Hu et al. EURASIP Journal on Audio, Speech, and Music Processing (2017) 2017:10
DOI 10.1186/s13636-017-0106-4

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Effective blind speech watermarking via adaptive mean modulation and package synchronization in DWT domain

Hwai-Tsu Hu[1*], Shiow-Jyu Lin[1] and Ling-Yuan Hsu[2]

Abstract

This paper outlines a package synchronization scheme for blind speech watermarking in the discrete wavelet transform (DWT) domain. Following two-level DWT decomposition, watermark bits and synchronization codes are embedded within selected frames in the second-level approximation and detail subbands, respectively. The embedded synchronization code is used for frame alignment and as a location indicator. Tagging voice active frames with sufficient intensity makes it possible to avoid ineffective watermarking during the silence segments commonly associated with speech utterances. We introduce a novel method referred to as adaptive mean modulation (AMM) to perform binary embedding of packaged information. The quantization steps used in mean modulation are recursively derived from previous DWT coefficients. The proposed formulation allows for the direct assignment of embedding strength. Experiment results show that the proposed DWT-AMM is able to preserve speech quality at a level comparable to that of two other DWT-based methods, which also operate at a payload capacity of 200 bits per second. DWT-AMM exhibits superior robustness in terms of bit error rates, as long as the recovery of adaptive quantization steps is secured.

Keywords: Blind speech watermarking, Adaptive mean modulation, Package synchronization, Discrete wavelet transform

1 Introduction

In the digital era, copyright protection of multimedia data (e.g., images, audios, and videos) is an important issue for content owners and service providers. Digital watermarking technology has received considerable attention owing to its potential application to the protection of intellectual property rights, content authentication, fingerprinting, and covert communications. Watermarking technology generally takes four factors (i.e., imperceptibility, security, robustness, and capacity) into consideration [1, 2]. An ideal watermarking algorithm minimizes perceptual distortion due to signal alteration while embedding a sufficient quantity of information within a host signal to ensure resistance against malicious attacks. The fact that the requirements of capacity, robustness, and imperceptibility are contradictory necessitates a tradeoff in the design of various watermarking schemes. For example, medical information systems are primarily intended to provide security and ensure the integrity of information, whereas payload capacity is of paramount importance in air traffic control systems. The annotation watermarking of music products emphasizes imperceptibility and robustness.

Robust watermarks are strongly resistant to attacks, whereas fragile watermarks are supposed to crumble under any attempt at tampering. There are numerous ways to categorize watermarking techniques. Depending on the requirement of the source material, watermarking schemes can be classified as blind, semi-blind, and non-blind. Blind watermarking is designed to recover an embedded watermark without the presence of the original source, while the non-blind approach can only be carried out using the original source. Semi-blind watermarking involves situations where information other than the source itself is required for watermark extraction.

Over the past two decades, numerous watermarking methods have been developed for images, audios, and

* Correspondence: hthu@mail.niu.edu.tw
[1]Department of Electronic Engineering, National Ilan University, Yi-Lan, Taiwan
Full list of author information is available at the end of the article

videos. Far less attention has been paid to the watermarking of speech signals. Speech is a specific form of audio signal; therefore, the techniques developed for audio watermarking are presumed to be applicable to speech watermarking. However, speech differs from typical audio signals with regard to spectral bandwidth, intensity distribution, signal continuity, and production modeling [3, 4]. The techniques developed for audio watermarking are not necessarily suitable for speech watermarking [5].

In [6], Hofbauer et al. exploited the fact that the ear is insensitive to the phase of a signal in non-voiced speech. They performed speech watermarking by replacing the excitation signal of an autoregressive representation in non-voiced segments. Chen and Liu [7] modified the position indices of selected excitation pulses in a watermarking scheme based on the codebook-excited linear prediction (CELP)-based speech codec. Coumou and Sharma [8] embedded data via pitch modification in voiced segments. The fact that multiple voiced segments may coalesce into a single voiced segment (or vice-versa) in a communication channel means that mismatches in voiced segments can lead to insertion, deletion, and substitution errors in the estimates of embedded data. They resorted to a concatenated coding scheme for synchronization and error recovery.

The vocal tract transfer function modeled by linear prediction (LP) has also been employed as an embedded target. Chen and Zhu [9] achieved robust watermarking by embedding watermark bits into codebook indices, while applying multistage vector quantization (MSVQ) to the derived LP coefficients. Yan and Guo [10] converted the LP coefficients to reflection coefficients, which were then transformed to inverse sine (IS) parameters. Watermark embedding was achieved by modifying the IS parameters using odd-even modulation [11].

Many watermarking algorithms applied to audio signals are implemented in the transform domain, such as discrete Fourier transform (DFT) [12–14], discrete cosine transform (DCT) [15–19], discrete wavelet transform (DWT) [15, 20–24], and cepstrum [25–27]. The objective is to take advantage of signal characteristics and/or auditory properties [28]. Among the transforms used to perform audio watermarking, DWT is currently the most popular due to its perfect reconstruction and good multi-resolution characteristics. The effectiveness of this approach in audio watermarking leads to conclude that it may also work for speech watermarking as well if speech characteristics can be adequately taken into account.

