Nematollahi *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:31
DOI 10.1186/s13636-015-0074-5

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

RESEARCH Open Access

Semi-fragile digital speech watermarking for online speaker recognition

Mohammad Ali Nematollahi[1*], Mohammad Ali Akhaee[2], S. A. R. Al-Haddad[1] and Hamurabi Gamboa-Rosales[3]

Abstract

In this paper, a semi-fragile and blind digital speech watermarking technique for online speaker recognition systems based on the discrete wavelet packet transform (DWPT) and quantization index modulation (QIM) has been proposed that enables embedding of the watermark within an angle of the wavelet's sub-bands. To minimize the degradation effects of the watermark, these sub-bands were selected from frequency ranges where little speaker-specific information was available (500–3500 Hz and 6000–7000 Hz). Experimental results on the TIMIT, MIT, and MOBIO speech databases show that the degradation results for speaker verification and identification are 0.39 and 0.97 %, respectively, which are negligible. In addition, the proposed watermark technique can provide the appropriate fragility required for different signal processing operations.

Keywords: Digital speech watermarking; Online speaker recognition; Discrete wavelet packet transform; Quantization index modulation

1 Introduction

Speaker recognition systems must have sufficient security and robustness to operate in real-world environments [1]. However, there are potential vulnerabilities that threaten the use of online speaker recognition systems. In [2], eight points of vulnerability in this type of online biometric system are discussed. These systems are vulnerable to attack because unsecured transmission channels are used. However, the systems can be protected and secured against these attacks by time stamps and watermarking. Recently, speech watermarking has been used to secure communication channels for speaker verification and identification against both intentional and unintentional attacks [3–6]. For this purpose, the watermark is embedded to verify both the authenticity of the transmitter (i.e., using sensor and feature extractors) and the integrity of the entire authentication mechanism. Basically, either reversible or irreversible watermarking can be applied to ensure authenticity and integrity. Invertible aspects are not usually required because spoken language is not very fragile when subjected to bit changes in the lower layers. However, invertibility is important when very small changes in speech can have an effect. This may be the case when a digital copy is made of an analog recording, and assumptions about cuts in the analog media are later made based on the digital copy. In addition, a semi-fragile speech watermark cannot be reversible through a channel because the scheme is highly fragile; the original signal can only be reproduced if there are no changes, which seems unlikely because of channel effects [7]. Therefore, the semi-fragile watermark should be tied intrinsically to the speaker biometrics for tamper detection, and any attempts to tamper with the speech should destroy the semi-fragile watermark. However, application of speech watermarking can seriously degrade the recognition performance. The main aim of speaker recognition technologies is to enhance the recognition performance, and use of watermarking technology in this context is thus questionable because of the potential degradation effects on recognition performance. Currently available speech watermarking techniques [8, 9] embed the watermarks in a specific frequency range or in the speech formants. However, these techniques can seriously degrade speaker recognition performance. Also, watermarking and speaker recognition systems have opposing goals whenever the signal-to-watermark ratio (SWR) is reduced and the robustness of the watermark is increased, and the speaker identification and verification

* Correspondence: greencomputinguae@gmail.com
[1] Department of Computer & Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, 43400 Selangor Darul Ehsan, Malaysia
Full list of author information is available at the end of the article

performance can be reduced [3–5, 10]. Some researchers therefore apply semi-fragile watermarking methods to reduce the effects on recognition performance [11]. Basically, speaker recognition is applied in forensic applications that need to recognize the owner of a speech signal from the speech signal itself. Semi-fragile speech watermarking can be applied to detect tampering in speech signals when passing through unsecured communication channels. Watermarking of specific spectral regions that are not dependent on the speaker voice characteristics is not in direct conflict with speaker biometric recognition processes and is thus a valid approach for speaker authentication.

In this paper, a novel digital speech watermarking technique that uses the discrete wavelet packet transform (DWPT) and quantization index modulation (QIM) is proposed for online speaker recognition systems. For this application, the watermark bits are embedded at locations where fewer speaker-specific sub-bands are available. Basically, the discriminative speaker features are contained within the low- and high-frequency bands: the glottis frequency range is between 100 and 400 Hz, the piriform fossa range is between 4 and 5 kHz, and the constriction of consonants occurs at 7.5 kHz [12–14].

The rest of this paper is organized as follows. First, the most relevant studies in speech watermarking are reviewed, and the applied methodology is then discussed. The proposed semi-fragile digital speech watermarking algorithm is then explained, and experimental results for the proposed digital speech watermarking method are evaluated. The effects of the proposed digital semi-fragile speech watermarking technique on speaker recognition performance are described, and a discussion of the semi-fragility property in communication channel transmission applications is presented. Finally, conclusions are drawn and future trends are discussed.

2 Literature review

While the prior art has been reviewed in previous studies [15, 16], it would be useful to compile a summary of the more relevant developments in speech watermarking for analog and digital media. Therefore, the main studies in speech watermarking technology are discussed in this section.

In [17], a high capacity speech watermark was proposed based on replacement of the linear predictive (LP) residual with a watermark pulse. This study was then extended to tackle the noise and synchronization issues that arise in aeronautical voice radio channels in [18]. In this approach, the watermark is embedded in the unvoiced parts of the narrowband speech signal by shaping of the LP-residual pulse. In addition, a simple structure with low-complexity spectral line bit synchronization has been developed for analog channels. Another study applied speech watermarking to enhance the intelligibility and quality of speech by extending the bandwidth of the public switched telephone network (PSTN), which is in the 200- to 3500-Hz range [19]. For this reason, imperceptible spectrum components of the PSTN have been removed to enable audible components to be embedded outside the PSTN bandwidth and thus extend the PSTN bandwidth. Therefore, each audible component is multiplied to produce a specific, orthogonal, and high-autocorrelation pseudo-noise (PN) code to spread out the hidden channel. In [20, 21], speech watermarks with synchronization have been proposed using outer q-ary low-density parity-check (LDPC) codes and an inner insertion, deletion, and substitution (IDS) code. For watermarking, the average pitch is modified by QIM, which is then incorporated in the watermarked speech signal by pitch synchronous overlap-add (PSOLA). Synchronization and error recovery are performed by IDS, which is separated from the embedding and extraction phases.

Apart from these robust speech watermarking approaches, few semi-fragile audio and speech watermarking studies have been conducted. In [22], a semi-fragile audio watermarking approach was developed based on the dual-tree complex wavelet transform (DT-CWT) and the discrete cosine transform. In this technique, the DC value of the low-frequency signal is quantized by QIM to carry the watermark bits. In [23, 24], semi-fragile speech watermarking based on manipulation of the bandwidth of the speech formants was proposed. For this approach, pairs of linear spectral frequencies (LSFs) were shifted. Also, the sharpest and second sharpest bandwidths of the speech formants were manipulated to carry the 0 and 1 watermark bits, respectively. In addition, another semi-fragile speech watermarking approach based on quantization of the linear prediction (LP) parameters has been proposed [25]. For this approach, the LP coefficients were converted into inverse sine (IS) coefficients, in which the watermark bits were embedded using QIM. To reduce the bit error rate (BER) of the developed approach due to the statistical nature of the LP parameters, the analysis by synthesis (AbS) method has been used. A genetic algorithm-based fragile audio watermarking algorithm has been developed in the time domain by substituting the watermark bits for the least significant bits [26]. Neither of these semi-fragile approaches can be successfully applied to speaker recognition because the watermark is embedded in the relevant speaker sub-bands; they are also unable to provide appropriate trade-offs among capacity, semi-fragility, and imperceptibility. Therefore, there is a need to develop an efficient semi-fragile speech watermarking technique that is not only tied intrinsically to the speaker biometrics for tamper detection to prevent intentional content manipulation through the communication channels but that also has negligible recognition performance degradation.

