# A Sparse Representation-Based Wavelet Domain Speech Steganography Method

Soodeh Ahani, Shahrokh Ghaemmaghami, and Z. Jane Wang

***Abstract*—In this paper, we present a novel speech steganography method using discrete wavelet transform and sparse decomposition to address the undetectability concern in speech steganography. The proposed speech steganography method exploits the sparse representation to embed secret messages into higher semantic levels of the cover signal, resulting in increased undetectability. The proposed method also yields improvements on both stego signal quality and embedding capacity, which are the two major requirements of a steganography algorithm. Our experimental results illustrate that the stego signals generated by the proposed method are perceptually indistinguishable from the original cover signals, quantified by both SNR and PESQ quality measures. When compared with two well-known steganography methods, the proposed method is shown to be superior on addressing major requirements of a steganography algorithm, imperceptibility, undetectability, and capacity.**

***Index Terms*—Data hiding, dictionary learning, discrete wavelet transform, sparse representation, speech steganography.**

## I. Introduction

STEGANOGRAPHY is the technology of covert communication via a digital cover media such as image, audio or video files over a public channel [1]. Increasing need of secure transmission of private information over Internet inspired researchers to make great efforts to explore steganography methods [1]. Different from cryptography which is about concealing the content of the messages, steganography is about concealing the existence of the secret message. The ultimate goal of a steganography method is to conceal the presence of secret messages embedded in the cover media.

In a steganography problem, the sender and the receiver intend to communicate in an innocent-looking way over a public channel so that an adversary, eavesdropping on the channel, cannot even detect the presence of the hidden messages [2]. There are two types of attack in the field of data hiding: passive attack and active attack. In steganography applications, passive attack is considered while its goal is to detect whether steganography is being used or not and it does not disturb the communication before detecting the suspicious signals [2]. Active attacks such as noise addition, MPEG audio coding, cropping, time shifting, filtering, resampling, requantization, *etc.* are investigated in watermarking applications [3]. One main goal of the watermarking systems is for copyright protection of digital media against unauthorized copy and redistribution, known as data piracy [3]. In the watermarking applications, the secret data, called the watermark, is embedded into the host signal and can be used as the user ID to completely characterize the person who applied it [3]. Active attacks try to remove or disturb the watermark [3]. In the steganography applications, the secret data is called stego which is inserted into the cover signal to generate the stego signal.

In steganography applications, it is generally assumed that the secret keys are shared between the sender and the receiver through a secure channel while the stego signal is transmitted through a public channel [4], [5]. There are different protocols for secure key exchange in steganography [4]. The public channel is supposed to be monitored by an adversary, called the steganalyzer, that aims to determine whether the transmitted file contains any hidden data [4]. Therefore the sender and the receiver in a steganography system should communicate in a way that nothing seems "unusual" to an adversary [2].

There are three main issues in designing a speech steganography method: undetectability, imperceptibility, and capacity. Undetectability ensures that the stego signal, containing the secret data, is indistinguishable from the original speech signal by means of either the human auditory system (HAS) or steganalysis methods. Imperceptibility means that the hidden data insertion process makes no audible distortion on the original speech signal, where capacity gives the relative amount of hidden data which can be inserted into the cover speech signal.

Generally, steganographic methods have been developed from two distinctive research origins which form two categories of steganography methods [6]. The first category is called "technical steganography" methods [9] (e.g., [1], [12], [13], [15], [17], [18], [19]) that use digital media data like images, audio and video as cover media to insert the message bits [6]. Technical steganography methods assume that the stego files are transmitted based on a reliable transportation protocol, such as TCP, which guarantees impeccable delivery of digital files [7]. On the other hand, the second category of steganography methods, called "network steganography" methods [8], uses network protocol fields as cover media to insert the message bits (e.g., [6], [8], [10], [11]). These steganography methods insert hidden bits by using specific Voice-over-IP (VoIP) protocol

Manuscript received April 12, 2014; revised August 22, 2014; accepted November 03, 2014. Date of publication November 20, 2014; date of current version January 14, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Woon-Seng Gan.

S. Ahani is with the Department of Electrical Engineering and Electrical Research Institute (ERI), Sharif University of Technology, Tehran, Iran, and also with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: s_ahani@ee.sharif.edu; soodeh@ece.ubc.ca).

S. Ghaemmaghami is with the Department of Electrical Engineering and Electrical Research Institute (ERI), Sharif University of Technology, Tehran, Iran (e-mail: ghaemmag@sharif.edu).

Z. J. Wang is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: zjanew@ece.ubc.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2014.2372313

fields [6]. The two categories use different covers (digital media data vs. network protocol fields) for hidden data insertion, and they are different in terms of requirements and applications. For instance, undetectability against steganalysis methods (attacks designed to detect the presence of hidden data) is one of the most important requirements for the first mentioned category of steganography methods, while the second category of steganography methods is used in real-time applications that their performances are evaluated in presence of common network transmission problems such as lag, jitter and packet loss [8], [9]. An acceptable audio quality has to be preserved in the first category steganography methods [9]. In this paper, our focus is on the first category - the technical steganography methods.

Several methods have been developed for hiding secret messages into the cover audio signals in the time domain. Least Significant Bit (LSB) Substitution is one of the earliest techniques used for the secret data embedding in audio signals and other media types [12]. Sridevi *et al.* in [13] presented a time domain audio steganography scheme which replaces the LSB of each cover audio sample with a secret bit. Although hiding secret data in LSBs of the cover audio is one of the simplest methods enabling a high rate of embedded information, different steganalysis algorithms have been proposed to challenge its undetectability. The phase coding steganography methods [14], [15], [16] are other time domain audio steganography schemes which hide secret data in the phase of the cover audio signals. Data embedding capacity of these methods is quite low when compared with other steganography algorithms [1].

Transform domain audio steganography methods are the ones in which various transform domains such as Fast Fourier Transform (FFT), Discrete Cosine Transform (DCT), and Discrete Wavelet Transform (DWT) are used to embed the secret data in the coefficients of the cover audio. The methods presented by Seppanen *et al.* in [17] and [18] employ the LSB technique to embed secret data in wavelet transform coefficients. Generally, the objective of such LSB based steganography methods is to increase the embedding capacity, while minimizing the quality degradation of the resulting stego signals [19]. Objective and subjective tests show that the methods of [17] and [18] outperform classical time domain LSB method [17]. Delforouzi and Pooyan in [19] proposed an audio steganography method based on the LSB Substitution of wavelet coefficients, by using the hearing threshold levels. Rekik *et al.* presented another transform domain method in [1] which embedded the secret data in the high frequency components of the cover signal, and the stego signals were perceptually indistinguishable from the original cover signal, because the low frequency components of the cover signal were remained intact [1].

The goal of this paper is to present a speech steganography method which is less detectable, so more secure, against steganalyzers than existing popular methods. The novelty of this work comes from the employment of sparse representation (SR) of wavelet coefficients for secret data embedding. Sparse representation makes it possible to embed the secret data into the higher semantic levels of the speech signal which leads to a more secure speech steganography method. To the best of our knowledge, there are no previous reports on the investigation of sparse representation for audio steganography yet.

Steganalyzers try to detect the hidden messages embedded within the cover signals. Generally, audio steganalyzers rely on the statistical properties of audio signals to determine whether a suspicious signal contains hidden data or not [20]. Semantic contents of the signal are usually disregarded in these statistical analysis. Therefore, the high semantic levels of the cover signal are good choices for hidden data insertion. Structural elements of the cover signal are its semantic levels, in which the proposed speech steganography method embeds the secret messages to increase its undetectability against steganalyzers comparing to existing methods. Sparse decomposition of a signal represents it as a linear combination of some structural elements, called atoms, and we embed the secret messages in the coefficients of such a representation.

Sparse representation is a powerful decomposition method which allows salient information within a signal to be represented as a linear combination of few elementary components (atoms) [21]. These atoms collectively form a dictionary. Some applications use pre-specified dictionaries to represent signals sparsely. In some other applications, the dictionary is estimated based on training data, so that the estimated atoms can better capture specific features of the signal of interest [22].

Jafari and Plumbley presented a fast greedy adaptive dictionary learning (GAD) algorithm for sparse decomposition of speech signals in [21]. The GAD algorithm estimates an orthogonal dictionary from speech signal frames which are non-overlapping segments of the audio signal [21]. The GAD algorithm constructs a signal-adaptive orthogonal dictionary whose atoms encode local properties of the signal. In [21], it is shown that the proposed GAD algorithm yields sparse atoms and a sparse signal representation.

The proposed steganography algorithm embeds the secret data in the frames whose energies are higher than a pre-specified threshold to decrease audible distortions. We use wavelet transform to separate the selected frames into high frequency and low frequency components. High frequency components are used to estimate a proper dictionary by using the GAD dictionary learning algorithm [21] over which the low frequency components will be represented sparsely. The secret messages are then embedded into nonzero elements of the sparse representation vectors of the low frequency wavelet sub-band over the estimated dictionary. We refer to the proposed algorithm as the DWT-SD speech steganography method, where the DWT and the SD stand for the discrete wavelet transform and the sparse decomposition concepts, respectively.

Wavelet analysis is capable of identifying special features of the signal, such as the trends, discontinuities in higher derivatives, breakdown points and self-similarity [1]. We use the wavelet transform to separate the speech signal frames into high and low frequency components. The high frequency sub-band is used for dictionary estimation and the low frequency sub-band is opted for data insertion. The high frequency band is left intact to allow for a blind speech steganography. The steganography method is blind when the hidden messages could be extracted at the receiver side without requiring the original cover signal.