In this study, we introduce a robust blind watermarking scheme for hiding two types of information (watermark bits and synchronization codes) within embeddable regions of DWT subbands designated as information packages. The position of the synchronization codes is used in frame alignment to indicate the start of packaged binary data. This scheme allows the watermark to be dissembled into parts during the embedding phase and reassembled during extraction.

The remainder of this paper is organized as follows. Section 2 describes the watermarking framework, whereby information bits and synchronization codes are embedded in selected DWT subbands. Section 3 discusses configuring the watermark to cope with speech signals. We also outline a package strategy used for information grouping and synchronization and delineate the complete watermarking process. Section 4 presents experiment results aimed at evaluating speech quality and watermark robustness against commonly encountered attacks. Conclusions are drawn in Section 5.

2 Mean modulation in the DWT domain

In this study, we embedded two types of binary data (watermark bits and synchronization codes) within the same time frame under the same framework. This was achieved using DWT to conduct signal decomposition, thus allowing the embedding of different types of binary data within separate DWT subbands. However, the detectability of the embedded binary information differs somewhat between the watermark bits and synchronization codes. Unlike watermarks, where each bit conveys individual information, the bit sequence of a synchronization code can be considered a distinct entity. The existence of the synchronization code depends on a certain number of bits being recognizable, which means that the synchronization code may have some tolerance for faults. Thus, the watermark bits in our design are inserted within the lowest subband, wherein the coefficient magnitudes are larger than that observed in the subband used for the insertion of synchronization codes. Larger coefficients enable the use of stronger strengths to embed watermark bits. This is conducive to the robustness of the watermark and helps to keep it perceptually imperceptible.

2.1 Watermarking by adaptive mean modulation in DWT domain

Assuming that a speech signal is sampled at a rate of 16 kHz with 16-bit resolution, a two-level one-dimensional (1-D) DWT is employed to decompose the speech signal into a single approximation subband and two detail subbands. Here, the second-level DWT is performed on the approximation coefficients obtained from the first-level DWT of the host speech signal. The Daubechies-8 basis [29] is used as a wavelet function. Thus, the resulting second-level approximation subband occupies a frequency range roughly between 0 and 2000 Hz, while the second-level detail subband spans 2000 to 4000 kHz. The spectral density of speech signals is normally concentrated below 4 kHz; therefore, these two subbands are considered suitable candidates for watermarking applications. Analogous

to most watermarking methods, we divide the selected DWT coefficients into frames in order to facilitate the embedding and detection of watermarks. Within each frame, l adjacent coefficients, termed $c(i)$'s, drawn from the selected subband are gathered as a subgroup for the implementation of binary embedding.

$$G_k = \{c(\kappa_k + 1),\ c(\kappa_k + 2),\ \cdots,c(\kappa_k + l)\}; \kappa_k = (k-1)l. \tag{1}$$

In this study, the embedding of a binary bit w_k within G_k is achieved by modulating the coefficient mean m_k, which is defined as follows, using quantization index modulation (QIM) [30]:

$$m_k = \frac{1}{l}\sum_{i=1}^{l} c(\kappa_k + i). \tag{2}$$

The formulation of the QIM can be expressed as

$$\widehat{m}_k = \begin{cases} \left\lfloor \frac{m_k}{\Delta_k} + 0.5 \right\rfloor \Delta_k, & \text{if } \ w_k = 0; \\ \left\lfloor \frac{m_k}{\Delta_k} \right\rfloor \Delta_k + \frac{\Delta_k}{2}, & \text{if } \ w_k = 1, \end{cases} \tag{3}$$

where $\lfloor \bullet \rfloor$ denotes the floor function and Δ_k represents a quantization step. In essence, Eq. (3) changes m_k to the nearest integer multiple of Δ_k if $w_k = 0$ and to the middle of two integers multiples of Δ_k whenever $w_k = 1$. We note that QIM can be regarded as a special case of dither modulation [30, 31], wherein dither noise is added first and then quantized. The distortion compensation technique introduced in [30, 32] may also be incorporated into the quantization realization. Distortion compensated QIM allows the adjustment of the quantization steps without introducing extra distortion while pursuing robustness. In our former studies [33, 34], we have shown that the incorporation of distortion compensation into QIM can successfully enhance the robustness of the watermark while the imperceptibility is still maintained.

In this study, the mean value of the coefficients in each subgroup is selected as the embedding target because this statistical property is less susceptible to intentional attacks and/or unintentional modifications. Besides the insusceptibility to probable perturbation, employing the mean value as the embedding target makes it fairly easy to control the signal-to-watermark ratio, while using mean modulation for binary embedding. The robustness of the embedded watermark generally depends on the number of involved coefficients (i.e., l) and the embedding strength characterized by the quantization step size (i.e., Δ_k). Our scheme shares some similarities with those in [35, 36], where the one in [35] changed the DWT coefficients based on the average of a relevant frame and the one in [36] took into account the average of the linear regression of fast Fourier transform (FFT) values.

