scrambler without residual intelligibility, IEEE Region 10 Conference, TENCON (2007), pp. 1-4.

- [67] L. Qiu-Hua, Y. Fui-Liang, M. Tie-Min, L. Hua-Lou, A speech encryption algorithm based on blind source separation, International Conference on Communications, Circuits and Systems, ICCAS (2004), pp. 1013-1017.
- [68] L. Qiu-Hua, Y. Fu-Liang, M. Tie-Min, L. Hualou, A blind source separation based method for speech encryption, IEEE Transactions on Circuits and Systems I: Regular Papers, 53 (2006) 1320-1328.
- [69] A. Mermoul, A. Belouchrani, A subspace-based method for speech encryption, 10th International Conference on Information Sciences Signal Processing and their Applications, ISSPA (2010), pp. 538-541.
- [70] A. Mermoul, An iterative speech encryption scheme based on subspace technique, 7th International Workshop on Systems, Signal Processing and their Applications, WOSSPA (2011), pp. 361-364.
- [71] A. Mermoul, A. Belouchrani, Subspace-based technique for speech encryption, Digital Signal Processing, 22 (2012) 298-303.
- [72] L. Shujun, L. Chengqing, L. Kwok-Tung, C. Guanrong, Cryptanalyzing an Encryption Scheme Based on Blind Source Separation, IEEE Transactions on Circuits and Systems I: Regular Papers, 55 (2008) 1055-1063.
- [73] H. Li, Z. Qin, L. Shao, S. Zhang, B. Wang, Variable Dimension Space Audio Scrambling Algorithm Against MP3 Compression, in: A. Hua, S.-L. Chang, (Eds.), Algorithms and Architectures for Parallel Processing, Springer Berlin Heidelberg, 2009, pp. 866-876.
- [74] A. Madain, A. Abu Dalhoum, H. Hiary, A. Ortega, M. Alfonseca, Audio scrambling technique based on cellular automata, Multimedia Tools and Applications (2012) 1-20.
- [75] C.E. Shannon, Communication Theory of Secrecy Systems, Bell System Technical Journal, vol. 28-4, 1949, pp. 656-715.

- [76] D.M. Ballesteros L, J.M. Moreno A, (Under Review) Speech scrambling based on the ability of adaptation of speech signals, Digital Signal Processing. (Submitted 27.10.12).
- [77] Servei de Llengües i Terminologia (SLT). Universitat Politècnica de Catalunya. Class-Talk: A University teaching phrasebook. (<http://www.upc.edu/slt/classtalk/>) (last checked 17.10.12).
- [78] J. Benesty, C. Jingdong, H. Yiteng, On the Importance of the Pearson Correlation Coefficient in Noise Reduction, IEEE Transactions on Audio, Speech, and Language Processing, 2008, pp. 757-765.
- [79] F. Djebbar, B. Ayad, H. Hamam, K. Abed-Meraim, A view on latest audio steganography techniques, International Conference on Innovations in Information Technology (IIT), 2011, pp. 409-414.
- [80] F. Djebbar, B. Ayad, K. Meraim, H. Hamam, Comparative study of digital audio steganography techniques, Eurasip Journal on Audio, Speech, and Music Processing 2012 (2012) 1-16.
- [81] M. Sheikhan, K. Asadollahi, R. Shahnazi, Improvement of embedding capacity and quality of DWT-based audio steganography systems, World Applied Sciences Journal 13 (2011) 507-516.
- [82] L. Jin, Z. Ke, T. Hui, Least-significant-digit steganography in low bitrate speech, IEEE International Conference on Communications (ICC), 2012, pp. 1133-1137.
- [83] T. Hui, Z. Ke, H. Yongfeng, F. Dan, L. Jin, A Covert Communication Model Based on Least Significant Bits Steganography in Voice over IP, The 9th International Conference for Young Computer Scientists, ICYCS 2008, pp. 647-652.
- [84] T. Hui, Z. Ke, J. Hong, H. Yongfeng, L. Jin, F. Dan, An adaptive steganography scheme for voice over IP, IEEE International Symposium on Circuits and Systems, ISCAS 2009, pp. 2922-2925.

