EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Speech steganography using wavelet and Fourier transforms

Siwar Rekik[1,2*], Driss Guerchi[2], Sid-Ahmed Selouani[3] and Habib Hamam[4]

Abstract

A new method to secure speech communication using the discrete wavelet transforms (DWT) and the fast Fourier transform is presented in this article. In the first phase of the hiding technique, we separate the speech high-frequency components from the low-frequency components using the DWT. In a second phase, we exploit the low-pass spectral proprieties of the speech spectrum to hide another secret speech signal in the low-amplitude high-frequency regions of the cover speech signal. The proposed method allows hiding a large amount of secret information while rendering the steganalysis more complex. Experimental results prove the efficiency of the proposed hiding technique since the stego signals are perceptually indistinguishable from the equivalent cover signal, while being able to recover the secret speech message with slight degradation in the quality.

Keywords: Audio steganography, Discrete wavelet transform, Fast Fourier transform, Data hiding, Speech steganography

Introduction

One of the concerns in the field of secure communication is the concept of information security. Today's reality is still showing that communication between two parties over long distances has always been subject to interception. Providing secure communication has driven researchers to develop several cryptography schemes. Cryptography methods achieve security in order to make the information unintelligible to guarantee exclusive access for authenticated recipients. Cryptography consists of making the signal look garbled to unauthorized people. Thus, cryptography indicates the existence of a cryptographic communication in progress, which makes eavesdroppers suspect the existence of valuable data. They are thus incited to intercept the transmitted message and to attempt to decipher the secret information. This may be seen as weakness in cryptography schemes. In contrast to cryptography, steganography allows secret communication by camouflaging the secret signal in another signal (named the cover signal), to avoid suspicion. This quality motivated the researchers to work on this burning field to develop schemes ensuring better resistance to hostile attackers.

The word steganography is derived from two Greek words: Stego (means cover) and graphy (means writing). The two combined words constitute steganography, which means covert writing, it is the art of hiding written communications. Several steganography techniques were used to send message secretly during wars through the territories of enemies. The use of steganography dates back to ancient time where it was used by romans and ancient Egyptians [1]. One technique according to Greek historian Herodotus was to shave the head of a slave, tattoo the message on the slave's scalp, and send him after his hair grew back. Another technique was to write the secret message underneath the wax of a writing tablet. A third one is to use invisible ink to write secret messages within covert letters [2].

Many techniques have been developed for hiding secret signals into other cover signals. Sridevi et al. [3] presented a method for audio steganography. It consists of substituting the least significant bit (LSB) of each sample of the cover speech signal with the secret data. While this method is easy to implement and can be used to hide larger secret messages, it cannot protect the hidden message from small modifications that can happen as a result of format conversion or compression. Hiding data in LSBs of audio samples in the time domain is one of the simplest algorithms enabling a very high data rate of inserted information. However,

* Correspondence: Siwar.Rekik@etudiant.univ-brest.fr
[1]Université de Bretagne Occidentale, Brest, France
[2]Canadian University of Dubai, Dubai, UAE
Full list of author information is available at the end of the article

several steganalysis algorithms have been developed to challenge the robustness of this method. Bender et al. [4] have presented a technique for data hiding based on phase coding. This method consists of substituting the phase of the first part of an audio segment by a reference phase that represents the data. In order to conserve the relative phase between segments, an adjustment must be made in the phase of the succeeding segment. The series of steps of phase coding is as follows: (i) The original audio signal is decomposed into smaller segments such that their length is equal to the size of the message to be encoded; (ii) A discrete Fourier transform (DFT) is then applied on each segment leading to a phase matrix; (iii) Compute the differences between the phase of each pair of the consecutive segments; (iv) Identify the phase shifts between the consecutive segments. Although, the absolute phases of the segments may change, the relative phase differences between the consecutive segments must remain unchanged; (v) Use the new phase of the first segment and the set of original phase differences to create a new phase matrix; (vi) Regenerate the audio signal with an inverse DFT and then connect the audio segments together. This step is based on the original magnitude matrix and the newly created phase matrix. The receiver determines the length of the secret message, then applies a DFT and extract the hidden message from the cover signal. A distinctive characteristic of phase coding is the low data transmission rate due to the fact that the secret data are encoded only in the first segment of the audio signal. Controversially, any enhancement in the length of the segment may result in shifting the phase relations among the frequency elements of the segment, leading therefore to an easier detection of the existence of a secret message. Thus, the phase coding algorithm is more efficient when hiding small amount of data. Kirovski and Malvar [5] have proposed a new steganographic scheme, called Spread Spectrum (SS) coding method. This method randomly spreads the bits of the secret data message across the frequency spectrum of the audio signal. However, in contrast to LSB coding, the SS coding scheme spreads the secret message using a code independent from the concrete cover signal. The SS coding technique may outperform the LSB coding and phase coding techniques by offering a good quality for medium data transmission rates while ensuring a high level of robustness against steganalysis. However, similarly to the LSB coding technique, the SS method may introduce noise to the audio file. This is presenting a weakness since it facilitates detection by steganalysis systems.

Huang and Yeo [6] have presented an information hiding method based on echo hiding. An echo is introduced into the discrete audio signal in order to embed secret information. Similar to the SS coding method, echo hiding is used to provide a better data transmission rate and higher robustness comparing to the noise-inducing techniques. To accomplish successfully the hiding process, three fundamental parameters need to be changed from the original signal: decay rate, offset (time delay), and amplitude. These three parameters are easily defined since they are located below the human audible threshold limit which is different from the echo. Also, the offset is altered to characterize the binary message to be hidden. The first and the second offsets represent a one (binary) and a zero (binary), respectively. Shirali and Shahreza [7] present an approach for hiding information in a speech signal. This method consists of detecting the silence intervals of a speech and the corresponding length of these intervals (number of samples) and changing them with the secret information. Hiding data in silent interval of the audio samples is one of the simplest algorithms enabling a very high data rate of inserted information. However, this method is already well known and several steganalysis algorithms have been developed to defeat the robustness of this method.

Speech steganography takes advantage of the recent advancements in speech compression and data hiding. Speech is a low-pass signal; its intelligibility is retained when preserving at least the first three formants of the magnitude spectrum. In this article, we will take advantage of these speech characteristics to propose an efficient speech-in-speech hiding method. Our speech steganography system consists of embedding the secret speech parameters in the high-frequency regions of the magnitude spectrum of the cover speech. Our aim is to ensure that the stego signal obtained from combining the original phase spectrum and the modified magnitude spectrum shows similar subjective quality to the cover signal. Theoretically, the resultant stego speech is expected to be perceptually indistinguishable from the cover speech since the pertinent low-frequency components will remain intact.

Potential applications of our speech hiding scheme are reduction of speech storage and transmission overhead in electronic voice mail applications and audio streaming, speech translation, data communication secrecy, and many other web-based applications.

Objectives

Our objective is to develop a high performance speech steganography system. The design of such system consists principally of the optimization of the following attributes:

- The hiding capacity, defined by the amount of the secret information (speech, text, or image) to be hidden in the cover speech signal.
- The impact of the hiding process on the cover speech quality. We hope to produce a stego signal that is perceptually indistinguishable from the cover signal.

- The complexity of the steganography system. Our aim is to render the steganalysis (the attempt to discover the existence of the secret message from the stego signal) by the opponent more complex.
- The accuracy with which the hidden message can be recovered at the receiver. Efficient techniques are to be developed to minimize the impact of the compression on the stego signal.

We choose a speech signal as secret information to be hidden in the cover speech. Since our objective in discrete wavelet transform-fast Fourier transform (DWT-FFT)-based hiding approach is secrecy, we propose to hide the secret information within the high-frequency of the wavelet components.