With QIM embedding, embedding strength is reflected in the quantization step size. The use of a large step size tends to increase robustness but impair quality by imposing more alterations to the speech signal. Making the embedded watermarks inaudible would require suppressing the distortion below the auditory masking threshold. Because the maximum tolerable noise level in each critical band is generally proportional to the short-time energy of the host speech, a sensible strategy involves adapting the quantization step according to the intensity of the speech segment. In other words, the quantization step is augmented when the energy of the DWT coefficients climbs, and it is reduced when the energy drops. By referring to the approach presented in [37], we derive the local energy level from previous coefficients in a recursive manner.

$$\overline{\rho}_k = (1-\alpha)\hat{\rho}_{k-1} + \alpha\overline{\rho}_{k-1}, \tag{4}$$

where $\hat{\rho}_{k-1} = \sum_{i=1}^{l} \hat{c}^2(\kappa_{k-1} + i)$ is the energy computed from the modified coefficients in the $(k-1)$th subgroup and $\overline{\rho}_k$ is the output of the first-order recursive low-pass filter. α is a positive controlling parameter deliberately rendering a unity DC gain. It should be pointed out that the coefficients in the current subgroup cannot be used for estimation, due to the fact that they are about to be modified by watermarking. Consequently, $\overline{\rho}_k$ can be regarded as an estimate of the short-time energy derivable from previous coefficients. Owing to the short-time stability of speech signals, the resulting $\overline{\rho}_k$ can be regarded as a smoothed version of $\hat{\rho}_k$.

The acquisition of short-time energy $\overline{\rho}_k$ makes it possible to regulate the signal-to-watermark ratio, which is defined as the energy ratio between the signal and watermarking perturbation measured in decibels. The relationship among $\overline{\rho}_k$, η, and Δ_k is expressed mathematically as follows:

$$\begin{aligned} 10^{\frac{\eta}{10}} &= \frac{\sum_{i=1}^{l} c^2(\kappa_k + i)}{E\left[\sum_{i=1}^{l} (\hat{c}(\kappa_k + i) - c(\kappa_k + i))^2\right]} \\ &\approx \frac{\sum_{i=1}^{l} c^2(\kappa_k + i)}{\frac{\Delta_k^2}{12}} \\ &\approx \frac{12\overline{\rho}_k}{\Delta_k^2}, \end{aligned} \tag{5}$$

where $E[\bullet]$ denotes the expected probability distribution. The term on the left-hand side of Eq. (5) is meant to convert a specific decibel value η to its linear magnitude, while the numerator and denominator of the fractional

expression on the right-hand side of Eq. (5) denote the energy levels of the signal and noise, respectively. The alteration due to the QIM in Eq. (3) is presumably distributed uniformly over $[-\Delta_k/2, \Delta_k/2]$. As a result, Δ_k can be computed directly once η is specified, as follows:

$$\Delta_k \approx \sqrt{12\overline{\rho}_k} \times 10^{-\frac{\eta}{20}}. \tag{6}$$

Following the acquisition of $\hat{m}_k$ as in Eq. (3), binary embedding in each subgroup is accomplished by modifying the corresponding DWT coefficients.

$$\hat{c}(\kappa_k + i) = c(\kappa_k + i) - m_k + \hat{m}_k. \tag{7}$$

After all of the watermark bits are embedded within designated DWT subbands, we take inverse DWT to obtain the watermarked speech signal using the modified subband coefficients.

Watermark extraction follows the basic procedure used in embedding. After applying two-level DWT to the watermarked file, the coefficients in the selected subband are divided into subgroups in order to derive the coefficient mean $\tilde{m}_k$ and quantization step $\tilde{\Delta}_k$ using Eqs. (2) and (6). The watermark bit $\tilde{w}_k$ residing in each subgroup is extracted based on standard QIM:

$$\tilde{w}_k = \begin{cases} 1, & \text{if } \left| \frac{\tilde{m}_k}{\tilde{\Delta}_k} - \left\lfloor \frac{\tilde{m}_k}{\tilde{\Delta}_k} \right\rfloor - 0.5 \right| \le 0.25; \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

where the tilde atop participating variables implies the effect of possible attacks.

2.2 Frame synchronization via DWT-AMM

The prerequisite for accurate watermark extraction using the abovementioned adaptive mean modulation (AMM) scheme is the perfect alignment of the boundary of each subgroup. A simple strategy by which to synchronize the locations used in watermark insertion and detection is to insert synchronization codes within the host signal. Actual watermark extraction begins after identifying the locations of the synchronization codes.

In fact, mean modulation with a fixed quantization step has previously been explored for the embedding of synchronization codes in the time domain for many audio watermarking algorithms [14, 15, 17, 23, 38]. In principle, watermark bits and synchronization codes are hidden in different segments of the audio file to avoid mutual interference. In this study, we propose embedding watermark bits and synchronization codes within separate DWT subbands. This arrangement offers additional advantages other than an increase in payload capacity. For example, the successful detection of synchronization codes in one subband can signify the presence of a data sequence in another subband.

