Frontiers of Information Technology & Electronic Engineering
www.zju.edu.cn/jzus; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A subband excitation substitute based scheme for narrowband speech watermarking*

Wei LIU, Ai-qun HU‡

(*School of Information Science and Engineering, Southeast University, Nanjing 210096, China*)

E-mail: weiliu@seu.edu.cn; aqhu@seu.edu.cn

Received Aug. 26, 2016; Revision accepted Mar. 3, 2017; Crosschecked Apr. 27, 2017

Abstract: We propose a new narrowband speech watermarking scheme by replacing part of the speech with a scaled and spectrally shaped hidden signal. Theoretically, it is proved that if a small amount of host speech is modified, then not only an ideal channel model for hidden communication can be established, but also high imperceptibility and good intelligibility can be achieved. Furthermore, a practical system implementation is proposed. At the embedder, the power normalization criterion is first imposed on a passband watermark signal by forcing its power level to be the same as the original passband excitation of the cover speech, and a synthesis filter is then used to spectrally shape the scaled watermark signal. At the extractor, a bandpass filter is first used to get rid of the out-of-band signal, and an analysis filter is then employed to compensate for the distortion introduced by the synthesis filter. Experimental results show that the data rate is as high as 400 bits/s with better bandwidth efficiency, and good imperceptibility is achieved. Moreover, this method is robust against various attacks existing in real applications.

Key words: Analysis filter; Linear prediction; Narrowband speech watermarking; Passband excitation replacement; Power normalization; Spectral envelope shaping; Synthesis filter

http://dx.doi.org/10.1631/FITEE.1601503 **CLC number:** TP309.2

1 Introduction

Digital watermarking is a technique for manipulating a work or data stream to deliberately embed an extra message that can hardly or impossibly be perceived by a human. The work or data stream to be modified is called the 'host signal' or simply the 'cover', and might be any multimedia signal such as a speech, audio, image, or video signal. The additional information to be embedded is named 'watermark', and the embedding is usually expected to result in minimal perceptual deterioration to the cover. Robustness is referred to as the ability of a designed watermarking scheme to reach a low bit error rate (BER) after detecting the channel-attacked watermarked signal. Watermark capacity is the achievable embedding bit rate under the constraints of the cover, the permissible perceptual distortion to the cover, and the robustness against various channel attacks. In most watermarking methods there is a tradeoff among capacity, imperceptibility, and robustness. In this paper, the case of blind watermarking in the narrowband speech is considered, where the original cover is not known to the watermark detector (Nematollahi and Al-Haddad, 2013).

Early applications of digital speech watermarking are usually for security concerns, such as speaker identification and verification (Faundez-Zanuy *et al.*, 2006; 2007; Nematollahi *et al.*, 2015a; 2015b), speech authentication (Park *et al.*, 2007), and speech forensic analysis (Faundez-Zanuy *et al.*, 2010). Recent applications, however, have been advanced to much broader areas, such as speech quality evaluation (Cai *et al.*, 2007), speech recovery (Sarreshtedari *et al.*, 2015), speech bandwidth extension (Chen *et al.*, 2013), speech enhancement (Chen *et al.*, 2007), and

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61571110)

ORCID: Wei LIU, http://orcid.org/0000-0002-7930-1943

legacy speech communication system enhancement (Hofbauer *et al.*, 2009).

This paper presents a novel watermarking algorithm that is aimed at systems that use narrowband speech with a frequency range of 0 to 4 kHz. In contrast to wideband speech (0–7 kHz), narrowband speech has been widely used in systems such as the public switched telephone network (PSTN), military or aeronautical voice radio, the voice over Internet protocol (VoIP), and the cellular infrastructure such as global system for mobiles (GSM) and code division multiple access (CDMA) based networks. In this research, the aim is to design a narrowband speech watermarking system that can work in or between these systems. Therefore, the transmission channel attacks are specified by these applications and can be modeled as additive white Gaussian noise (AWGN), flat-fading, re-quantization, compression, transcoding, and bandpass filtering attacks. It should also be noted that the designed watermarking system is usually expected to provide a high capacity for embedding the message and maintain a good perceptual fidelity even under these attacks.

In current speech watermarking algorithms, there are three main steps to generate a watermarked speech. The first step is to preprocess the host speech. Commonly, three frequently used methods are employed. The first one is to keep the time domain speech signal unchanged during embedding. The next one is to first transform the speech into other domains and then embed the watermark to the speech represented in those domains. Typical transformations include discrete Fourier transform (DFT) (Nematollahi *et al.*, 2017), discrete cosine transform (DCT) (Eslami *et al.*, 2006), discrete Hadamard transform (DHT) (Chen *et al.*, 2013), and wavelet (Nematollahi *et al.*, 2015a; 2015b). The third one uses the analysis-by-synthesis (AbS) technique based on the canonical linear prediction (LP) theory. By analyzing the speech before embedding, the speech can be decomposed into the excitation component and the vocal tract component. Both of them are available for embedding the watermark. For example, the excitation signal of the unvoiced speech can be replaced by the watermark (Hofbauer *et al.*, 2009). In another example, by using the property that the vocal tract can be expressed as either an all-pole model with predictor coefficients or the equivalent line spectrum pairs (LSPs) which represent the speech formants, the watermark can be embedded into the speech by adjusting locations, peaks, and valleys of the formants (Wang and Unoki, 2015).

The second step is referred to as mapping of the watermark bits to the watermark signal, for which human perception characteristics are often used. By doing so, minimal perceptual distortion to the speech (Cheng and Sorensen, 2001) can be obtained, while the communication capability is maintained.

The third step is concerned with the approaches to combine watermark and cover speech to form the watermarked signal. There are two types of commonly used embedding methods. One is to directly add the watermark to the cover, including the least significant bit (LSB) method (Chen *et al.*, 2013) and spread spectrum method (Chen and Leung, 2006). The other is to modulate the speech signal or one of its parameters according to the watermark, including the quantization index modulation (QIM) method (Nematollahi *et al.*, 2015a; 2015b) and its more general version (Hofbauer *et al.*, 2009).

The above-mentioned algorithms are limited in terms of either their embedding capacity or their robustness against the transmission channel attacks described before. Specifically, for narrowband speech watermarking, most achieved bit rates are only from several bits to dozens of bits per second, which limits speech watermarking to some real applications where larger capacity is usually needed. In addition, most transmission channels used in previous studies are assumed to be digital channels, but some practical channels as mentioned above are not considered. Nevertheless, several researchers have provided some heuristic techniques. In Hofbauer *et al.* (2009), for instance, speech watermarking was designed for the analog flat-fading channel, and the capacity is much higher when compared to other speech watermarking methods. This method, however, has some disadvantages. First, it uses only unvoiced speech segments to embed the watermark, so the speech watermarking complexity is increased in the sense that using additional unvoiced/voiced (UV) speech detection methods becomes a must. Even worse, the detection results after extraction are likely to be different from those before embedding due to channel attacks, which will further make recovering the watermark more complicated. Second, watermark generation is very intricate because it is based on the

theories of non-uniform sampling, and sampling and interpolation of band-limited signals. Third, the watermarking method is assumed to be used in only radio systems, but several other channel attacks such as transcoding existing in cellular networks are not considered, which obviously limits its applications.

It is claimed that capacity remains unused by not considering the specific properties of the narrowband speech and the way it is perceived. In the proposed approach, the watermarking theory presented before is combined with a principle of speech perception, leading to a substantial improvement in capacity and a simplification in watermarking algorithm complexity. From this, a practical watermarking scheme is proposed. This scheme contains two key concepts. The first one is to replace perceptually insensitive components of the speech signal by the band-limited watermark signal in both unvoiced and voiced segments, and the second one is to use common speech coding and digital communication techniques.

It is a long-known fact that, for the speech signal, one can replace certain speech components by a perceptually similar watermark signal while preserving the imperceptibility of the watermark and the intelligibility of the watermarked speech. Example methods have been proposed that exchange time-frequency components of audio signals above 5 kHz that have strong noise-like properties by a spread spectrum sequence. However, this algorithm cannot be directly used for the narrowband speech due to the bandwidth limitation and the fact that human ears are more sensitive to distortions below 4 kHz. However, motivated by this idea and by considering the specific perceptual properties of the narrowband speech, it is found that if only a part of narrowband speech signals in the high-frequency region are exchanged by a shaped data signal that carries watermark information and is perceptually similar to the original speech components, high capacity, good intelligibility of the watermarked speech, and good imperceptibility of the watermark signal can be achieved. Our main contributions are as follows.