3 Methodology

Figure 1 shows the critical bands that have been chosen to embed the watermark. As shown in Fig. 1, the selected bands contain less speaker-specific information, which has thus led to reduced recognition performance degradation for online speaker recognition systems. In this approach, the speech signal has been decomposed into 16 critical bands by applying the DWPT. Then, eight critical bands (with numbers 2, 3, 4, 5, 6, 7, 13, and 14) where the *F*-ratio level is low were chosen to produce minimum degradation of the speaker-specific information. The *F*-ratio curve shown in Fig. 1 was captured specifically from previous work [12, 27, 28].

3.1 Digital speech watermarking technique

In this section, a semi-fragile speech watermarking technique is proposed based on angle quantization of the energy ratio between two blocks, which is highly sensitive to any manipulation. The proposed semi-fragile speech watermarking technique can provide authentication over an unknown channel and can provide imperceptible watermarking. Manipulation of the watermark signal will destroy the watermark bits, which are changed into random bit streams. Any minor manipulation of the speech signal can seriously change the angles of the signal; quantization of the signal's angles is therefore a good candidate technique for semi-fragile speech watermarking.

To apply angle quantization, each watermark bit is embedded into two sets of the original signal. For this purpose, two sets of the original signal (designated x_1 and x_2) have been selected to provide a space in a two-dimensional coordinate system. Then, the polar coordinates of (x_1, x_2) are calculated based on Eqs. (1) and (2), as shown in Fig. 2:

$$\theta = \arctan\left(\frac{x_2}{x_1}\right), \tag{1}$$

$$r = \sqrt{x_1^2 + x_2^2}. \tag{2}$$

In angle quantization, θ is quantized to embed the watermark bit. However, this technique is very fragile because, even without any attack, the watermark bits cannot be extracted and thus can cause serious errors. To overcome this problem, the watermark bits are embedded via quantization of the ratio between two energy blocks of the original signal. Only one bit is embedded into each frame by the semi-fragile digital speech watermarking technique. However, each watermark bit is repeatedly embedded into a frame to reduce the potential error. Therefore, each frame is divided into blocks with length L_b, and two sets designated X and Y are selected. Then, θ is calculated as shown in Eq. (3):

$$\theta = \arctan\left(\frac{\sum_{i=1}^{L_b/2} y_i^2}{\sum_{i=1}^{L_b/2} x_i^2}\right) \tag{3}$$

After angle quantization, the variation for Y must be estimated. In this study, the Lagrange method has been used to estimate the coefficients after angle quantization. The Lagrange method can reduce the effects of watermark distortion after angle quantization. Therefore, each watermarked coefficient is estimated by solving an

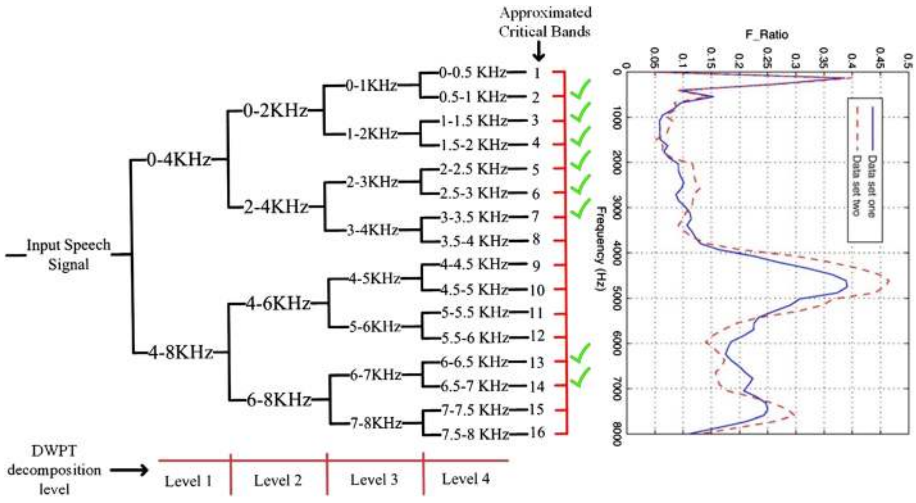

Fig. 1 Eight selected critical bands (2, 3, 4, 5, 6, 7, 13, and 14) where reduced speaker-specific information is available for watermarking by application of DWPT decomposition

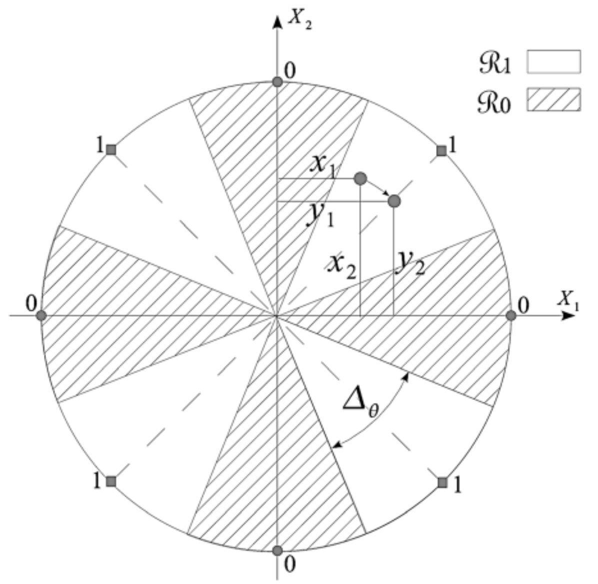

Fig. 2 Embedding of watermark by angle quantization [38]

optimization problem, which is formulated as shown in Eq. (4):

$$\begin{cases} \text{Cost: } J(Y) = \sum_{i=1}^{L_b/2} \left(y_i^{Q} - y_i\right)^2 \\ \text{Condition: } C(X) = \sum_{i=1}^{L_b/2} \left(y_i^{Q}\right)^2 - \theta^{Q} \times E_X = 0 \end{cases} . \tag{4}$$

To solve this optimization problem, the Lagrange method must be used to estimate the optimized values of the equation system, as shown in Eq. (5):

$$\nabla J(Y) = \lambda \, \nabla C(X). \tag{5}$$

These optimized values are computed simply by solving Eqs. (6) and (7):

$$y_i^{Q,\mathrm{Opt}} = \frac{y_i}{1 - \lambda_{\mathrm{Opt}}}, \tag{6}$$

$$\lambda_{\mathrm{Opt}} = 1 - \sqrt{\frac{E_Y}{\theta^{Q} \times E_X}}. \tag{7}$$

3.2 Semi-fragile digital speech watermarking algorithm

As discussed earlier, the watermark bits are embedded in specific frequency sub-bands of the DWPT. Details of the embedding and extraction process are presented in the following algorithms:

Embedding process

a) Segment the original speech signal into frames Fi with lengths of *N*.
b) Apply DWPT to each frame with *L* levels to compute the different sub-bands.
c) Select the specific frequency sub-bands in the last level, and arrange them into a data sequence.
d) Divide the data sequence into blocks with length L_b. Then, divide each block into two sets, *X* and *Y*, with equal lengths of $N/2$ for each set.
e) Compute the energy ratio for *X* and *Y* using $\frac{E_Y}{E_X}$.
f) Embed the watermark bit repeatedly into all blocks in a frame based on Eq. (8):

$$\theta^{Q} = \frac{\theta + m_i \times \Delta}{2\Delta} \times 2\Delta + m_i \times \Delta, \tag{8}$$

where Δ corresponds to the quantization step, m_i is the angle of the energy ratio, and θ^{Q} is the modified angle of the energy ratio. Use of small quantization steps provides greater imperceptibility but reduced robustness and vice versa.

g) Apply the Lagrange method to the *Y* set to make the required changes to minimize the watermarked distortion.
h) Apply the inverse DWPT to reconstruct the watermarked signal.

Figure 3 shows a block diagram of the embedding process for the proposed semi-fragile speech watermarking technique.

The extraction process is to be performed via segmentation of the watermarked speech signal. However, segmentation cannot occur when arbitrary differences in the data occur between the transmitter and the receiver. Therefore, this process can only proceed when a synchronization method is used to align the received data with the transmitted data. However, in this study, the watermarked speech signal is assumed to be always synchronized. The sizes of the frames, the quantization parameters, and the threshold value are all known at the receiver. In addition, it would be possible to use state-of-the-art synchronization techniques for this purpose.