Assume that the second-level detail subband has been selected as the embedding target. The derivation of a detail coefficient sequence can be imagined as a process of high-pass filtering and subsequent downsampling; hence, the spectral orientation is reversed for the detail coefficients. To flip the spectrum back to its normal direction, we simply alter the sign of the odd index coefficients, as follows:

$$\hat{c}_d^{(2)}(\kappa_k + i) = (-1)^i c_d^{(2)}(\kappa_k + i), \tag{9}$$

where $c_d^{(2)}(\kappa_k + i)$ denotes the ith coefficient in the kth subgroup of the second-level detail subband. The DWT level is specified in the superscript alongside the coefficient variable. The subscript "d" denotes the initial of the word "detail." Note that the spectral energy of speech signals is normally concentrated in low frequencies, and AMM tends to track low-frequency variations. Once Eq. (9) restores the energy distribution in the low frequencies, spectral flipping enables the rendering of larger quantization steps that eventually enhances robustness.

In our design, the synchronization code is a random binary sequence $\varphi(i) \in \{0, 1\}$ of length L_{code}. Each $\varphi(i)$ bit is inserted into l_{sync} coefficients in the second-level detail subband. We tentatively choose $L_{\text{code}} = 120$, $l_{\text{sync}} = 4$, and $\eta = 10$ to provide adequate resistance against possible attacks. The overall length of the synchronization code covers an interval of $l_{\text{sync}} \times L_{\text{code}}$ (=480) coefficients. Variable α used in the recursive filter is set to 0.9 in order to render a slowly varying estimate of the short-time energy. The search for subgroup demarcation is on a sample-by-sample basis. Because the second-level detail coefficients are derived from a signal of which the length is four times the coefficient amount, we conduct four times of DWT decomposition respectively from the first to fourth position of the speech signal and then attach the resultant coefficient sequence to every sample location. While detecting the presence of synchronization codes, we switch around the four sequences as the process proceeds from sample to sample. Specifically, for every replacement of a new coefficient in one of the four sequences, we regroup l_{sync} coefficients and recompute the short-time energy. After obtaining the coefficient mean and quantization step for each subgroup, we acquire binary bit $b_w(i)$ using Eq. (8). The synchronization code is then detected using a matched filter. The entire computation proceeds through three steps. First, the information bit sequence and synchronization code are both converted to bipolar form. Second, the extracted bipolar stream is convolved with the reversed version of the bipolar-converted synchronization code. Third, the presence of the synchronization code is presumed whenever the filter output $y(i)$ exceeds threshold T, which is set

as $0.45L_{\text{code}}$. The following inequality summarizes the aforementioned three steps.

$$y(i)=\sum_{n=0}^{L_{\text{code}}-1}(2\varphi(L_{\text{code}}-1-n)-1)\left(2b_w\left(i-l_{\text{sync}}n\right)-1\right) \overset{?}{\geq} T=0.45L_{\text{code}}. \tag{10}$$

As shown on the left side of Eq. (10), the bipolar stream is decimated by l_{sync} during convolution due to the fact that each bit stems from every l_{sync} coefficient. There can be two types of error during the search for synchronization codes. A false-positive error (FPE) involves declaring a non-embedded speech signal as an embedded one, whereas a false-negative error (FNE) involves classifying an embedded speech signal as a non-embedded one.

Assuming that the extracted watermark bits are independent random variables with probability P_e, then FPE P_{fp} can be computed as follows:

$$P_{fp}=\sum_{k=T'}^{L_{\text{code}}}\binom{L_{\text{code}}}{k}(P_e)^k(1-P_e)^{L_{\text{code}}-k}, \tag{11}$$

where k denotes the number of matched bits in a total of L_{code} bits. $\binom{L_{\text{code}}}{k}$ represents the binomial coefficient. The threshold T and the number of matched bits T' hold the relationship as $T' = (L_{\text{code}} - T)/2$, because T is the summed result of T' matched bits and $L_{\text{code}} - T'$ unmatched bits.

$$T=(+1)\times T'+(-1)\times\left(L_{code}-T'\right), \tag{12}$$

where a matched bit corresponds to +1 and an unmatched bit corresponds to −1. Since non-embedded bits are either 0 or 1 with pure randomness, P_e is assumed to be 0.5. Thus, Eq. (11) can be further simplified as

$$P_{fp}=\frac{1}{2^{L_{\text{code}}}}\sum_{k=T}^{L_{\text{code}}}\binom{L_{\text{code}}}{k}. \tag{13}$$

Given that $L_{\text{code}} = 120$ and $T = 0.45L_{\text{code}} = 54$, P_{fp} turns out to be 4.34×10^{-7}, which implies that FPE rarely happens while using the presumed parameter setting.

