IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 18, 2019, accepted October 16, 2019, date of publication October 22, 2019, date of current version November 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948946

Robust Speech Steganography Using Differential SVD

YIMING XUE[1], (Member, IEEE), KAI MU[1], YUZHU WANG[1], YAO CHEN[2], PING ZHONG[3], AND JUAN WEN[1]

[1]College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
[2]Beijing Zhongdian Huada Electronic Design Company Ltd., Beijing 102209, China
[3]College of Science, China Agricultural University, Beijing 100083, China

Corresponding author: Yiming Xue (xueym@cau.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61872368 and Grant 61802410.

ABSTRACT The speech signal is different from the typical audio in terms of spectral bandwidth, intensity distribution, and signal continuity, thus how to achieve high imperceptibility and strong robustness for speech steganography is a big challenge. In this paper, we present a speech steganography scheme based on the parity-segmented method and the differential singular value decomposition (SVD). The selected discrete cosine transform (DCT) coefficients are divided into two segments according to parity order. In this way, the energy of the paired segments is approximately equal, therefore the changes in the singular values caused by data embedding are reduced, and high imperceptibility is achieved. Unlike the common SVD-based steganography, the differential SVD scheme can effectively remove the impact of amplitude scaling attack by embedding the secret message into the difference between the singular values. Experimental results show that the proposed method achieves high imperceptibility and strong robustness while resisting the state-of-the-art steganalytic methods.

INDEX TERMS Steganography, differential SVD, paired segments, imperceptibility, amplitude scaling.

I. INTRODUCTION

Steganography is an important way of secure communication via digital cover media such as image, audio and video [1]–[3]. In recent years, steganography has received considerable attention due to the growing necessity for data security, and a lot of steganography methods have been proposed to embed secret message into cover objects and transmit through public channels.

With the rapid development of advanced communication technology, mobile wireless and Voice over IP (VoIP) are widely used around the world, and speech steganography has increasingly high value covering the secure and covert communication. Speech is a special case of audio signals, and it is different from the typical audio signals in terms of spectral bandwidth, intensity distribution, and signal continuity [4]–[6]. In general, methods designed for audio steganography are not suitable for speech steganography because those methods take the media object as continuous signal and do not consider the speech characteristics. In addition, because the spectral bandwidth of speech signal is too narrow, the imperceptibility would be decreased seriously after embedding the secret message. Besides the strict imperceptibility requirement for speech steganography, the speech data would suffer from the typical digital speech processing such as mp3 compression and amplitude scaling, which means robustness should also be carefully studied. Therefore, it is a big challenge to embed the secret message in speech data with high imperceptibility and strong robustness.

The existing speech steganography methods can be generally classified as temporal domain methods and transform domain methods [7], [8]. The temporal domain methods can be further divided into echo-based solutions [9], [10] and least significant bit (LSB) substitution solutions [11]. Generally, the temporal domain steganography methods can achieve high payload capacity with low computational complexity, but they are highly vulnerable to attacks [12]–[14].

The transform domain methods have become more favored by researchers for their good imperceptibility and robustness. Rekik *et al.* [15] put forward a discrete wavelet transformation (DWT) steganography method in which the speech high-frequency components are separated from the low-frequency

The associate editor coordinating the review of this manuscript and approving it for publication was Zhu Han.

components, and the secret data are embedded into the low amplitude and high frequency regions of the cover signal, thus the stego signal is perceptually indistinguishable from the original speech signal. Ahani *et al.* [16] use discrete wavelet transform and sparse decomposition to address the imperceptibility and undetectability in speech steganography, and both stego signal quality and embedding capacity have been improved. In order to design an efficient speech steganographic scheme with robustness to signal processing attacks, Spread Spectrum (SS) and Quantization Index Modulation (QIM) technique are commonly used [17], [18]. The additive and multiplicative data hiding methods based on spread spectrum techniques are proposed [19], [20]. These methods utilize the key-dependent pseudorandom sequence. Compared to the spread spectrum methods, the QIM scheme has capacity advantage for data embedding, and it is easy to be implemented by using various sets of quantizers to embed message [18], but the main weakness of QIM-base scheme is its vulnerability against the amplitude scaling attack. Rational Dither Modulation (RDM) is the improvement to QIM, and the quantization step is derived recursively from the previous vector-norms, thus gain-invariant adaptive quantization step-size can be achieved, which will effectively resist the amplitude scaling attack [21], [22]. However, this method is suitable for audio signals but does not work well for speech data since speech is not the continuous signal.

Recently, Singular value decomposition (SVD) has been applied to speech steganography due to its two important properties [23]: 1). The singular values are determined by the signal energy, which contributes to the stability of the singular values and the data hiding. 2). Slight changing singular values does not affect the quality of the signal. Kanhe and Aghila [24] propose a DCT-SVD-based speech steganography method which embeds the secret message in voiced frames. Dhar and Shimamura [25] propose a DCT-SVD-based scheme using entropy and log-polar transformation (LPT), and data is embedded by quantizing the Cartesian component of highest singular value obtained from the DCT sub band with highest entropy value. Nematollah *et al.* [26] design a scheme which utilizes linear predictive analysis (LPA) with SVD and QIM by applying the ability of LPA for modeling quasi-stationary part of the signal. Although the imperceptibility is improved, these SVD-based steganographic methods are still not robust enough to amplitude scaling attack. Recently, Hwang *et al.* [23] introduce an SVD-QIM-based algorithm that has strong robustness to amplitude scaling attack by utilizing the SVD of stereo audio signals. But this approach can not be applied to data embedding on mono audio and speech signals because the matrix for SVD transformation must be constructed with two channels.

In this paper, instead of modifying the singular value in transform domain directly, we propose a differential SVD steganographic method that can achieve high imperceptibility and resist common attacks, especially for amplitude scaling attack. Firstly, the speech signal is partitioned into speech frames. Secondly, DCT is applied to the speech frame to get the DCT coefficients $X_i(k)$. The DCT coefficients $X_i(k)$ are divided into a pair of segments according to the parity of k. The aim of this segmentation method is to ensure that the paired segments is approximately equal. In this way, the data modification can be decreased and the speech distortion can be reduced accordingly. Finally, the DCT coefficients in each segment are sorted to form a matrix. SVD is employed to decompose the matrices and the difference of the largest singular values from the paired segments are adopted for data embedding. The embedding thresholds are adjusted adaptively to maintain a satisfactory balance between robustness and imperceptibility. Data extraction follows the similar procedure to data embedding. The proposed scheme is verified by the state-of-the-art steganalytic methods and tested with several common attacks. The considerable imperceptibility, security and robustness are achieved.

The main contributions of the paper can be summarized as follows.