The rest of the article is organized as follows: in the following section, we introduce our DWT-FFT-based approach dedicated for the steganography task. Section "Secret speech parameterization" will describe the secret speech analysis including the linear predictive coding (LPC) analysis and the line spectral frequencies (LSF) extraction procedure. In Section "Speech hiding algorithm", we proceed with the description of the used speech hiding algorithm. The general step to retrieve the secret speech signal is also included in this section. Then a description of the speech signals database used for our simulations, the parameters of our experiments, the evaluation and discussion of the results of our proposed DWT-FFT hiding approach are presented in Section "Evaluation". Finally, we conclude and suggest directions for further research in Section "Conclusions".

DWT-FFT-based approach

Speech DWT

The wavelet transform can be considered as transforming the signal from the time domain to the wavelet domain. This new domain contains more complicated basis functions called wavelets, mother wavelets, or analyzing wavelets [8]. The fundamental idea behind wavelets is to analyze according to scale. Any signal can then be represented by translated and scaled versions of the mother wavelet. Wavelet analysis is capable of enlightening aspects of data that other signal analysis techniques are unable to perform, aspects like trends, and discontinuities in higher derivatives, breakdown points, and self-similarity.

The basic idea of DWT for one-dimensional signals is shortly described. The wavelet analysis allows the split of a signal into two parts, usually the high- and the low-frequency parts. This process is called decomposition. The edge components of the signal are largely limited to the high-frequency part. The signal is passed through a series of high-pass filters to analyze the high frequencies, and it is passed through a series of low-pass filters to analyze the low frequencies. Filters of different cutoff frequencies are used to analyze the signal at different resolutions [9,10].

The DWT involves choosing scales and positions based on powers of two, the so-called dyadic scales and positions. The mother wavelet is rescaled by powers of two and transformed by integers. Specifically, a function $f(t) \in L^2(R)$ (defines space of square integrable functions) can be represented as:

$$f(t) = \sum_{j=1}^{L} \sum_{k=-\infty}^{\infty} d(j,k)\psi(2^{-j}t-k) + \sum_{k=-\infty}^{\infty} a(L,k)\phi(2^{-L}t-k) \tag{1}$$

The function $\psi(t)$ is known as the mother wavelet, while $\phi(t)$ is known as the scaling function. The set of function $\left\{\sqrt{2^{-L}}\phi(2^{-L}t-k), \sqrt{2^{-j}}\psi(2^{-j}t-k) \middle| j \le L, j, k, L \in Z\right\}$, where Z is the set of integers in an orthonormal basis for $L^2(R)$. The numbers $a(L,k)$ are known as the approximation coefficients at scale L, while $d(j,k)$ are identified as the detail coefficients at scale j. The approximation and detail coefficients can be expressed consecutively as:

$$a(L,k) = \frac{1}{\sqrt{2^L}} \int_{-\infty}^{\infty} f(t)\phi(2^{-L}t-k)dt \tag{2}$$

$$d(j,k) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} f(t)\psi(2^{-j}t-k)dt \tag{3}$$

For a better understanding of the above coefficients, let's consider a projection $f_l(t)$ of the function $f(t)$ that provides the best approximation (in the sense of minimum error energy) to $f(t)$ at a scale l. This projection can be constructed from the coefficients $a(L,k)$, using the equation:

$$f_l(t) = \sum_{k=-\infty}^{\infty} a(l,k)\phi(2^{-l}t-k) \tag{4}$$

As the scale l decreases, the approximation becomes finer, converging to $f(t)$ as $l \rightarrow 0$. The difference between the approximation at scale $l+1$ and that at l, $f_{l+1}(t) - f_l(t)$, is totally defined by the coefficients $d(j,k)$ using the equation of decomposition and can mathematically be expressed as follows:

$$f_{l+1}(t) - f_l(t) = \sum_{k=-\infty}^{\infty} d(l,k)\psi(2^{-l}t-k) \tag{5}$$

These given relations, $a(L,k)$ and $\{d(j,k)|j \le L\}$, are useful for building the approximation at any scale. Hence, the wavelet transform breaks the signal up into a coarse approximation $f_L(t)$ (given $a(L, k)$) and a number of layers of detail coefficients $\{f_{j+1} - f_j(t) | j < L\}$ (given by $\{d(j, k) | j \leq L\}$). As each layer of detail is added, the approximation at the next higher scale is achieved. The original signal can be reconstructed using the Inverse DWT (IDWT), following the above procedures in the reverse order. The synthesis starts with the approximation and detail coefficients cA_j and cD_j, and then reconstructs cA_{j-1} by up sampling and filtering with the reconstruction filters [11,12].

Speech Fourier transform

Since speech is processed on a time-frame basis, the speech spectrum is evaluated using the DFT. The DFT of a signal $s(n)$ defined for $0 \leq n \leq M - 1$ is given by

$$S(k) = \sum_{n=0}^{M-1} s(n) e^{-j2\pi kn/M}, 0 \leq \mathrm{k} \leq M - 1 \qquad (6)$$

In general, $S(k)$ is a complex function of the variable k and can be expressed in polar coordinates as:

$$S(k) = |S(k)| e^{j\phi}(k) \qquad (7)$$

The sequence $S(k)$ has the same number of elements as $s(n)$. However, the last $M/2$ elements of the DFT are conjugates of the first $M/2$ elements, in inverse order. Consequently, the magnitude spectrum $|S(k)|$ could be defined uniquely by the first $M/2$ frequency components since it satisfies the following symmetry:

$$|S(k)| = |S(M - k)| \qquad (8)$$

This equation represents one of the DFT properties that must be maintained when hiding a message in the magnitudes. This feature is used in the fast Fourier transform (FFT) algorithm to reduce the DFT computational complexity [13]. For simplicity, we will adopt in the subsequent sections the following notations:

$$S = \mathrm{fft}(s) \qquad (9)$$

and

$$s = \mathrm{ifft}(S) \qquad (10)$$

where ifft, the inverse FFT, calculates the inverse DFT.

Speech spectrum characteristics

Speech is a baseband signal with most of the pertinent intelligibility-preserving frequency components confined to a bandwidth of 4 and 7 kHz for narrowband and wideband speech, respectively [14]. The distribution of the first three speech formants represents the primary cues to the English vowels. Most of the vowel energy is condensed below 1 kHz and decays at about –6 db/oct with frequency [15]. Figure 1 shows the wideband speech spectrum for both a liquid frame and an unvoiced fricative frame. In all vowels and most of the voiced consonants, the magnitude spectrum shows very week components at high frequencies.