Analogous to the discussion on the derivation of FPE, the FNE P_{fn} can be computed as

$$\begin{aligned}P_{fn}&=\sum_{k=0}^{T-1}\binom{L_{\text{code}}}{k}(1-P_{\text{BER}})^k\times(P_{\text{BER}})^{L_{\text{code}}-k}\\&=\sum_{k=T}^{L_{\text{code}}}\binom{L_{\text{code}}}{k}(P_{\text{BER}})^k\times(1-P_{\text{BER}})^{L_{\text{code}}-k},\end{aligned} \tag{14}$$

where P_{BER} denotes the error rate for each bit. According to Eq. (14), P_{fn} remains below 0.982 even if P_{BER} is as high as 0.2.

3 Watermarking with package synchronization

In Sections 2 and 3, we discuss the DWT-AMM framework and its application to watermark synchronization. Further considerations must be taken into account in the development of a practical speech watermarking system. It has been pointed out in the introduction that a speech signal exhibits several distinct acoustic characteristics, which differentiate the speech from other types of audio. Unlike most music signals, silent segments commonly occur in speech utterances. The insertion of watermark bits into silent segments would render them vulnerable to noise perturbation and susceptible to attacks through the simple removal of silence. Thus, we developed an energy-based scheme to enable the selection of frames for the embedding of watermarks and synchronization codes. One general principle in watermarking is to hide information among large coefficients in the transformed domain, because this enables the employment of stronger embedding to resist attacks with less concern for imperceptibility.

The energy of a speech signal is normally concentrated below 4 kHz. To make the best use of DWT decomposition, we selected the second-level approximation subband for the embedding of binary information and reserved the second-level detail subband for frame synchronization on condition that the speech is sampled at 16 kHz. After taking two-level DWT of the host signal, the coefficients in the second-level approximation and detail subbands are both partitioned into non-overlapping frames of size L_f. In this study, L_f is tentatively set to 160 to facilitate subsequent scheme development. Then, we calculate the root-mean-square (RMS) values, termed $\sigma_a(t)$ and $\sigma_d(t)$, respectively, for the second-level approximation and detail subbands.

$$\sigma_a(t)=\sqrt{\frac{1}{L_f}\sum_{i=1}^{L_f}\left(c_a^{(2)}(i;\ t)\right)^2}; \tag{15}$$

$$\sigma_d(t)=\sqrt{\frac{1}{L_f}\sum_{i=1}^{L_f}\left(c_d^{(2)}(i;\ t)\right)^2}, \tag{16}$$

where $c_{a^{(2)}}(i;\ t)$ and $c_{d^{(2)}}(i;\ t)$ are respectively the ith second-level approximation and detail coefficients in the tth frame. Let ψ_a and ψ_d be the corresponding thresholds, which are assigned as ratios proportional to the maximum values. The frames with RMS values exceeding pre-specified thresholds are selected for watermarking. This type of frame selection can be expressed as follows:

$$\Lambda(t) = \begin{cases} \text{"embeddable"}, & \text{If} \quad \sigma_a(t) \geq \psi_a \ \& \ \sigma_d(t) \geq \psi_d; \\ \text{"non-embeddable"}, & \text{otherwise}, \end{cases} \tag{17}$$

with

$$\psi_a = 0.035 \max\{\sigma_a(t)\}; \tag{18}$$

$$\psi_d = 0.04 \max\{\sigma_d(t)\}. \tag{19}$$

The frame attribute $\Lambda(t)$ is categorized as "embeddable" if both $\sigma_a(t)$ and $\sigma_d(t)$ surpass their respective thresholds.

Figure 1 illustrates the process of searching embeddable frames within a speech signal. In Fig. 1e, the frame is categorized as embeddable, as long as the corresponding approximation and detail coefficients are of sufficiently large magnitude to allow the embedding of watermark bits and synchronization codes. The insertion and detection of synchronization codes are illustrated in Fig. 2. In this example, the synchronization code is embedded in four places, each spanning an interval of three frames. The four embedding segments are rendered in red in (b). As indicated by the four sharp peaks precisely at the ends of the red areas, the output of the matched filter is sufficient to identify the synchronization code. For a segment comprising consecutive embeddable frames, the synchronization code is inserted only within the first three frames in the second-level detail subband, whereas binary embedding is applied to every embeddable frame in the approximation subband.

Depending on the number of frames available for data hiding, we divide the watermark bits into several packages, each containing a header in conjunction with a series of watermark bytes. The implantation of a complete synchronization code requires an interval stretching