1) DCT is applied to the speech frame to get the DCT coefficients, and the selected DCT coefficients are divided into two segments such that the energy of the paired segments is approximately equal.
2) The adaptive embedding thresholds are adopted to improve the robustness against attacks, and the values of the thresholds are adjusted adaptively by the largest singular values of the two segments, respectively.
3) A robust differential SVD steganographic scheme is proposed. The secret message is embedded into the difference of the two largest singular values of the paired DCT coefficients segments.

The remainder of the paper is organized as follows. Section II provides background information related to SVD. Section III presents the embedding and extraction processes of the proposed method. The experimental results are shown in Section IV. Section V concludes the paper.

II. TECHNICAL BACKGROUND

A. SINGULAR VALUE DECOMPOSITION

SVD, a mathematical tool for matrix analysis, has become a generally used means in the information hiding field owing to its unique characteristics. A typical SVD is used to decompose an $m \times n$ matrix A as follows.

$$A = USV^T = U \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N \end{bmatrix} V^T \tag{1}$$

where U is a left singular vector matrix, V is a right singular vector matrix, and they are orthogonal matrices. S is a diagonal matrix with $\sigma_i \geq 0, i = 1, \cdots r$. The diagonal entries of S are called the singular values of A. Singular values are related to the energy of the signal. U and V can be evaluated through the eigenvalue decomposition of AA^T and A^TA as follows:

$$\begin{aligned} AA^T &= US^2U^T \\ A^TA &= VS^2V^T \end{aligned} \tag{2}$$

In addition, S can be evaluated by taking the square root of the eigenvalues to either AA^T or A^TA.

B. RELATIONSHIP BETWEEN SINGULAR VALUES AND SIGNAL ENERGY

Let x be a sequence of speech signal samples. To get the matrix for SVD transformation, we divide the signal samples into N segments signal vectors x_i, where $i = 1, 2, ..., N$.

Then we concatenate vectors to form a matrix

$$A = [x_1, x_2, ..., x_N] \tag{3}$$

Since N vectors are seldom linearly dependent, it is reasonable to assume that the rank of the matrix A is N. We use SVD to decompose the matrix A as

$$A = USV^T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_N u_N v_N^T \tag{4}$$

where u_i and v_i are the *ith* column vectors of U and V, respectively. σ_i is the singular values of A. From the Eq.(2), we know that σ_i can be calculated by

$$|A^TA - \sigma^2 I| = 0 \tag{5}$$

where A^TA can be expressed as

$$\begin{aligned} A^TA &= \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \\ &= \begin{bmatrix} E_1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & E_2 & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & E_N \end{bmatrix} \end{aligned} \tag{6}$$

where $E_i = \|x_i\|^2$ ($i = 1, 2, ..., N$), E_i represents the energy of the segment x_i, $\rho_{ij} = x_i^T x_j$, and $\rho_{ij} = \rho_{ji}$, $i \neq j$.

Eq.(5) can be expressed as

$$|A^TA - \sigma_i^2 I| = \begin{vmatrix} E_1 - \sigma_i^2 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & E_2 - \sigma_i^2 & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & E_N - \sigma_i^2 \end{vmatrix} = 0 \tag{7}$$

where $\sigma_i^2 (i = 1, 2, \cdots, N)$ is the eigenvalue of the square matrix, so we get

$$\sigma_1^2 + \sigma_2^2 + \cdots \sigma_N^2 = E_1 + E_2 + \cdots E_N \tag{8}$$

Based on the Eq.(8), it can be clearly seen that σ_i is determined by the energy of the speech signal.

III. PROPOSED METHOD

In this section, we give further insights into the cause of weakness of the QIM-based method with respect to the amplitude scaling at first, then we propose a differential SVD steganographic scheme in detail.

A. THE PROBLEM OF QIM

In the QIM-based steganography method, the secret message is embedded into the host speech signal via quantizer:

$$y = Q(x; b_k; \Delta) \tag{9}$$

where x is the original speech signal, y is the stego-speech signal, $y = x + b_k$, b_k is the secret message, and Δ is the quantization step size. Based on the Eq.(9), the procedure of embedding can be expressed as

$$y = \begin{cases} \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor \Delta, & b_k = 0 \\ \left\lfloor \frac{x}{\Delta} \right\rfloor \Delta + \frac{\Delta}{2}, & b_k = 1 \end{cases} \tag{10}$$

where $\lfloor \cdot \rfloor$ denotes the floor function.

In the extraction process, a minimum distance decoder is used to extract the secret message $\hat{b}_k$:

$$d_{min} = \arg\min_{b_k} |y - \lfloor Q(y; b_k; \Delta) \rfloor| = |\lambda| \tag{11}$$

$$\hat{b}_k = \begin{cases} 1, & |\lambda| - \frac{1}{2}\Delta \leq \frac{1}{4}\Delta \\ 0, & otherwise \end{cases} \tag{12}$$

where Δ is equal to the quantization step size of embedding process.

Now we assume that z is a stego-signal after amplitude scaling attack, $z = \rho y$, which is equivalent to scaling the output of encoder by ρ. Then, we substitute $z = \rho y$ into Eq.(11), and Eq.(11) can be expressed as

$$d_{min} = \arg\min_{b_k} |\rho y - \lfloor Q(\rho y; b_k; \Delta) \rfloor| = \rho |\lambda| \tag{13}$$

Then the secret message is extracted as

$$\hat{b}_k = \begin{cases} 1, & \rho |\lambda| - \frac{1}{2}\Delta \leq \frac{1}{4}\Delta \\ 0, & otherwise \end{cases} \tag{14}$$

Obviously, the minimum distance is scaled by ρ after the amplitude scaling attack, but the quantization step size at the decoder is not scaled accordingly. It will lead to a mismatch between encoder and decoder, which seriously affects performance in the extraction process. Therefore, the QIM-based watermarking method are generally weak against the amplitude scaling attack.

B. PROPOSED APPROACHES

1) PARITY-SEGMENTED METHOD FOR DCT COEFFICIENTS

First, we split the speech data into N_s frames and the DCT coefficients of the speech frames are calculated. Let $x(n)$ be the speech signal of the current frame with length N. The DCT coefficients are denoted by $X(k)$, and it can be expressed as

$$X(k) = w(k) \sum_{n=0}^{N-1} x(n) \cos\left(\frac{\pi (2n+1) k}{2N} \right), \tag{15}$$

where $k = 0, 1,...,N-1$, and

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 0 \\ \sqrt{\frac{2}{N}}, & 1 \le k \le N-1 \end{cases} \tag{16}$$

Since the low and high frequency components are vulnerable to the attacks such as filtering and compression, we only select DCT coefficients $X(k)$ corresponding to a certain frequency range $[f_l, f_h]$, where f_l and f_h are determined experimentally.