By selecting a simple technique for the embedding process, the reverse process for extraction of the watermark is also made simple, as described in the following:

Extraction process

a) Segment the original speech signal into frames Fi with length *N*.
b) Apply the DWPT to each frame with *L* levels to compute the different sub-bands.
c) Select the specific frequency sub-bands in the last level, and arrange them into a data sequence.

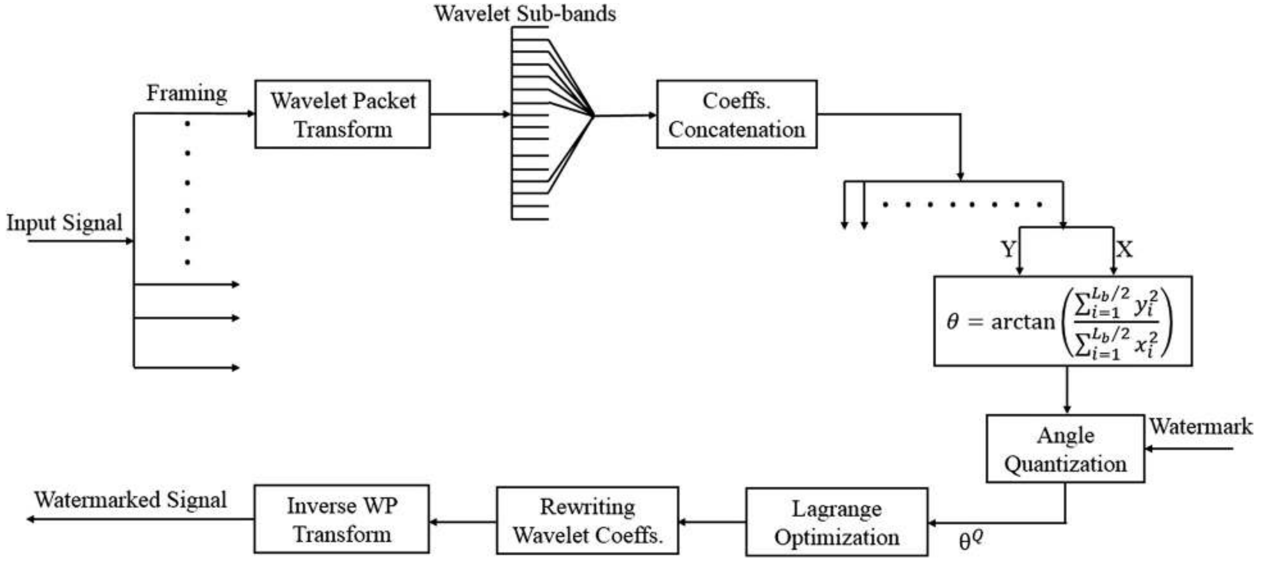

Fig. 3 Block diagram of embedding process for the proposed semi-fragile digital speech watermarking technique

d) Divide the data sequence into different blocks with length L_b. Then, divide each block into two sets, X and Y, with equal lengths of $N/2$ for each set.
e) Compute the energy ratio for X and Y, i.e., $\frac{E_Y}{E_X}$.
f) Extract the binary watermark bit from the angle θ, which is the nearest quantization step to this angle according to Eq. (9):

$$\hat{b}_\mathrm{k} = \mathrm{argmin}_{b_\mathrm{k}=\{0,1\}} \left| r_\mathrm{k} - \mathrm{Q}_{b_k}(r_k) \right|, \tag{9}$$

where r_k is the angle of the energy ratio of the received signal, and Q_{b_k} is the quantization function when meeting the watermark bits $b_\mathrm{k} = \{0,1\}$.

g) Perform steps e and f repeatedly for all blocks in the frame.

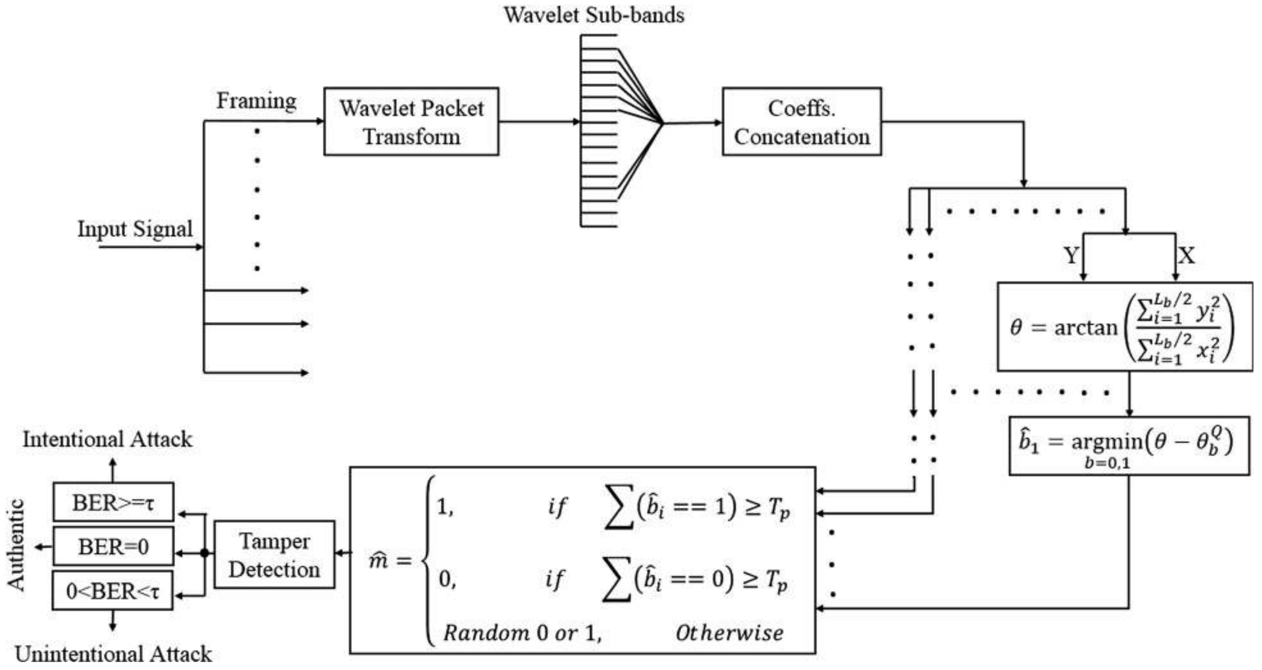
Wavelet Sub-bands
Framing
Wavelet Packet Transform
Coeffs. Concatenation
Input Signal
Y
X
$\theta = \arctan\left(\frac{\sum_{i=1}^{L_b/2} y_i^2}{\sum_{i=1}^{L_b/2} x_i^2}\right)$
$\hat{b}_1 = \underset{b=0,1}{\mathrm{argmin}}(\theta - \theta_b^Q)$
Intentional Attack
BER>=τ
BER=0
0<BER<τ
Authentic
Unintentional Attack
Tamper Detection
$\hat{m} = \begin{cases} 1, & if \quad \sum(\hat{b}_i == 1) \geq T_p \\ 0, & if \quad \sum(\hat{b}_i == 0) \geq T_p \\ Random\ 0\ or\ 1, & Otherwise \end{cases}$

Fig. 4 Block diagram of extraction process for the proposed semi-fragile digital speech watermarking technique