To obtain a perceptually transparent and practical implementation of our approach, an autoregressive (AR) speech signal model is assumed. When embedding a watermark signal, a host speech segment is first decomposed into an excitation signal and a vocal tract filter using the LP analysis technique. The vocal tract is constrained to remain unchanged, but it allows a subband component of the excitation signal to be changed (replaced). In the original excitation signal, its frequency range is divided into two parts, i.e., a lower part from 0 to 2.5 kHz and a higher part from 2.5 to 4 kHz. By applying the concept of replacing certain excitation signal components, a subband component in the high-frequency region is then exchanged by a watermark signal, which is a subband bandpass signal in nature, and a new excitation signal conveying the watermark information is created. Finally, this excitation signal is shaped by the same vocal tract filter to form the watermarked signal, and this step can be seen as a speech synthesis process. Perceptual transparence is guaranteed by forcing the watermark signal and the replaced excitation signal to have the same averaged power level.

Another contribution of this work is that, we have proved that the vocal tract can be recovered at the extractor if the distortion of the host speech introduced by embedding the watermark signal is small and the channel attacks are not severe. This fundamental result provides us with a theoretical basis, and therefore sets up the rationality of our proposed watermarking scheme. This can be explained as follows. From the communication point of view, the received watermark signal of the proposed algorithm has endured two cascade filters, i.e., a synthesis filter at the embedder and an analysis filter at the extractor. Using the result mentioned above, the overall frequency response of these two filters is actually flat or ideal in the subband of the watermark signal. This means that the channel can be perfectly compensated for transmitting and receiving the watermark signal.

Several experiments have been conducted to test and evaluate the proposed speech watermarking algorithm. First, the autocorrelation sequence of the received channel attacked watermarked signal is compared to that of the original host speech signal, since the analysis filter and synthesis filter have a very close relationship to the autocorrelation sequence. Our results show that, under the AWGN attack, the autocorrelation sequence can be recovered using the channel attacked watermarked signal at the extractor. This indicates that the analysis filter at the extractor can perfectly compensate for the synthesis filter at the embedder, and therefore is consistent with the theoretical result mentioned above. Second, it is

demonstrated that the original speech and watermarked speech are very close to each other in both the time domain and the frequency domain. The listening test results in terms of objective tests and subjective tests are also provided. All these results show that both intelligibility of the host speech and imperceptibility of the watermark signal are obtained. Then the robustness of the proposed algorithm against several channel attacks listed before is tested. The results show that the capacity reaches 200 bits/s with the bit error rate below 0.001, indicating that the proposed algorithm can be used in real applications. This data rate can be further increased to 400 bits/s if the passband width is increased at the cost of a lower bit error rate (BER). Finally, our algorithm is compared to state-of-the-art speech watermarking methods. The results show that the proposed algorithm is more efficient in terms of bandwidth usage than most of other algorithms.

2 Theory

2.1 Speech signal model

A speech frame of 20–30 ms can be seen as a stationary process. Suppose that a speech frame is denoted by $x(n)$, n=0, 1, …, N−1. Using the analysis process in LP, a vocal tract in terms of $A(z)=\sigma^{-1}\left(1+\sum_{l=1}^{p}a_l z^{-l}\right)$ can be obtained, where σ is the gain scalar and $\{a_l\}_{l=1}^{p}$ is a set of LP coefficients with order p. Moreover, by passing $x(n)$ through the vocal tract $A(z)$, the residual or the excitation $e(n)$ is obtained:

$$e(n)\sigma = x(n) + \sum_{l=1}^{p} a_l x(n-l). \tag{1}$$

This is called the analysis process. On the other hand, by passing $e(n)$ through the vocal tract $A^{-1}(z)$, the speech can be synthesized:

$$x(n) = e(n)\sigma + \sum_{l=1}^{p} a_l x(n-l). \tag{2}$$

This is called the synthesis process. To compute the coefficients $\{a_l\}$ in Eqs. (1) and (2), the Yule-Walker method can be used to find the solution to a set of normal equations:

$$\sum_{l=1}^{p} a_l r_x(k-l) = -r_x(k), \quad k=1, 2, ..., p, \tag{3}$$

where $r_x(k)$ is the autocorrelation sequence of $x(n)$, and $r_x(k)=N^{-1}\sum_{n=0}^{N-1}x(n)x(n-k)$.

2.2 Ideal channel model for hidden communication

In conventional speech applications such as speech enhancement and speech coding, the analysis process is usually first used to obtain the vocal tract and the excitation, which can further be used to remove the noise or the redundancy components, and then the synthesis process is used to recover the original speech. However, in this study we use the analysis process and synthesis process in a different way. The synthesis process is first used to shape the excitation and the analysis process is then used to equalize the channel distortion to recover the original excitation. This model is shown in Fig. 1. The key idea is that the vocal tract information can be kept if the autocorrelation sequence changes only a little. In previous works this idea was often directly used by embedding the watermark into the excitation, but the rationality behind this was not provided. In our work, the reason for the use of the excitation signal for watermarking or hidden communication will be given in the following.

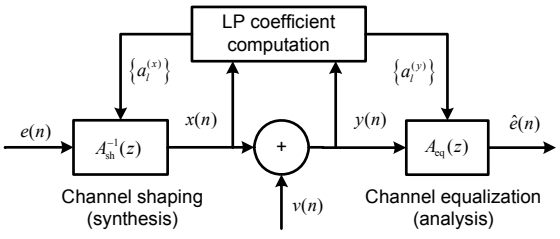

Fig. 1 Principle of using the synthesis process as the channel shaping step and the analysis process as the channel equalization step for communication

According to the model shown in Fig. 1, it is assumed that an excitation $e(n)$ first passes through a shaping filter $A_{sh}^{-1}(z)$ to obtain a speech segment $x(n)$, and then is distorted by an AWGN $v(n)$ with variance σ_v^2 to output a noisy speech signal $y(n)$=

$x(n)+v(n)$. Thereafter, this noisy speech further passes through an equalization filter $A_{eq}(z)$, resulting in an estimate of $e(n)$, denoted by $\hat{e}(n)$.

In addition, if it is assumed that $x(n)$ and $v(n)$ are not correlated, then $r_y(k)=r_x(k)+\sigma_v^2\delta(k)$ can be obtained, where $\delta(n)$ is the Dirac function. Taking one step further, if variance σ_v^2 is very small when compared to $r_x(k)$, then it can be derived that $r_y(k)\approx r_x(k)$.

As mentioned earlier, coefficients $\{a_l\}$ can be computed using the Yule-Walker method, in which an autocorrelation sequence is used to construct a Toeplitz autocorrelation matrix and a cross-correlation vector. Since $r_y(k)\approx r_x(k)$, it can be derived that $a_l^{(y)}\approx a_l^{(x)}$ and $A_{sh}(z)\approx A_{eq}(z)$, where $a_l^{(x)}$ is calculated using $r_x(k)$ and $a_l^{(y)}$ is calculated using $r_y(k)$. Based on this, it is obtained that

$$A_{sh}^{-1}(z)A_{eq}(z)\approx 1. \tag{4}$$

Eq. (4) indicates that the channel distortion can be nearly equalized using the proposed model, and therefore the excitation $e(n)$ can be recovered, i.e., $\hat{e}(n)\approx e(n)$.

As a consequence, the proposed model can be used for watermarking or covert communication purpose. Specifically, if the watermark signal is embedded into the original excitation $e(n)$ at the embedder, then the estimated excitation $\hat{e}(n)$ can be obtained at the extractor, and finally the watermark bits can be detected using $\hat{e}(n)$.

2.3 Imperceptibility principle

As mentioned in Section 1, replacing some or all components of the speech in the high-frequency subband with a hidden signal conveying the secret message can accomplish covert communication. To achieve this, the host speech is indirectly changed by exchanging the original excitation with the watermark signal, and then resynthesizing the watermarked speech. Specifically, two filters are used to divide the excitation $e(n)$—now denoted as a full band signal $e_{fb}(n)$—into the passband component and the stopband component. The first filter is a bandpass filter. By passing $e_{fb}(n)$ through it, a passband excitation $e_{fb}(n)$ can be obtained, and it can be used to synthesize the passband speech $x_{pb}(n)$. For covert communication purpose and for the proposed algorithm, a hidden watermark signal $\hat{e}_{pb}(n)$, instead of the original $e_{fb}(n)$, is used to synthesize a passband speech $\hat{x}_{pb}(n)$. The second filter is a bandstop filter. After bandstop filtering the excitation $e_{fb}(n)$, a stopband excitation $e_{sb}(n)$ can be obtained, which is kept unchanged to synthesize the stopband speech $x_{sb}(n)$. The relationships between the above-mentioned signals are described as follows.