Then, the selected DCT coefficients of the *ith* frame are denoted as

$$X_i(k) = [X_i(0), X_i(1), \cdots, X_i(2M-1)], \quad i = 1, 2, ..., N_s \tag{17}$$

where $2M$ is the number of selected DCT coefficients.

Furthermore, the $X_i(k)$ is partitioned into the two segments. To achieve the high imperceptibility, the proposed embedding algorithm requires the energy of the paired segments approximately equal. In general, the energy varies significantly if two segments are partitioned according to the sequence order of the DCT coefficients (We name this method as sequence-segmented method). As the energy contour becomes smoother with the samples increasing [27], we would adopt a parity-segmented method as

$$\begin{cases} X_{i,1}(k) = [X_i(0), X_i(2), \cdots, X_i(2M-2)] \\ X_{i,2}(k) = [X_i(1), X_i(3), \cdots, X_i(2M-1)] \end{cases} \tag{18}$$

where $X_{i,1}(k)$ is the first segment of $X_i(k)$, and $X_{i,2}(k)$ is the second segment of $X_i(k)$.

To evaluate the energy difference between the paired segments, we compare the different segmentation strategies: parity-segmented method and sequence-segmented method. The distribution of energy difference between the paired segments on TIMIT database is shown in Fig.1. It can be seen that the energy difference using the parity-segmented method is relatively small. From the analysis in section II-B, singular values are determined by the energy of the speech signal, thus the largest singular values of the paired segments are approximately equal using the parity-segmented method, which will be helpful in achieving high imperceptibility performance.

2) DIFFERENTIAL SVD SCHEME

To overcome the limitation of QIM-based method, we propose a differential SVD steganographic scheme.

The DCT coefficients in the two segments are sorted to form a matrix, respectively. Then we use the Eq.(1) to decompose the each matrix, and select the two largest singular values of segments to embed secret message. The largest singular values of the first and second segment are called σ_{11} and σ_{21}, respectively. The procedure of embedding can be expressed as

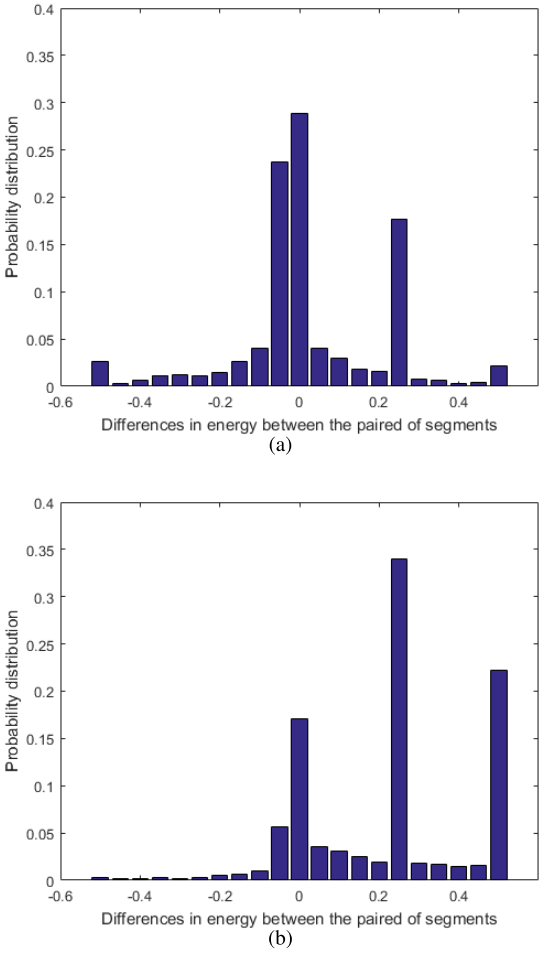


FIGURE 1. Distribution of energy difference between the paired segments; (a) parity-segmented method; (b) sequence-segmented method.

Embedding of message bit "0":

$$\hat{\sigma}_{11} = \begin{cases} \sigma_{11}, & if\,(\sigma_{11} - \sigma_{21}) \ge T_{bu,1} \\ \sigma_{21} + T_{bu,1}, & else \end{cases} \tag{19}$$

$$\hat{\sigma}_{21} = \sigma_{21} \tag{20}$$

Embedding of message bit "1":

$$\hat{\sigma}_{11} = \sigma_{11} \tag{21}$$

$$\hat{\sigma}_{21} = \begin{cases} \sigma_{21}, & if\,(\sigma_{21} - \sigma_{11}) \ge T_{bu,2} \\ \sigma_{21} + T_{bu,2}, & else \end{cases} \tag{22}$$

where $\hat{\sigma}_{11}$ and $\hat{\sigma}_{21}$ are the two largest singular values after embedding data. $T_{bu,1} = \alpha_1 * \sigma_{21}$, $T_{bu,2} = \alpha_2 * \sigma_{11}$, and they represent the embedding strength.

Because σ_{11} and σ_{21} are approximately equal with the parity-segmented method, the changes caused by data embedding are relatively small. Thus, the higher imperceptibility can be achieved.

In the extraction process, we get the singular values according to Eq.(1), then the difference between the two largest

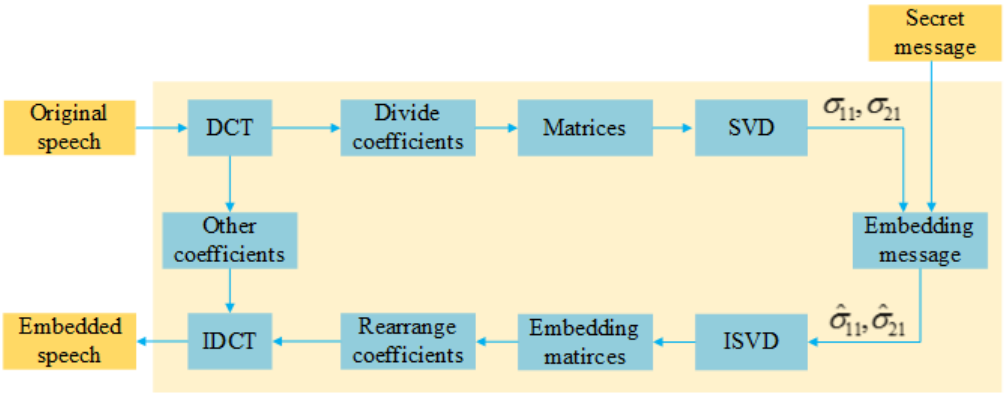


FIGURE 2. Embedding process of the proposed method.

singular values $\hat{\sigma}_{11}$ and $\hat{\sigma}_{21}$ can be calculated as

$$d_j = \hat{\sigma}_{11} - \hat{\sigma}_{21} \tag{23}$$

The secret message extraction decision criterion is as follows.

if $d_j = T_{bu,1} > 0$, the message bit $\hat{b}_k = 0$.

if $d_j = -T_{bu,2} < 0$, the message bit $\hat{b}_k = 1$.