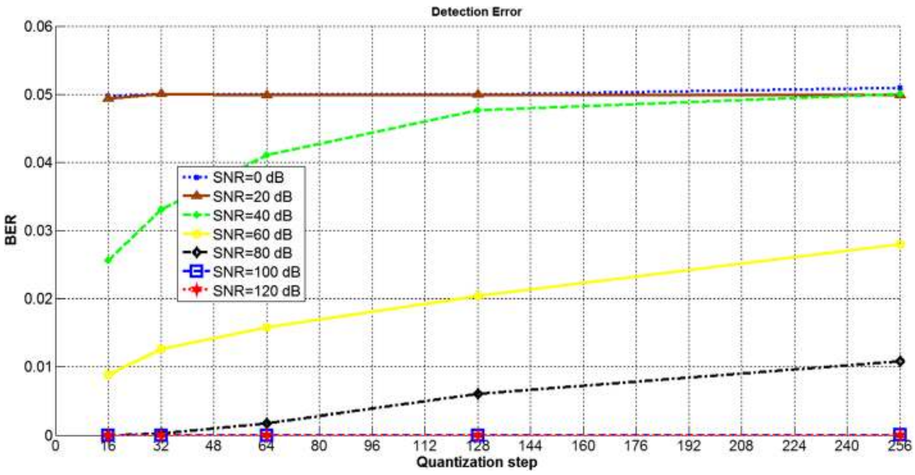

Fig. 5 Effects of quantization steps with respect to the probability of a watermark detection error for different SNRs in AWGN channels (where the quantization step is normalized by dividing by pi and reversing the denominator)

h) By embedding the same watermark bit in each block of a frame, different bits are extracted from the frame that must then be made into one bit. For this reason, a threshold has been considered to decide about the extracted bit. When the number of the extracted bits for 1 is higher than the threshold value, the extracted watermark bit is 1. Otherwise, the number of 0 bits must be higher than the number of bits for 1, and the extracted watermark bit is thus 0. Whenever the threshold is considered to be close to 1, the fragility of the developed semi-fragile system is greater. However, when this threshold is near 0.5, the robustness of the developed semi-fragile system is greater.

Figure 4 shows the block diagram of the extraction process for the proposed semi-fragile speech watermarking technique.

As shown in the diagram, a predefined threshold τ can be used to distinguish between intentional and unintentional attacks. If the extracted watermark has a BER that is higher than τ, it can be inferred that the speech signal has been modified maliciously. Otherwise, it can be inferred that the speech signal has been modified accidentally.

4 Experimental setup

Simulations were performed to evaluate the fragility of the performance of the developed semi-fragile speech watermarking technique. Therefore, the watermarking

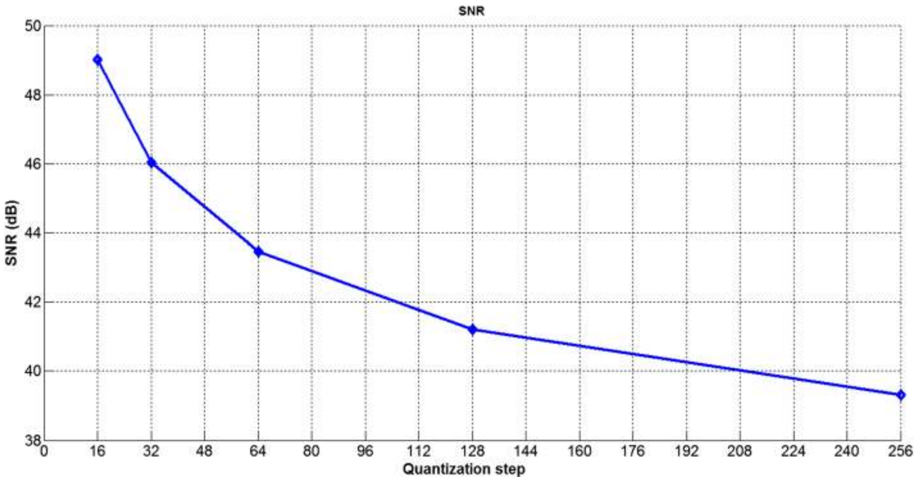

Fig. 6 Effects of quantization steps with respect to SNR (where the quantization step is normalized by dividing by pi and reversing the denominator)

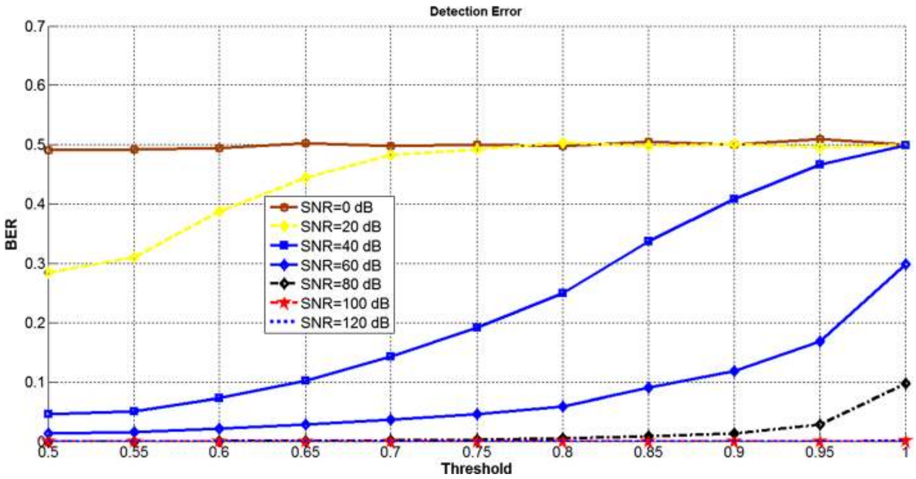

Fig. 7 Probability of watermark detection error with respect to the threshold for different SNRs of AWGN channels

technique was applied separately to evaluate its performance. The simulation results were reported based on the average results obtained for the TIMIT speech signals [29]. The simulation parameters used were as follows:

a) The size of each frame was 32 ms, which was equivalent to $F_s \times 0.032 = 512$ samples.
b) The level of the wavelet was 4. The selected sub-bands for watermarking were explained in Fig. 1. The Daubechies wavelet function was also used for the DWPT.
c) The size of each block in the frame was considered to be 8 and was equally divided such that the size of each set of X and Y in the block was 4.
d) Figures 5 and 6 show the effects of the quantization step $(\Delta = {}^{\pi}/_{256}\, to\, {}^{\pi}/_{16})$ on the BER and the signal-to-noise ratio (SNR), respectively. As shown in the figures, whenever the quantization step increased, the fragility of the watermark decreased. Also, increasing the quantization step could reduce the imperceptibility of the speech signals in terms of their SNR. To preserve both the fragility and the imperceptibility, the quantization step was assumed to be $\Delta = {}^{\pi}/_{64}$, which can be selected arbitrarily and depends on the usage of the quantization. If the usage is for data copyright protection, then a system with more quantization steps offers a more suitable model. However, if the usage is for a forensic application that needs to determine the owner of the speech signal, then a more suitable model is a system with fewer quantization steps. However, this assumption was experimentally selected to provide reasonable robustness for content preservation during attacks (such as normalize, invert, and amplify) and provide appropriate fragility for content manipulation (such

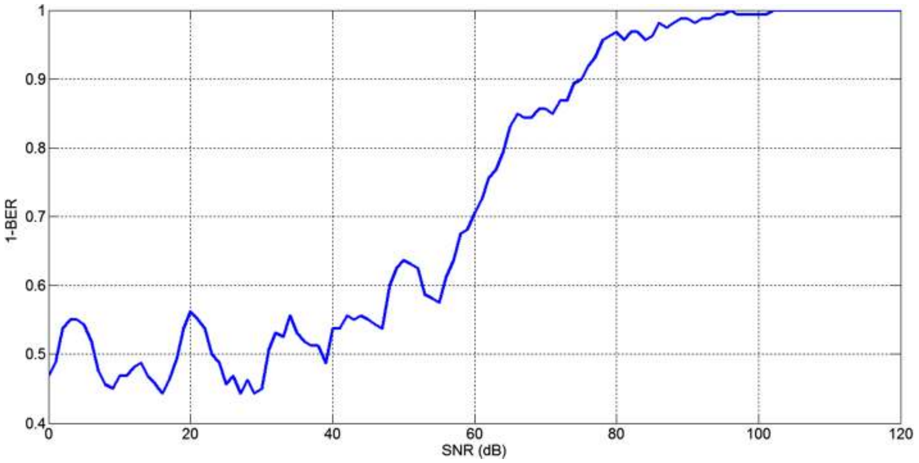

Fig. 8 Probability of correct watermark detection for different SNRs under AWGN attack