The full band host speech $x_{fb}(n)$ in the frequency range from 0 to 4 kHz can be divided into the stopband component $x_{sb}(n)$ and the passband component $x_{pb}(n)$, i.e., $x_{fb}(n)=x_{sb}(n)+x_{pb}(n)$. The full band watermarked speech $\hat{x}_{fb}(n)$, which is an approximation to $x_{fb}(n)$, can be divided into the same stopband component $x_{sb}(n)$ as before and the new passband component $\hat{x}_{pb}(n)$, i.e., $\hat{x}_{fb}(n)=x_{sb}(n)+\hat{x}_{pb}(n)$. The excitation signals can be obtained by using Eq. (1):

$$e_{fb}(n)=e_{sb}(n)+e_{pb}(n), \tag{5}$$

$$\hat{e}_{fb}(n)=e_{sb}(n)+\hat{e}_{pb}(n). \tag{6}$$

If the original passband excitation signal is formulized as the new passband excitation $\hat{e}_{pb}(n)$ with a perturbation $\Delta e_{pb}(n)$, or

$$e_{pb}(n)=\hat{e}_{pb}(n)+\Delta e_{pb}(n), \tag{7}$$

then by substituting Eqs. (6) and (7) into Eq. (5) it is derived that

$$e_{fb}(n)=\hat{e}_{fb}(n)+\Delta e_{pb}(n). \tag{8}$$

Further, if this variation in Eq. (8) is assumed to be very small, then we have

$$e_{pb}(n)\approx\hat{e}_{pb}(n). \tag{9}$$

By substituting Eq. (9) into Eqs. (5) and (6), it is obtained that

$$e_{fb}(n)\approx\hat{e}_{fb}(n). \tag{10}$$

Eq. (10) indicates that changing the original passband excitation slightly has negligible influence on the integrity of the full band excitation. Finally, by

passing $\hat{e}_{\mathrm{fb}}(n)$ through the synthesis filter and using Eq. (10), it is derived that

$$\begin{aligned}\hat{x}_{\mathrm{fb}}(n) &= \hat{e}_{\mathrm{fb}}(n) * Z^{-1}[A^{-1}(z)] \\ &\approx e_{\mathrm{fb}}(n) * Z^{-1}[A^{-1}(z)] = x_{\mathrm{fb}}(n),\end{aligned} \tag{11}$$

where '*' denotes convolution and $Z^{-1}[\cdot]$ represents the inverse Z-transform in Eq. (11).

To summarize, if the original passband excitation is replaced by a passband watermark signal and this replacement does not change the full band excitation a lot, then the resynthesized full band watermarked speech will be very close to the original full band host speech. In other words, they must have similar, if not the same, intelligibility and quality. In addition, the watermark signal can hardly be perceived by human ears.

2.4 Passband communication for transmitting and receiving watermark bits

As mentioned in Section 2.3, if the original passband excitation signal is replaced by the passband watermark signal, then both intelligibility and imperceptibility can be achieved. There are several ways to generate a passband watermark signal. One possible way is to use the canonical passband communication theory to first generate a band-limited baseband signal and then up-convert it into the passband. Specifically, the band-limited requirement can be satisfied by passing the watermark bits through the pulse shaping filter, and the passband constraint can be fulfilled by forcing the sinusoidal carriers to modulate the amplitude of the baseband signal. This idea is shown in more detail in Fig. 2, and is explained as follows.

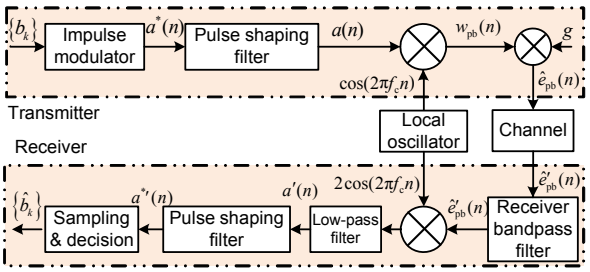

Fig. 2 The principle of the hidden communication system

Here the binary sequence $b_k \in \{+1, -1\}$ is used, for simplicity, to represent the secret message. First, b_k passes through the impulse modulator, resulting in the signal $a^*(n) = \sum_{k=-\infty}^{+\infty} b_k \delta(n - kT_{\mathrm{b}})$, where T_{b} is the bit interval. This signal then passes through the baseband transmit shaping filter with impulse response $p(n)$. In this study we use the square root raised cosine (SRRC) filters at both the transmitter and the receiver. The baseband shaping filter is low pass in nature and its cutoff frequency is somewhat greater than $1/(2T_{\mathrm{b}})$. The output of the transmitter shaping filter is $a(n) = \sum_{k=-\infty}^{+\infty} b_k p(n - kT_{\mathrm{b}})$.

To translate its spectra up to the passband of the bandpass channel, $a(n)$ is amplitude modulated by the carrier $\cos(2\pi f_{\mathrm{c}} n)$, where f_{c} is the carrier center frequency, resulting in the passband watermark signal

$$w_{\mathrm{pb}}(n) = a(n)\cos(2\pi f_{\mathrm{c}} n). \tag{12}$$

To better improve the imperceptibility, this signal is then scaled to output the new passband excitation signal $\hat{e}_{\mathrm{pb}}(n) \approx g \cdot w_{\mathrm{pb}}(n)$. Thereafter, $\hat{e}_{\mathrm{pb}}(n)$ passes into the channel.

As mentioned in Sections 1 and 2, this hidden signal will endure channel attacks such as synthesis processing, analysis processing, and transmission channel distorting. These attacks will be discussed in detail in Section 3. As a whole, however, it has been proved that if the transmission channel corruption is moderate, then the overall frequency response is approximately ideal. This result indicates that the input of the receiver can be expressed as $\hat{e}'_{\mathrm{pb}}(n) \approx \hat{e}_{\mathrm{pb}}(n)$.

At the receiver, the channel output $\hat{e}'_{\mathrm{pb}}(n)$ first passes through an ideal receiver bandpass filter with its center frequency at f_{c} and its cutoff frequency near $f_{\mathrm{c}} \pm 1/(2T_{\mathrm{b}})$, resulting in the passband signal $\hat{w}_{\mathrm{pb}}(n)$. According to the above analysis, it can be inferred that $\hat{w}_{\mathrm{pb}}(n) \approx g \cdot w_{\mathrm{pb}}(n)$ since the receiver filter is considered to be flat.

If carrier synchronization is already achieved, then the received baseband signal $a'(n)$ is obtained, which is the estimate of $a(n)$ through low-pass filtering the down-converted version of $\hat{w}_{\mathrm{pb}}(n)$, or

$$a'(n) = \mathrm{LPF}[2\cos(2\pi f_{\mathrm{c}} n)\hat{w}_{\mathrm{pb}}(n)] \approx g \cdot a(n), \tag{13}$$

where LPF[·] denotes the low-pass filtering operation. Eq. (13) shows that the waveform of $a(n)$ is kept after a series of processing. So, there is no information lost, specifically for binary communication. Then $\hat{a}(n)$ passes through the same shaping filter as in the transmitter to obtain $\hat{a}^*(n)$, which is the estimate of $a^*(n)$. By sampling this signal at the instant of the multiple of T_b and making a decision, the following maximum likelihood (ML) detector is obtained:

$$\hat{b}_k = \text{sign}\left[a^*(n)\big|_{n=kT_b} \right] = \begin{cases} +1, & a^*(kT_b) \geq 0, \\ -1, & a^*(kT_b) < 0, \end{cases} \quad (14)$$

where sign[x] returns the sign of x.

3 System implementation and analysis

3.1 Overall system design

Based on the analysis in Section 2, a speech watermarking scheme is illustrated in Fig. 3. It consists of a watermark bits embedder, a watermark bits extractor, and a transmission channel. In the following sections we will discuss them in greater detail. Overall, the key techniques that guarantee the intelligibility of the speech and the imperceptibility of the watermark signal aim to adjust the power of the passband watermark signal and to shape the scaled passband watermark signal using the synthesis filter to generate the desired watermarked speech.

3.2 Watermark bits embedding

3.2.1 Host speech preprocessing

As shown in Fig. 3, the full band narrowband host speech $x_{fb}(n)$ in one segment is used as the cover in the following ways. First, the LP coefficients are obtained by using LP techniques such as the Levinson-Durbin algorithm. These coefficients are then used to form the analysis filter $A(z)$ and the synthesis filter $A^{-1}(z)$.

Second, after passing $x_{fb}(n)$ through an analysis filter $A(z)$, a full band excitation signal $e_{fb}(n)$ is obtained. This signal is divided into a passband excitation signal $e_{pb}(n)$ and a stopband excitation signal $e_{sb}(n)$ by using a bandpass filter and a bandstop filter, respectively.

Note that there are two gaps or guarding bands between the transition areas of these two filters. By doing so, the passband watermark signal can be recovered without inter channel interference (ICI).