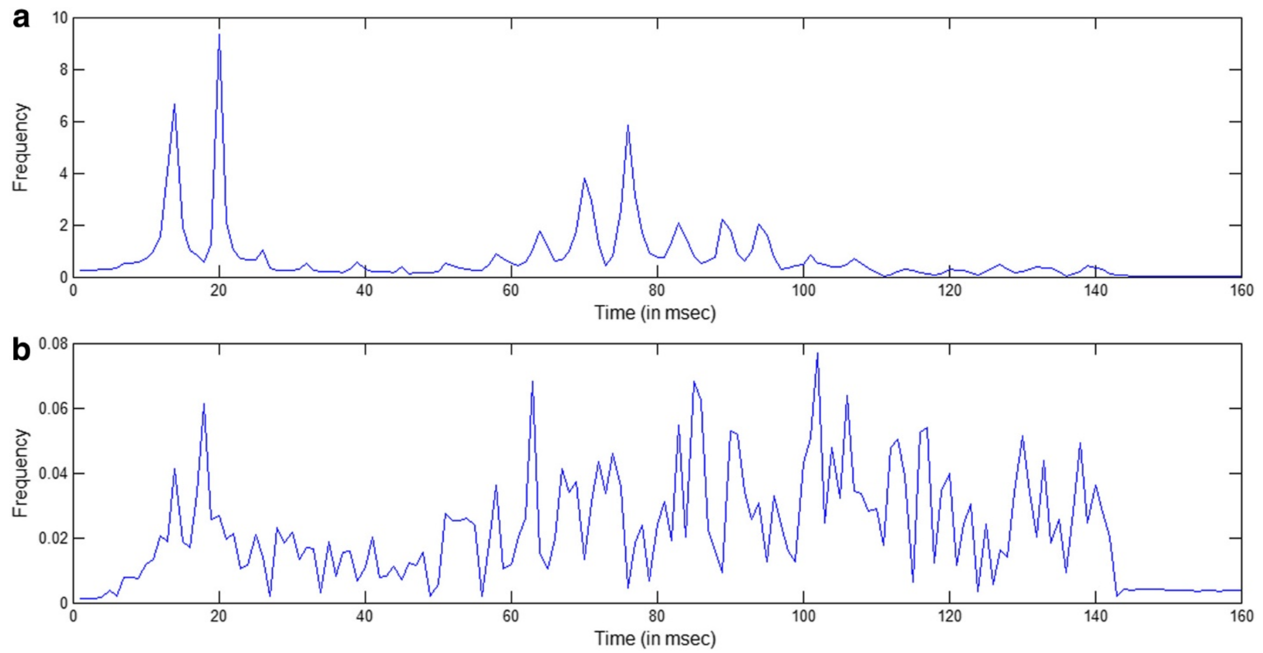

Figure 1 Magnitude spectrum of (a) a voiced frame, (b) unvoiced frame.

Even though few unvoiced fricative consonants, such as /s/, present large magnitudes at high frequencies, the intelligibility of the speech signal is negligibly affected if we do not model accurately these frequency components [14]. On the other hand, even for wideband unvoiced fricative consonants, frequencies above 7 kHz do not contribute considerably to the speech spectrum content. These two facts have motivated us to embed a separate signal in the low amplitude high-frequencies of the cover signal.

Secret speech parameterization

Many factors require the parameterization of the secret speech message before the hiding process. Among these factors, we cite the restricted number of the hiding locations in the narrowband cover speech. Speech parameterization termed as *speech analysis* is generally used in different research areas, such as automatic speech recognition and speech coding. In speech coding, the original signal is subject to a speech analysis algorithm to extract the pertinent speech parameters. In order to recreate a copy of the original signal, an inverse algorithm known as speech synthesis is used. Most of the speech analysis schemes are based on the human speech production model [15]. In this speech production model, a sequential excitation of two filters is used to produce a speech signal, a linear prediction (LP) filter is used to model the vocal tract, produces a short-term correlation present in all types of speech and a pitch filter to represent the periodicity created to the vibration of the vocal cords in voiced segments. A basic diagram of the speech production model is shown in Figure 2. The LPC is based on this diagram. The LPC schemes are usually used in the field of speech coding. For example in transmission, the speech frames are represented with a restricted number of parameters. These parameters in the receiver side are used to reconstruct a synthetic-quality speech signal. The speech analysis algorithm is based on two phases: an LP analysis to obtain p LP coefficients, $a_i(i = 1, \ldots, p)$ and a pitch analysis to extract the pitch gain g and the pitch delay d. The LP filter and the pitch filter are constructed using the LP parameters and the pitch, respectively. In the LPC model, for the unvoiced speech signal, an LP filter is used since there is no periodicity in this class of speech. The pitch filter is used for the voiced frames. Details about the speech analysis procedure are given in [16]. The LP coefficients (LPC) must be transformed to a more improved representation before any processing, since the LPC are very susceptible to errors and their direct quantization might generate an unbalanced LP filter. One of the most used representations is the LSF [17]. In this study, we adopted this representation, in the hiding process p magnitude locations are replaced by p LSF coefficients of the secret speech.

Secret speech analysis

To perform the secret speech analysis, we will use the LP speech production model. In this model, the speech signal is subject to an LP analysis followed by pitch analysis.

LP analysis

The LP analysis is performed every L-ms (for M = L × Fs samples), for a sampling frequency of Fs kHz, to extract p LP coefficients. These coefficients represent the vocal-track poles (or formants). To smooth the inter-frame variation of the spectral parameters, the analysis window contains more samples than the analysis frame. In addition to the current speech frame, the analysis window contains 5 ms from past speech and 5 ms from future speech. In the LP analysis, we adopt a tapered rectangular window with three parts [18]. The first part is the first half of a hamming window, the second part is a rectangular window, and the third part is the second half of a Hamming window. This window produces a narrower main lobe than the asymmetric window used in G.729 and G722.2 codec standards.

$$w(n) = \begin{cases} 54 - .46\cos\left(\dfrac{2\pi n}{M-1}\right), & n = 0, \ldots, \dfrac{M}{2} - 1 \\ 1, & n = \dfrac{M}{2}, \ldots, \dfrac{3M}{2} - 1 \\ .54 - .46\cos\left(\dfrac{2\pi\left(n - \dfrac{M}{2}\right)}{M-1}\right), & n = \dfrac{3M}{2}, \ldots, 2M \end{cases} \tag{11}$$

The existence of a short-term correlation in speech signals motivates us to adopt the LP analysis. This correlation

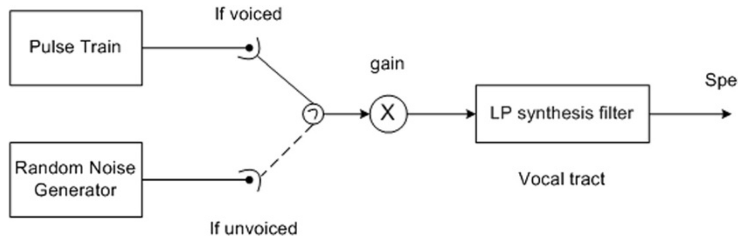

Figure 2 LP model of speech production.

is helpful to predict a speech sample $s_2(n)$ at time n from its previous p samples $s_2(n-i)$. For each speech frame, a 10-order predictor ($p = 10$) is employed on the windowed speech, $s_2(n)$, to estimate the spectral envelope. The predicted signal $\hat{s}(n)$ is given by

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{12}$$

The LP coefficients $a_i(i = 1, \ldots, p)$ are predicted from the minimization (by autocorrelation method) of the error between the windowed sample $s_2(n)$ and the predicted sample $\hat{s}_2(n)$. Since the pitch and excitation analysis phases are completed in a closed-loop manner, the LP synthesis filter is required in order to reduce the error between the original speech and the synthesized speech candidates. The LP synthesis filter in the Z-domain, $H(z)$, is connected to the LPC vector by

$$H(z) = \frac{1}{\sum_{i=1}^{p} a_i z^{-1}} \tag{13}$$

The filter $H(z)$ is represented in the time domain by the impulse response function $h(n)$.

Pitch analysis

Due to the vocal cords vibration, the voice speech segments show some long-term correlation. The vibration frequency, named pitch, is reflected in the quasi-periodicity behavior of the time domain speech waveform. An autocorrelation scheme is used to calculate the pitch lag (the inverse of the pitch frequency). Since the LP analysis frame may contain more than one pitch period, the pitch analysis is performed on sub-frame basis to extract one pitch gain and one pitch delay. One pitch gain and one pitch lag are used to represent consequently the periodicity in each speech frame [19]. In the pitch analysis algorithm, an open-loop analysis is first applied to each speech frame to estimate the pitch period. Open-loop pitch estimation is based on the weighted speech signal $s_w(n)$ which is obtained by filtering the input speech signal through the perceptual weighting filter, s_w is given by:

$$W(z) = \frac{A(Z/y_1)}{A(Z/y_2)} = \frac{1 + \sum_{i=1}^{10} y_1^i a_i z^{-1}}{1 + \sum_{i=1}^{10} y_2^i a_i z^{-1}} \tag{14}$$

That is, in a frame of size L, the weighted speech is given by:

$$S_w(n) = s(n) + \sum_{i=1}^{10} a_i y_1^i s(n-i) - \sum_{i=1}^{10} a_i y_2^i s_w(n-i), n = 0, \ldots, L-1 \tag{15}$$

Residual excitation

The signal e(n) after removing the long-term and short term redundancies has a noise-like shape with a flat spectrum. Figure 3 shows the residual signal after removing the long and short correlations. This signal could be modulated by a random signal. Since the random signal has no correlation, this residual will be generated at the receiver side using a random signal generator. By this, we reduce the amount of information to be hidden in the cover signal. As mentioned below, the speech analysis algorithm is based on two phases: an LP analysis to obtain p LP coefficients, $a_i(i = 1, \ldots, p)$ and a pitch analysis phase to extract the pitch gain g and the pitch delay d. Table 1 shows the used parameters of the LP-model for narrowband speech.