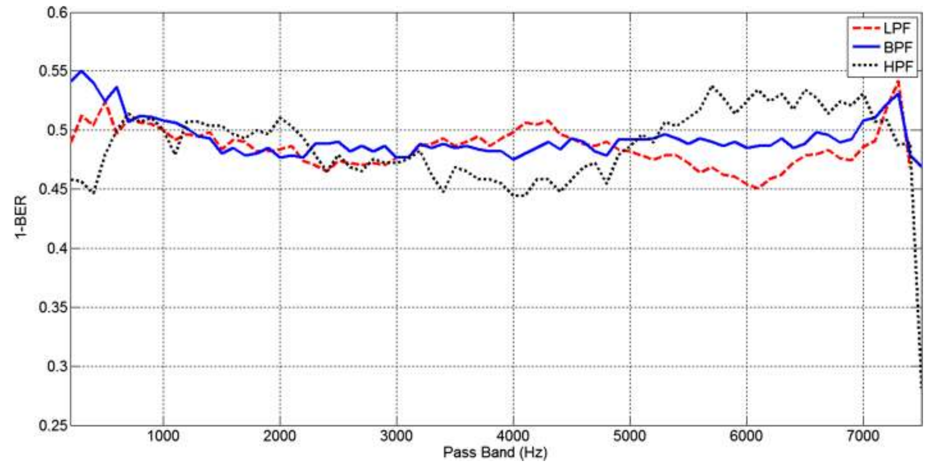

Fig. 9 Probability of correct watermark detection under different pass-bands for LPF, BPF, and HPF attacks

as filters, addition, humming, and removal). Also, it cannot degrade the recognition performance of online speaker recognition systems.

e) The decision threshold for extraction of the watermark bits was assumed to be 0.9. Figure 7 shows the effect of changing the threshold with respect to the probability of a watermark detection error for different additive white Gaussian noise (AWGN) channels. As shown, whenever the threshold was increased to 1, the fragility of the developed semi-fragile system also increased. However, if this threshold was reduced to 0.5, the robustness of the developed semi-fragile system then increased. For serious noise (i.e., where SNR = 0 dB), it emerged that the threshold could not affect the fragility of the watermark because the watermark bits were extracted in a random sequence.

Figures 5 and 6 show that the quantization step $\Delta = {}^{\pi}/_{64}$ cannot only provide semi-fragility with a reasonable BER for many AWGN channels but also can provide a high SNR, which in turn caused the high imperceptibility.

In the following, the robustness, capacity, and imperceptibility of the proposed semi-fragile speech watermarking are discussed.

4.1 Robustness

To evaluate the fragility property of the proposed semi-fragile digital speech watermarking technique, some attacks were designed, including AWGN, low-pass filter (LPF), band-pass filter (BPF), high-pass filter (HPF), median filter, and resampling attacks. Without any applied attack, the BER of the watermark was 0.

a) AWGN attack

For the AWGN attack, the watermarked speech signals were passed through the AWGN channel with different SNRs. Figure 8 shows the probability of correct detection (1 – BER) of the watermark, which ranges from

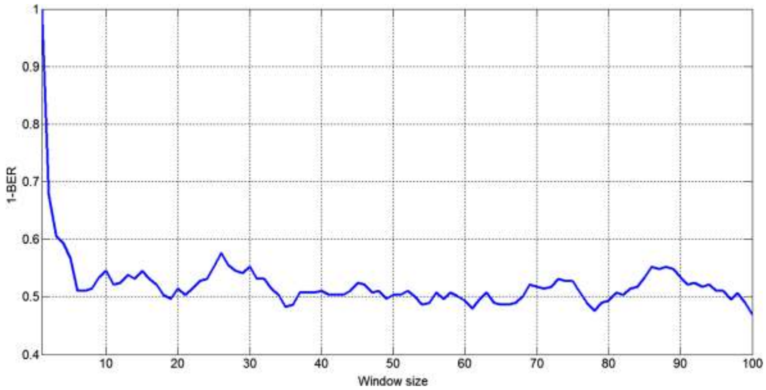

Fig. 10 Probability of correct watermark detection for various window sizes under median filter attack

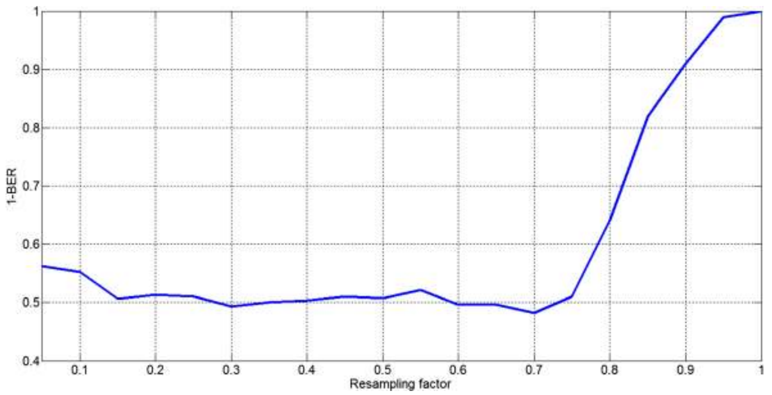

Fig. 11 Probability of correct watermark detection for different sampling factors under resampling attack

0 to 120 dB. As shown, the BER was less than 10 % for SNR = 75 dB. In addition, the watermark was extracted without any errors for all SNRs that were higher than 104 dB.

b) LPF attack

For the LPF attack, the watermarked speech signals were passed through an LPF with different pass-bands within the range from 100 to 7500 Hz. Figure 9 shows the probability of correct watermark detection for the various pass-bands. For all pass-bands, the correct detection probabilities were less than 50 %. Therefore, any manipulation via an LPF attack can be detected.

c) BPF attack

For the BPF attack, the watermarked speech signals were passed through a BPF with a bandwidth ranging from 100 to 7500 Hz and a central frequency of 4 kHz. The watermarked speech signals were filtered by changing the BPF bandwidth. Then, the watermark bits were extracted. Figure 9 shows the random nature of the extracted watermark under BPF attack.

d) HPF attack

For the HPF attack, the watermarked speech signals were passed through an HPF with a bandwidth range from 200 to 7500 Hz by selecting various bandwidths. Figure 9 shows the correct watermark detection probability of around 50 % for all bandwidths.

Table 1 BERs for various fragile speech watermarking techniques under Stirmark® attacks

Attack	Semi-fragile DWPT-QIM (proposed)	AbS [25]	LSF [23, 24]	DT-CWT [22]	Genetic [26]
No attack	0	0.03	0.05	0	0.07
AddBrumm	0.45	0.50	0.56	0.12	0.53
AddSinus	0.43	0.54	0.54	0.13	0.41
AddNoise	0.39	0.52	0.55	0.25	0.48
Stat1 (statistical distortion)	0.54	0.73	0.67	0.11	0.58
Stat2 (statistical distortion)	0.51	0.41	0.34	0.23	0.64
Smooth1 (simple smoothing)	0.61	0.59	0.49	0.32	0.49
Smooth2 (simple smoothing)	0.55	0.53	0.61	0.10	0.42
Amplify (increases amplitude)	0.02	0.42	0.57	0.13	0.36
Invert (180° phase shift)	0.58	0.49	0.43	0.04	0.54
Exchange (swap samples)	0.63	0.54	0.65	0.03	0.23
CutSamples (7 samples per 1000)	0.41	0.51	0.71	0.32	0.45
LSBZero (resets LSBs)	0.37	0.44	0.51	0.02	0.65
ZeroCross (resets samples)	0.56	0.47	0.49	0.08	0.48
ZeroRemove (removes 0 samples)	0.5	0.45	0.52	0.17	0.44
Average	0.4367	0.4780	0.5127	0.1367	0.4513

Table 2 Capacity relative to robustness and recognition rate

Capacity (bps)	SNR (dB)	Robustness (BER)
1000	41.43	0.03
500	52.98	0.01
250	62.76	0
125	85.37	0
31.25	103.58	0

e) Median filter attack

For the median filter attack, the watermarked speech signals were passed through a median filter with window sizes ranging from 1 to 100. Figure 10 shows the correct watermark detection probability for all window sizes. Apart from the case where the window size was 1, the watermark bits were extracted randomly.

f) Resampling attack

For the resampling attack, the watermarked speech signals were initially downsampled using a specific sampling factor. Then, the signals were upsampled using the previous sampling factor. Figure 11 presents the correct watermark detection probability for the resampling factor range from 1 to $\frac{1}{20}$. Apart from the smaller sampling factors, the sampling factors generally changed the extracted watermark bits randomly. This shows that the resampling factor can seriously affect the semi-fragile watermark.