The rest of this paper is organized as follows. In Section II, a brief introduction to the sparse representation concept is given and the GAD dictionary learning algorithm is described. The embedding and extraction procedures of the proposed speech steganography method are explained in Section III. Section IV is dedicated to the description of the speech signal database used for our simulations. The experiments conducted to evaluate the

performance of the proposed method are reported in Section IV and a conclusion is drawn in Section V.

## II. Sparse Representation and the GAD Dictionary learning Algorithm

Recently there has been increasing interest in sparse representations of signals, i.e., representing a signal as a linear combination of atoms, where most of the coefficients are zero [23]. It is shown that sparse representation is a powerful tool for analysis and processing of audio signals [23]. The salient information within a signal could be represented efficiently by using some structural elementary components, called atoms, which collectively form a dictionary [21]. Sparse representations of signals have been successfully employed in many applications, such as compression [23], [24], audio denoising [23], [25], and blind source separation [23].

Audio signals are usually made by physical impacts or by resonant systems, or both of them [23]. Resonant systems produce sounds that are dominated by a small number of frequency components, allowing a sparse representation of the signal in the frequency domain. Impacts produce sounds that are concentrated in time, allowing a sparse representation of the signal in both the time domain and the wavelet transform domain [23]. Therefore, using a sparse representation of the audio signals seems to be a promising approach for audio representation in different applications, especially in audio steganography where such representation can be employed for data embedding.

To explain the sparse decomposition (SD) problem, let us consider the following representation:

$$\boldsymbol{x} \simeq \sum_{i=1}^{m} \varphi_i \boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{\varphi} \qquad \boldsymbol{\varphi} \triangleq (\varphi_1, \cdots, \varphi_m)^T, \tag{1}$$

where the signal $\boldsymbol{x}$ with size $n \times 1$ is represented approximately as a linear combination of $m$ predefined atoms, and $\boldsymbol{d}_i$ is the $i$-th atom ($i$-th column) of the dictionary $\boldsymbol{D}$ that is an $n \times m$ matrix. The coefficient vector, $\boldsymbol{\varphi}$, satisfies:

$$\|\boldsymbol{\varphi}\|_0 \ll m, \tag{2}$$

in which $\|.\|_0$ stands for the $\ell_0$-norm of a vector, i.e., the number of non-zero entries in the vector. The expression in Eq. (2) defines the decomposition as a sparse one if $\|\boldsymbol{\varphi}\|_0$ is small.

Let $\varphi_i$ denote the specific weight of the $i$-th atom. Many $\varphi_i$ s are zero in sparse decomposition problems. It should be mentioned that sparse representation of a signal represents an approximation of the signal $\boldsymbol{x}$ as shown in Eq. (1).

Successful application of sparse decomposition depends on the dictionary used and whether it matches the content of the signal properly [21]. Using a pre-specified dictionary leads to simple and fast sparse decompositions, though the success of such dictionaries in applications depends on how suitable they are to sparsely represent the signals [26]. Dictionary learning methods usually use an alternating optimization approach, in which the sparse representation of the training signals is obtained over a fixed dictionary; then the dictionary atoms are updated while the sparse representation assumed to be fixed [21]. One example is the KSVD method proposed by Aharon *et al.* in [26].

The GAD dictionary learning method [21] is a computationally efficient algorithm that finds sparse atoms from audio signals. It is motivated by the observation that sparsity in the dictionary and sparsity in the decomposition appear to be related for certain types of signals such as audio signals [21]. The GAD algorithm has been employed in speech representation and speech denoising [21]. We use the GAD algorithm to estimate a proper dictionary for the sparse representation of wavelet coefficients of speech signals in the proposed steganography method.

The GAD algorithm is a deterministic algorithm which gets a one dimensional signal as the input and learns the dictionary by providing a sparse representation for the signal's frames [21]. The GAD algorithm divides the input audio signal into frames. The frames are non-overlapping segments of the audio signal, as defined in Eq. (7), and collectively form a new matrix $\boldsymbol{X}$[21]. The residual matrix is initialized to $\boldsymbol{X}$ and updated at each iteration. The dictionary is then built by selecting a residual vector (a column of residual matrix) that has the lowest sparsity index [21]. It uses the following sparsity index proposed in [27] to extract new atoms from the residual vectors to form the dictionary:

$$\zeta = \frac{\|\boldsymbol{r}_i\|_1}{\|\boldsymbol{r}_i\|_2}, \tag{3}$$

where $\boldsymbol{r}_i$ is the $i$-th residual vector ($i$-th column of the residual matrix) and $\|.\|_1$ and $\|.\|_2$ denote the $\ell_1$-norm and $\ell_2$-norm, respectively. The GAD algorithm initializes the residual matrix by the audio signal frames at the first iteration and selects a residual vector with the lowest sparsity index as a new atom to add to the dictionary. Then, each column of the residual matrix is updated as:

$$\boldsymbol{r}_i^{l+1} = \boldsymbol{r}_i^l - \boldsymbol{d}^l \langle \boldsymbol{d}^l \cdot \boldsymbol{r}_i^l \rangle, \tag{4}$$

where $\boldsymbol{r}_i^l$ is the $i$-th column of the residual matrix in the $l$-th iteration and $\boldsymbol{d}^l$ is the atom added to the dictionary in this iteration. This procedure is repeated until the predefined termination rule is satisfied.

There are two termination rules: 1) the pre-determined number of atoms is reached and 2) the reconstruction error is below a predefined threshold. The reconstruction error is given as:

$$\epsilon = \|\tilde{x}(t) - x(t)\|_2 \leq \sigma, \tag{5}$$

where $\sigma$ is the maximum reconstruction error which should be set by the user and $\tilde{x}(t)$ is an approximation to the input audio signal $x(t)$ obtained from its sparse representation in the following way:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{D}\boldsymbol{D}^T\boldsymbol{X}, \tag{6}$$

in which $\boldsymbol{D}$ is the estimated dictionary and $\widetilde{\boldsymbol{X}}$ is the sparsely represented version of $\boldsymbol{X}$. The columns of $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$ are the frames of $x(t)$ and $\tilde{x}(t)$, respectively.

## III. Proposed DWT-SD Speech Steganography Method

We propose a speech steganography method which hides the secret data into non-zero coefficients of the sparse representation of the low frequency (approximation) wavelet sub-band. We now explain the embedding and extraction procedures of the proposed algorithm.

### A. Embedding procedure

The proposed DWT-SD speech steganography algorithm is a transform domain method. As mentioned in Section II, the audio

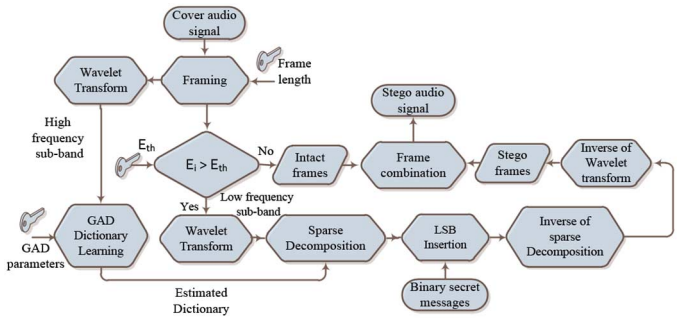



Fig. 1. Embedding procedure: The block diagram describes the major steps of the proposed DWT-SD algorithm to embed the secret message bit stream into the cover signal.

signals are sparse in the wavelet domain [23]. The DWT-SD algorithm uses this fact to insert the stego-message bit stream $\boldsymbol{p}$ into the sparse representation coefficients of the low frequency (approximation) wavelet sub-band. The block diagram of the embedding procedure is shown in Fig. 1. The step-by-step description of each block is given as follows:

**Step 1 (framing):** We divide the speech cover signal $x(t)$ into $K$ non-overlapping frames $\boldsymbol{x}_k$'s, each with $L$ samples. E.g., the $k$-th frame, $\boldsymbol{x}_k$, is represented as:

$$\boldsymbol{x}_k = [x((k-1)L+1), \ldots, x(kL)]^T, \tag{7}$$

where $k \in \{1, \ldots, K\}$. These frames collectively form the new data matrix $\boldsymbol{X}$, where its $k$-th column is $\boldsymbol{x}_k$. The length of the frames, $L$, is a secret key of the algorithm and it is assumed to be available at the receiver side.

**Step 2 (frame selection):** We calculate the energy of the frames in this step. Suppose that $E_k$ stands for the energy of the $k$-th frame $\boldsymbol{x}_k$. The frames whose energies are higher than a specific threshold are selected and used further for data embedding. We have:

$$f = \{k | \forall k \in \{1, \ldots, K\}, E_k \geq E_{th}\}, \tag{8}$$

where $f$ is the index set of the selected frames. The threshold $E_{th}$ is set as a specific fraction of the average energy:

$$E_{th} = \alpha \times \left( \frac{1}{K} \sum_{i=1}^{K} E_i \right), \tag{9}$$

where $\alpha$ is the predetermined parameter and it is available at the receiver side as one of the secret keys of the algorithm.

We select these high energy frames for data hiding because data embedding in these frames makes less audible distortion compared to using low energy frames. The proposed steganography algorithm embeds the secret bits in the coefficients of the sparse representation of wavelet coefficients, and the embedding procedure slightly changes the effect of the structural elements in the signal representation. The ratio of the energy variation due to embedding procedure to the primitive energy level is higher for the frames with lower primitive energy compared to the frames with higher primitive energy. In other words, slight modifications caused by the hidden data insertion lead to less audible distortions in high energy frames.