According to the extraction decision criterion, to successfully extract message bit "0" (or message bit "1"), we simply demand $d_j \geq 0$ (or $d_j < 0$). Now we assume that z is the stego DCT coefficient vector after amplitude scaling attack, $z = \rho y$. We sort z to form an $m \times n$ matrix A'.

$$A' = \rho A \tag{24}$$

We substitute A' into Eq.(1), and Eq.(1) can be expressed as

$$A' = \rho A = US'V^T = U \begin{bmatrix} \rho\hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho\hat{\sigma}_N \end{bmatrix} V^T \tag{25}$$

Then we use the two largest singular values $\rho\hat{\sigma}_{11}$ and $\rho\hat{\sigma}_{21}$ to extract the secret message.

$$d'_j = \rho\hat{\sigma}_{11} - \rho\hat{\sigma}_{21} = \rho d_j. \tag{26}$$

According to the Eq.(26), the sign of d'_j is the same as that of d_j, so the proposed differential method is resistant to the amplitude scaling attack.

Besides, the proposed method can also reduce the impact of additive interference. As the embedding rule mentioned above, we have $d_j \geq T_{bu,1}$ if we embed the message bit "0". Similarly, we have $d_j \leq -T_{bu,2}$ if we embed the message bit "1". Because of the impact of additive interference, $\hat{\sigma}_{11}$ becomes to $\hat{\sigma}_{11} + \varepsilon_1$, $\hat{\sigma}_{21}$ becomes to $\hat{\sigma}_{21} + \varepsilon_2$, and d_j becomes to $d_j + \varepsilon$, where distortion $\varepsilon = \varepsilon_1 - \varepsilon_2$. Generally, ε is less than the minimum of ε_1 and ε_2. As long as the value of $abs(\varepsilon) < T_{bu,1}$ or $T_{bu,2}$, the embedded message bit "0" or "1" can be extracted correctly. What's more, $T_{bu,1}$ and $T_{bu,2}$ are adaptively determined by the two largest singular values, which can further reduce the influence of attacks. Therefore, the proposed differential SVD scheme can reduce the impact of additive interference.

C. EMBEDDING AND EXTRACTION PROCESS

1) EMBEDDING PROCESS

The original speech signal is divided into frames firstly, then we apply DCT to the frame, and the DCT coefficients are divided into a pair of segments using parity-segmented method. We sort DCT coefficients of each segment to form a matrix, and SVD is applied to decompose each matrix. Fig.2 illustrates the embedding process based on DCT and SVD. The detailed procedure of data embedding is as follows.

Step 1: Split the original speech data into multiple frames, and $x(n)$ is the signal of one frame with length N.

Step 2: Apply DCT to $x(n)$ by Eq.(15).

Step 3: Select those DCT coefficients $X_i(k)$ in a certain frequency range $[f_l, f_h]$.

Step 4: Divide $X_i(k)$ into a pair of segments $X_{i,1}(n)$ and $X_{i,2}(n)$ by Eq.(18).

Step 5: Sort $X_{i,1}(n)$ and $X_{i,2}(n)$ to form a matrix, respectively. SVD is applied to decompose each matrix by Eq.(1).

Step 6: Select the two largest singular values σ_{11} and σ_{21} to embed secret message b_k as Eq.(19), Eq.(20), Eq.(21), Eq.(22)

Step 7: Use $\hat{\sigma}_{11}$ and $\hat{\sigma}_{21}$ to replace σ_{11} and σ_{21}, then apply inverse SVD to compose the new matrices and get $X'_{i,1}(n)$ and $X'_{i,2}(n)$.

Step 8: Rearrange $X'_{i,1}(n)$ and $X'_{i,2}(n)$ back to their original positions and reconstruct the stego-speech frame $x'(n)$ by inverse DCT.

2) EXTRACTION PROCESS

Secret message extraction follows the similar procedure to data embedding. The detailed procedure of data extraction is shown as follows.

Step 1: Split the stego speech data into multiple frames, and $x'(n)$ is the signal of one frame with length N.

Step 2: Apply DCT to $x'(n)$ by Eq.(15).

Step 3: Select those DCT coefficients $X'_i(k)$ in a certain frequency range $[f_l, f_h]$.

Step 4: Divide $X'_i(k)$ into a pair of segments $X'_{i,1}(n)$ and $X'_{i,2}(n)$ by Eq.(18).

Step 5: Sort $X'_{i,1}(n)$ and $X'_{i,2}(n)$ to form a matrix, respectively. Then SVD is applied to decompose each matrix by Eq.(1).

Step 6: Extract the secret message bit $\hat{b}_k$ from the two largest singular values $\tilde{\sigma}_{11}$ and $\tilde{\sigma}_{21}$ as follows.

Calculate $d_j = \tilde{\sigma}_{11} - \tilde{\sigma}_{12}$. if $d_j > 0$, $\hat{b}_k = 0$. Otherwise, $\hat{b}_k = 1$.

IV. EXPERIMENTS

In this section, we implement the methods in [20], [22], [25], [26], to compare the performance with the proposed method. These methods cover the representative methods for speech steganography: SS, RDM, SVD algorithms.

The performance of the proposed algorithm is evaluated on TIMIT database. The TIMIT is a well-known speech database which contains 6300 utterances recorded from 630 adult speakers of eight major dialects of American English, each read ten sentences. The length of the speech file varies between 1.4 and 5.04s [28].The sentences are originally sampled at 16kHz and down-sampled to 8 kHz.

Through our experiments and analysis, we set $f_l = 1.5\ kHz, f_h = 2.5\ kHz, \alpha_1 = \alpha_2 = 0.2$. To ensure fairness in the evaluation of performance, the embedding rates of methods in [22], [25], [26] and the proposed method are identical at 25 bps, while the embedding rate of method in [20] is 12.5 bps for facilitating the use of spread spectrum in steganography. And we compare the performance from the imperceptibility, robustness, and security.

A. IMPERCEPTIBILITY

Imperceptibility or inaudibility means that the secret message embedded into the host signal is inaudible. In this section, we used various metrics to assess the quality of the stego-speech.