As shown in Figs. 8, 9, 10, and 11, the random nature of the extracted watermark bits demonstrated the fragility property of the proposed semi-fragile digital speech watermarking technique. Therefore, any manipulation (only conventional signal processing operations were used here) of the watermarked speech signal was detected by the developed semi-fragile digital speech watermarking technique.

Because of the wide variety of attacks, the well-known and effective Stirmark® package was also used to evaluate the fragility of the proposed semi-fragile technique [30, 31]. Table 1 presents the BERs for the various semi-fragile speech watermarking techniques. Apart from "No attack" and "Amplify attack," which are considered to be unintentional attacks, the developed semi-fragile DWPT-QIM approach seems to be fragile with respect to the other intentional attacks. Even without any attack, it seems that the AbS, LSF, and genetic semi-fragile watermarking techniques always extracted watermark bits with errors. While the DT-CWT technique could extract the watermark without any errors under "No attack" conditions, it cannot be used as a semi-fragile technique because of its low average BER with respect to the other intentional attacks.

4.2 Capacity

The capacity or payload is defined as the amount of information carried by a watermarked signal for a specific amount of time. It is measured in bits per second (bps). The data capacity for this scheme is computed as shown in Eq. (11):

$$C = N_{\mathrm{sb}} \times \frac{F\mathrm{s}}{L_{\mathrm{F}}} = (1 \text{ to } 8) \times \frac{16,000}{128 \text{ to } 512} \qquad (11)$$
$$= 31.25 \text{ to } 1000\ (\mathrm{bps})\ ,$$

where C is the capacity, N_{sb} is the number of selected DWPT sub-bands for embedding, F_{s} is the sampling frequency, and L_{F} is the frame length. Whenever the size of the host speech signal and the number of frames for embedding are increased, the capacity is increased as a consequence.

While it may be possible to use error correction and repetition coding to improve data recovery, this paper focuses solely on the embedding and extraction of raw watermark binary bits. Table 2 presents the effects of the capacity on imperceptibility and robustness. As the results indicate, low bit-rate embedding can improve both imperceptibility and robustness because of the reduction in watermark bit distortion.

4.3 Imperceptibility

To compare the postulated imperceptibility with substantiated values, objective and subjective validations of the imperceptibility were performed to enable analysis of the perceptual quality of the watermarked speech signal. In this experiment, the mean opinion score (MOS) was

Table 3 MOS grades [32]

MOS	Quality	Quality scale	Effort required to understand meaning scale
5	Excellent	Imperceptible	No effort required
4	Good	Perceptible, but not annoying	No appreciable effort required
3	Fair	Slightly annoying	Moderate effort required
2	Poor	Annoying	Considerable effort required
1	Bad	Very annoying	No meaning was understood

Table 4 Comparison of different watermarking techniques in terms of their objective (SNR) and subjective (MOS) measurements, capacity (bps), EER (%), and identification rate (%)

Watermark techniques	MOS speech[b]	Speech SNR (dB)	MOS music[a]	Capacity (bps)	EER (%)	Identification rate (%)
Semi-fragile DWPT-QIM	5	43.39	5	31.25–1000	20.23	66.43
AbS [25]	4.11	28.08	3.22	33.33–50	27.23	52.98
LSF [23, 24]	4.67	30.32	3.10	33.33–50	23.14	62.76
DT-CWT [22]	4.88	31.36	3.45	15.66–976.56	39.58	46.37
Genetic algorithm [26]	4.51	29.30	3.56	N/A	42.23	49.38

[a] Effort required to understand meaning scale was applied
[b] Quality scale was applied

used because of its simplicity and availability. The International Telecommunications Union (ITU-T) [32] method for subjective measurement of speech quality based on MOS, which is presented in Table 3, was used.

In the MOS evaluation method, 10 subjects were asked to listen blindly to the original and watermarked speech signals. They then reported the differences between the quality of the original and that of the watermarked speech signals. Their levels of understanding of the speech signals are described using the terms noted in Table 3, and results for the average values of these reports on dissimilarities were computed for MOS music and MOS speech and are presented in Table 4.

5 Effects of semi-fragile watermarking on speaker recognition system performance

In this section, the effects of the proposed semi-fragile digital speech watermarking method on speaker verification performance were evaluated using two speaker verification systems: the i-vector [33, 34] and GMM-UBM [35] systems. Figures 12, 13, and 14 show the performance levels of the different speaker verification systems for the TIMIT [29], MIT [36], and MOBIO [37] speech databases, respectively.

As shown in Figs. 12 and 13, the equal error rates (EERs) for both systems and for the databases were approximately the same before and after application of semi-fragile digital speech watermarking, i.e., the performance of the speaker recognition systems decreased only slightly with semi-fragile digital speech watermarking. Also, the performance of the i-vector speaker verification system was better than that of the GMM-UBM system. This is because the i-vector system used low-dimensional feature vectors, unlike the GMM-UBM system. Also, although the recognition performance when using the LP-residual cepsrum coefficients (LPRC) was worse than that when using the mel frequency cepstral coefficients (MFCC), the LPRC performance was more

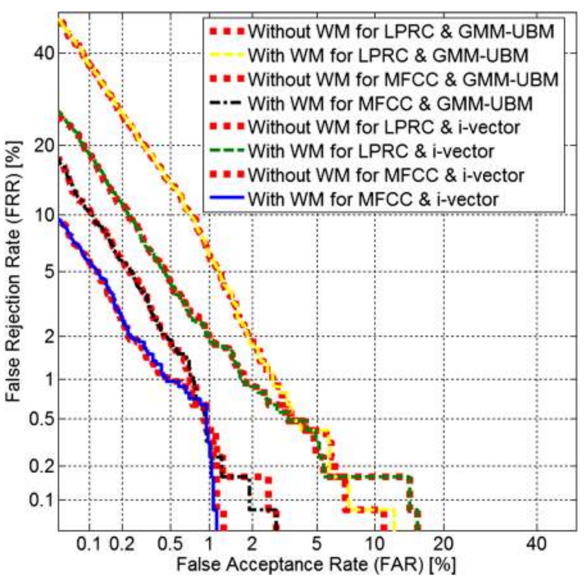

Fig. 12 Effects of semi-fragile watermarking on speaker verification performance for different features and systems in the TIMIT speech database

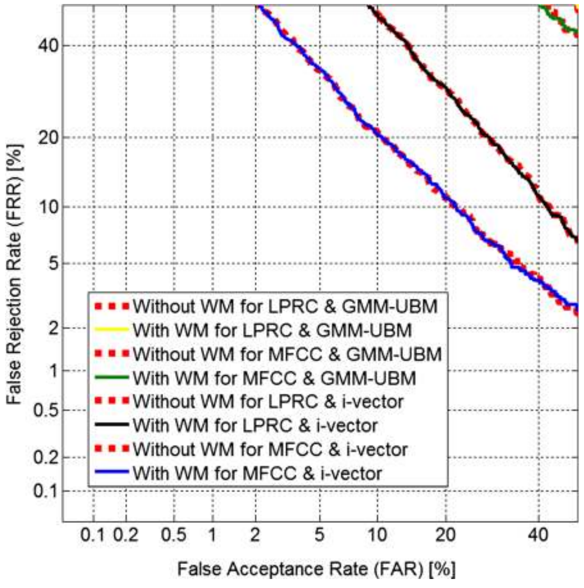

Fig. 13 Effects of semi-fragile watermarking on speaker verification performance for different features and systems in the MIT speech database