The parameter $\alpha$, or equivalently $E_{th}$, affects the trade-off between capacity and imperceptibility. A larger $\alpha$ results in a larger $E_{th}$, but decreases the capacity because fewer frames will exceed the threshold, while data insertion in fewer frames generally leads to a higher imperceptibility.

**Step 3 (wavelet transform):** Invertible transforms such as the DFT, the discrete cosine transform (DCT), or the discrete wavelets transform (DWT) can be easily calculated [23]. The DWT decomposes the cover signal frames into the low and high frequency components. The low frequency component of the signal is the most important part of the speech perception system, while the high frequency component influences the nuance of the signal that may include the noise [1].

Removing high frequency components of the human voice signal leads to a different voice sound but it is still quite understandable [1]. However, removing sufficient amount of low frequency components leads to a gibberish voice sound. For this reason, we use high frequency component to estimate a proper dictionary, represent the low frequency wavelet component sparsely over the atoms, and then embed the secret data in the sparse representation coefficients. Different wavelet transforms such as Haar or Daubechies family can be used. Here we use the Haar wavelet due to its simplicity.

For the frames belonging to set $f$, approximation (low frequency) and detail (high frequency) wavelet sub-bands are calculated. The high and low frequency components of the $k$-th frame, $\boldsymbol{x}_k$, are denoted as $\boldsymbol{y}_h^k$ and $\boldsymbol{y}_l^k$, respectively. Approximation wavelet vectors, $\{\boldsymbol{y}_l^k | \forall k \in f\}$, collectively form the approximation wavelet coefficient matrix $\boldsymbol{Y}_l$. The detail wavelet coefficient matrix $\boldsymbol{Y}_h$ is generated from $\{\boldsymbol{y}_h^k | \forall k \in f\}$.

**Step 4 (Dictionary estimation):** We use the high frequency components of all frames of the speech signal to estimate a proper dictionary to represent the low frequency component sparsely over it. The basic idea is to use the dependencies between the high and low frequency wavelet sub-bands to represent the low frequency wavelet sub-band sparsely over the dictionary estimated from the high frequency wavelet sub-band.

As mentioned in Section II, we employ the GAD dictionary learning algorithm in this step. The GAD algorithm finds the dictionary elements that are sparse. We use the detail wavelet coefficient matrix as the training set to estimate the dictionary matrix $\boldsymbol{D}$. The output of the GAD algorithm is a dictionary matrix $\boldsymbol{D}$ where the train set could be sparsely represented over it. We also use the estimated dictionary $\boldsymbol{D}$ to represent the low frequency wavelet sub-band sparsely.

As mentioned earlier, GAD is a deterministic algorithm. It reconstructs the dictionary at the receiver side, in the same way as that at the sender side, by using the same input signal, frame size and the termination rule, since the high frequency wavelet sub-band remains intact in the proposed DWT-SD algorithm. It is worth mentioning that although modifications of the low frequency wavelet sub-band may generally lead to changes of the high frequency sub-band, slight modifications due to data embedding by the DWT-SD method in the low frequency wavelet sub-band have a negligible effect on the high frequency wavelet sub-band. Thus, the dictionary is well reconstructed at the receiver side. In our preliminary experiments over 600 audio files, we note that the $\ell_2$-norm of the difference between dictionaries reconstructed at the receiver and the sender sides is in the order of $10^{-13}$.

One of the two termination rules explained in Section II, the pre-determined number of atoms or the highest reconstruction error value allowed, can be used in the GAD dictionary learning algorithm. The termination rule and its required value is another key of the proposed DWT-SD algorithm. We set the highest

reconstruction error value as the termination rule of the GAD algorithm in our experiments.

**Step 5 (sparse decomposition):** After finding a dictionary, the low frequency wavelet sub-band is represented sparsely over the estimated dictionary $\boldsymbol{D}$. It is shown in [21] that in addition to the advantage of producing atoms that are directly relevant to the data, the estimated dictionary by the GAD algorithm is orthogonal. Due to this orthogonality, inverse of the GAD's dictionary exists and is equal to its transpose [21].

The low frequency wavelet coefficient matrix $\boldsymbol{Y}_l$ is sparsely represented over the estimated orthogonal dictionary $\boldsymbol{D}$:

$$\widetilde{\boldsymbol{Y}}_l \simeq \boldsymbol{D}\boldsymbol{D}^{\dagger}\boldsymbol{Y}_l, \tag{10}$$

and

$$\boldsymbol{\Phi} = \boldsymbol{D}^{\dagger}\boldsymbol{Y}_l, \tag{11}$$

where $\widetilde{\boldsymbol{Y}_l}$ is an approximation of the low frequency wavelet sub-band matrix $\boldsymbol{Y}_l$ obtained from its sparse decomposition over the dictionary $\boldsymbol{D}$ and $\boldsymbol{D}^{\dagger}$ is the right pseudo-inverse of $\boldsymbol{D}$. $\boldsymbol{\Phi}$ is the sparse representation coefficient matrix where the secret messages will be inserted in its non-zero elements.

**Step 6 (secret data embedding):** Sparse decomposition of a signal represents it as a linear combination of a set of elementary structural elements. The non-zero elements of the sparse representation of a signal represent the importance of each structural element in signal reconstruction. Secret data inserted into the non-zero coefficients slightly changes the effect of the related elements in signal reconstruction. The zero coefficients are not used for data insertion, because any changes in them may lead to audible distortions in the stego signal.

Different methods can be used for data insertion into the non-zero coefficients. We use the repetition code, one of basic error-correcting codes, for data embedding. We use the $N_b$ LSB substitution method in which a secret bit is encoded $N_b$ times in the $N_b$ LSBs of the related coefficient. If the number of LSBs used for data embedding increases, the hidden data extraction errors that may occur at the receiver side decrease, while may lead to a higher distortion. $N_b$ LSBs of the fractional part of the non-zero coefficients are substituted with a secret bit in the proposed DWT-SD method. In our experiments, $N_b$ is set as 2 without loss of generality. We do not change the integer part of the non-zero coefficients, so that less distortion is introduced into the cover signal.

Suppose that $\boldsymbol{\Phi}(i,j)$ is a non-zero element of the sparse representation coefficient matrix defined in Eq. (11). The binary representation of $\boldsymbol{\Phi}(i,j)$ is:

$$\boldsymbol{\Phi}(i,j) = (b_7 \dots b_0.b_{-1}b_{-2}\dots b_{-8})_2. \tag{12}$$

We insert the $n$-th bit of the secret message, $\boldsymbol{p}_n$, in $\boldsymbol{\Phi}(i,j)$ and we get:

$$\hat{\boldsymbol{\Phi}}(i,j) = \begin{cases} \hat{b}_\ell = \boldsymbol{p}_n & -N_b \le \ell \le -1 \\ \hat{b}_\ell = b_\ell & else, \end{cases} \tag{13}$$

where $\hat{\boldsymbol{\Phi}}$ is the sparse representation coefficient matrix of the low frequency wavelet sub-band of the stego signal frames.

As mentioned earlier, the secret bits are embedded in the non-zero coefficients and they could be decoded at the receiver side. However during the data insertion procedure, some non-zero coefficients could be set to zero by embedding a zero secret bit. In this case, desyncronization between the transmitter and the receiver may happen, since the receiver tracks the positions of the non-zero coefficients to extract the hidden message bits. To prevent such desyncronization, the mentioned coefficients should be set to a non-zero value. We set the $(N_b+1)$-th bit of the fractional part of the mentioned coefficient to 1 and change the stego coefficient to a non-zero value as:

$$\text{if} \quad \hat{\boldsymbol{\Phi}}(i,j) = 0 \quad \text{and} \quad \boldsymbol{\Phi}(i,j) \neq 0$$
$$\hat{\boldsymbol{\Phi}}(i,j) = \begin{cases} \hat{b}_\ell = 1; & \ell = -(N_b+1) \\ \hat{b}_\ell = \hat{b}_\ell; & else, \end{cases} \tag{14}$$

Therefore the receiver will not lose the positions of the coefficients carrying hidden bits.


Fig. 2. The bit stream inserted into the cover signal by using the proposed DWT-SD algorithm.

Fig. 2 shows the bit stream embedded in the cover speech signal by using the DWT-SD algorithm. The first embedded byte is the number of bytes of the message length ($L_p$). The next $L_p$ bytes define the length of the secret message ($M_L$) and the next $M_L$ bytes are the main secret messages which should be transmitted. Since, the number of overhead bytes, $L_p+1$, is much smaller than the length of the secret message, $M_L$, its effect on the capacity is negligible.

**Step 7 (inverse of sparse decomposition):** Following the hidden data insertion stage, the low frequency wavelet sub-band of the stego signal is reconstructed from its sparse representation coefficient matrix as:

$$\hat{\boldsymbol{Y}}_l = \boldsymbol{D}\hat{\boldsymbol{\Phi}}, \tag{15}$$

where $\hat{\boldsymbol{Y}}_l$ is the low frequency wavelet sub-band of the stego signal, $\boldsymbol{D}$ is the estimated dictionary and $\hat{\boldsymbol{\Phi}}$ is the stego sparse representation coefficient matrix given by Eq. (13).