The first metric is the signal-to-noise ratio (SNR) defined as

$$SNR = 10\log\left(\frac{\sum_{n=1}^{K} x^2(n)}{\sum_{n=1}^{K}\left(x(n) - x'(n)\right)^2}\right) \tag{27}$$

where $x(n)$ and $x'(n)$ are the original and stego speech in time domain respectively, and K is the length of the signal.

The SNR test can only give a general evaluation without taking into account the specific characteristics of the human auditory system. Thus, the test is further conducted using the perceptual evaluation of speech quality (PESQ). The PESQ is the effective method for objective speech quality assessment described in ITU-T Recommendation P.862, and a cognition model is adopted to predict the perceived speech quality of the degraded speech signal. The PESQ assesses speech quality on a range of [−0.5,4.5], which means from annoyance to imperceptibility.

We also employ one of the most popular methods called mean opinion score (MOS) as a better measurement of imperceptibility based on human perception. Ten subjects are asked to classify the difference between the original and the stego speech in terms of 5-points.

TABLE 1. Comparison of SNR, PESQ and MOS.

Algorithm	SNR	PESQ	MOS
Method in [20]	23.1418	2.9234	4.45
Method in [22]	-4.6986	3.0888	4.07
Method in [25]	5.8904	2.3055	4.15
Method in [26]	23.4663	3.1060	4.61
Proposed	24.3813	3.5681	4.90

Table 1 shows the values of different metrics for speech signal resulting from the proposed method and the other compared methods. The randomly generated secret message is embedded into the files of the TIMIT database, and then the average SNR, PESQ and MOS values of the all stego files obtained under 25bps are listed. We can see that our method achieves the largest SNR value among these five methods. It should be noted that the coefficients are modified to embed synchronous signal repetitively in [22], which results in the obvious difference between the original speech and stego speech, thus the negative SNR value occurs. In addition, we also observe that the average PESQ and MOS values of our method are 3.5681 and 4.9 respectively, which indicates that original and stego speech are perceptually indistinguishable.

Fig.3 presents the time waveform of the original, stego speech and the differences between them, and it can be seen that the differences are quite small using the proposed method. Fig.4 shows that the spectrogram of the original and stego speech signal, and there are no obvious changes.

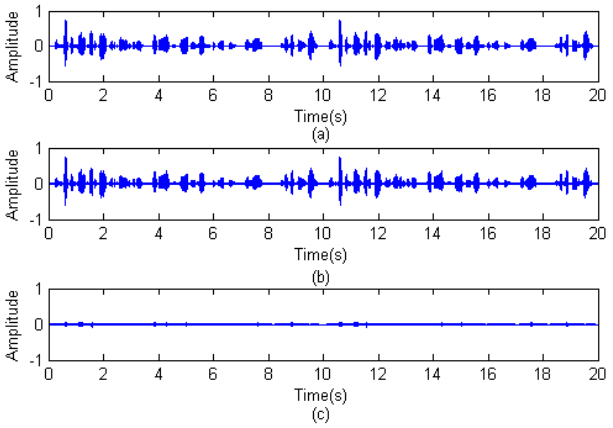
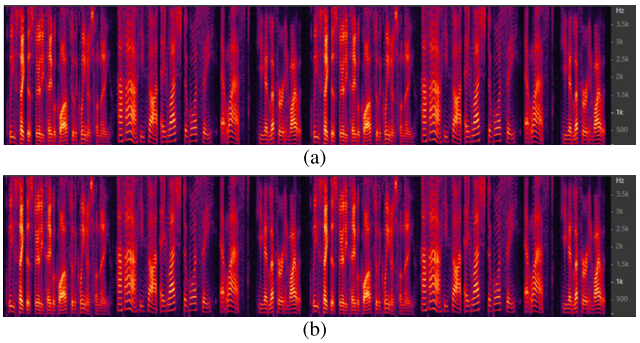


FIGURE 3. Waveform of the original, stego speech and difference between them.

In order to further verify the imperceptibility performance of the proposed scheme, we attempt to compare SNR and PESQ under the different embedding rates (range from 12.5 to 33.3bps). The proposed method and the other four steganography methods embed the secrete message into signal files of the TIMIT database, and then the average SNR and PESQ values of the all stego files obtained at each embedding rate are shown in Fig.5 and Fig.6, respectively. As shown in Fig.5 and Fig.6, both SNR and PESQ of the proposed method are higher than the other methods, which



(a) (b)

FIGURE 4. Spectrogram of the original and stego speech using the proposed method; (a) original speech; (b) stego speech.

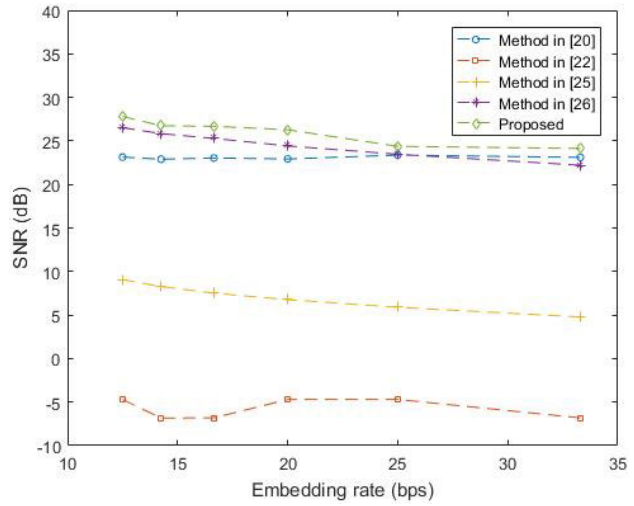


FIGURE 5. SNR values under different embedding rates.

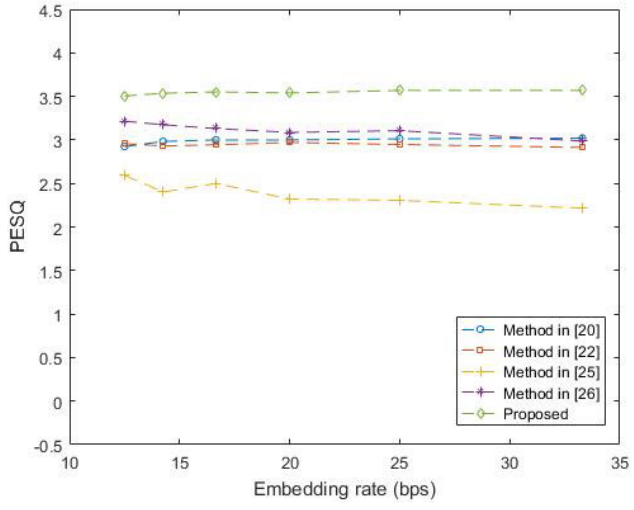


FIGURE 6. PESQ values under different embedding rates.

indicates that it outperforms the other methods in terms of imperceptibility.