LP-model parameters adjustment

The spectral amplitudes must always be positive due to the absolute value applied to the speech spectrum. Direct embedding of the LP coefficients C in the magnitude spectrum will drastically destroy the cover signal since the LP parameters could have negative values. To accommodate this problem, we propose to convert the LP coefficients C to one of their frequency representations, such as LSF. As shown in the following equation, the LSF parameters w_i are ordered and are all positive.

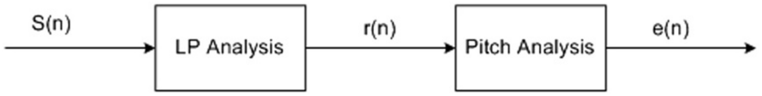

Figure 3 **Residual signal after removing the long and short correlations.**

Table 1 The LP model parameters

Model parameters	Symbol	Number of parameters per frame
Pitch lag	d	1
Pitch gain	g	1
LP coefficients	$a_1, a_2, \ldots, a_p$	p
Voice/unvoiced decision	V/UV	1
Total		$p+3$

$$0 \le w\ 1 \le w_2 \le \cdots \le w_p \le \pi \tag{16}$$

Since the pitch delay varies from 20 to 147 samples, direct embedding of the pitch delay in the cover speech spectrum will affect the high-frequencies small-amplitudes cover spectrum components. Hence, the need to normalize the pitch delay is by 147, the maximum pitch delay, before the hiding process. The normalized pitch delay will have a value ranging from 0 to 1. For this reason, the best location to hide these parameters is the last cover speech spectrum location since the amplitude of this last component is very small.

LSF cues

Itakura [20] has proposed the LSF to represent the LPC. They have been demonstrated to acquire different advantageous proprieties like bounded range, sequential ordering, and ability of constancy verification [21]. Moreover, the LSFs coefficients facilitate the integration of human observation system proprieties in the frequency domain representation. According to the ITU-T Recommendation G.723.1, the extraction of the LSFs parameters is recommended in case of need to convert the LPC parameters to LSFs [22]. In LPC, the mean squared error between the original and the predicted speech is minimized over a short time interval to produce distinctive set of LP coefficients. The transfer function of the LPC filter is given by

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} \tag{17}$$

where P the prediction order, G is the gain, and a_k is the LPC filter coefficients. The poles of this transfer function contain the poles of the vocal tract as well as those of the voice source. Solving for roots of the denominator of the transfer function gives both the formant frequencies and the poles corresponding to the voice source. Two transfer functions $Q_{p+1}(z)$ and $P_{p+1}(z)$, respectively, called difference and sum polynomials can be resulting from $H(z)$. The difference polynomial is given by:

$$Q_{p+1} = A_p(z) - z^{-(p+1)})A_p\left(z^{(-1)}\right) \tag{18}$$

and the sum polynomial is given by:

$$P_{p+1} = A_p(z) + z^{-(p+1)} A_p\left(z^{-1}\right) \tag{19}$$

where $A_p(z)$ is the denominator of $H(z)$. The polynomials contain trivial zeros for even values of p at $z = -1$ and at $z = 1$. These roots can be removed in order to obtain the following quantities:

$$\hat{Q}(z) = \frac{Q_{p+1}(z)}{(1+z)} = \beta_0 z^p + \beta_1 z^{p-1} + \cdots + \beta_p, \tag{20}$$

and

$$\hat{P}(z) = \frac{P_{p+1}(z)}{(1+z)} = \alpha_0 z^p + \alpha_1 z^{p-1} + \cdots + \alpha_p. \tag{21}$$

The LSFs are the roots of $\hat{Q}(z)$ and $\hat{P}(z)$ and alternate with each other on the unit circle. Note that $Q_{p+1}(z)$ is an antisymmetric polynomial and $P_{p+1}(z)$ is a symmetric polynomial. The polynomials $\hat{Q}(z)$ and $\hat{P}(z)$ derived from $Q_{p+1}(z)$ and $P_{p+1}(z)$ are symmetrical. Therefore, for even values of p we can derive the following property:

$$\alpha_i = \alpha(p-i), 0 \le i \le \frac{p}{2} \tag{22}$$

Consequently (20) and (21) can be written as follows:

$$\hat{Q}(z) = z^{p/2}\left[\beta_0\left(z^{p/2} + z^{-p/2}\right) + \beta_1\left(z^{p/2-1} + z^{-(p/2-1)}\right) + \cdots + \beta_{p/2}\right], \tag{23}$$

and

$$\hat{P}(z) = z^{p/2}\left[\alpha_0\left(z^{p/2} + z^{-p/2}\right) + \alpha_1\left(z^{p/2-1} + z^{-(p/2-1)}\right) + \cdots + \alpha_{p/2}\right] \tag{24}$$

By putting $z = e^{jw}$ and then $z + z^{-1} = 2\cos(w)$, we obtain the equations to be solved in order to find the LSFs according to the real root scheme ITU-T Recommendation G.723.1:

$$\hat{Q}\left(e^{jw}\right) = 2e^{jpw/2}\left[\beta_0 \cos\left(\frac{p}{2}w\right) + \beta_1 \cos\left(\frac{p-2}{2}w\right) + \cdots + \frac{1}{2}\beta_{p/2}\right] \tag{25}$$

and

$$\hat{P}\left(e^{jw}\right) = 2e^{jpw/2}\left[\alpha_0\cos\left(\frac{p}{2}w\right) + \alpha_1\cos\left(\frac{p-2}{2}w\right) + \cdots + \frac{1}{2}\alpha_{p/2}\right] \tag{26}$$

Input speech is segmented to different frames. Additionally, each frame is subdivided into four sub-frames. On these sub-frames, the LPC analysis is performed. The conversion of the p LPC coefficients into their p corresponding LSFs is performed in the last sub-frame. For the three of the sub-frames, the LSFs are obtained by executing linear interpolation between the LSFs of the current and the previous frame.

To achieve this purpose, the unit circle is then divided into 512 equal intervals, each of length $\pi/256$. The roots (LSFs) of $Q(z)$ and $P(z)$ polynomials are searched along the unit circle from 0 to π. A linear interpolation is performed on intervals where a sign change is observed in order to find the zeros of the polynomials. According to [20], if a sign change appears between intervals l and $l-1$, a first-order interpolation is executed as follows:

$$\hat{l} = l - 1 + \frac{\left|P(z)_{l-1}\right|}{\left|P(z)_{l-1}\right| + P(z)_l} \tag{27}$$

where $\hat{l}$ is the interpolated solution index, $|P(z)_l|$ is the absolute magnitude of the result of sum polynomial evaluation at interval l (similarly for $l-1$). Since the LSFs are interlacing in the region from 0 to π, only one zero is evaluated on $P(z)$ at each step. The search for the next solution is performed by evaluating the different polynomial Q(z), starting from the current solution [23,24]. Therefore, two main reasons motivated our choice to consider the LSFs representation. The first reason is related to the fact that LP coefficients are very sensitive to errors. The direct quantization of these coefficients might produce an unstable LP filter. The second reason is related to the fact that LSFs are widely used in conventional coding schemes. This avoids the incorporation of new parameters that may require significant and costly modifications to current devices and codecs.