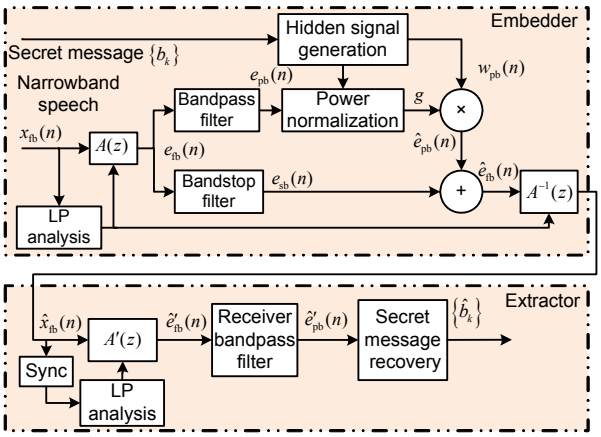

Fig. 3 Overall system design for narrowband speech watermarking

3.2.2 Watermark signal generation

If the binary sequence of the watermark is denoted as $b_k \in \{+1, -1\}$, then the passband watermark signal $w_{pb}(n)$ can be obtained using the method introduced in Section 2.4.

3.2.3 Power adjustment of the passband watermark signal

The passband watermark signal $w_{pb}(n)$ can be very different in terms of the waveform from the original passband excitation signal $e_{pb}(n)$. If it is directly used to synthesize the passband watermarked speech, then neither intelligibility nor imperceptibility can be achieved. Therefore, a power adjustment algorithm is proposed for scaling $w_{pb}(n)$ to a more suitable level before it is used to synthesize the watermarked speech. After power adjustment, a gain factor g is obtained, which can be used to multiply the original passband watermark signal $w_{pb}(n)$ to output the new passband excitation signal $\hat{e}_{pb}(n)$. This can be explained as follows.

To calculate g, the passband watermark signal and original passband excitation are forced to have the same power, or $\sum_{n=0}^{N-1} \left|\hat{e}_{pb}(n)\right|^2 = \sum_{n=0}^{N-1} \left|e_{pb}(n)\right|^2$. Then it can be derived that

$$g=\left(\sum_{n=0}^{N-1}\left|e_{\mathrm{pb}}(n)\right|^2\Big/\sum_{n=0}^{N-1}\left|w_{\mathrm{pb}}(n)\right|^2\right)^{1/2}. \quad (15)$$

Finally, by using Eq. (15) the new passband excitation signal is obtained:

$$\hat{e}_{\mathrm{pb}}(n)=g\cdot w_{\mathrm{pb}}(n). \quad (16)$$

3.2.4 Watermarked speech generation

As mentioned earlier, the secret message is conveyed by the hidden signal $w_{\mathrm{pb}}(n)$. It can be directly added to the original stopband speech to form the full band watermarked speech, or $\hat{x}_{\mathrm{fb}}(n)=x_{\mathrm{sb}}(n)+w_{\mathrm{pb}}(n)$. However, the passband watermark signal can be perceived by human ears or seen by human eyes in terms of the spectrogram graph. This is due to the fact that the passband watermark actually has fixed power levels, while the speech signal is the long-term nonstationary process and has a much wider dynamic range. Therefore, the power level of the hidden signal and the original replaced passband speech may differ a lot. Moreover, the envelope of the hidden signal is not guaranteed to be the same as that of the original replaced passband speech. To summarize, covert communication cannot be accomplished by directly adding the passband watermark signal to the stopband host speech signal.

The power level of the passband watermark signal can be adjusted according to Eq. (16). In this way, the level of the replaced passband speech is increased or decreased to match that of the original passband speech. To better improve the intelligibility of the speech and the imperceptibility of the watermark signal, the scaled hidden signal is further shaped by the shaping filter $A^{-1}(z)$ to output the passband replaced speech $\hat{x}_{\mathrm{pb}}(n)=\hat{e}_{\mathrm{pb}}(n)*Z^{-1}[A^{-1}(z)]$, where $\hat{e}_{\mathrm{pb}}(n)$ is defined in Eq. (6). By doing so, this signal will have the same envelope as that of the original passband speech $x_{\mathrm{pb}}(n)$. Consequently, the intelligibility and quality of the full band watermarked speech and the imperceptibility of the watermark signal are well kept. It can also be derived that

$$\hat{x}_{\mathrm{fb}}(n)=x_{\mathrm{sb}}(n)+\hat{x}_{\mathrm{pb}}(n). \quad (17)$$

Eq. (17) indicates that the watermarked signal consists of the original stopband signal and the new passband signal.

Finally, according to these embedding steps, sample pseudo code of the proposed watermark embedding algorithm is given in Fig. 4.

Input: the nth frame speech signal $\boldsymbol{x}_n$, the nth frame watermark signal $\boldsymbol{w}_n$.
Output: the nth frame watermarked signal $\boldsymbol{x}_{\mathrm{fb},n}$.
For n=1:N // embedding frame by frame
  $\boldsymbol{x}_{\mathrm{sb},n}$←SBF($\boldsymbol{x}_n$) // stopband filtering
  $\boldsymbol{x}_n$←Hann($\boldsymbol{x}_n$) // windowing
  σ_N^2, $\boldsymbol{a}_n$←LPC($\boldsymbol{x}_n$) // LP analysis
  $\boldsymbol{e}_{\mathrm{pb},n}$←Filter($\sigma_n$, $\boldsymbol{a}_n$, $\boldsymbol{w}_n$) // passband excitation
  g←$||\boldsymbol{e}_{\mathrm{pb},n}||/||\boldsymbol{w}_n||$ // scaling factor
  $\boldsymbol{w}_n$←$g\cdot\boldsymbol{w}_n$ // power normalization
  $\boldsymbol{x}_{\mathrm{pb},n}$←Filter($\sigma_n$, $\boldsymbol{a}_n$, $\mathbf{w}_n$) // new passband signal
  $\boldsymbol{x}_{\mathrm{fb},n}$←$\boldsymbol{x}_{\mathrm{sb},n}$+$\boldsymbol{x}_{\mathrm{pb},n}$ // watermarked signal
End

Fig. 4 Pseudo code of the proposed watermark embedding algorithm

3.3 Watermark bits extraction

As shown in Fig. 3, the watermark extractor consists of synchronization, watermark signal recovery, and watermark bits detection. Before the watermark bits are really extracted, the received watermarked speech should be synchronized to determine the start point of embedding. In this study, a 2-s unvoiced speech excerpt is used for timing synchronization or sampling clock recovery. This is not the main focus of this study, so this matter is not further treated and hereafter a perfect timing synchronization is assumed.

After synchronization, the watermark signal can be recovered in two steps. First, the excitation signal is obtained using the estimated LP coefficients to equalize the channel distortion. The passband watermark signal is then recovered by using the bandpass filter to remove the useless out-of-band components of the excitation signal.

Finally, this recovered passband watermark signal is used for watermark bits detection using the method described in Section 2.

3.4 Attacks

It is known that the speech watermarking scheme should be robust against specified attacks. In our research, system-inherent attacks and transmission

channel attacks are considered. The former are introduced by the speech watermarking algorithm itself, including time-variant gain, time-variant all-pole filtering, additive stopband speech signal, and imperfect channel compensation. The latter are determined by practical applications, including resampling, AWGN, bandpass filtering, magnitude and phase distortion, and transcoding. These attacks are listed in Table 1. All these attacks can increase the bit error rate (BER).

Table 1 System-inherent and transmission channel attacks

Type	Attacks
System-inherent attack	Time-variant gain, time-variant all-pole filtering, time-variant equalization, additive stopband signal
Transmission channel attack	Resampling, AWGN, filtering, magnitude and phase distortion, transcoding

3.5 Filter designs

For demonstration purpose, in this study the bit rate of the secret message is set to be 200 bits/s, and every bit can by represented by 40 samples. According to the classical communication theory, the bandwidth needed for this secret message is at least 200 Hz for binary communication. On the other hand, the carrier center frequency is set to be 3000 Hz, which indicates that the passband ranges from 2900 Hz to 3100 Hz.

Under these constraints, the settings of several filters in the proposed system are summarized in Table 2. At the embedder, a bandpass filter, a bandstop filter, and a pulse shaping filter are needed. At the extractor, a bandpass filter, a low-pass filter, and a pulse shaping filter must be employed.

Besides these settings, the magnitude responses for these filters are presented (Fig. 5). As can be seen, the stopband width of the bandstop filter in Fig. 5a is set to be equal to or greater than the passband width of the bandpass filter at the embedder, and the passband width of the extractor bandpass filter should be equal to or less than that of the bandpass filter in the embedder, and greater than that of the watermark signal. By doing so, the watermark signal can avoid the out-of-band interference at the embedder, and the extractor bandpass filter can filter out the out-of-band signal and let the watermark signal pass through without any harm.