B. ROBUSTNESS

Robustness is a measure of secret messages against attempts to eliminate or corrupt it, intentionally or unintentionally, by different kinds of digital signal processing. For the evaluation of robustness, we examine the bit error rates (BER) between the original secret message and the extracted secret message. BER is defined as

$$BER = \frac{B_{ERR}}{N} \times 100\% \tag{28}$$

where B_{ERR} is the number of erroneous bits and N is the total number of bits.

In the following experiments, we attack the stego speech signal separately using the following typical signal processing manipulations:

1) Amplitude scaling: the amplitudes of the stego speech are rescaled by ±30%;
2) Re-sampling attack: the stego speech is up-sampled to 16 kHz and then down-sampled back to 8 kHz;
3) Low-pass filtering (LPF): low-pass filter is applied to the stego speech, where the cut-off frequency is 3.5 kHz;
4) High-pass filtering (HPF): high-pass filter is applied to the stego speech, where the cut-off frequency is 500 Hz;
5) MP3 compression: MPEG-1 Layer-III compression is applied to the stego speech signal, where the compression bit rates are 128 kbps and 96 kbps;
6) Noise addition: random noise is added to the stego speech signal, where the signal-to-noise rate (SNR) is 30 dB.

Table 2 shows the BER of five steganographic methods after these different attacks. Methods in [22], [25] and [26] provide little resistance to HPF attack. Method in [25] and method in [26] fail severely in the amplitude scaling attack, which are based on SVD-LSB and SVD-QIM technique respectively. Although method in [22] can improve the robustness to the amplitude attack by using RDM, but it is still inferior to our method while embedding data into the non-voiced, salient frames in the speech signal. In general, the SS-based embedding method in [20] performs well on the varies of attacks but the embedding rate is relatively low. It can be seen that the proposed method provides stronger robustness against common signal processing attacks compared with the other four methods, especially for the amplitude scaling attack. These results verify the good robust performance of the differential SVD scheme.

C. SECURITY

The steganographic security (statistical undetectability) is the import criteria to evaluate the steganographic systems. In the experiments, four steganalysis methods [29], [30] are used to test the security of the proposed method. Accordingly, four kinds of features are extracted for steganalysis, which include derivative-based high-frequency spectrum (DHS), derivative-based mel-cepstrum (DMC), wavelet-based mel-cepstrum (WMC) and reversed-Mel energy (RME). We use ensemble classifier to identify stego speech according to the different features. Then, we randomly select 1000 cover files and 1000 stego files as the training set and the remaining

TABLE 2. Comparison robustness (BER) among different methods.

Attack	Method in [20]	Method in [22]	Method in [25]	Method in [26]	Proposed
Amplitude(0.7)	0.0063	0.0389	0.0647	0.3802	0
Amplitude(1.3)	0.0063	0.0389	0.5432	0.3684	0
Re-sampling	0.0132	0.0379	0.0148	0.1145	0.0032
LPF(3.5kHz)	0.0135	0.0387	0.0257	0.0913	0.0030
HPF(500Hz)	0.0270	0.3086	0.5060	0.3533	0.0056
MP3(128kbps)	0.0064	0.0386	0.0243	0.0095	0
MP3(96kbps)	0.0064	0.0386	0.0251	0.0095	0
Noise(30dB)	0.0567	0.0830	0.0136	0.0113	0.1473

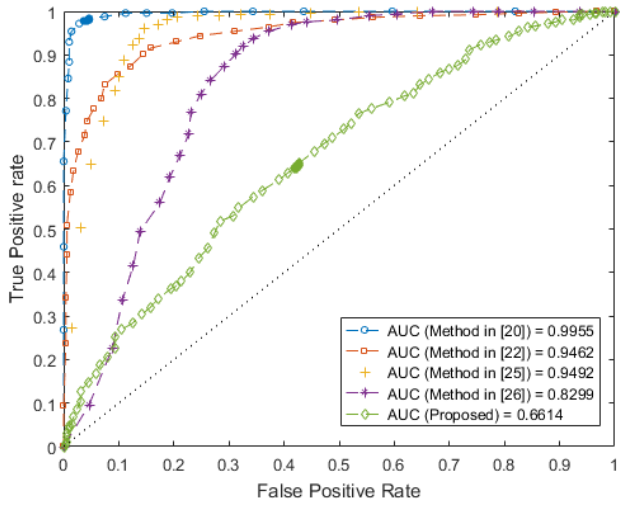


FIGURE 7. ROC curve of the DHS steganalysis method.

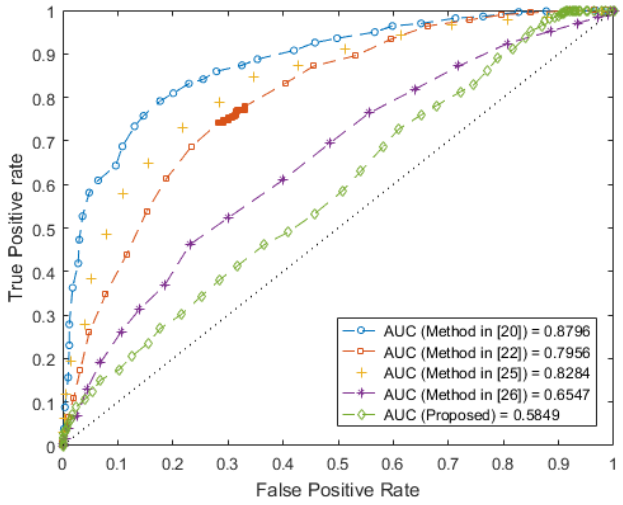


FIGURE 8. ROC curve of the DMC steganalysis method.

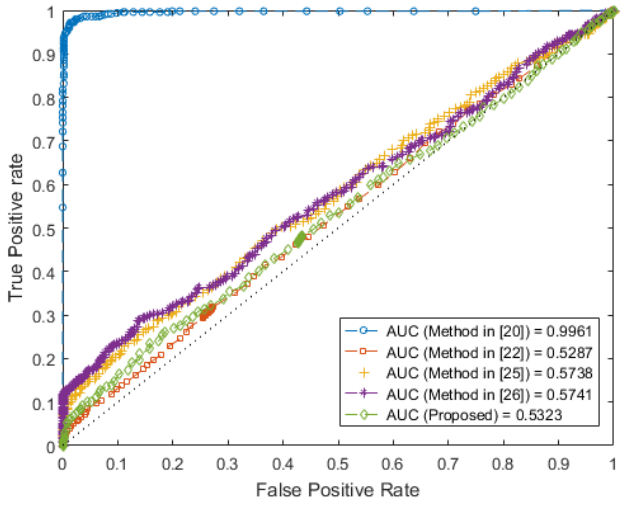


FIGURE 9. ROC curve of the WMC steganalysis method.