Speech hiding algorithm

We propose a new method for speech signal steganography, the secret speech signal is embedded into the coefficients in the wavelet domain. The DWT decomposes the cover speech signal into low- and high-frequency components. For speech signals, the low-frequency component is the most significant part for speech perception. On the other hand, the high-frequency component impacts flavor or nuance (noise) to the signals. Let's consider the human voice. If we remove the high-frequency components, the voice sounds different, but we can still tell what's being said. However, if we remove sufficient amount of the low-frequency components, we hear gibberish and we cannot understand what's being said. For this reason, we decide to hide information in the high-frequency in the wavelet domain. Furthermore, in wavelet analysis, we can divide the speech signal in approximations and details. The approximations are the high-scale, low-frequency components of the signal. The details are the low-scale, high-frequency components. As shown in Figure 4 after passing through two complementary filters, two signals emerge from the original signal.

A variety of wavelets can be used depending on the expected results. Each family of wavelets (such as Haar or Daubechies family) are wavelet subclasses distinguished by the number of filter coefficients and the level of iteration. In steganography, whatever the used algorithm for hiding data, we need to reconstruct the speech signals after embedding the message in the original signal. After that, performance measure can be used to compare the original speech signal and the stego-speech. In our method, after using the DWT to decompose the speech signals for hiding a message speech signals, we use the IDWT to reconstruct the signal. The speech-in-speech hiding algorithm is illustrated in Figure 5. Both of secret and cover speech must be pre-processed in order to facilitate the hiding process. The cover speech is partitioned into L-ms frames. The DFT of each time-frame s_1(m) defined for $0 \leq m \leq m-1$ is computed using the DWT-FFT method. The obtained speech spectrum is decomposed into magnitude and phase spectra. Each L-ms of the secret message s_2(m) is embedded in the low-amplitude high-frequency region of the magnitude spectrum of the cover signal.

Secret speech hiding

In order to hide the secret speech, the DWT is applied to the speech cover speech frame to separate the high-

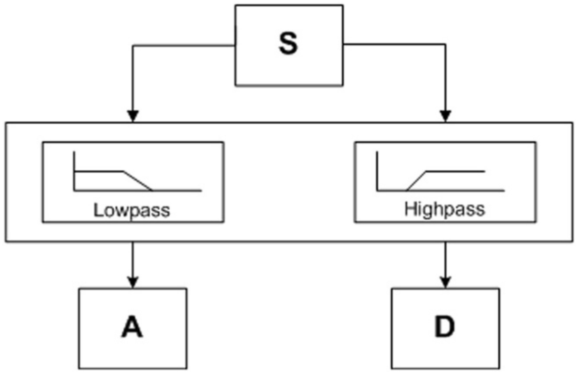

Figure 4 Level 1 decomposition of DWT coefficients.

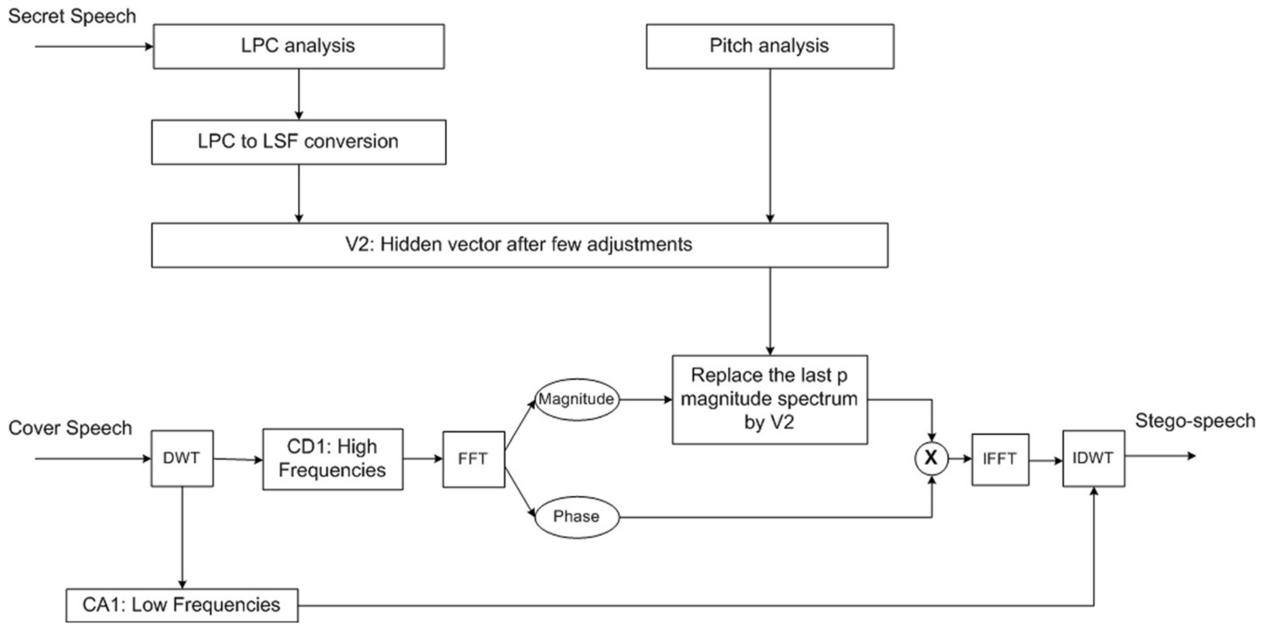

Figure 5 Block diagram showing the general steps to embed a secret speech signal s_2 inside another cover speech signal s_1 to produce a stego-speech signal s_3.

and the low-frequency regions. Then the FFT is applied to the high-frequency wavelets part producing a spectrum $S_1(k)(k = 0,\ldots, M - 1)$. The spectrum is decomposed into magnitude spectrum $|S_1(k)|$ and phase spectrum $\phi_1(k)$.

The magnitude spectrum is symmetric. The hiding process consists of representing the L last elements of the first half of $|S_1(k)|$ by the LP parameters V_2 of the secret speech$s_2(m)$.

The resulting magnitude spectrum, denoted by $|S_3(k)|$, is defined by the following expressions:

$$|S_3(k)| = \begin{cases} |S_1(k)|, & k = 0,\ldots,\dfrac{M}{2} - p - 3 \\ V_2\left(k - \dfrac{M}{2} - p - 2\right), & k = \dfrac{M}{2} - p - 2,\ldots,\dfrac{M}{2} - 1 \\ V_2\left(\dfrac{M}{2} + p + 1 - k\right), & k = \dfrac{M}{2},\ldots,\dfrac{M}{2} + p + 1 \\ |S_1(k)|, & k = \dfrac{M}{2} + p + 2,\cdots, M - 1 \end{cases} \tag{28}$$

The third right-hand term in the above equation is included to preserve the DFT symmetry. These modifications lead to a new speech signal s_3. Its spectrum is a simple combination of the magnitude spectrum $|S_3(k)|$ and the cover phase spectrum $\phi_1(k)$,

$$S_3(k) = |S_3(k)|e^{j\phi 1(k)}k = 0,\ldots, M - 1 \tag{29}$$

The time-frame composite (stego) signal $s_3(m), m = 0,\ldots, M - 1$, is obtained by the IDWT,

$$s_3 = \text{IDWT}(S_3) \tag{30}$$

The stego signal $s_3(m)$ is a composite signal since it contains the L-ms cover speech $s_1(m)$ and the L-ms secret signal $s_2(m)$.

Energy normalization

In order to improve the speech quality, we preserved the speech energy by normalizing all the hidden parameters by the total energy of the original spectrum magnitudes. However, the energy preservation requires the hiding of the energy as side information. At the receiver, this energy will be used to rescale the hidden information to its original values. The scaling coefficient a is given by

$$a = \sqrt{\frac{E_c}{1 + E_{\text{LSF}\,secret}}} \tag{31}$$

where E_c is the energy of the cover speech spectrum and E_{LSF} is the energy of the LSF vector.