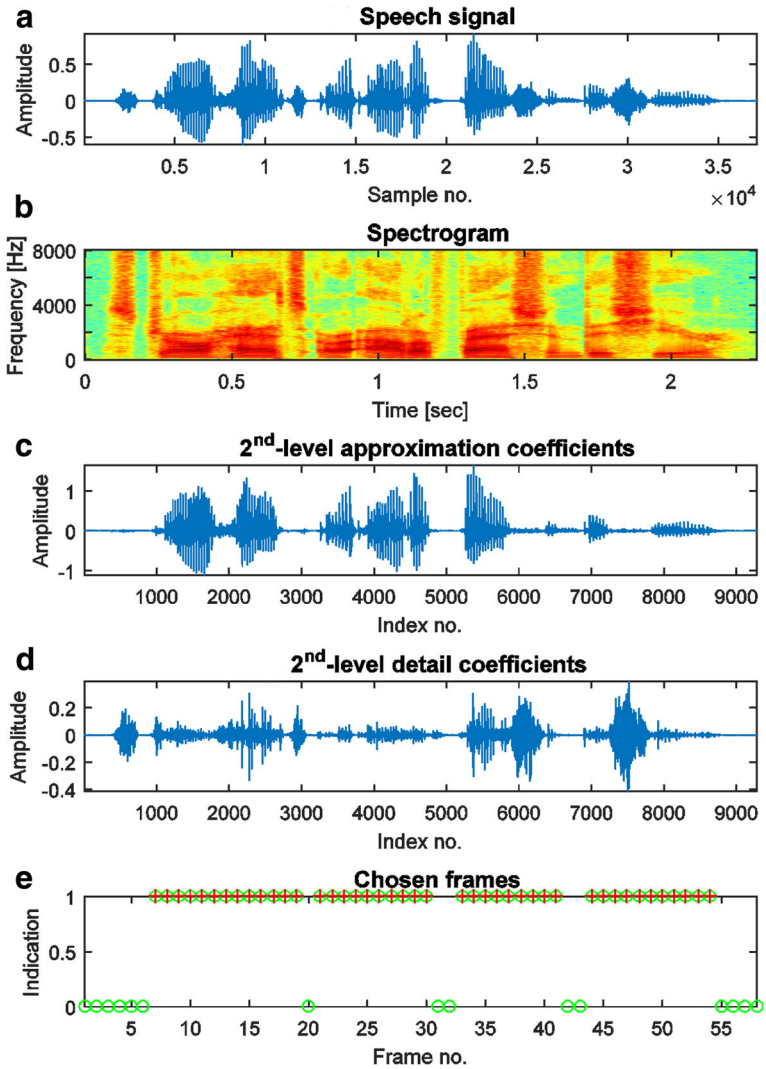

Fig. 1 a–e Illustration of searching for embedding frames. The embeddable frames are indicated by symbol "⊕"

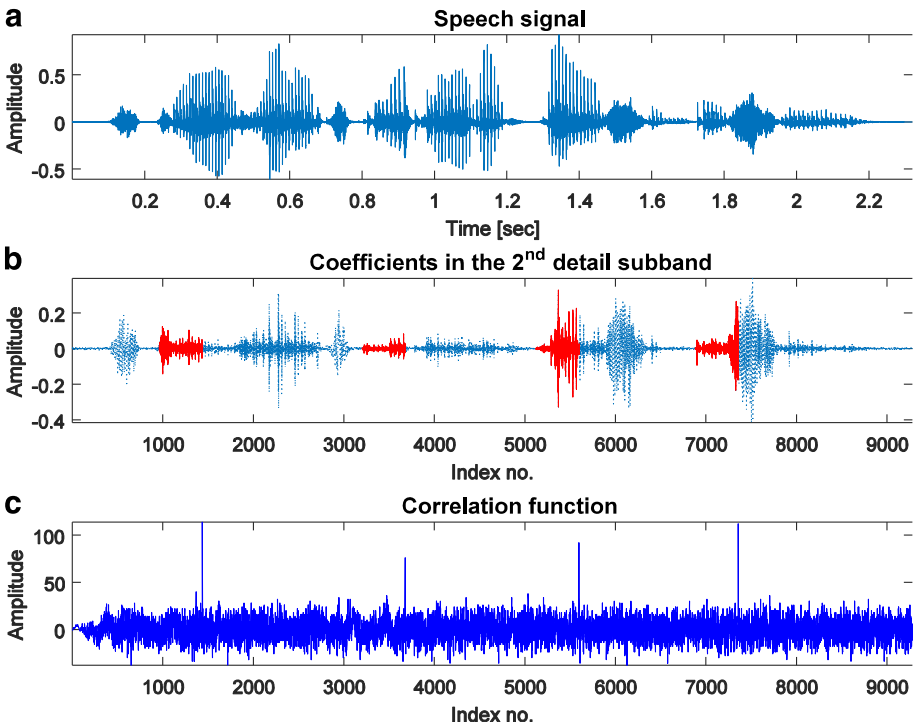

Fig. 2 **a**, **c** Detection of the synchronization code from a noise-corrupted speech signal with SNR = 30 dB. The leading three frames in each embeddable segment are particularly drawn in *red* in **b**

480 coefficients; therefore, only the speech segments extending beyond three consecutive frames are used as watermark packages. During the embedding phase, the DWT-AMM settings are $l_a^{(2)} = 20$ and $\eta_a^{(2)} = 20$ for watermark embedding in the approximation subband and $l_d^{(2)} = 4$ and $\eta_d^{(2)} = 10$ for synchronization in the detail subband. The subscripts "a" and "d" alongside the variables represent subband attributes. In accordance with these specifications, each frame in the approximation subband carries 8 bits of information. Thus, for a wideband speech signal sampled at 16 kHz, the maximum payload capacity would be 200 (=16000/($2^2 \times 20$)) bits per second (bps).