**Step 8 (stego signal reconstruction):** The low frequency wavelet coefficient matrix of the stego signal is calculated in Eq. (15). As mentioned in Step-4, the high frequency wavelet coefficient matrix of the cover signal, $\boldsymbol{Y}_h$, is left intact. The frames of the stego signal, belonging to the set $f$, are reconstructed as:

$$\hat{\boldsymbol{X}}^f = \text{IDWT}(\hat{\boldsymbol{Y}}_l, \boldsymbol{Y}_h), \tag{16}$$

where $\hat{\boldsymbol{X}}^f$ is a matrix containing the stego signal frames which their indices belong to the set $f$. IDWT$(\cdot)$ means the column-wise inverse discrete wavelet transform. The set $f$, defined in Eq. (8), contains the indices of the frames carrying the secret data. Other frames of the cover signal are retained unchanged in the stego signal. The stego signal $x^s(t)$ is produced by reversing the framing process. In summary, the DWT-SD algorithm replaces the high energy frames of the cover signal, whose indices are saved in the set $f$, with the stego frames, which are the columns of $\hat{\boldsymbol{X}}^f$ given by Eq. (16).

### *B. Extraction Procedure*

At the receiver side, the following information is assumed available: the frame length $L$, the threshold value $E_{th}$ or the pa-

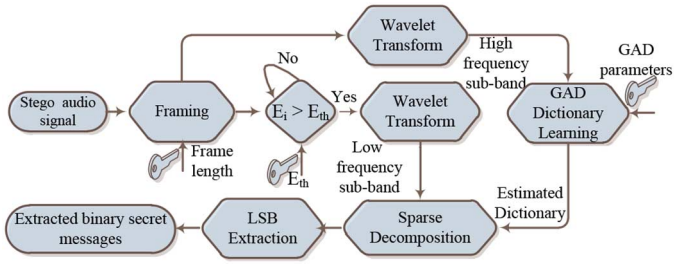



Fig. 3. Block diagram of the extraction procedure of the proposed DWT-SD algorithm.

rameter $\alpha$ used for the threshold calculation, the termination rule of the GAD step and its related parameter (reconstruction error $\sigma$ or pre-determined number of atoms). The embedded messages can be extracted based on the received stego signal and the secret keys.

The block diagram of the extraction procedure is shown in Fig. 3, with detailed descriptions as follows:

**Step 1 (framing):** Same as Step-1 of the embedding procedure, the stego signal $x^s(t)$ is divided into $K$ non-overlapping frames $\boldsymbol{x}_k^s$, each of length $L$ samples. It is to be noted that the frame length, $L$, is used as a secret key that is available to the receiver.

**Step 2 (frame selection):** The frames which are used for data insertion at the transmitter side should be selected in this step to extract the secret data. The energy threshold used for the frame selection, $E'_{th}$, is calculated by using the stego signal frames, as expressed by Eq. (9). The frames of the cover signal are separated into two distinct sets: the frames in the set $f$ whose energies are higher than the threshold $E_{th}$ and the remaining frames in the set $f^c$.

The frames belonging to $f^c$ are retained intact in the embedding procedure of the DWT-SD algorithm and their energies remain unchanged. The frames belonging to set $f$ are changed slightly because of the secret data insertion that might have altered their energies. Therefore, the threshold $E'_{th}$, calculated by using the stego signal frames at the receiver side, may be different from $E_{th}$ calculated from the cover signal frames in Eq. (9) at the transmitter side. So, theoretically, it is possible that $E'_{th}$ results in a different set of the high energy frames $f'$ that may lead to the hidden bit extraction errors at the receiver side.

To address this concern, the parameter $\alpha$, which is used as the secret key to calculate $E'_{th}$ at the receiver side, is replaced by the following threshold value as the secret key:

$$E'_{th} = \min_{i \in f} \{E'_i | E'_i > E_M^{f^c}\}, \tag{17}$$

where $E'_i$ is the energy of the $i$-th frame of the stego signal belonging to set $f$ and $E_M^{f^c}$ is the maximum energy of the frames belonging to set $f^c$, which is calculated as:

$$E_M^{f^c} = \max_{j \in f^c} \{E_j\}, \tag{18}$$

where $E_j$ is the energy of the $j$-th frame of the cover signal belonging to set $f^c$. Utilizing the threshold defined in Eq. (17) as the secret key, instead of $\alpha$, ensures the correct restoration of the set $f$ at the receiver side. During the embedding procedure, the energy of each stego frame is calculated and compared to $E_M^{f^c}$. In the case that data insertion in the $\ell$-th frame causes its energy to be $E'_\ell \leq E_M^{f^c}$, the sender will assume that the $\ell$-th frame does not carry any hidden bits (even though hidden data is inserted into the $\ell$-th frame) and the associated hidden bits will be inserted again in the next eligible frame. Therefore, at the sender side, the sender will not consider the energy of the $\ell$-th frame to set the secret key (the threshold $E'_{th}$) as defined in Eq. (17). While at the receiver side, such a $\ell$-th frame will be correctly classified as a frame carrying no hidden bits.

Although a secure channel is assumed to transmit the keys, the threshold could be set to a fixed $\ddot{E}_{th}$ in the case that there is no secure channel available for the key exchange. A training set of cover signals can be used for setting a proper pre-established $\ddot{E}_{th}$. The sender can calculate the threshold values of a training set by using Eq. (9) and set their average as $\ddot{E}_{th}$. Using a fixed $\ddot{E}_{th}$ instead of the threshold calculated for the cover signal, $E_{th}$ in Eq. (9), may lead to capacity or imperceptibility degradation. In the case where $\ddot{E}_{th}$ is lower than $E_{th}$ of the cover signal, hidden data insertion in these frames $f' = \{k | \forall k \in \{1, \ldots, K\}, \ddot{E}_{th} \leq E_k \leq E_{th}\}$ may lead to imperceptibility violation. In the case where $\ddot{E}_{th}$ is higher than $E_{th}$, these frames $f' = \{k | \forall k \in \{1, \ldots, K\}, E_{th} \leq E_k \leq \ddot{E}_{th}\}$ are not utilized for data insertion by using $\ddot{E}_{th}$, while they were exploited for data insertion by using $E_{th}$. Thus, a bit of the available capacity in the cover signal may be lost by using a fixed $\ddot{E}_{th}$ instead of $E_{th}$.

**Step 3 (wavelet transform):** Low and high frequency wavelet components of the stego frames $\boldsymbol{x}_k^s$, belonging to the set $f$, are calculated by using the Haar DWT similar to the embedding procedure. Approximation and detail wavelet sub-bands of the $k$-th frame are denoted as $\boldsymbol{y}_l^{s^k}$ and $\boldsymbol{y}_h^{s^k}$, respectively. The stego approximation wavelet coefficient matrix $\boldsymbol{Y}_l^s$ is reconstructed from the stego approximation wavelet vectors, $\{\boldsymbol{y}_{l^k}^s | \forall k \in f\}$. The stego detail wavelet coefficient matrix $\boldsymbol{Y}_h^s$ is reconstructed in the same way from $\{\boldsymbol{y}_{h^k}^s | \forall k \in f\}$.

**Step 4 (dictionary restoration):** The detail wavelet coefficient matrix $\boldsymbol{Y}_h$ is retained unchanged during the embedding procedure. We therefore use the detail wavelet coefficient matrix of all frames as the train set of the GAD to restore the dictionary $\boldsymbol{D}$. As described in Step-4 of the embedding procedure, the termination rule of the GAD algorithm and its related parameters, the pre-specified number of the atoms or the highest reconstruction error value, are assumed available as secret keys of the algorithm at the receiver side.

**Step 5 (sparse decomposition):** Following the dictionary restoration step, the approximation wavelet matrix $\boldsymbol{Y}_l^s$ is represented sparsely over the obtained dictionary $\boldsymbol{D}$ and we have the sparse representation coefficient matrix $\boldsymbol{\Phi}_l^s$ as:

$$\boldsymbol{\Phi}_l^s = \boldsymbol{D}^\dagger \boldsymbol{Y}_l^s. \tag{19}$$

Non-zero coefficients of $\boldsymbol{\Phi}_l^s$ contain the secret data.

**Step 6 (LSB extraction):** The $n$-th bit of the hidden message $\boldsymbol{p}_n$ is extracted from the $n$-th non-zero element of $\boldsymbol{\Phi}_l^s$. The hidden data extraction is done based on majority vote: the extracted bit is set to the most frequent bit in the $N_b$ LSBs of the fractional part of the related non-zero coefficient. The frequency of ones may be equal to the frequency of zeros when $N_b$ is an even number. In this case, the decision is made according to the first LSB of the fractional part (i.e., the most significant bit of the $N_b$ LSBs containing the secret data).

TABLE I
COMPARING THE SNR VALUES OF THE STEGO SIGNALS OBTAINED BY THE PROPOSED DWT-SD METHOD FOR DIFFERENT PAYLOADS AND FRAME LENGTH VALUES $L$

| payload | frame length (L) | | | |
|---|---|---|---|---|
| | L=16 | L=32 | L=64 | L=128 |
| 0.4 kbps | 36.12 dB | 36.03 dB | 36.08 dB | 36.01 dB |
| 0.75 kbps | 32.70 dB | 33.02 dB | 33.08 dB | 33.20 dB |
| 1.1 kbps | 31.02 dB | 31.18 dB | 31.22 dB | 31.09 dB |
| 1.8 kbps | 29.81 dB | 29.94 dB | 29.96 dB | 30.03 dB |

## IV. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed DWT-SD speech steganography algorithm on TIMIT [25], [28], [29], [30], [31], [32], [33] and NOIZEUS [1], [34], [35], [36] databases. The TIMIT is a well-known database of speech signals containing 6300 utterances recorded from 630 adult speakers of eight major dialects of American English, each read ten sentences, where 70% of the speakers are male and 30% are female [28]. The length of the speech file varies between 1.4 and 5.04 seconds. The speech signals are sampled at 16 kHz, where each sample is quantized by 16 bits. The NOIZEUS database, the database used in [1], contains 30 sentences from the IEEE sentence database, recordings of three male and three female speakers [34], [35], [36]. 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 seconds. We evaluate the proposed DWT-SD algorithm based on three main attributes of steganography: imperceptibility, capacity, and undetectability.