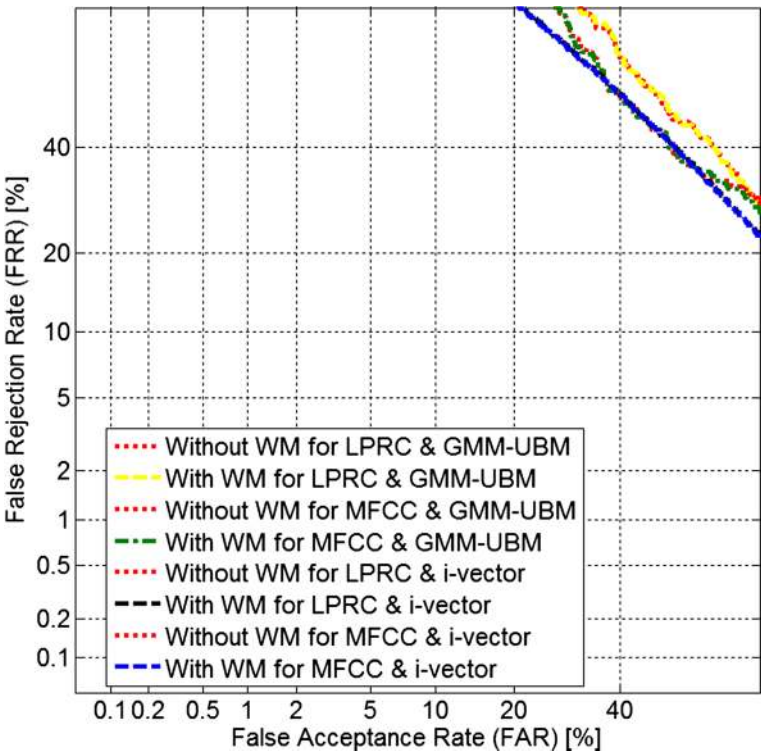

Fig. 14 Effects on semi-fragile watermarking on speaker verification performance for different features and systems in the MOBIO speech database

robust than that of the MFCC when using semi-fragile watermarking. Because a minority of the MFCC features were extracted from the frequency area where semi-fragile watermarking had already been applied, the MFCC performance was not seriously degraded. In addition, the linear predictor coefficients (LPCs) vary even under clean conditions, which can affect the LPRC features. However, semi-fragile watermarking does not affect the LP residual seriously because the LPRC features are extracted from the LP residual.

As shown in Fig. 14, the EERs for both systems and for the database were approximately the same before and after application of semi-fragile digital speech watermarking, i.e., the performance of the speaker recognition systems again decreased only slightly with semi-fragile digital speech watermarking. Also, the performance of the i-vector's speaker verification system was again better than that of the GMM-UBM system. This is because the i-vector system used low-dimensional feature vectors, unlike the GMM-UBM system. Also, although the recognition performance when using the LPRC was worse than that when using the MFCC, the LPRC was more robust than the MFCC when using semi-fragile watermarking. Because a minority of the MFCC features were extracted in the frequency area where semi-fragile watermarking had already been applied, the MFCC performance was not significantly degraded. In addition, the LPCs vary even under clean conditions, which can affect the LPRC features. However, semi-fragile watermarking again does not affect the LP residual seriously because the LPRC features are extracted from the LP residual.

Table 5 shows the effect of semi-fragile digital speech watermarking on the performance of the different

Table 5 Effects of semi-fragile watermarking on speaker verification performance for different speech databases

Database	System	EER (%) without fragile WM	EER (%) with fragile WM
TIMIT	i-vector + MFCC	0.71	0.74
	i-vector + LPRC	1.45	1.46
	GMM-UBM + MFCC	0.79	0.80
	GMM-UBM + LPRC	1.90	1.90
MIT	i-vector + MFCC	15.04	15.17
	i-vector + LPRC	24.20	24.26
	GMM-UBM + MFCC	46	46.02
	GMM-UBM + LPRC	49.43	49.93
MOBIO	i-vector + MFCC	45.96	45.96
	i-vector + LPRC	46	46
	GMM-UBM + MFCC	46.66	46.66
	GMM-UBM + LPRC	47	47

Table 6 Effect of semi-fragile watermarking on speaker identification performance for different speech databases

Feature	Speech database	Recognition rate (%) without fragile WM	Recognition rate (%) with fragile WM
MFCC	TIMIT	94.36	93.57
	MIT	54.42	53.13
	MOBIO	51.32	51.30
LPRC	TIMIT	88.80	86.98
	MIT	47.56	46.64
	MOBIO	45.87	45.86

speaker verification systems for the different speech databases.

As shown in Table 5, the best results have been reported for the TIMIT speech database, which is a clean speech database. Because of mismatches in the channel, the microphone, and the environment, the other databases demonstrated poorer performance levels than TIMIT. Table 5 also shows that i-vector with MFCC outperformed the other speaker verification systems. Also, semi-fragile speech watermarking has a negligible effect on LPRC, which is the source feature. From Table 5, the total effect of semi-fragile digital speech watermarking on the EER was calculated to be 0.39 %, as shown in the calculation below. This amount is very small and shows that the semi-fragile digital speech watermarking method has negligible degradation effects on the performance of online speaker recognition systems.

$$\frac{\begin{array}{r}|0.71-0.74|+|1.45-1.46|+|0.79-0.80|+|15.04-15.17|\\ +|24.20-24.26|+|46-46.02|+|49.43-49.93|\end{array}}{7}=0.3914\,\%$$

Table 6 presents the effects of the developed semi-fragile digital speech watermarking system on the performance of the GMM speaker identification system in terms of recognition rate. As shown in the table, the best recognition rates were reported for the TIMIT speech database, which is a clean speech database. Because of mismatches in the channel, the microphone, and the environment, the other databases had poorer recognition rates than TIMIT. In addition, the MFCC outperformed the LPRC. From Table 6, the total degradation effect of semi-fragile digital speech watermarking on the recognition rate was calculated to be 0.97 %, as shown in the calculation below. Therefore, the degradation effects of semi-fragile digital speech watermarking are negligible.

$$\frac{\begin{array}{r}|94.3651-93.57|+|54.42-53.13|+|51.32-51.30|\\ +|88.80-86.98|+|47.56-46.64|\end{array}}{5}=\ 0.9690\ \%$$

6 Discussion

Online speaker recognition systems are used in channels full of lossy compression, such as GSM (Global System for Mobile Communications), MPEG (Moving Picture Experts Group), or at least adaptive differential pulse-code modulation (ADPCM) channels. A semi-fragile watermark that is not broken in a normal distribution channel is highly desirable for improved communication channel security. Therefore, the quantization step (Δ), the threshold (T_{P}), and the block length (L_{b}) of the semi-fragile speech watermarking system should be selected such that they provide a tradeoff between communication channel security and recognition performance for the speaker recognition system. Table 7 presents BER data for each of the watermark parameters for various communication channels. As shown in the table, whenever Δ is increased and T_{P} and L_{b} are reduced, the robustness of the watermark is increased. These parameters can change the functionality of the proposed watermark from fragile to robust. However, the best watermarking approach, which provides security against communication channel attack, authentication, and tamper detection, is the semi-fragile speech watermarking approach. The semi-fragile speech watermark can survive in the channel if the watermarking parameters are set appropriately.