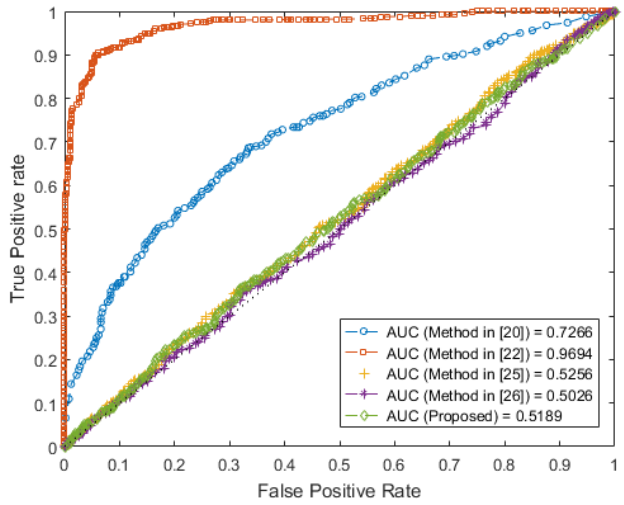


FIGURE 10. ROC curve of the RME steganalysis method.

as the testing set to obtain the ROC (Receiver Operating Characteristic) curves.

As a general assessment of the steganalysis results, a larger area under the ROC curve (AUC) indicates a higher detection accuracy of the steganalyzer. Conversely, a ROC curve closer to the bisector line shows less detectability of stego signal by the steganalyzer. Fig. 7, 8, 9, and 10 vividly show the comparison of ROC curves between five different steganography methods.

We further test the undetectability performance at different embedding rates (range from 12.5 to 33.3bps), and the detection errors of five different steganography methods against four steganalysis methods are reported in Table 3.

From the steganalysis results, we can see only the proposed steganographic method can effectively resist the steganalysis methods based on DHS and DMC features, and the other

TABLE 3. Detection errors at different embedding rate.

Embedding rate	Steganalysis method	Method in [20]	Method in [22]	Method in [25]	Method in [26]	Proposed
12.5	DHS	0.0407±0.0086	0.2320±0.0190	0.1311±0.0150	0.2335±0.0119	0.3615±0.0139
	DMC	0.2047±0.0167	0.3924±0.0130	0.2772±0.0207	0.3605±0.0138	0.4561±0.0063
	WMC	0.0267±0.0052	0.4983±0.0045	0.4774±0.0064	0.4728±0.0075	0.4677±0.0065
	RME	0.3391±0.0101	0.4165±0.0131	0.4922±0.0030	0.4964±0.0055	0.4855±0.0038
14.25	DHS	0.0367±0.0071	0.1536±0.0181	0.1446±0.0093	0.2312±0.0073	0.3783±0.0132
	DMC	0.2275±0.0133	0.3846±0.0191	0.2752±0.0151	0.3652±0.0148	0.4582±0.0064
	WMC	0.0176±0.0048	0.4977±0.0037	0.4824±0.0052	0.4553±0.0065	0.4692±0.0058
	RME	0.3297±0.0061	0.4207±0.0064	0.4909±0.0054	0.4963±0.0044	0.4846±0.0067
16.65	DHS	0.0354±0.0073	0.1070±0.0066	0.1353±0.0102	0.1973±0.0109	0.3908±0.0073
	DMC	0.2374±0.0148	0.3254±0.0134	0.2676±0.0152	0.3560±0.0219	0.4539±0.0088
	WMC	0.0128±0.0020	0.4972±0.0046	0.4799±0.0097	0.4660±0.0059	0.4711±0.0024
	RME	0.2425±0.0080	0.2315±0.0380	0.4872±0.0064	0.4975±0.0048	0.4805±0.0041
20	DHS	0.0348±0.0062	0.1253±0.0113	0.1231±0.0135	0.2037±0.0104	0.3842±0.0103
	DMC	0.2279±0.0112	0.2940±0.0183	0.2717±0.0368	0.3650±0.0154	0.4568±0.0052
	WMC	0.0200±0.0045	0.4883±0.0051	0.4631±0.0058	0.4524±0.0044	0.4662±0.0063
	RME	0.1067±0.0061	0.0604±0.0077	0.4590±0.0063	0.4998±0.0040	0.4844±0.0061
25	DHS	0.0362±0.0070	0.1398±0.0167	0.0857±0.0088	0.2277±0.0192	0.3854±0.0093
	DMC	0.2356±0.0246	0.2956±0.0224	0.2436±0.0151	0.3613±0.0138	0.4594±0.0054
	WMC	0.0286±0.0050	0.4843±0.0025	0.4645±0.0101	0.4593±0.0065	0.4754±0.0038
	RME	0.3157±0.0115	0.0789±0.0095	0.4830±0.0055	0.4960±0.0066	0.4813±0.0086
33.3	DHS	0.0208±0.0040	0.1015±0.0069	0.0863±0.0132	0.1793±0.0109	0.3868±0.0222
	DMC	0.2469±0.0114	0.2838±0.0106	0.2173±0.0140	0.3394±0.0109	0.4607±0.0066
	WMC	0.0205±0.0031	0.4979±0.0067	0.4410±0.0101	0.4563±0.0090	0.4768±0.0055
	RME	0.3409±0.0085	0.0371±0.0067	0.4396±0.0079	0.4968±0.0044	0.4814±0.0061

four steganographic methods can be detected easily. The WMC steganalyzer is not able to detect the stego speech generated by the different kinds of steganographic methods except for the method in [20]. Regarding the RME steganalyzer, SVD based steganographic methods including the proposed method, methods in [25] and [26] perform well in anti-detecting when compared to the SS method [20] and RDM method [22]. It can be observed that only the proposed method can resist the above four steganalysis methods, so we conclude that the proposed method can achieve higher security than the other four steganographic methods.

V. CONCLUSION

In this paper, we have proposed a robust speech steganographic method that utilizes the characteristics of the differential SVD. We adopt DCT transform and divide the DCT coefficients into a pair of segments. The two segments are further split with equal energy approximately, and the changes in the singular values caused by data embedding are reduced. The difference of the two largest singular values is modified to embed the secret message, and the adaptive embedding thresholds are determined by the two largest singular values. The experimental results show that compared with existing methods, the proposed method achieves higher imperceptibility, stronger robustness and better security. In the near future, the embedding capacity and security performance will be optimized, and the synchronization mechanism will be explored to improve the robustness further.