We first evaluate the quality of the stego speech signal $x^s(t)$ when compared with the cover signal $x(t)$, which is a performance measure of steganographic schemes. In this paper, we use two objective performance measures to verify imperceptibility of the proposed method. The first objective measure is the signal-to-noise ratio (SNR). To determine the imperceptibility of the proposed method, we use the DWT-SD to embed different payload values in 100 randomly selected cover signals. The embedding is done for different frame length values, $L$. Table I compares the average SNR values of the stego signals, when applying the DWT-SD with various frame lengths $L$, at different embedding rates. In all these experiments, the parameter $\alpha$ used for calculating the threshold $E_{th}$ is set to be 0.5 without loss of generality.

The termination rule of the GAD dictionary learning is based on the highest reconstruction error defined in Eq. (5). The lower $\sigma$ results in a dictionary which represents the cover signal more accurately. Parameter $\sigma$ is obtained by searching for a predetermined range of values and finding the value which results in a desirable PESQ for the approximated cover signal. The initial $\sigma$ value and the desirable PESQ are set to 0.1 and 4 in our experiments. The parameter $\sigma$ is decreased gradually and the average quality of five approximated random cover signals is calculated. The reduction procedure is terminated when the average PESQ reaches its desirable value. The highest reconstruction error, $\sigma$, is set to be 0.01 in all the experiments based on the searching process. Table I shows that the SNR values of the proposed DWT-SD method are in the acceptable SNR range at different embedding rates and frame length values according to [37].

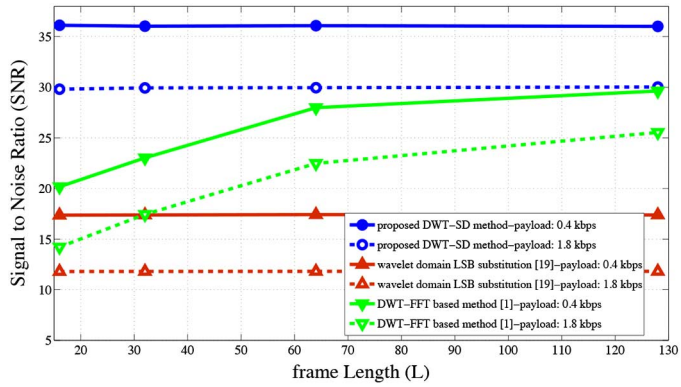



Fig. 4. Comparisons of the SNR values of the stego signals obtained by the proposed DWT-SD method, the DWT-FFT based method [1], and the wavelet domain LSB substitution method [19], as a function of the frame length ($L$) at two embedding rates.

We then compare the imperceptibility of the proposed DWT-SD method with that of the methods in [1] and [19]. In the experiment, 100 audio covers are selected randomly from the database and segmented into non-overlapping frames. The proposed DWT-SD steganography method, the DWT-FFT based speech steganography method [1] and the wavelet LSB Substitution method [19] are used to insert secret data at two embedding rates 0.4, and 1.8 kbps, into the randomly selected cover signals. For each method, the average of the SNR values of the 100 obtained stego signals which contain specified payload are calculated versus the frame length $L$. The parameters of the DWT-SD algorithm are the same as those in the previous experiment. Fig. 4 compares the SNR values of the proposed DWT-SD method to those of the two other methods.

We select the DWT-FFT based method [1] for comparison because it allows hiding a large amount of secret data while keeps it undetectable by steganalysis [1]. It is shown that the stego signals made by the DWT-FFT method are indistinguishable from their original cover signals [1]. In addition, the DWT-FFT method [1] is also a wavelet transform based method which makes it a good choice for a fair comparison with the proposed method. The DWT-FFT based method [1] hides the secret data in the magnitude of the FFT transform of the high frequency wavelet sub-band, while the proposed DWT-SD uses the sparse representation of the low frequency wavelet sub-band for data embedding. The wavelet domain LSB Substitution method [19] is selected for comparison because it is a LSB Substitution based method such as the proposed method, which uses the wavelet transform domain to insert the secret data. As shown in Fig. 4, the average SNR value of the proposed method, at each embedding rate, is significantly larger than the values obtained by the methods in [1] and [19].

The second objective measure is the perceptual evaluation of speech quality (PESQ) [38], to evaluate the imperceptibility of the proposed method. Subjective listening test is the most accurate and reliable method to evaluate the quality of audio signals, but it is costly and time consuming [38]. The PESQ measure has been shown to be highly correlated with the subjective listening tests [38]. Therefore we use the PESQ to evaluate the quality of the stego signals.

In general, the PESQ measure is a score between 0.5 and 4.5, where a higher score denotes a better quality [1]. The same as in previous experiments, four different payload rates are selected for embedding data into 100 randomly selected cover signals by

TABLE II
COMPARISONS OF THE PESQ VALUES OF THE STEGO SIGNALS OBTAINED BY THE DWT-SD METHOD FOR DIFFERENT PAYLOADS AND FRAME LENGTH VALUES, $L$

| payload | frame length (L) | | | |
|---|---|---|---|---|
| | L=16 | L=32 | L=64 | L=128 |
| 0.4 kbps | 4.13 | 4.15 | 4.14 | 4.16 |
| 0.75 kbps | 3.87 | 3.90 | 3.94 | 3.92 |
| 1.1 kbps | 3.74 | 3.76 | 3.78 | 3.80 |
| 1.8 kbps | 3.68 | 3.69 | 3.71 | 3.73 |

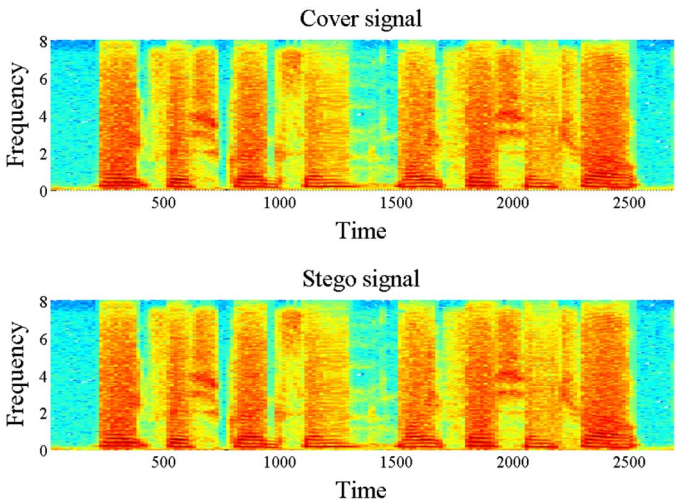



Fig. 5. Spectrograms of the cover signal (top) and the stego signal (bottom). The payload is 1.8 kbps.

using the proposed DWT-SD method. The average PESQ values of the stego signals obtained by the DWT-SD method are presented in Table II. The parameters of the DWT-SD algorithm are set to be the same as those in previous experiments and average PESQ values are reported for different frame length values ($L$).

Table II shows that the PESQ values of the proposed DWT-SD is higher than 3.7 and the values obtained for different frame lengths are close to each other. The PESQ analysis shows that the stego signals made by the DWT-SD and the cover signals yield similar subjective quality. The results reported in Tables I and II confirm the imperceptibility of the proposed DWT-SD method. In other words, the proposed DWT-SD steganography algorithm generates no audible distortion and thus does not attract suspicion about the existence of a secret data in the stego signal. This result is supported by the resemblance between the cover and stego signal spectrograms in Fig. 5.

Fig. 6 compares the quality of the stego signals obtained by the proposed DWT-SD method, the DWT-FFT based method [1], and the wavelet domain LSB Substitution method [19] in terms of the PESQ measure. Each steganography method embeds data at two payload values (0.4, and 1.8kbps) into the 100 randomly selected cover signals, and then the average of the PESQ values of 100 stego signals obtained at each embedding rate is reported in Fig. 6. Comparisons between the PESQ curves of three methods for each payload value show that the proposed DWT-SD method outperforms both of the DWT-FFT [1] and the wavelet domain LSB Substitution [19] methods in terms of the stego signal quality.

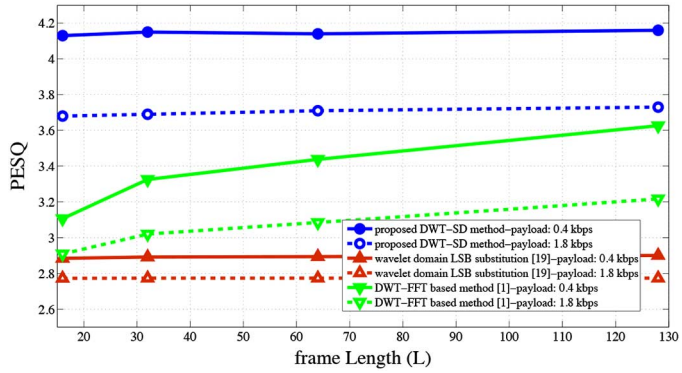



Fig. 6. Comparisons of the PESQ values of the stego signals obtained by the DWT-SD method, the DWT-FFT based method [1], and the wavelet domain LSB substitution method [19] as a function of the frame length ($L$).