The header of each package consists of a 15-bit message produced by a [11, 15] BCH encoder [39]. The message contains information in two parts: 7 bits indicating the allocated position and 4 bits specifying the total length. This means that there are 2^7 starting positions that could be assigned. The length of data allowable in each package stretches from 1 to 16 bytes. The maximum size of watermark bits that can be accommodated is 8×2^7. Through the BCH encoder, the 11-bit message is appended with a parity symbol to form a code of length 15. The resulting BCH code is capable of correcting 1 bit error.

Figure 3 illustrates the means by which embeddable frames are configured for various lengths of data. The start

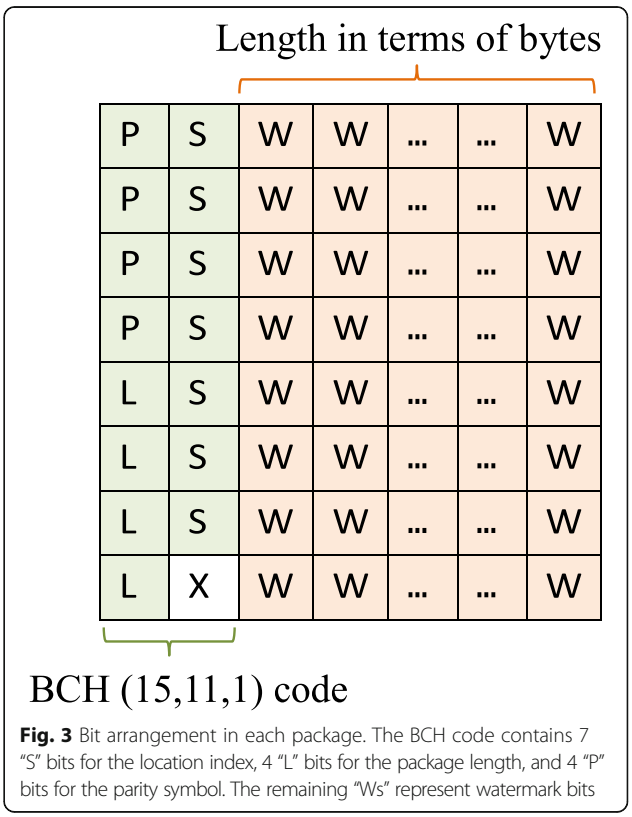

Fig. 3 Bit arrangement in each package. The BCH code contains 7 "S" bits for the location index, 4 "L" bits for the package length, and 4 "P" bits for the parity symbol. The remaining "Ws" represent watermark bits

locations of embeddable segments implicitly synchronize the time windows for the embedding and extraction of the watermark. The watermark is tentatively selected as a binary image logo of size 32×32 with an equal number of "1s" and "0s". To reinforce security, we scrambled the watermark using the Arnold transform [40] and then converted it to a 1-D bit sequence. The bit sequence was then divided into packages of various size matching the lengths of the embeddable segments in different locations. Multiple watermarks can be embedded as long as the speech file is of sufficient length. When reconstructing the watermark, we employ a majority voting scheme to verify each retrieved bit in cases where multiple copies are received.

To conclude this section, Fig. 4 outlines the processing flow of the proposed watermarking method. The required steps are summarized as follows:

Step 0 Scramble the watermark logo using an encryption key and convert the results to a bit stream.

Step 1 Decompose the host speech signal using two-level DWT.

Step 2 Seek embeddable segments.

Step 3 Implant the synchronization code into the first three frames of an embeddable segment in the second-level detail subband.

Step 4 Partition the watermark bit sequence into packages in accordance with the size of the embeddable segment. The location and size of each watermark package are saved as a 15-bit message using a (15,11) BCH encoder, and the resulting BCH code is combined with scrambled watermark bits to form a package.

Step 5 For each embeddable segment, the packaged bits are embedded within the approximation coefficients using AMM.