REFERENCES

[1] A. Cheddad, J. Condell, K. Curran, and P. M. Kevitt, "Digital image steganography: Survey and analysis of current methods," *Signal Process.*, vol. 90, no. 3, pp. 727–752, Mar. 2010.

[2] K. Gopalan, "Audio steganography by cepstrum modification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, Mar. 2005, pp. v-481–v484.

[3] R. J. Mstafa, K. M. Elleithy, and E. Abdelfattah, "A robust and secure video steganography method in DWT-DCT domains based on multiple object tracking and ECC," *IEEE Access*, vol. 5, pp. 5354–5365, 2017.

[4] A. Sehgal and N. Kehtarnavaz, "A convolutional neural network smartphone app for real-time voice activity detection," *IEEE Access*, vol. 6, pp. 9017–9026, 2018.

[5] H. T. Hu, S.-J. Lin, and L.-Y. Hsu, "Effective blind speech watermarking via adaptive mean modulation and package synchronization in DWT domain," *EURASIP J. Audio Speech Music Process.*, vol. 2017, no. 1, 2017, Art. no. 10.

[6] H.-T. Hu and T.-T. Lee, "Frame-synchronized blind speech watermarking via improved adaptive mean modulation and perceptual-based additive modulation in DWT domain," *Digit. Signal Process.*, vol. 87, pp. 75–85, Apr. 2019.

[7] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, "Techniques for data hiding," *IBM Syst. J.*, vol. 35, nos. 3–4, pp. 313–336, 1996.

[8] Y. Huang, C. Liu, S. Tang, and S. Bai, "Steganography integration into a low-bit rate speech codec," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1865–1875, Dec. 2012.

[9] 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.

[10] B. Tabara, J. Wojtuń, and Z. Piotrowski, "Data hiding method in speech using echo embedding and voicing correction," in *Proc. Signal Process. Symp. (SPSympo)*, Sep. 2017, pp. 1–6.

[11] J. Liu, K. Zhou, and H. Tian, "Least-significant-digit steganography in low bitrate speech," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 1133–1137.

[12] M. A. Nematollahi and S. A. R. Al-Haddad, "An overview of digital speech watermarking," *Int. J. Speech Technol.*, vol. 16, no. 4, pp. 471–488, 2013.

[13] K. Gopalan and Q. Shi, "Audio steganography using bit modification—A tradeoff on perceptibility and data robustness for large payload audio embedding," in *Proc. Int. Conf. Comput. Commun. Netw.*, Aug. 2010, pp. 1–6.

[14] F. Djebbar, B. Ayad, K. A. Meraim, and H. Hamam, "Comparative study of digital audio steganography techniques," *EURASIP J. Audio Speech Music Process.*, vol. 2012, no. 1, 2012, Art. no. 25.

[15] S. Rekik, D. Guerchi, S.-A. Selouani, and H. Hamam, "Speech steganography using wavelet and Fourier transforms," *EURASIP J. Audio Speech Music Process.*, vol. 2012, no. 1, 2012, Art. no. 20.
[16] S. Ahani, S. Ghaemmaghami, and Z. J. Wang, "A sparse representation-based wavelet domain speech steganography method," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 80–91, Jan. 2015.
[17] Q. Cheng and J. Sorensen, "Spread spectrum signaling for speech watermarking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, May 2001, pp. 1337–1340.
[18] B. Chen and G. W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.
[19] R. Kazemi, F. Pérez-González, M. A. Akhaee, and F. Behnia, "Data hiding robust to mobile communication vocoders," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2345–2357, Dec. 2016.
[20] R. Kazemi, R. Rezaei, M. A. Akhaee, and F. Behnia, "Covert communications through mobile voice channels," *IET Inf. Secur.*, vol. 10, no. 3, pp. 156–164, 2016.
[21] F. Pérez-González, C. Mosquera, M. Barni, and A. Abrardo, "Rational dither modulation: A high-rate data-hiding method invariant to gain attacks," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3960–3975, Oct. 2005.
[22] H.-T. Hu and J.-R. Chang, "Efficient and robust frame-synchronized blind audio watermarking by featuring multilevel DWT and DCT," *Cluster Comput.*, vol. 20, no. 1, pp. 805–816, 2017.
[23] M.-J. Hwang, J. Lee, M. Lee, and H.-G. Kang, "SVD-based adaptive QIM watermarking on stereo audio signals," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 45–54, Jan. 2017.
[24] A. Kanhe and G. Aghila, "A DCT–SVD-based speech steganography in voiced frames," *Circuits Syst. Signal Process.*, vol. 37, no. 11, pp. 5049–5068, 2018.
[25] P. K. Dhar and T. Shimamura, "Blind SVD-based audio watermarking using entropy and log-polar transformation," *J. Inf. Secur. Appl.*, vol. 20, pp. 74–83, Feb. 2015.
[26] M. A. Nematollahi, C. Vorakulpipat, H. Gamboa-Rosales, F. J. Martinez-Ruiz, and I. Jose, "Digital speech watermarking based on linear predictive analysis and singular value decomposition," *Proc. Nat. Acad. Sci., India A, Phys. Sci.*, vol. 87, no. 3, pp. 433–446, 2017.
[27] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Signal Processing*, vol. 64. Upper Saddle River, NJ, USA, Pearson, 2011.
[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic–Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
[29] 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.
[30] H. Ghasemzadeh and M. Arjmandi, "Universal audio steganalysis based on calibration and reversed frequency resolution of human auditory system," *IET Signal Process.*, vol. 11, no. 8, pp. 916–922, 2017.

YIMING XUE is currently an Associate Professor with the College of Information and Electrical Engineering, China Agricultural University. His research interests include multimedia processing, multimedia security, and VLSI design.

KAI MU is currently pursuing the M.E. degree in computer science and technology with the College of Information and Electrical Engineering, China Agricultural University. His research interest includes network and information security.

YUZHU WANG is currently pursuing the M.E. degree in computer technology with the College of Information and Electrical Engineering, China Agricultural University. Her research interest includes information hiding.

YAO CHEN received the M.E. degree in signal and information processing from the College of Information and Electrical Engineering, China Agricultural University. He is currently an Engineer with Beijing Zhongdian Huada Electronic Design Company Ltd. His research interests include information hiding and VLSI design.

PING ZHONG received the Ph.D. degree from China Agricultural University. She is currently a Professor with the College of Science, China Agricultural University. Her research interests include machine learning and information hiding.

JUAN WEN received the B.E. degree in information engineering and Ph.D. degree in signal and information processing from the Beijing University of Post and Telecommunication. She is currently a Lecturer with the College of Information and Electrical Engineering, China Agricultural University. Her research interests include artificial intelligence, information hiding, and natural language processing.