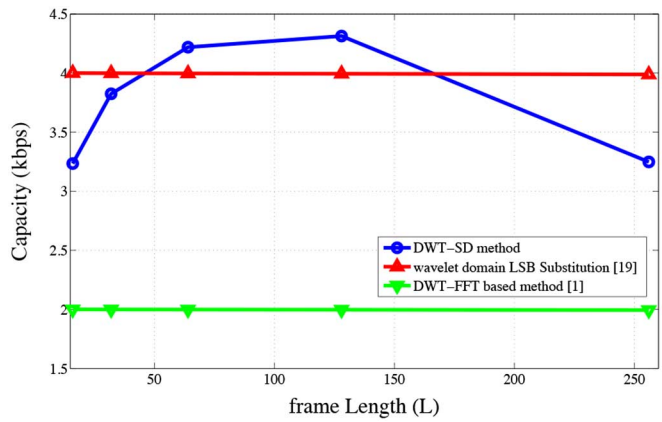



Fig. 7. Comparison between the embedding capacities of the DWT-SD method, the DWT-FFT method [1], and the wavelet domain LSB substitution method [19].

The capacity of a steganography method is the amount of secret data bits which could be embedded into the cover signal and is one of the main concerns of steganographic algorithms. We evaluate the capacity of the proposed DWT-SD algorithm when compared with the methods given in [1] and [19].

By using the same parameters as those in previous experiments, the number of locations available for embedding the secret data in the cover speech signal is calculated. The average of this number as a function of the frame length, $L$, is taken as a measure of capacity. The average value for each frame length $L$ is obtained from the tests over 100 randomly selected cover signals. Fig. 7 shows that the proposed DWT-SD method outperforms the DWT-FFT method [1] in the sense of the capacity. Although the capacity of the wavelet domain LSB substitution method [19] is higher than that of the proposed method for some frame length values, it is worth noting that the quality of the stego signals obtained by the method [19] is significantly lower than that of the proposed method, as shown in Figs. 4 and 6.

From Fig. 7, it is noted that an optimum frame length value may exist, i.e., yielding the maximum embedding capacity for the DWT-SD method. The optimum frame length value is 128 in our experiments. Since the GAD algorithm estimates an orthogonal dictionary, increasing the frame length is equal to the increase of the number of atoms $m$. Therefore the upper bound of the $\ell_0$-norm, explained in Eq. (2), of the sparse representation coefficient vector $\varphi$, defined in Eq. (1), increases. For frame length values lower than the optimum value, increasing the number of atoms leads to more non-zero elements in sparse

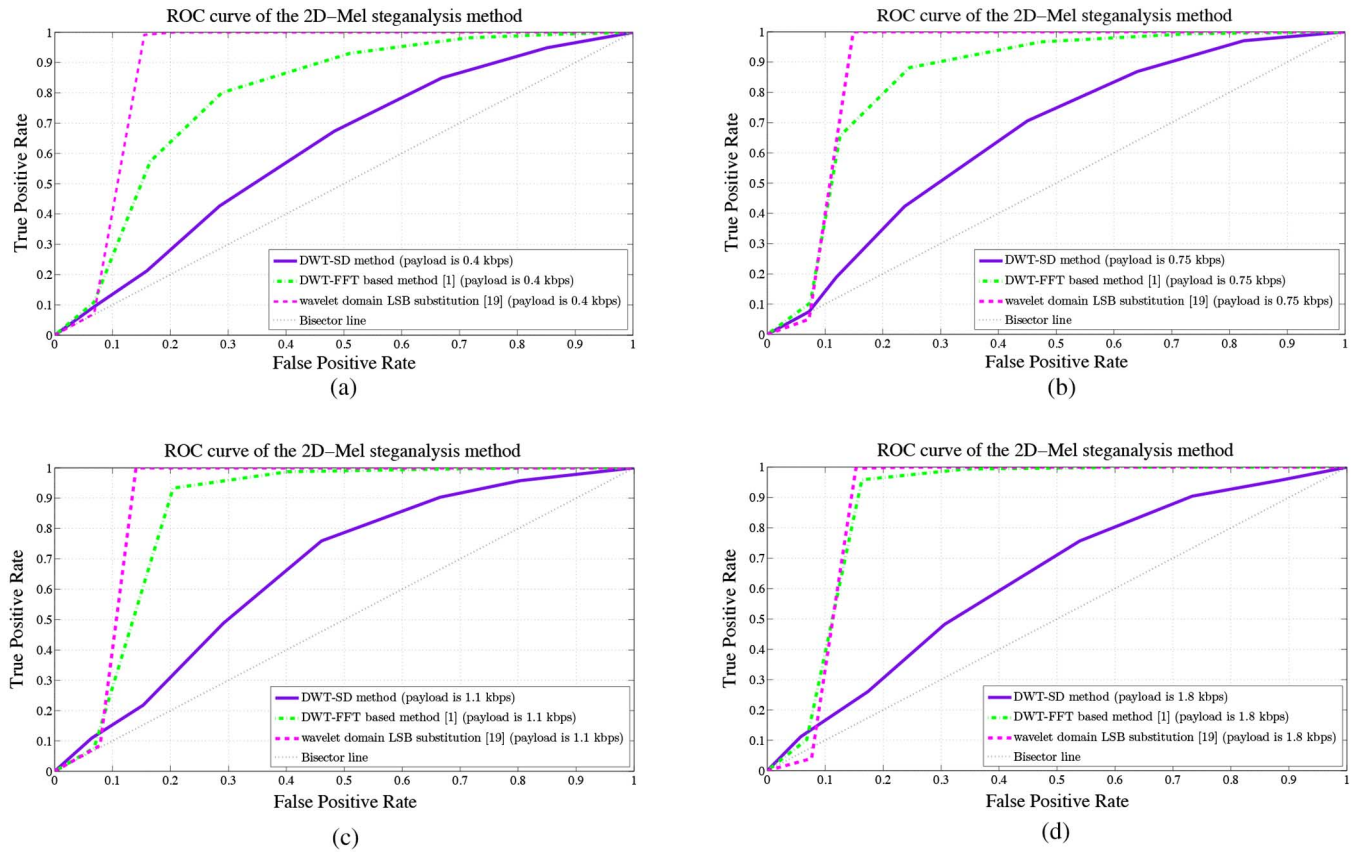



Fig. 8. ROC curves of the 2D-Mel steganalysis method [39] on the stego speech files generated by three different steganography methods, the proposed DWT-SD, the DWT-FFT based [1] and the wavelet domain LSB substitution methods [19] (a) ROC curves of the 2D-Mel steganalysis method [39]. The payload is 0.4 kbps for different steganography methods. (b) ROC curves of the 2D-Mel steganalysis method [39]. The payload is 0.75 kbps for different steganography methods. (c) ROC curves of the 2D-Mel steganalysis method [39]. The payload is 1.1 kbps for different steganography methods. (d) ROC curves of the 2D-Mel steganalysis method [39]. The payload is 1.8 kbps for different steganography methods.

representation of signals used for data insertion in the DWT-SD method. The reason is that more atoms are required in the sparse representation to represent the signals more accurately. Increasing the number of atoms (frame length), larger than the optimum value, has an opposite effect on the capacity. This is because with more atoms, the chance for finding a smaller subset of the atoms representing signals sparsely increases. Consequently, the capacity of the proposed method, which is the number of non-zero elements in the sparse representation of the cover signal, decreases.

Steganalysis is the countermeasure of steganography which tries to identify the presence of hidden data in the suspicious digital signals [39]. Steganalysis methods are effective tools to evaluate the undetectability performance of a steganography method [40]. We use two successful audio steganalysis methods to evaluate the undetectability of the proposed DWT-SD method, when compared with the methods introduced in [1] and [19]. Generally, for stego signals created using the same steganography method, a higher detection accuracy is achieved for a higher embedding ratio [39]. We use the DWT-SD algorithm and the methods given in [1] and [19] to generate stego signals at different embedding rates.

Liu *et al.* presented the second-order derivative-based Mel-cepstrum (2D-Mel) audio steganalysis method [39]. Their results demonstrated that the 2D-Mel steganalysis method significantly improved the state of the art in audio steganalysis [39]. The Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound [39]. The 2D-Mel steganalysis method extracts 58 features from the Mel-frequency cepstrum of the second-order derivative of the audio signals and uses support vector machines (SVM) for classification [39]. We use 600 randomly selected cover signals and 600 stego signals, generated by the DWT-SD at the 0.4 kbps payload rate, to obtain the ROC (Receiver Operating Characteristic) curve of the 2D-Mel steganalysis of the DWT-SD method in Fig. 8(a). The ROC curves of the 2D-Mel steganalysis of the DWT-FFT based method [1] and the wavelet domain LSB substitution method [19] are obtained in the same way and utilized for comparison. We used six different frame length values (16, 32, 64, 128, 256, and 512) to generate the stego files by using different steganography algorithms.

In all the experiments, the SVM classifier uses randomly selected 20% of the stego signals and 20% of the cover signals as the training set. The remaining 80% of the stego and the cover signals are used as the test set. Fig. 8 compares the ROC curves of the 2D-Mel steganalysis [39] of three different steganography methods, the proposed method and the methods of [1] and [19], for different payloads.