- [85] F. Xihui, D. Wenlong, Audio information hiding algorithm based on energy compare, IEEE International Conference on Computer Science and Automation Engineering (CSAE), 2012, pp. 491-493.
- [86] S. Shokri, M. Ismail, N. Zainal, Voice quality in speech watermarking using spread spectrum technique, International Conference on Computer and Communication Engineering (ICCCE) 2012, pp. 169-173.
- [87] Z. Qingquan, G. Wei, A speech information hiding algorithm based on the energy difference between the frequency band, Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on, 2012, pp. 3078-3081.
- [88] S. Rekik, D. Guerchi, S.A. Selouani, H. Hamam, Speech steganography using wavelet and Fourier transforms, Eurasip Journal on Audio, Speech, and Music Processing 2012 (2012).
- [89] H.M.D. Kabir, S.B. Alam, Hardware based realtime, fast and highly secured speech communication using FPGA, IEEE International Conference on Information Theory and Information Security (ICITIS), 2010, pp. 452-457.
- [90] S. Arora, S. Emmanuel, Real-time adaptive speech watermarking scheme for mobile applications, Joint Conference of the Fourth International Conference on Communications and Signal Processing and Fourth Pacific Rim Conference on Multimedia, pp. 1153-1157 vol.1152.
- [91] P. Karthigaikumar, K. Jaraline Kirubavathy, K. Baskaran, FPGA based audio watermarking—Covert communication, Microelectronics Journal 42 (2011) 778-784.
- [92] D.M. Ballesteros L, J.M. Moreno A, Real-time, speech-in-speech hiding scheme based on least significant bit substitution and adaptive key, Comput Electr Eng, 2013, <http://dx.doi.org/10.1016/j.compeleceng.2013.02.006>

- [93] D.M. Ballesteros L, J.M. Moreno A, Wavelet-denoising on hardware devices with Perfect Reconstruction, low latency and adaptive thresholding, *Comput Electr Eng*, 2013, <http://dx.doi.org/10.1016/j.compeleceng.2013.03.005>
- [94] G. Strang ,T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Second Edition, 1997.
- [95] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, Third Edition, 2008.
- [96] R. Perez-Andrade, R. Cumplido, C. Feregrino-Urbe, F. Martin Del Campo, A versatile linear insertion sorter based on an FIFO scheme, *Microelectronics Journal* 40 (2009) 1705-1713.
- [97] J. Ortiz, D. Andrews, A configurable high-throughput linear sorter system, *IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*, 2010, pp. 1-8.
- [98] Lee, D.-U., Kim, L.-W., Villasenor, J.D. Precision-aware self-quantizing hardware architectures for the discrete wavelet transform. *IEEE Transactions on Image Processing*, vol. 21 No. 2 (2012) 768-777.
- [99] Yeong-Kang Lai, Lien-Fei Chen, Yui-Chih Shih . A High-Performance and Memory-Efficient VLSI Architecture with Parallel Scanning Method for 2-D Lifting-Based Discrete Wavelet Transform. *IEEE Transactions on Consumer Electronics*, Vol. 55, No. 2, pp. 400-407. May 2009.
- [100] K.A. Kotteri, A.E. Bell, J.E. Carletta, Design of multiplierless, high-performance, wavelet filter banks with image compression applications, *IEEE Transactions on Circuits and Systems I: Regular Papers*, 51 (2004) 483-494.
- [101] K.A. Kotteri, S. Barua, A.E. Bell, J.E. Carletta, A comparison of hardware implementations of the biorthogonal 9/7 DWT: convolution versus lifting, *IEEE Transactions on Circuits and Systems II: Express Briefs*, 52 (2005) 256-260.

- [102] K.A. Kotteri, A.E. Bell, J.E. Carletta, Multiplierless Filter Bank Design: Methods that Improve Both Hardware and Image Compression Performance, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 16, No. 6, June 2006. Pp. 776-780.
- [103] A. Abbas, T.D. Tran, Multiplierless Design of Biorthogonal Dual-Tree Complex Wavelet Transform using Lifting Scheme, *IEEE International Conference on Image Processing*, 2006, pp. 1605-1608.
- [104] A. Abbas, T.D. Tran, Rational Coefficient Dual-Tree Complex Wavelet Transform: Design and Implementation, *IEEE Transactions on Signal Processing*, 56 (2008) 3523-3534.
- [105] Z. Ming, D. Rangyu, M. Zhuo, Z. Minxuan, A FPGA-based low-cost real-time wavelet packet denoising system, *International Conference on Electronics and Optoelectronics (ICEOE)*, pp. V2-350-V352-353.
- [106] H Kim, A Parallel Algorithm for the Biorthogonal Wavelet Transform Without Multiplication, *Lecture Notes in Computer Science*, 2004.