Secret speech reconstruction

The secret speech is reconstructed from the stego speech by subsequent the hiding algorithm in overturn order. Figure 6 illustrates the pursued steps to extract the hidden information and reconstruct the secret speech message. The first step consists of performing the DWT. Transforming by FFT the high frequencies obtained with the DWT to its corresponding spectrum. The magnitude spectrum is then acquired from the speech spectrum. The secret speech

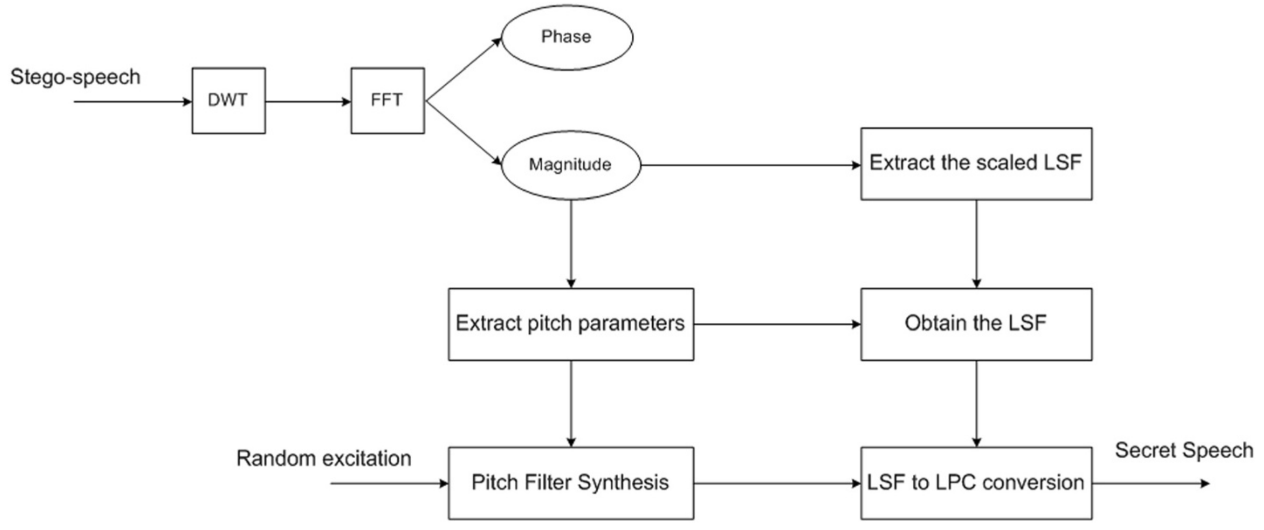

Figure 6 **Block diagram showing the general steps to retrieve the secret speech signal S_2 from the stego-speech S_3.**

parameters are extracted from the same locations they were embedded in the spectral magnitude of the stego speech signal. The LSF vector is converted back to a *P*-order LPC vector $(a_1, \ldots, a_p)$ to build the LP synthesis filter $H(z)$.

$$H(z) = \frac{1}{1 - \sum_{i=1}^{10} a_{\mathrm{i}}\, z^{-i}} \tag{32}$$

A random excitation signal $e(n)$ is applied to the series of the pitch and LP synthesis filters. The signal $\hat{s}(n)$, at the output of the LP synthesis filter, is a reproduction of the original secret message $s(n)$. Since the LPC-model parameter values that are extracted from the stego speech have approximately the same exact values as the embedded parameters, the reconstructed secret speech signal is not affected by the hiding process. The minor degradations noticed in this signal, when compared with the original secret signal, are resulting from the LPC model and the LSF conversion.

Evaluation

Experimental setup

To evaluate the performance of the proposed hiding technique, we conducted several simulations using NOIZEUS database [25,26,27]. This corpus contains 30 sentences from the IEEE sentence database, recorded in a sound-proof booth using Tucker Davis Technologies recording equipment. The sentences are produced by three male and female speakers. The 30 sentences: 15 male and 15 female speakers include all phonemes in the American English language. The sentences were originally sampled at 25 kHz and down-sampled to 8 kHz. The length of the speech file varies between 0.02 and 0.03 ms. In the comparative evaluation, we conducted four sets of tests. In the first set of simulations, we embedded each of the 15 male speech files in each of the 15 female speech files. In the second set of tests, we hide each of the 15 female speech files in each of the 15 male speech files. In the third set of tests, we embedded each of the 15 male speech signals in the remaining 14 male speech files. In the last sets of tests, we hide each of the 15 female speech segments in the remaining same gender speech files. Each set is iterated for five different wavelet families (Haar, Daubechies, Symlets, Coiflets, and

Table 2 SNR of the DWT-FFT-based hiding approach

Cover signals	Secret signals	SegSNR (dB)
Female	Male	31.86
Male	Female	32.70
Male	Male	34.45
Female	Female	31.13
Average		32.54

Table 3 SNR of FFT-based hiding approach

Cover signals	Secret signals	SegSNR (dB)
Female	Male	51.46
Male	Female	52.62
Male	Male	54.37
Female	Female	51.09
Average		52.39

Table 4 Different wavelets results of DWT-FFT-based steganography systems

Wavelet name		Haar	Daubechies (db1)	Symlets (sym1)	Coiflets (coif1)	BiorSplines (bior1.1)
Cover signals	Secret signals	SegSNR (dB)	SegSNR (dB)	SegSNR (dB)	SegSNR (dB)	SegSNR (dB)
Female	Male	31.53	31.86	31.48	31.41	31.39
Male	Female	31.98	32.70	31.86	31.96	31.91
Male	Male	34.12	34.35	34.08	34.08	34.04
Female	Female	30.79	31.13	30.68	30.76	30.71
Average		32.11	32.51	32.03	32.05	32.01

BiorSplines). In total, we conducted 4,210 computer simulations ((15*15*2 + 14*14*2)*5).

In order to evaluate the impact of the DWT-FFT technique, we conducted two different comparative experiments using DWT-FFT method and then using FFT only.

Evaluation outcomes

One of the performance measures of any steganographic system is the comparison between the cover and the stego signals. In this study, we used subjective and objective performance measures. In the subjective measures, we conducted several informal listening comparative tests. In these simulations, we played in a random order the cover speech $s_1(m)$ and the stego signal $s_3(m)$ to several listeners. Each listener had to identify the better quality speech file among the cover and the stego signals. The majority of listeners could not distinguish between the two speech files. As an objective measure, we used the segmental signal-to-noise ratio (SegSNR) and the perceptual evaluation of speech quality (PESQ). PESQ measurement provides an objective and automated method for speech quality assessment. The SegSNR is defined by

$$\mathrm{SegSNR(dB)} = 10\log_{10}\left(\frac{\sum_{m=0}^{159}[s_1(m)]^2}{\sum_{\mathrm{m}=0}^{159}[s_1(m)-s_3(m)]^2}\right) \tag{33}$$

where s_1 and s_3 are the cover and the stego speech files, respectively. In this study, we segmented the speech files into frames of 20 ms ($L = 20$) (or 160 samples ($\mathrm{M} = 160$)). In Table 2, we present the average SegSNR values for each of the four different sets of tests using DWT-FFT algorithm. In Table 3, we present the average SegSNR of the same set of tests using the FFT only. The quality of the stego signal produced by the FFT is better than the one produced by the DWT-FFT. However, the DWT-FFT increases the robustness of the hiding algorithm against steganalysis techniques. We used some of the existing wavelets to compare the impact of the different wavelet on the speech quality. The decomposition of all used wavelets is done with one level. Table 4 shows the result of different wavelets for the four different sets of tests. As can be noticed, different wavelets have almost similar results; therefore, this method is not depending on a particular type of wavelet. The SegSNR value did not differ a lot for different wavelets. The SegSNR is just an indicative performance measure. The PESQ is a more reliable method to assess the performance of our hiding technique. The PESQ measurement provides an objective and automated technique for speech quality evaluation. The degradation of the speech sample can be predicted using the PESQ algorithm with subjective opinion score. In general, the PESQ returns a score from 0.5 to 4.5, with higher scores signifying better quality [28,29]. The PESQ method is used in our experiments to evaluate the stego speech. The reference signal refers to an original (cover) signal and the degraded signal refers to the stego signal with the hidden secret message. In Table 5, we present the average PESQ values for male and female speakers obtained by the two hiding techniques (using DWT-FFT and FFT only). Figure 7 shows variations of PESQ for 20 speech signals of the 2 hiding approach. The hiding method achieves 3.68 and 4.14 PESQ average for DWT-FFT and FFT algorithms, respectively. Figure 8 shows the magnitude spectrum of the cover signal and the corresponding of stego speech after hiding the LPC parameters of the secret signal. The PESQ analysis shows that the stego and cover speech provide similar subjective quality. This result is supported by the resemblance between the cover