Table 2 Parametric settings for filters

Location	Type	Parameters
Embedder	Bandpass filter	Stop frequency 1: 2850 Hz Pass frequency 1: 2900 Hz Pass frequency 2: 3100 Hz Stop frequency 2: 3150 Hz
Embedder	Bandstop filter	Pass frequency 1: 2800 Hz Stop frequency 1: 2850 Hz Stop frequency 2: 3150 Hz Pass frequency 2: 3200 Hz
Extractor	Bandpass filter	The passband width is equal to or less than that in the transmitter, and greater than that of the hidden signal
Extractor	Low-pass filter	Pass frequency: 100 Hz Stop frequency: 150 Hz
Embedder/ Extractor	Pulse shaping filter	Roll-off factor: 0.32 Upsampling: 40 samples Group delay: $4T_b$

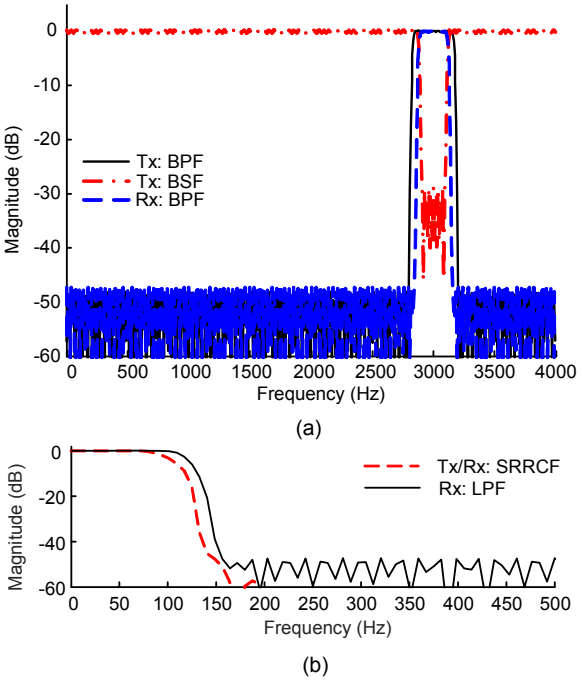

Fig. 5 The magnitude responses for the bandpass filter and the bandstop filter at the embedder, and the bandpass filter at the extractor (a) and the pulse shaping filter at the embedder and extractor, and the low-pass filter at the extractor (b)

Fig. 5b shows that the bandwidth of the low-pass filter should be equal to or greater than that of the pulse shaping filter, since the bandwidth of the hidden

signal is determined by that of the pulse shaping filter.

It is emphasized that, however, these parameter settings are only for demonstrating how the designed system works, and thereby are not optimized.

3.6 Capacity

According to the binary phase shift keying (BPSK) modulation method described in Section 2, the theoretical capacity can be expressed as

$$C = W\log[1 + P/(N_0 W)] \text{ bits/s.} \quad (18)$$

In Eq. (18) the channel is band-limited to $[-W, W]$, the noise is Gaussian with a two-sided power spectral density of $N_0/2$, and the channel input signal has a power constraint of P. On the other hand, the BER is $p_e = Q(\sqrt{2E/N_0})$, where E is the signal energy and $Q(x) = \int_{-\infty}^{-x} e^{-x^2/2} dx \Big/ \sqrt{2\pi}$. Given the transmission rate and BER, it is possible to express the achievable watermark capacity by considering the channel as a memoryless binary symmetric channel (BSC). The channel capacity R is

$$R = C[1 + p_e \log_2 p_e + (1 - p_e)\log_2(1 - p_e)] \text{ bits/s.} \quad (19)$$

This rate in Eq. (19) can be asymptotically achievable with appropriate channel coding.

4 Experiments

4.1 Parametric settings

In the following experiments, 40 randomly selected speech utterances from the IEEE corpus are used, and the segments that contain no voices but noise or silence are removed, resulting in a total of 34 s of speech. These speeches are resampled to 8000 Hz, and analyzed with a linear predictor with its order p=10. When using the LP technique, the speech frames are non-overlapping, and the duration of each frame is 30 ms.

The experiments are designed as follows. The hidden channel quality is first tested to verify that the channel can be compensated frame by frame. Then single- and multi-frame results of the watermarked speech are shown from both time domain and frequency domain points of view, followed by subjective and objective listening tests. The evaluations of robustness under several attacks are also provided. Finally, the proposed speech watermarking scheme is compared to some state-of-the-art algorithms in terms of their capacities.

4.2 Hidden channel quality tests

According to the channel model in Section 2, the hidden channel quality relies heavily on the recovered autocorrelation sequence at the extractor. This is due to the fact that the synthesized watermarked speech is in fact different from the original host speech. For the watermarking system proposed in Section 3, however, one requirement is that their differences should be as small as possible to obtain the more accurately estimated vocal tracts for compensating the channel distortion. In the following, the autocorrelation sequences of the watermarked speech are compared to those of the original speech with and without AWGN channel attacks. By doing so, it can be directly determined whether the overall channel distortion introduced by our model can be neglected or not.

For comparison, the autocorrelation sequence in one frame is normalized by forcing its maximum value to be 1, or

$$\bar{r}^{(m)}(k) = r^{(m)}(k) \Big/ \max(|r^{(m)}(k)|), k = 0, 1, \ldots, p, \quad (20)$$

where m denotes the current frame number. The original and recovered autocorrelation sequences of the 22nd frame defined in Eq. (20) are illustrated in Fig. 6. As can be seen, they are very close to each other in this frame, indicating that the autocorrelation sequence at the extractor is well estimated, thus providing the possibility of recovering the corresponding vocal tract response.

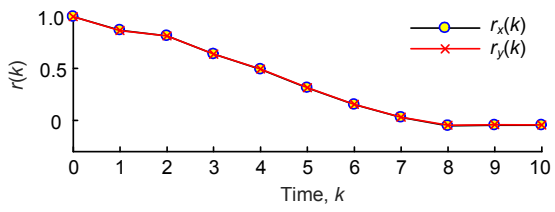

Fig. 6 The original and recovered autocorrelation sequences of the 22nd frame

For multiple frames, on the other hand, the autocorrelation sequence norm ratio (ASNR) criterion is used to evaluate the hidden channel quality. ASNR is defined in Eq. (21), being the ratio of the norm of the autocorrelation sequence vector $\bar{\boldsymbol{r}}_x^{(m)}$ of the original speech $x(n)$ to the norm of the autocorrelation sequence error vector $\Delta\bar{\boldsymbol{r}}_x^{(m)}$ between the original $x(n)$ and the synthesized speech $y(n)$:

$$\mathrm{ASNR}_m = 10\lg(\|\bar{\boldsymbol{r}}_x^{(m)}\|^2 / \|\Delta\bar{\boldsymbol{r}}_x^{(m)}\|^2), \tag{21}$$

where $\bar{\boldsymbol{r}}_x^{(m)}=[\bar{r}_x^{(m)}(0),..,\bar{r}_x^{(m)}(p)]^{\mathrm{T}}$, and the difference vector $\Delta\bar{\boldsymbol{r}}_x^{(m)}=\left[\bar{r}_x^{(m)}(0)-\bar{r}_y^{(m)}(0),...,\bar{r}_x^{(m)}(p)-\bar{r}_y^{(m)}(p)\right]^{\mathrm{T}}$.

This criterion is reasonable since it reflects the similarity between two different autocorrelation sequences from the estimation point of view. If the estimation error is small, then ASNR will be large, and vice versa. Fig. 7 shows the results of the ASNR values for the first 100 frames. As can be seen, these values range from 10 dB to almost 70 dB.

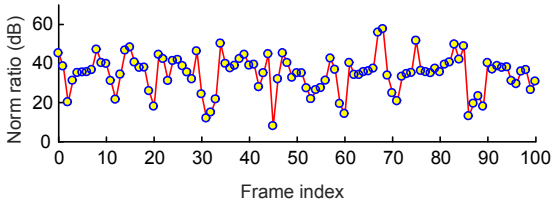

Fig. 7 The autocorrelation sequence norm ratio of the first 100 frames

Furthermore, by averaging the ASNR in each frame, we have

$$\overline{\mathrm{ASNR}} = \frac{1}{M}\sum_{m=0}^{M-1}\mathrm{ASNR}_m. \tag{22}$$

The indicator in Eq. (22) represents the overall estimation performance of the recovered autocorrelation sequences without any attacks between the embedder and the extractor. For this experiment setting, the averaged ASNR value is close to 40 dB for 1000 frames, meaning that the recovered autocorrelation sequences are very close to the original ones.