Table 4 presents a comparison of recent semi-fragile watermarking techniques in terms of their average subjective, objective, capacity, and recognition performances for the TIMIT, MIT, and MOBIO speech databases. The proposed semi-fragile speech watermarking technique is shown to be more efficient than the other techniques in

Table 7 BER data for each watermark parameter for various communication channels

Communication channel	Quantization step (Δ)			Threshold (T_{P})			Block length (L_{b})		
	$\Delta={}^{\pi}/_{16}$	$\Delta={}^{\pi}/_{64}$	$\Delta={}^{\pi}/_{256}$	$T_{\mathrm{P}}=0.5$	$T_{\mathrm{P}}=0.75$	$T_{\mathrm{P}}=1$	$L_{\mathrm{b}}=4$	$L_{\mathrm{b}}=8$	$L_{\mathrm{b}}=16$
16 bit PCM	0	0	0	0	0	0	0	0	0
G.711 A-law	0	0	0	0	0	0	0	0	0
G.711 μ-law	0	0	0	0	0	0	0	0	0
ADPCM	0	0	0	0	0	0	0	0	0
GSM	0.04	0.13	0.16	0.17	0.33	0.42	0.03	0.24	0.39
MPEG	0.14	0.27	0.35	0.08	0.11	0.28	0.08	0.10	0.34

terms of imperceptibility, recognition rate, and capacity. In addition, the imperceptibility was high after embedding. Therefore, the semi-fragile digital speech watermark does not degrade the speaker recognition performance.

7 Conclusions

In this study, a new semi-fragile digital speech watermarking technique was implemented by application of DWPT and angle quantization. This watermarking technique is fragile against various attacks, including filtering, additive noise, cut sampling, and compression attacks. The degradation effect on the recognition performance of this watermarking technique is negligible. In addition, any intentional or unintentional tampering with the watermarked speech signal can easily be detected via a tampering threshold because the watermark is embedded in the least speaker-specific of the speech sub-bands.

Future work in this area is likely to include a study of new adaptive quantization techniques. Also, a synchronization technique for this approach could also improve the watermark extraction process.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments that enabled us to improve the quality of this paper.

Author details

[1]Department of Computer & Communication Systems Engineering, Faculty of Engineering, Universiti Putra Malaysia, UPM Serdang, 43400 Selangor Darul Ehsan, Malaysia. [2]Department of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran. [3]Faculty of Electrical Engineering, Autonomous University of Zacatecas, Zacatecas, Mexico.

Received: 27 May 2015 Accepted: 1 October 2015
Published online: 21 October 2015

References

1. MA Nematollahi, S Al-Haddad, Distant speaker recognition: an overview. Int. J. Humanoid Rob. **12**(03), 1–45 (2015)
2. Z Wu et al., Spoofing and countermeasures for speaker verification: a survey. Speech Comm. **66**, 130–153 (2015)
3. M Faundez-Zanuy, M Hagmüller, G Kubin, Speaker identification security improvement by means of speech watermarking. Pattern Recognit. **40**(11), 3027–3034 (2007)
4. M Faundez-Zanuy, M Hagmüller, G Kubin, Speaker verification security improvement by means of speech watermarking. Speech Comm. **48**(12), 1608–1619 (2006)
5. W Al-Nuaimy et al., An SVD audio watermarking approach using chaotic encrypted images. Digital Signal Process. **21**(6), 764–779 (2011)
6. SARS AL-HADDAD, M Iqbal, AR RAMLI, MA NEMATOLLAHI, *A Method for Speech Watermarking in Speaker Verification*, 2015. Google Patents
7. M Steinebach, J Dittmann, Watermarking-based digital audio data authentication. EURASIP Journal on Applied signal processing **2003**, 1001–1015 (2003)
8. M Faundez-Zanuy, JJ Lucena-Molina, M Hagmüller, Speech watermarking: an approach for the forensic analysis of digital telephonic recordings. J. Forensic Sci. **55**(4), 1080–1087 (2010)
9. K Hofbauer, G Kubin, WB Kleijn, Speech watermarking for analog flat-fading bandpass channels. IEEE Trans. Audio Speech Lang. Process. **17**(8), 1624–1637 (2009)
10. Baroughi, AF and Craver S (2014). Additive attacks on speaker recognition. in IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics.
11. Hämmerle-Uhl J, Raab K, and Uhl A (2011). Watermarking as a means to enhance biometric systems: a critical survey. in Information Hiding. Springer.
12. X Lu, J Dang, An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Comm. **50**(4), 312–322 (2008)
13. S Hyon, An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution, in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012* (IEEE, Asia-Pacific, 2012)
14. L Besacier, J-F Bonastre, C Fredouille, Localization and selection of speaker-specific information with statistical modeling. Speech Comm. **31**(2), 89–106 (2000)
15. MA Nematollahi, S Al-Haddad, An overview of digital speech watermarking. Int. J. Speech Technol. **16**(4), 471–488 (2013)
16. F Djebbar et al., Comparative study of digital audio steganography techniques. EURASIP Journal on Audio, Speech, and Music Processing **2012**(1), 1–16 (2012)
17. K Hofbauer, G Kubin, High-rate data embedding in unvoiced speech, in *INTERSPEECH*, 2006
18. Hofbauer, K and Hering, H (2007). Noise robust speech watermarking with bit synchronisation for the aeronautical radio. in Information Hiding. Springer.
19. S Chen, H Leung, H Ding, Telephony speech enhancement by data hiding. IEEE Trans. Instrum. Meas. on **56**(1), 63–74 (2007)
20. Coumou, DJ and Sharma, G (2006). Watermark synchronization for feature-based embedding: application to speech. in Conference on Multimedia and Expo, 2006 IEEE International. IEEE.
21. DJ Coumou, G Sharma, Insertion, deletion codes with feature-based embedding: a new paradigm for watermark synchronization with applications to speech watermarking. IEEE Trans. Inf. Forensics Secur. **3**(2), 153–165 (2008)
22. M-Q Fan et al., 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 (2013)
23. W Shengbei, M Unoki, Speech watermarking method based on formant tuning. IEICE TRANSACTIONS on Information and Systems **98**(1), 29–37 (2015)
24. S Wang, M Unoki, NS Kim, Formant enhancement based speech watermarking for tampering detection, in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014
25. B Yan, Y-J Guo, Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization. Multimedia tools and applications **67**(2), 383–405 (2013)
26. Zamani, M and Manaf, A.B.A. (2014). Genetic algorithm for fragile audio watermarking. Telecommunication Systems, p. 1–14.
27. Nematollahi MA, Gamboa-Rosales H, Akhaee MA, Al-Haddad S.A.R (2015) Robust digital speech watermarking for online speaker recognition. Mathematical Problem in Engineering, Hindawi.
28. Mohammad Ali Nematollahi, S.A.R.A.-H., Shyamala Doraisamy, Speaker frame selection for digital speech watermarking. National Academy Science Letters.(In press)
29. Garofolo, JS and Consortium, LD (1993). TIMIT: acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.
30. Steinebach, M., et al. (2001). StirMark benchmark: audio watermarking attacks. in Information Technology: Coding and Computing, 2001. Proceedings. International Conference on. IEEE.
31. A Al-Haj, An imperceptible and robust audio watermarking algorithm. EURASIP Journal on Audio, Speech, and Music Processing **2014**(1), 1–12 (2014)
32. I Rec, *P. 800: Methods for Subjective Determination of Transmission Quality* (International Telecommunication Union, Geneva, 1996)
33. N Dehak et al., Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on **19**(4), 788–798 (2011)
34. P Kenny, A small foot-print i-vector extractor, in *Proc. Odyssey*, 2012
35. DA Reynolds, TF Quatieri, RB Dunn, Speaker verification using adapted Gaussian mixture models. Digital Signal Process. **10**(1), 19–41 (2000)

36. Woo, RH, Park, A and Hazen, T.J (2006). The MIT mobile device speaker verification corpus: data collection and preliminary experiments. in Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The. IEEE.
37. McCool, C, et al. (2012). Bi-modal person recognition on a mobile phone: using mobile phone data. in Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on. IEEE.
38. Coria, L, Nasiopoulos, P and Ward, R (2009). A region-specific QIM-based watermarking scheme for digital images. in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2009. BMSB'09. IEEE.