As a general assessment of the steganalysis result, a larger area under the ROC curve (AUC) indicates higher detection accuracy of the steganalyzer and vice versa. Conversely, a ROC curve closer to the bisector line shows less Detectability of stego signals by the steganalyzer. Fig. 8 shows that 2D-Mel steganalysis method [39] is less successful in detecting the secret data

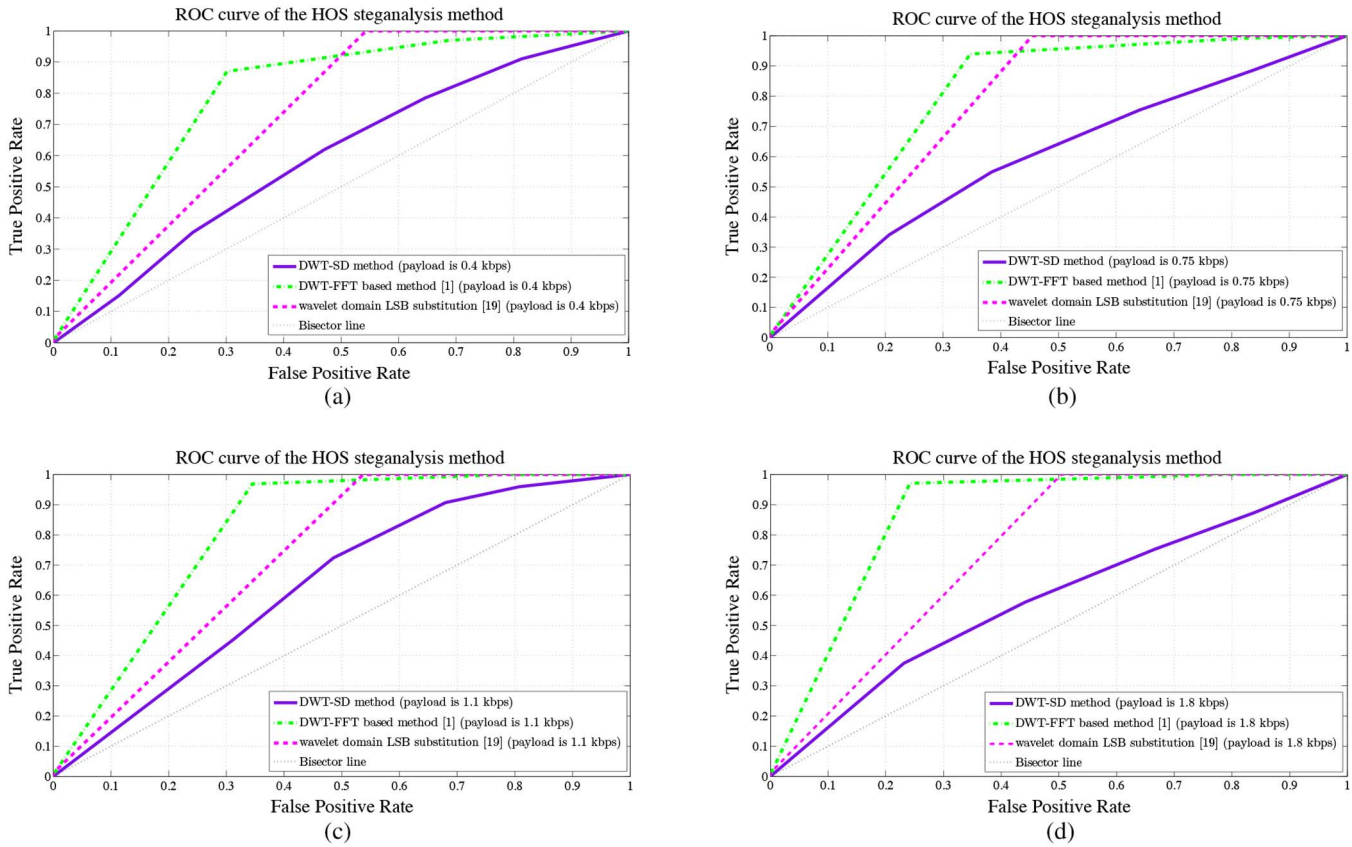



Fig. 9. ROC curves of the HOS steganalysis method [40] on the stego files obtained by using three different steganography methods, the proposed DWT-SD, the DWT-FFT based [1] and the wavelet domain LSB Substitution methods [19] (a) ROC curves of the HOS steganalysis method [40]. The payload is 0.4 kbps for different steganography methods. (b) ROC curves of the HOS steganalysis method [40]. The payload is 0.75 kbps for different steganography methods. (c) ROC curves of the HOS steganalysis method [40]. The payload is 1.1 kbps for different steganography methods. (d) ROC curves of the HOS steganalysis method [40]. The payload is 1.8 kbps for different steganography methods.

inserted by the proposed DWT-SD method when compared to the methods in [1] and [19]. From Fig. 8(a), (b), (c), and (d), the 2D-Mel steganalysis method is slightly more successful in detecting the DWT-FFT method [1] and the wavelet domain LSB substitution method [19] for higher payloads, while it is not more successful in detecting the DWT-SD stego files even for higher payloads.

We use the steganalysis method proposed in [40] as the second steganalyzer to evaluate the security of the proposed steganography method. In [40], the authors proposed a distortion metric based on Hausdorff distance to measure the distortion caused by the hidden data inserted into the audio signals [40]. The distortion measurements are gained at various wavelet decomposition levels and their high-order statistics are used as the features of a SVM classifier to determine the existence of hidden data in an audio signal [40]. We call the steganalysis method presented in [40] as the HOS steganalysis method throughout the paper where HOS abbreviates the high order statistics. Extensive experimental results proved that the HOS steganalysis method has a strong discriminatory ability for the LSB substitution based steganography methods [40]. So, we use the HOS steganalysis method [40] to detect the stego files generated by the proposed DWT-SD method, the DWT-FFT method [1] and the wavelet domain LSB substitution method [19].

As in previous steganalysis experiments, each steganography method is utilized to insert the hidden data into 600 randomly selected speech files, for each embedding rate (0.4, 0.75, 1.1, and 1.8 kbps). The training and test sets of the SVM classifier are generated in the same way as done in previous experiments. The number of wavelet decomposition levels used by the HOS steganalyzer [40] is set to be 2 in our experiments. The number of the higher order statistics is set to be 5 as in the experiments conducted in [40].

We see in Fig. 9 that the HOS steganalyzer [40] is not able to detect the stego files generated by the proposed DWT-SD method. Performances of the HOS steganalyzer in detecting different steganography methods are compared in Fig. 9(a), (b), (c), and (d) for different embedding rates. Fig. 9 indicates that the HOS steganalyzer performs better in detecting the stego files generated by the DWT-FFT based method [1] and the wavelet domain LSB substitution method [19] when compared with the proposed method.

We so far have compared our method with the DWT-FFT based [1] and the wavelet domain LSB Substitution methods [19] in terms of imperceptibility, capacity, and undetectability. Now we also investigate the computational complexity of the proposed method. Table III shows the amounts of time taken to execute the MATLAB codes of the embedding and the extraction phases of the proposed DWT-SD, the DWT-FFT [1], and the wavelet domain LSB Substitution [19] methods, on a machine with Intel Core 2 Duo at 2 GHz processor. The execution time of the proposed DWT-SD and that of the DWT-FFT in [1] are quite close. As shown in the table, the wavelet domain LSB Substitution method [19] is slightly faster than the DWT-SD method,

TABLE III
AVERAGE EXECUTION TIME OF THE DATA EMBEDDING AND DATA EXTRACTION PHASES OF THE PROPOSED DWT-SD, THE DWT-FFT BASED [1] AND THE WAVELET DOMAIN LSB SUBSTITUTION METHODS [19]. THE PAYLOAD IS 1.8 KBPS

| method | embedding phase | extraction phase |
|---|---|---|
| proposed DWT-SD | 2.21 sec | 1.81 sec |
| DWT-FFT method [1] | 2.08 sec | 1.89 sec |
| wavelet domain LSB Substitution method [19] | 1.79 sec | 1.35 sec |

while it is outperformed by the proposed method in the terms of undetectability and imperceptibility, two main issues of a steganography system.

As mentioned in Section III, the keys of the proposed DWT-SD method are the frame length $L$, the threshold value $E_{th}$ or the parameter $\alpha$ used for the threshold calculation, the termination rule of the GAD step and its related parameter ($\sigma$ or the number of the atoms). The DWT-FFT based method [1] also shares the keys between the sender and the receiver. The keys of the DWT-FFT based method [1] are the frame length $L$ and a scaling coefficient $a$ which is a function of the energy of the cover speech and the energy of the secret data. At the receiver side, the scaling coefficient $a$ is used to rescale the secret data to its original values [1]. The wavelet domain LSB Substitution method [19] uses the keys to increase the security of the method. Both communication partners share and use the keys as the seeds to a pseudo-random number (PN) generator, so they can create the same random sequences. The wavelet domain LSB Substitution method [19] uses the sequences made by the PN generator for randomizing wavelet coefficients used in data embedding and also combining the sequence with the secret data to make it a random sequence. It is supposed that sender and receiver have knowledge of the PN generator and they share the seeds needed for sequence generation as the secret keys [19].

## V. CONCLUSION

We have proposed a novel wavelet domain speech steganography method, called the DWT-SD method, by exploring the sparse representation of wavelet coefficients. The proposed DWT-SD method hides the secret data in speech frames whose energy is higher than a specified threshold to avoid audible distortions due to data hiding. The low frequency wavelet sub-band is sparsely represented over the dictionary estimated from the high frequency wavelet sub-band. The secret messages are embedded into non-zero elements of the sparse representation.

The experiments conducted show that the quality of the stego signals generated by the proposed DWT-SD method is much better than the quality of the stego signals generated by two popular steganography methods, the DWT-FFT based method [1] and the wavelet domain LSB substitution method [19], in terms of both the average SNR and PESQ of the stego signals.

In terms of the embedding capacity, it has been shown that the proposed DWT-SD method yields a higher embedding capacity than that of the DWT-FFT method [1] for various values of the frame length. Although embedding capacity of the wavelet domain LSB substitution method [19] is higher than our method's capacity for some frame length values, the quality of the stego signals generated by the method [19] is much worse than that of the proposed method.