Table 5 PESQ of DWT-FFT and FFT-based hiding approach

Speaker	PESQ	
	DWT-FFT	FFT
Female	3.58	4.12
Male	3.78	4.16
Average	3.68	4.14

Table 6 Impact of the hiding process on the secret speech in terms of SegSNR

Speaker	SegSNR(dB)	
	DWT-FFT	FFT
Female	21.76	24.64
Male	23.89	26.28
Average	22.83	25.46

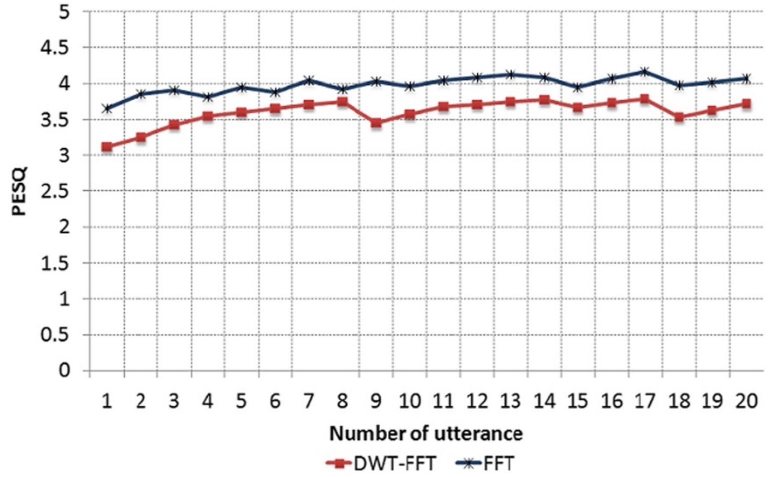

Figure 7 Comparison of the PESQ variations between DWT-FFT and FFT-based steganography systems: PESQ scores of cover speech $s_1(m)$ and the stego signal $s_3(m)$ utterances using DWT-FFT and FFT hiding approach.

and stego speech spectrograms in Figure 9. The objective and subjective performance measures show that the proposed hiding technique attracts no suspicion about the existence of a hidden message in the stego speech, while being able to recover an intelligible copy of the original secret message at the receiver side. The informal listening test to the original and the reassembled secret speech message advocate the result of the other objective performance measurement. The reconstructed secret speech $\hat{s}(n)$ (from both DWT-FFT and FFT hiding approaches) still completely comprehensible, even some perceptual distortions are simply noticeable. What concerns us is the speech intelligibility since the objective is to convey the secret message to the intended receiver. Table 6 shows the impact of the hiding algorithms on the secret speech in terms of the SegSNR.

Conclusions

In this article, we presented a new steganography system for secrecy applications. The proposed hiding method produces stego speech files that are indistinguishable from their equivalent cover speech files. Moreover, the complexity of our hiding technique is so high any eavesdropper cannot extract the hidden information even after suspecting

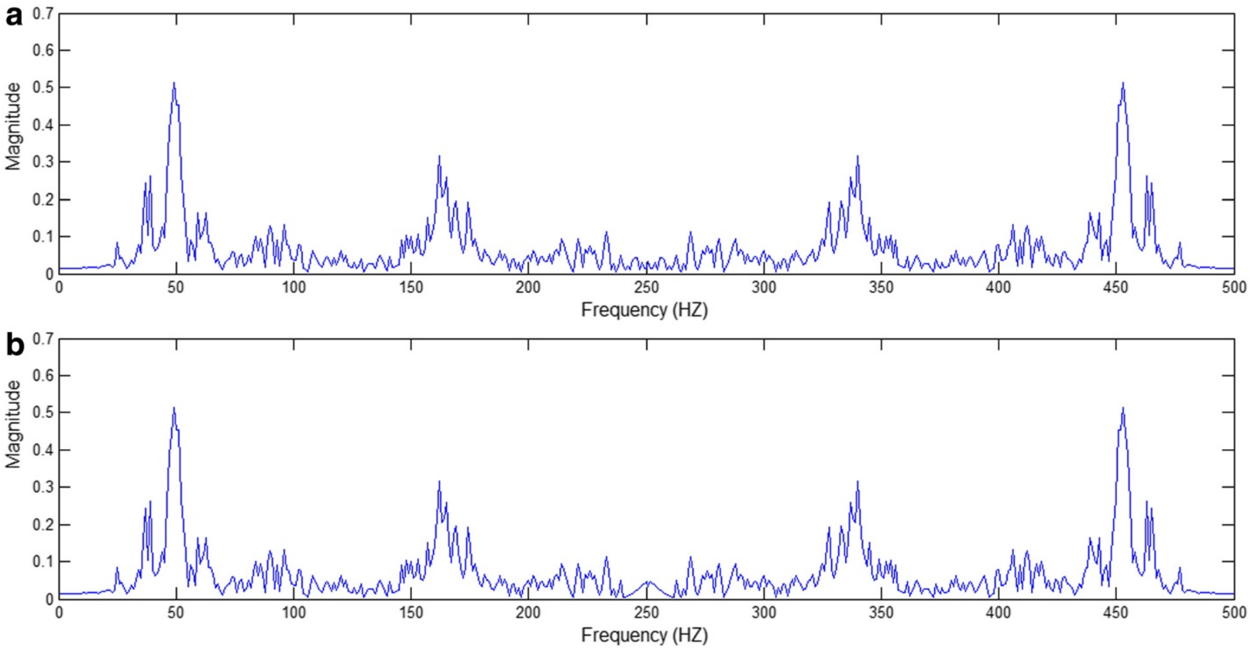

Figure 8 Magnitude spectrum of (a) the cover speech $s_1(m)$, (b) the stego speech.