Furthermore, the averaged ASNRs under the AWGN attacks with different signal-to-noise ratios (SNRs) are tested (Fig. 8). From the estimation theory point of view, the autocorrelation sequences can be recovered even with moderate additive noises (SNR>20 dB). The reason behind this is that the spectrum of the speech is not uniformly distributed in the frequency range; i.e., the low-frequency components usually have much more energy than the high-frequency components, and this is especially true for voiced speech. The low-frequency components can be better preserved than the high-frequency components when uniformly distributed noise is added, thus facilitating autocorrelation sequence calculation. For severe additive noises (SNR<20 dB), which are very common in noisy autocorrelation estimation problems, some iterative methods can be used to further improve the estimation accuracy (Zheng, 2005).

Fig. 8 Averaged auto-correlation sequence norm ratios (ASNRs) with and without AWGN

4.3 Demonstration of watermarked speech

For covert communication purpose, the synthesized watermarked speech must be as close as possible to the original speech, in order to avoid deliberate observation and detection. Therefore, both the time domain waveform and the magnitude spectrum of the watermarked speech should be similar to that of the original speech. The results of the synthesized watermarked speech are given in terms of one frame and multiple frames, respectively.

4.3.1 Single-frame watermarked speech

In the first place, the temporal waveforms of the synthesized watermarked speech and the original speech in a frame are shown in Fig. 9. As can be seen, they are very close to each other, so the quality of the synthesized speech is maintained.

In the spectral domain, the magnitude spectrum of a frame can be used to show the synthesized speech (Fig. 10). For comparison, Fig. 10 also plots the

original magnitude spectrum and the original vocal tract. It can be seen that, in the stopband the spectra of the original and the synthesized speech are the same, while in the passband they are different because the original passband speech is replaced by the hidden signal. Their spectral envelopes, however, are the same. As a result, the intelligibility represented by the spectral envelope of the synthesized speech is maintained.

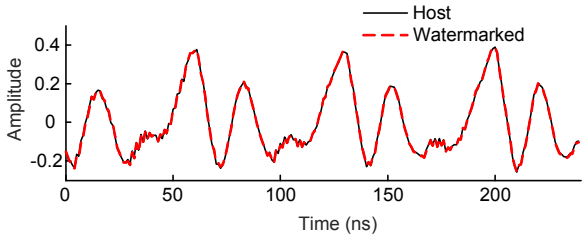

Fig. 9 The temporal waveforms of the original and synthesized speeches of a frame

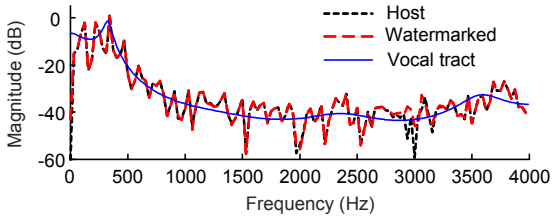

Fig. 10 Magnitude spectrum of the original speech and the synthesized watermarked speech in a frame

4.3.2 Multi-frame watermarked speech

For the watermarked speech in multiple frames, the time domain waveforms and spectrograms of the original and the synthesized watermarked speeches are shown in Fig. 11. As can be seen, the differences between the original and the synthesized watermarked speech waveforms in the time domain are very small. As for their spectrograms, which contain the intelligibility and the quality information, it is obvious that the formants, the fundamental frequencies and harmonics of the two signals are nearly the same. This means that the intelligibility of the synthesized speech is maintained. In addition, only a little distortion is introduced in the passband with its center surrounding the carrier frequency of 3 kHz, indicating that not only the quality of synthesized speech is kept but also the imperceptibility of the watermark signal is achieved.

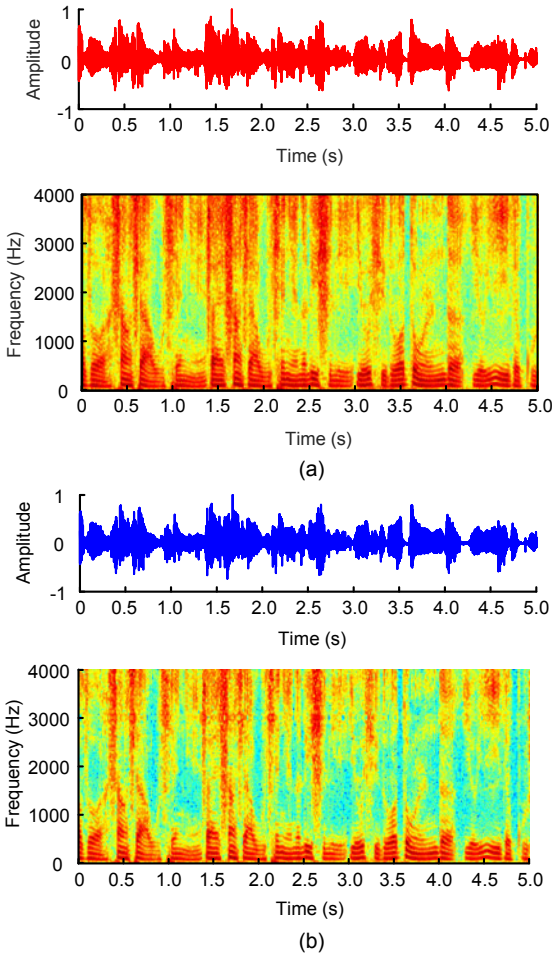

Fig. 11 Time domain waveforms and spectrograms of the original full band speech (a) and the synthesized full band speech (b)

4.4 Listening tests

4.4.1 Subjective tests

The speech quality in the output is evaluated by the mean opinion score (MOS), ranging from 1 to 5, and by using the International Telecommunication Union (ITU) recommendation P.800, which provides methods for conducting listening tests. In our experiment, 20 responders are included for this test. The averaged scores of the original and the watermarked speeches are 5 and 4.90, respectively. This result indicates that the speech quality is well kept and the watermark signal cannot be perceived by the listeners.

4.4.2 Objective tests

There are several measures to objectively test the watermarked speech. They are frequently used in

applications such as speech enhancement and speech coding. Here, segmental SNR (SegSNR), log likelihood ratio (LLR), Itakura-Saito (IS) distortion, and perceptual evaluation of speech quality (PESQ) are used as the measures. Their expressions are summarized in Table 3.

Table 3 Criteria for evaluating the speech quality

Criterion	Expression
SegSNR	$\mathrm{SegSNR}_m=10\lg\dfrac{\sum_{n=0}^{N-1}(x_{\mathrm{fb}}^{(m)}(n))^2}{\sum_{n=0}^{N-1}\left[x_{\mathrm{fb}}^{(m)}(n)-\hat{x}_{\mathrm{fb}}^{(m)}(n)\right]^2}$
LLR	$\mathrm{LLR}_m=\log(\hat{\boldsymbol{a}}^{(m)\mathrm{T}}\boldsymbol{R}_x^{(m)}\hat{\boldsymbol{a}}^{(m)})/(\boldsymbol{a}^{(m)\mathrm{T}}\boldsymbol{R}_x^{(m)}\boldsymbol{a}^{(m)})$
IS	$\mathrm{IS}_m=\dfrac{(\sigma_x^{(m)})^2}{(\hat{\sigma}_x^{(m)})^2}\dfrac{\hat{\boldsymbol{a}}^{(m)\mathrm{T}}\boldsymbol{R}_x^{(m)}\hat{\boldsymbol{a}}^{(m)}}{\boldsymbol{a}^{(m)\mathrm{T}}\boldsymbol{R}_x^{(m)}\boldsymbol{a}^{(m)}}+\log\dfrac{(\sigma_x^{(m)})^2}{(\hat{\sigma}_x^{(m)})^2}-1$
PESQ	$\mathrm{PESQ}=4.5-0.1d_{\mathrm{sym}}-0.0309d_{\mathrm{asym}}$

In this test, five females and five males are selected from the speech corpus to evaluate the speech quality. The speech quality results are listed in Table 4. The averaged SegSNR is 20.9 dB. In speech watermarking, if this value is not less than 20 dB, then imperceptibility can be guaranteed. Smaller values of LLR and IS represent better intelligibility of the synthesized watermarked speech. In our test, the averaged LLR value is 0.0078 and the averaged IS value is 0.013, indicating that the quality of the synthesized speech is very close to that of the original speech. The averaged PESQ value of the watermarked speech is 4.16, which is very close to the value for the original cover speech, indicating that the speech quality is high. These results can be expected, because only a small amount of host speech is modified in the proposed watermarking algorithm, and most of the low-frequency components are maintained to further improve the watermarked speech quality. To summarize, the results of the objective listening tests are satisfactory.

4.5 Robustness tests

The watermarked speech signal is subjected to various channel attacks. Motivated by typical applications in PSTN, cellular circuit-switch networks, VoIP, and radio communication, the following transmission channel attacks listed in Table 5 are tested. The ideal channel without any attack is also tested, which can be used as a benchmark.