The security of the proposed speech steganography method has been tested against two well-known, powerful steganalyzers, the 2D-Mel audio steganalysis method [39] and the HOS steganalyzer [40]. It is shown that the proposed DWT-SD algorithm is greatly less detectable by both of the 2D-Mel and HOS steganalyzers, when compared with the DWT-FFT [1] and the wavelet domain LSB substitution steganography methods [19].

As a future work, using the frames with more perceptually masked spectral components for data insertion may strengthen imperceptibility of the proposed method. We will explore the idea of using a number of perceptually masked spectral components for the frame selection and investigate its effects on the imperceptibility, undetectability and capacity of the proposed steganography method.

## REFERENCES


[1] S. Rekik, D. Guerchi, S. A. Selouani, and H. Hamam, "Speech steganography using wavelet and fourier transforms," *EURASIP J. Audio, Speech, Music Process.*, no. 1, pp. 1–14, Aug. 2012.

[2] L. Von Ahn and N. J. Hopper, "Public-key steganography," in *Advances in Cryptology-EUROCRYPT 2004*. Berlin, Germany: Springer-Verlag, 2004, vol. 3027, pp. 323–341.

[3] P. Bassia, I. Pitas, and N. Nikolaidis, "Robust audio watermarking in the time domain," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 232–241, Jun. 2001.

[4] A. D. Ker, P. Bas, R. Bohme, R. Cogranne, S. Craver, T. Filler, J. Fridrich, and T. Pevny, "Moving steganography and steganalysis from the laboratory into the real world," in *Proc. 1st ACM Workshop Inf. Hiding Multimedia Security*, Montpellier, France, Jun. 2013.

[5] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 4, pp. 474–481, May 1998.

[6] W. Mazurczyk, "Lost audio packets steganography: The first practical evaluation," *Security Commun. Netw.*, vol. 5, no. 12, pp. 1394–1403, 2012.

[7] A. S. Tanenbaum, *Computer Networks*. Upper Saddle River, NJ, USA: Prentice-Hall, 2003.

[8] W. Mazurczyk, K. Cabaj, and K. Szczypiorski, "What are suspicious VoIP delays," *Multimedia Tools Applicat.*, vol. 57, no. 1, pp. 109–126, 2012.

[9] M. Nutzinger, "Real-time attacks on audio steganography," *J. Inf. Hiding Multimedia Signal Process.*, vol. 3, no. 1, pp. 47–65, 2012.

[10] S. Murdoch and S. Lewis, "Embedding covert channels into TCP/IP," in *Proc. Inf. Hiding*, Jul. 2005, pp. 247–266.

[11] J. Lubacz, W. Mazurczyk, and K. Szczypiorski, "Vice over IP," *IEEE Spectrum*, vol. 47, no. 2, pp. 40–45, Feb. 2010.

[12] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam, "Comparative study of digital audio steganography techniques," *EURASIP J. Audio, Speech, Music Process.*, no. 1, pp. 1–16, Oct. 2012.

[13] R. Sridevi, A. Damodaram, and S. V. L. Narasimham, "Efficient method of audio steganography by modified LSB algorithm and strong encryption key with enhanced security," *J. Theoret. Appl. Inf. Technol.*, vol. 5, no. 6, pp. 768–771, Jun. 2009.

[14] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 35, no. 3, pp. 313–336, Jan. 1996.

[15] K. Gopalan, S. J. Wenndt, S. F. Adams, and D. M. Haddad, "Audio steganography by amplitude or phase modification," in *Proc. 15th Annu. Symp. Electron. Imag.- Security, Steganography, Watermarking of Multimedia Contents V*, San Jose, CA, USA, Jun. 2003.

[16] A. Takahashi, R. Nishimura, and Y. Suzuki, "Multiple watermarks for stereo audio signals using phase-modulation techniques," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 806–815, Feb. 2005.

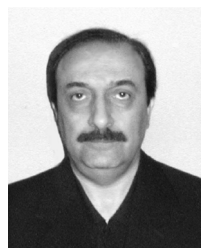
[17] N. Cvejic and T. Seppanen, "A wavelet domain LSB insertion algorithm for high capacity audio steganography," in *Proc. IEEE 10th Digital Signal Process. Workshop and 2nd Signal Process. Educ. Workshop*, Pine Mountain, GA, USA, Oct. 2002.

[18] N. Cvejic and T. Seppanen, "Increasing robustness of LSB audio steganography using a novel embedding method," in *Proc. IEEE Int. Conf. Inf. Technol.: Coding Comput. (ITCC)*, Las Vegas, NV, USA, Apr. 2004.

[19] A. Delforouzi and M. Pooyan, "Adaptive digital audio steganography based on integer wavelet transform," *Circuits, Syst. Signal Process.*, vol. 27, no. 2, pp. 247–259, Apr. 2008.
[20] Q. Liu, A. H. Sung, and M. Qiao, "Temporal derivative-based spectrum and mel-cepstrum audio steganalysis," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 359–368, Sep. 2009.
[21] M. G. Jafari and M. D. Plumbley, "Fast dictionary learning for sparse representations of speech signals," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 1025–1031, Sep. 2011.
[22] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
[23] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: From coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, Jun. 2010.
[24] F. Ghido and I. Tabus, "Sparse modeling for lossless audio compression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 1, pp. 14–28, Jan. 2013.
[25] Y. He, J. Han, S. Deng, T. Zheng, and G. Zheng, "A solution to residual noise in speech denoising with sparse representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4653–4656.
[26] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[27] V. Y. F. Tan and C. Fevotte, "A study of the effect of source sparsity for various transforms on blind audio source separation performance," in *Proc. Workshop Signal Process. Adaptive Sparse Structured Represent. (SPARS)*, Rennes, France, Nov. 2005.
[28] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Linguistic Data Consortium, Philadelphia, PA, USA, 1993.
[29] W. Mazurczyk, "Lost audio packets steganography: The first practical evaluation," *Security Commun. Netw.*, vol. 5, no. 12, pp. 1394–1403, Dec. 2012.
[30] K. Gopalan and Q. Shi, "Audio steganography using bit modification-A tradeoff on perceptibility and data robustness for large payload audio embedding," in *Proc. 19th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Zurich, Switzerland, Aug. 2010.
[31] K. Gopalan and S. Wenndt, "Audio steganography for covert data transmission by imperceptible tone insertion," in *Proc. IASTED Int. Conf. Commun. Syst. Applicat. (CSA)*, Banff, AB, Canada, 2004.
[32] A. Shahbazi, E. Soltanmohammadi, A. H. Rezaei, A. Sayadiyan, and S. Mosayyebpour, "Content dependent data hiding on GSM full rate encoded speech," in *Proc. IEEE Int. Conf. Signal Acquis. Process. (ICSAP'10)*, 2010, pp. 68–72.
[33] E. T. Su, "Robust data embedding based probabilistic global search in MDCT domain," in *Informatics Engineering and Information Science*. Berlin/Heidelberg, Germany: Springer , 2011, pp. 290–300.
[34] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7, pp. 588–601, Jul. 2007.
[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[36] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.
[37] Y. Xiang, I. Natgunanathan, D. Peng, W. Zhou, and S. Yu, "A dual-channel time-spread echo method for audio watermarking," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 383–392, Apr. 2012.
[38] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[39] Q. Liu, A. H. Sung, and M. Qiao, "Derivative-based audio steganalysis," *ACM Trans. Multimedia Computing, Commun., Applicat. (TOMCCAP)*, vol. 7, no. 3, pp. 1–19, Aug. 2011.
[40] Y. Liu, K. Chiang, C. Corbett, R. Archibald, B. Mukherjee, and D. Ghosal, "A novel audio steganalysis based on high-order statistics of a distortion measure with hausdorff distance," in *Information Security*. Berlin/Heidelberg, Germany: Springer , 2008, pp. 487–501.

**Soodeh Ahani** received the B.S. degree from Amirkabir University of Technology, Tehran, Iran, in 2006 in biomedical engineering. She received the M.Sc. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2008, where she is currently a Ph.D. candidate.

She joined the Electronics Research Institute at Sharif University of Technology in 2008 as a Research Assistant. Since October 2014, she has worked as a Visiting Researcher at the ECE Department, University of British Columbia, Vancouver, Canada. Her research interests are in sparse decomposition based audio, image and video analysis, including steganography, watermarking, coding, enhancement, and compression.



**Shahrokh Ghaemmaghami** is an Associate Professor of signal processing at Sharif University of Technology (SUT), Tehran, Iran. He is a member of Communications and System Engineering Group in the Electrical Engineering Department, a member of the Center of Excellence for Information Security, and Director of the Electronics Research Institute at the SUT. He received his B.S. and M.S. degrees from the Electrical Engineering Department of SUT and his Ph.D. from Queensland University of Technology, Brisbane, Australia. He has been leading many large research projects in speech, image, and video processing and teaching courses in speech processing, including coding, recognition, and synthesis, and also in information hiding, e.g., watermarking, steganography, and steganalysis. He is a member of the IEEE ComSoc, IEEE Computer Society, and the ACM and has served as a member of technical committees and advisory boards of several international conferences and journals.

**Z. Jane Wang** received the B.Sc. degree from Tsinghua University, China, in 1996 and the M.Sc. and Ph.D. degrees from the University of Connecticut in 2000 and 2002, respectively, all in electrical engineering. She has been a Research Associate of the ECE Department at the University of Maryland, College Park. Since 2004, she has been with the University of British Columbia, Canada, and is currently a Professor. Her research interests are in the broad areas of statistical signal processing theory and applications.