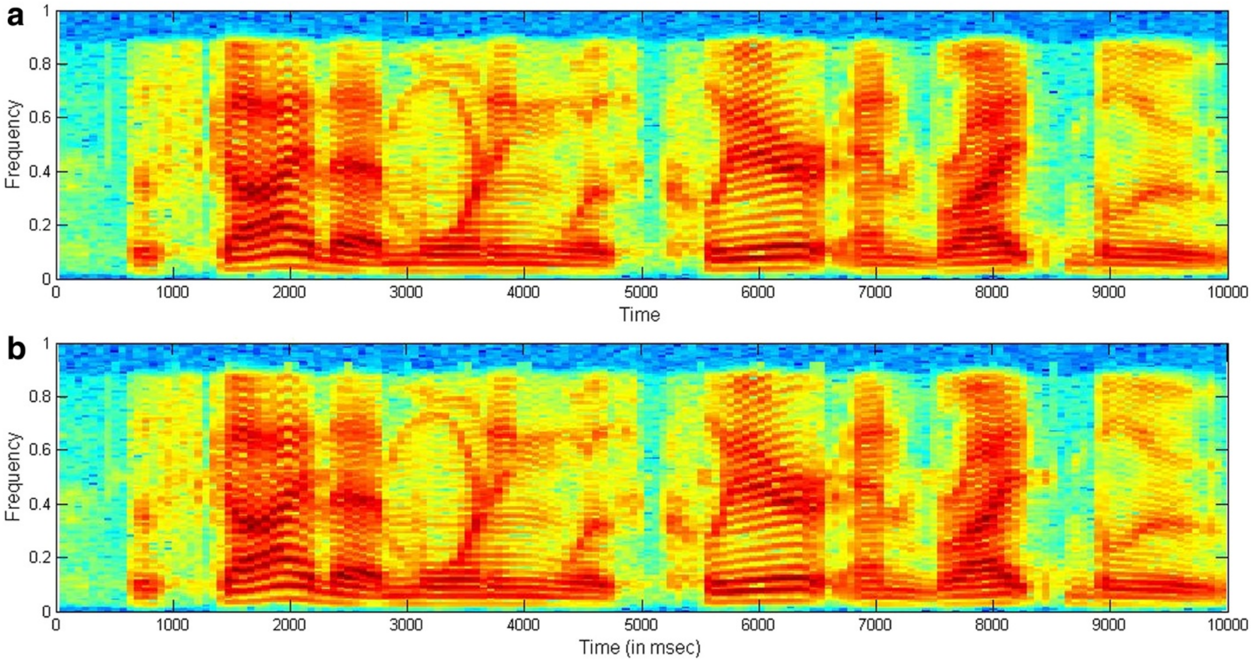

Figure 9 Spectrogram of (a) the cover speech, (b) the corresponding stego speech after hiding a male secret message within the cover speech.

the existence of a secret message. Since our aim is to render the steganalysis (the attempt to extract the secret message from the stego signal) by the opponent more complex. Our method exploits first the high frequencies using a DWT, then exploits the low-pass spectral properties of the speech magnitude spectrum to hide another speech signal in the low-amplitude high-frequencies region of the cover speech signal. Experimental simulations on both female and male speakers showed that our approach is capable of producing a stego speech that is indistinguishable from the cover speech. The receiver is still able to recover an intelligible copy of the secret speech message. In the future work, we will endeavor to extend our approach to applications involving Voice-over IP speech secrecy, which involves compressing the stego speech before transmission. This opens up the issue of preserving the secret speech after decoding the compressed stego speech.

Abbreviations

DFT: Discrete Fourier transform; DWT: Discrete wavelet transforms; DWT: FFT Discrete wavelet transform-fast Fourier transform; FFT: Fast Fourier transform; IDWT: Inverse discrete wavelet transform; IFFT: Inverse fast Fourier transform; LP: Linear prediction; LPC: Linear predictive coding; LSB: Least significant bit; LSF: Line spectral frequencies; PESQ: Perceptual evaluation of speech quality; SegSNR: Segmental signal-to-noise ratio; SS: Spread spectrum.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Université de Bretagne Occidentale, Brest, France. [2]Canadian University of Dubai, Dubai, UAE. [3]University of Moncton, Shippagan, NB, Canada. [4]University of Moncton, Moncton, NB, Canada.

Received: 13 August 2011 Accepted: 19 June 2012
Published: 8 August 2012

References

1. D Kahn, *The History of Steganography. Lecture Notes in Computer Science*, 1174th edn. (Springer, New York, 1996), p. 11023
2. NF Johnson, S Jajodia, Exploring steganography: seeing the unseen. IEEE Comput. **31**(2), 26–34 (1998)
3. R Sridevi, A Damodaram, SVL Narasimham, Efficient method of audio steganography by modified LSB algorithm and strong encryption key with enhanced security. J. Theor. Appl. Inf. Technol. **5**(6), 768–771 (2009)
4. W Bender, D Gruhl, N Morimoto, Techniques for data hiding. IBM Syst. J. **35**(3), 313–336 (1996)
5. D Kirovski, H Malvar, Spread-spectrum watermarking of audio signals. IEEE Trans. Signal Process. **51**(4), 1020–1033 (2003)
6. D Huang, T Yeo, *Robust and Inaudible Multi-Echo Audio Watermarking, in Proceedings of the Third IEEE Pacific-Rim Conference on Multimedia, Advances in Multimedia Information Processing Taipei* (China, 2002), pp. 615–622
7. S Shirali-Shahreza, M Shirali-Shahreza, *Steganography in Silence Intervals of Speech, Proceedings of the Fourth IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2008)* (Harbin, China, August 15–17, 2008), pp. 605–607
8. M Misiti, Y Misiti, G Oppenheim, JM Poggi, *Matlab Wavelet Toolbox (Version 4.0): Tutorial and Reference Guide The Mathworks* (Natick, USA, janv 2007)
9. B Lin, B Nguyen, ET Olsen, in *Orthogonal Wavelets and Signal Processing*, ed. by PM Clarkson, H Stark. Signal Processing Methods for Audio, Images and Telecommunications (Academic, London, 1995), pp. 1–70
10. S Mallat, *A Wavelet Tour of Signal Processing* (Academic, San Diego, CA, 1998)
11. Y Nievergelt, *Wavelets Made Easy* (Birkhäuser, Boston, 1999)
12. J Ooi, V Viswanathan, Applications of Wavelets to Speech Processing, in, ed. by RP Ramachandran, R Mammone. Modern Methods of Speech Processing (Kluwer Academic Publishers, Boston, 1995), pp. 449–464
13. DF Elliott, KR Rao, *Fast Transforms: Algorithms* (Analyses, Applications (Academic, New York, 1982)
14. S Andreas, PT Ed, A Venkatraman, *Audio Signal Processing and Coding* (Wiley-Interscience Publication, USA, 2006). ISBN 978-0-471-79147-8, TK5102.92.S73

15. W Strange, TR Edman, JJ Jenkins, Acoustic and phonological factors in vowel identification. J. Exp. Psychol. Hum. Percept. Perform. **5**(4), 643–656 (1979)
16. CY Espy-Wilson, Acoustic measures for linguistic features distinguishing the semivowels in American English. J. Acoust. Soc. Am **92**, 736–757 (1992)
17. DG Childers, M Hahn, JN Larar, Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech. IEEE Trans. ASSP **37**(11), 1771–1774 (1989)
18. D O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd edn. (Wiley-IEEE Press, New York, NY, 1999). ISBN 9780780334496
19. J Makhoul, Linear prediction: a tutorial review. Proc. IEEE **63**(5), 561–580 (1975)
20. F Itakura, Line spectrum representation of linear predictive coefficients of speech signals. J. Acoust. Soc. Am **57**(1), S35 (1975)
21. AV Oppenheim, WR Schafer, AJ Buck, *Discrete-Time Signal Processing* (Prentice Hall, Upper Saddle River, NJ, 1999), pp. 468–471. ISBN 0-13-754920-2
22. W Hess, *Pitch Determination of Speech Signals* (Springer, Berlin, 1983)
23. F Soong, B Juang, *Line spectrum pair (LSP) and speech data compression. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '84)* (San Diego, Calif, USA 9, March 1984), pp. 37–40
24. ITU-T, *Recommendation G. 723.1. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*, 1996
25. Y Hu, P Loizou, Subjective evaluation and comparison of speech enhancement algorithms. Speech Commun **49**, 588–601 (2007)
26. Y Hu, P Loizou, Evaluation of objective quality measures for speech enhancement. IEEE Trans. Speech Audio Process. **16**(1), 229–238 (2008)
27. J Ma, Y Hu, P Loizou, Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. **125**(5), 3387–3405 (2009)
28. ITU, *Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs, ITU-T Recommendation 862*, 2000
29. ITU-T Recommendation, *Methods for Subjective Determination of Speech Quality International Telecommunication Union* (Geneva, 2003), p. 800

doi:10.1186/1687-4722-2012-20
Cite this article as: Rekik *et al.*: Speech steganography using wavelet and Fourier transforms. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:20.