Table 4 Objective listening test results

Testee	SegSNR (dB)	LLR	IS	PESQ	PESQ (cover)
Female 1	21.5	0.0066	0.012	4.23	4.28
Female 2	22.2	0.0062	0.012	4.28	4.32
Female 3	22.0	0.0075	0.011	3.99	4.02
Female 4	21.7	0.0074	0.012	4.23	4.27
Female 5	21.5	0.0071	0.013	4.27	4.29
Male 1	21.4	0.0083	0.012	4.03	4.08
Male 2	19.8	0.0086	0.015	4.17	4.21
Male 3	19.9	0.0079	0.014	3.84	3.88
Male 4	19.7	0.0087	0.015	4.23	4.26
Male 5	19.3	0.0092	0.016	4.28	4.29
Average	20.9	0.0078	0.013	4.16	4.19

Table 5 Transmission channel attacks for simulations

No.	Channel attack	PSTN	Cellular	VoIP	Radio
1	Ideal channel	×	×	×	×
2	Bandpass filtering	√	√	√	√
3	Flat-fading	×	×	×	√
4	AWGN	√	√	√	√
5	FIR linear phase filtering	√	×	×	√
6	IIR non-linear, non-minimum phase filtering	√	√	√	√
7	Transcoding	×	√	×	×

The parametric settings of these attacks are given as follows. The FIR bandpass filter has an order of N=200 and a passband ranging from 300 to 3400 Hz. The flat-fading (sinusoidal amplitude modulation) has a modulation frequency of 3 Hz and a modulation index of 0.5. The AWGN is set to be 30 dB lower than that of the original signal. The finite impulse response (FIR) linear phase filter has an order of N=150, and an equalization filter is used at the extractor. The infinite impulse response (IIR) all-pass filter has nonlinear phase and is unstable. The transcoding is assumed to occur between the enhanced full rate (EFR) vocoder of the GSM system and the pulse code modulation (PCM) coder with the A-law of the ITU-T G.711 standard.

Under these attacks, the robustness of the overall watermarking system is evaluated. The robustness in terms of the bit error ratio (BER) in the detected watermark bits is measured, without any error correction

coding being applied. The results are shown in Fig. 12. As can be seen, the BER values are mostly around or below 0.001, meaning that the proposed watermarking system is robust against various channel attacks. The BER performance under the attacks of bandpass filtering, AWGN, FIR linear phase filtering, and transcoding is very close to that of the ideal channel without any attack. The flat-fading attack and the IIR nonlinear, non-minimum phase filtering can worsen the performance slightly, because the speech signal distortion introduced by them cannot be recovered. Among the four real applications, the VoIP case has the best performance while the radio case has the worst performance because more attacks are imposed on the watermarked speech.

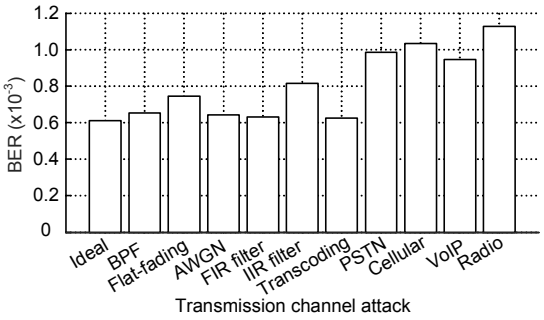

Fig. 12 Overall system robustness tests in the presence of various transmission channel attacks at a bit rate of 200 bits/s

4.6 Discussion

4.6.1 System performance for different passband widths

For binary waveform communication, the data rate is closely related to the bandwidth used. In this experiment, BERs with different passband widths are tested (Fig. 13). As can be seen, BER decreases when the data rate increases. This can be explained as follows. When the data rate becomes higher, more original passband speech is replaced by the hidden watermark signal, thereby introducing more distortions to the speech and making it more difficult to estimate the vocal tract.

Moreover, it is shown in this figure that the BER values are all around or below 0.001, indicating that reliable communication at the rate of below 400 bits/s can be accomplished.

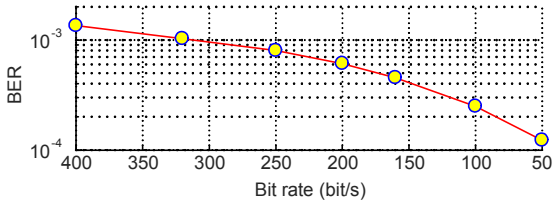

Fig. 13 BER performance at different data rates

4.6.2 Comparison with current algorithms

In the speech watermarking scenario, the main concerns are the channel capacity, the imperceptibility of the watermarked speech, the robustness against channel attacks, and the algorithm complexity. For a numerical comparison of our channel capacity results with the reported performance of other methods, R/W is used as a measure, which is based on the number of embedded watermark bits (in terms of R using the reported BER and Eq. (19)) and the channel bandwidth W. Some commonly used speech watermarking methods, including spread spectrum (SS), phase modulation (PM), and QIM, are used here for comparison. For narrowband speech watermarking, most of their achieved bit rates range from several bits to dozens of bits per second (Suzuki *et al.*, 1997; Hofbauer *et al.*, 2005; Faundez-Zanuy *et al.*, 2010; Chen *et al.*, 2013). The results of the channel capacity comparison are listed in Table 6, indicating that our method outperforms most of the state-of-the-art speech watermarking methods in terms of the data rate and channel capacity. Although the methods proposed by Hofbauer and Kubin (2006) and Hofbauer and Hering (2007) have higher data rates, their bandwidth efficiency is less than ours. Furthermore, under the constraint of the same data rate, the performance of QPSK outperforms that of BPSK in two aspects. First, lower BERs are obtained in QPSK. In this case, only a half passband is replaced when compared to BPSK. By doing so, the cover speech is less distorted, thereby resulting in better vocal tract estimates. Second, when compared to BPSK, the bandwidth efficiency in QPSK is nearly doubled. This is due to the fact that, if data symbol with higher dimensionality is used, then the transmission bandwidth needed is less.

As for imperceptibility, time and frequency domain waveforms of the synthesized watermarked speech are shown respectively. The speech signal can have one frame or multiple frames. The testing results

Table 6 Channel capacity performance comparisons

Method	Reference	C (bit/s)	BER (%)	R (bit/s)	W (kHz)	R/W
SS	Suzuki *et al.* (1997)	3	0.1	3	3.0	1
SS	Hofbauer *et al.* (2005)	24	0.1	24	2.8	8
QIM	Faundez-Zanuy *et al.* (2010)	36	0	36	4.0	9
QIM	Chen *et al.* (2013)	66	0	66	4.0	16.5
PM	Cheng and Sorensen (2001)	800	28	111	4.0	28
PM	Hofbauer and Kubin (2006)	2000	2.5	1663	4.0	416
PM	Hofbauer and Hering (2007)	2000	0.8	1866	4.0	466
Proposed (BPSK)		200	0.061	199	0.20	992
		300	0.103	316	0.32	988
		400	0.135	394	0.40	496
Proposed (QPSK)		200	0.053	198	0.10	1987
		300	0.092	317	0.16	1979
		400	0.107	395	0.20	1976

demonstrate that our method can make the watermarked speech very similar to the original speech. These results can rarely be seen in other studies. Furthermore, the watermarked speech is tested using MOS with values from zero to five. As mentioned earlier, 20 subjects are asked to listen blindly to the original and watermarked speech signals. They then report the differences between the quality of the original and that of the watermarked speech signals. The results for the average values of these reports in terms of dissimilarities, as well as some published results in the state-of-the-art literature, are presented in Table 7. As can be seen, the proposed algorithms have better imperceptibility performance than others, and the watermark signal can hardly be perceived by humans in the watermarked speech. Moreover, when the data rate increases from 200 bits/s with BPSK to 400 bits/s with QPSK, the MOS value only slightly decreases from 4.92 to 4.90.

Table 7 Subjective comparisons of different watermarking techniques

Watermark technique	Reference	MOS
AbS	Yan and Guo (2013)	4.11
Genetic algorithm	Zamani and Manaf (2015)	4.51
LSF	Wang and Unoki (2015)	4.67
DT-CWT	Fan *et al.* (2013)	4.88
Proposed (BPSK, 200 bits/s)	–	4.92
Proposed (QPSK, 400 bits/s)	–	4.90

Third, most transmission channels used in previous works are assumed to be digital channels, but some practical channels have not been considered before, except for the speech watermarking scheme designed for the analog flat-fading channel in Hofbauer *et al.* (2009). Our research is more application-specific, and four real application scenarios are considered according to their attacks on the watermarked speech, including PSTN, radio, cellular, and VoIP.

Finally, the SS- and QIM-based speech watermarking algorithms usually use the orthogonal transformations such as DFT, DCT, and DHT (Chen *et al.*, 2013; Nematollahi *et al.*, 2015a; 2015b; Sarreshtedari *et al.*, 2015), because these transformations can be implemented by fast computing methods. In our method, the heaviest burden is introduced by the computation of the LP parameters, but these can be implemented by the Levinson-Durbin method and is not a problem now for hardware or software accomplishment. For example, the radix-2 fast Fourier transform (FFT) requires $\log_2 N$ stages, $N\log_2(N/2)$ complex multiplications, and $N\log_2 N$ complex additions. On the other hand, when the LP coefficients are computed, the most popular method is the Levinson-Durbin algorithm, and its complexity is $p(p+2)$, where p is the prediction order (Malepati, 2010). To compare their complexities, N=256 and p=10 are chosen because these values are very common in engineering usage. In this case, the radix-2 FFT method needs 3072 multiply-and-accumulate (MAC) operations, while the Levinson-Durbin method needs only 120 MAC operations.

When compared to other LP-based watermarking algorithms, for example, that proposed by Hofbauer *et al.* (2009) where the LP technique is also used, our proposed method is more advantageous and has lower computation complexity. First, Hofbauer *et al.* (2009) used only the unvoiced speech segments to embed the watermark, so the speech watermarking complexity is increased in that using additional unvoiced/voiced (UV) speech detection methods becomes a must. Even worse, the detection results during the extraction process are likely to be different from those present before embedding due to channel attacks, which will further make recovering the watermark more complicated. In our work, both voiced and unvoiced speech segments are used, which obviously decreases the complexity of the algorithm. Second, the watermark generation in Hofbauer *et al.* (2009) is very intricate because it is based on the theories of non-uniform sampling, and sampling and interpolation of band-limited signals. In our method, the conventional passband communication theory is introduced, so it is much easier to implement.

5 Conclusions

We have proposed a new speech watermarking scheme by replacing the high-frequency parts of the host speech and keeping the low-frequency parts of the speech unchanged. At the embedder, the watermark bits are mapped into a passband watermark signal. This signal is first constrained to have the same energy as the passband excitation of the host speech using the masking thresholds, which ensures perceptual transparency of the watermark signal, and then this signal is pulse-shaped using the LP coefficients. The watermarked speech is synthesized using the stopband speech and the passband scaled and shaped watermark signal. At the receiver, the watermark signal is equalized using the vocal tract information derived from the watermarked speech, and the watermark bits are detected thereafter. Experimental results verify that the proposed method can obtain high embedding capacity, maintain the imperceptibility of the watermark signal, and resist several typical attacks in real applications. In the future we will consider the generation of the passband watermark signal using the multi-carrier modulation technique to further improve the capacity.

References

Cai, L.B., Tu, R.H., Zhao, J.Y., *et al.*, 2007. Speech quality evaluation: a new application of digital watermarking. *IEEE Trans. Instrum. Meas.*, **56**(1):45-55. http://dx.doi.org/10.1109/TIM.2006.887773

Chen, S., Leung, H., 2006. Concurrent data transmission through PSTN by CDMA. IEEE Int. Symp. on Circuits and Systems, p.3001-3004. http://dx.doi.org/10.1109/ISCAS.2006.1693256

Chen, S., Leung, H., Ding, H., 2007. Telephony speech enhancement by data hiding. *IEEE Trans. Instrum. Meas.*, **56**(1):63-74. http://dx.doi.org/10.1109/TIM.2006.887409

Chen, Z., Zhao, C., Geng, G., *et al.*, 2013. An audio watermark-based speech bandwidth extension method. *EURASIP J. Audio Speech Music Process.*, **2013**(1):1-8. http://dx.doi.org/10.1186/1687-4722-2013-10

Cheng, Q., Sorensen, J., 2001. Spread spectrum signaling for speech watermarking. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.1337-1340. http://dx.doi.org/10.1109/ICASSP.2001.941175

Eslami, R., Deller, J.R.Jr, Radha, H., 2006. On the detection of multiplicative watermarks for speech signals in the wavelet and DCT domains. IEEE Int. Conf. on Multimedia and Expo, p.1369-1372. http://dx.doi.org/10.1109/ICME.2006.262793

Fan, M.Q., Liu, P.P., Wang, H.X., *et al.*, 2013. A semi-fragile watermarking scheme for authenticating audio signal based on dual-tree complex wavelet transform and discrete cosine transform. *Int. J. Comput. Math.*, **90**(12): 2588-2602. http://dx.doi.org/10.1080/00207160.2013.805752

Faundez-Zanuy, M., Hagmüller, M., Kubin, G., 2006. Speaker verification security improvement by means of speech watermarking. *Speech Commun.*, **48**(12):1608-1619. http://dx.doi.org/10.1016/j.specom.2006.06.010

Faundez-Zanuy, M., Hagmüller, M., Kubin, G., 2007. Speaker identification security improvement by means of speech watermarking. *Patt. Recogn.*, **40**(11):3027-3034. http://dx.doi.org/10.1016/j.patcog.2007.02.016

Faundez-Zanuy, M., Lucena-Molina, J.J., Hagmüller, M., 2010. Speech watermarking: an approach for the forensic analysis of digital telephonic recordings. *J. Forens. Sci.*, **55**(4):1080-1087. http://dx.doi.org/10.1111/j.1556-4029.2010.01395.x

Malepati, H., 2010. Digital Media Processing: DSP Algorithms Using C. Elsevier, Burlington, USA, p.416-431. http://dx.doi.org/10.1016/B978-1-85617-678-1.00008-9

Hofbauer, K., Hering, H., 2007. Noise robust speech watermarking with bit synchronisation for the aeronautical radio. *LNCS*, **4567**:252-266. http://dx.doi.org/10.1007/978-3-540-77370-2_17

Hofbauer, K., Kubin, G., 2006. High-rate data embedding in unvoiced speech. INTERSPEECH, p.241-244.

Hofbauer, K., Hering, H., Kubin, G., 2005. Speech watermarking for the VHF radio channel. EUROCONTROL Innovative Research Workshop and Exhibition:

Envisioning the Future, p.215-220.

Hofbauer, K., Kubin, G., Kleijn, W.B., 2009. Speech watermarking for analog flat-fading bandpass channels. *IEEE Trans. Audio Speech Lang. Process.*, **17**(8):1624-1637. http://dx.doi.org/10.1109/TASL.2009.2021543

Nematollahi, M.A., Al-Haddad, S.A.R., 2013. An overview of digital speech watermarking. *Int. J. Speech Technol.*, **16**(4):471-488. http://dx.doi.org/10.1007/s10772-013-9192-6

Nematollahi, M.A., Gamboa-Rosales, H., Akhaee, M.A., *et al.*, 2015a. Robust digital speech watermarking for online speaker recognition. *Math. Probl. Eng.*, **2015**:372398. http://dx.doi.org/10.1155/2015/372398

Nematollahi, M.A., Akhaee, M.A., Al-Haddad, S.A.R., *et al.*, 2015b. Semi-fragile digital speech watermarking for online speaker recognition. *EURASIP J. Audio Speech Music Process.*, **2015**(1):1-15. http://dx.doi.org/10.1186/s13636-015-0074-5

Nematollahi, M.A., Vorakulpipat, C., Rosales, H.G., 2017. Digital Watermarking: Techniques and Trends. Springer, Singapore, p.39-51. http://dx.doi.org/10.1007/978-981-10-2095-7

Park, C.M., Thapa, D., Wang, G.N., 2007. Speech authentication system using digital watermarking and pattern recovery. *Patt. Recogn. Lett.*, **28**(8):931-938. http://dx.doi.org/10.1016/j.patrec.2006.12.010

Sarreshtedari, S., Akhaee, M.A., Abbasfar, A., 2015. A watermarking method for digital speech self-recovery. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **23**(11):1917-1925. http://dx.doi.org/10.1109/TASLP.2015.2456431

Suzuki, J., Hingdi, B., Yashima, H., 1997. Transmission of data on analog speech channel by spread spectrum modulation. IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing, p.697-700. http://dx.doi.org/10.1109/PACRIM.1997.620355

Wang, S.B., Unoki, M., 2015. Speech watermarking method based on formant tuning. *IEICE Trans. Inform. Syst.*, **E98D**(1):29-37. http://dx.doi.org/10.1587/TRANSINF.2014MUP0009

Yan, B., Guo, Y.J., 2013. Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization. *Multim. Tools Appl.*, **67**(2):383-405. http://dx.doi.org/10.1007/s11042-011-0861-7

Zamani, M., Manaf, A.B.A., 2015. Genetic algorithm for fragile audio watermarking. *Telecommun. Syst.*, **59**(3):291-304. http://dx.doi.org/10.1007/s11235-014-9936-x

Zheng, W.X., 2005. Fast identification of autoregressive signals from noisy observations. *IEEE Trans. Circ. Syst. II*, **52**(1):43-48. http://dx.doi.org/10.1109/TCSII.2004.838435