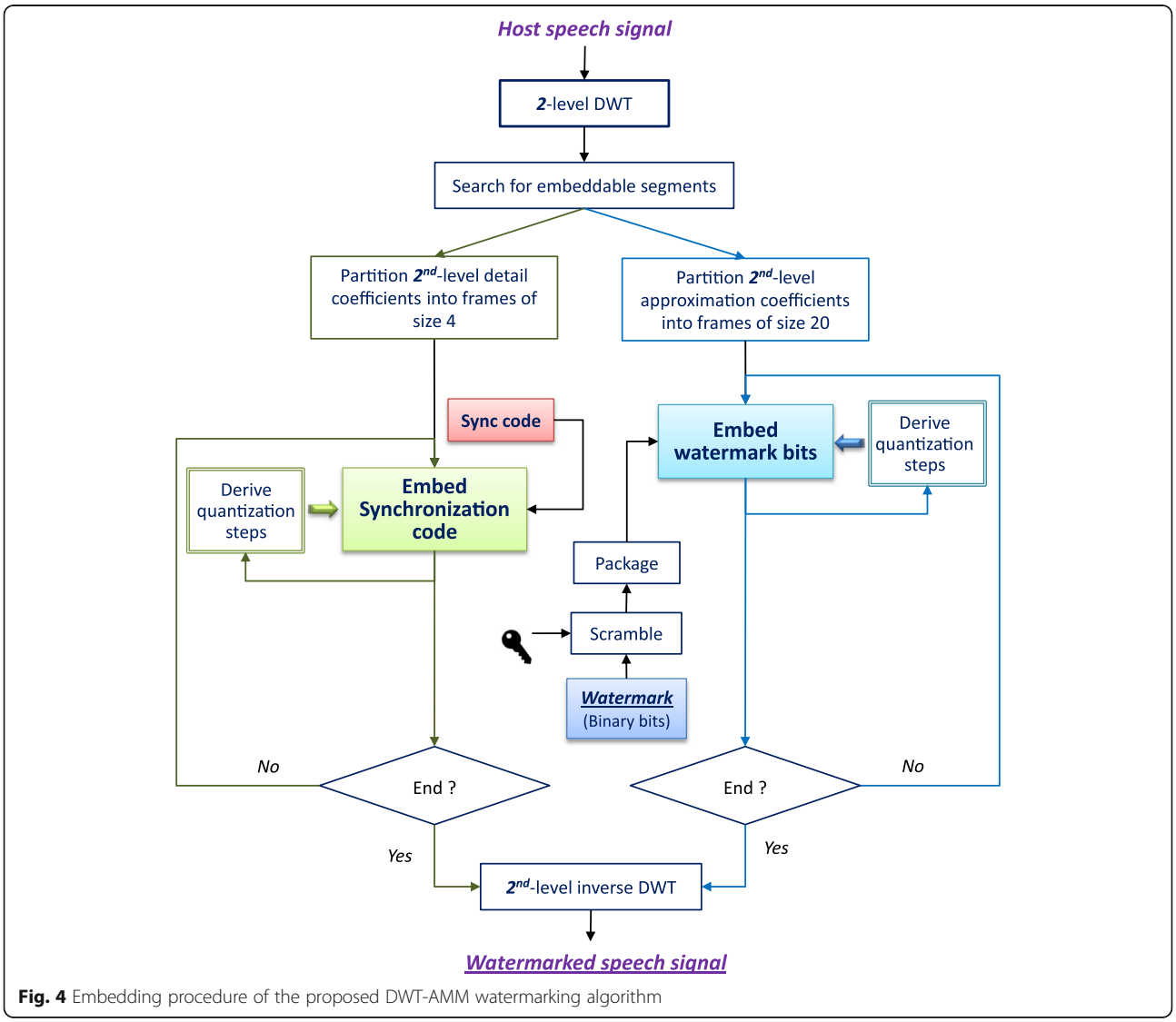

Fig. 4 Embedding procedure of the proposed DWT-AMM watermarking algorithm

Step 6 Repeat steps 3~5 if the end of the file is reached; otherwise, perform a two-level inverse DWT to attain a watermarked speech signal with synchronization information inside.

Watermark extraction follows the same procedure as that used in embedding. Figure 5 provides an illustrative depiction of the process, as briefly outlined in the following:

Step 1 Decompose the host speech signal using two-level DWT. To take every sample shift into account, we need to perform two-level DWT four times starting from the first to fourth position.
Inspect the segment beginning with current sample i. Detect the synchronization code using the technique developed in Section 4. If the synchronization code is present, go to step 3; otherwise, move one sample forward ($i \leftarrow i + 1$) and repeat step 2.
Step 3 Extract the bits residing in each package using AMM. The located position of the retrieved watermark bits is resolved from the BCH decoder. Update the current index i to the new position.
Step 4 If the index reaches the end, go to step 5. Otherwise, go to step 2.
Step 5 Adopt the majority vote strategy to determine the ultimate value of each bit.
Step 6 Convert the 1-D bit sequence to a matrix and apply the inverse Arnold transform to descramble the matrix using the correct key.

4 Performance evaluation

The test materials consisted of 192 sentences uttered by 24 speakers (16 males and 8 females) drawn from the core set of the TIMIT database [41]. Speech utterances were recorded at 16 kHz with 16-bit resolution. For the convenience of computer simulation, speech files belonging to the same dialect region were concatenated to form a longer file. Since each speech utterance was recorded separately, the maximum amplitude of each file was uniformly rescaled to an identical level to maintain consistent intensity. The watermark bits for the test were a series of alternate 1s and 0s of sufficient length to cover the entire host signal.

4.1 Smoothing factor for the recursive filter

Our initial concern lies in the choice of an appropriate value for variable α used in the recursive filtering (i.e., Eq. (4)) of the DWT-AMM framework. The recursive filter is meant to render a smooth estimate of short-time energy. To understand the influence of variable α, we conducted a pilot test examining the watermarked speech in the presence of white Gaussian noise with signal-to-noise set at 20 dB. The testing set included an arithmetic

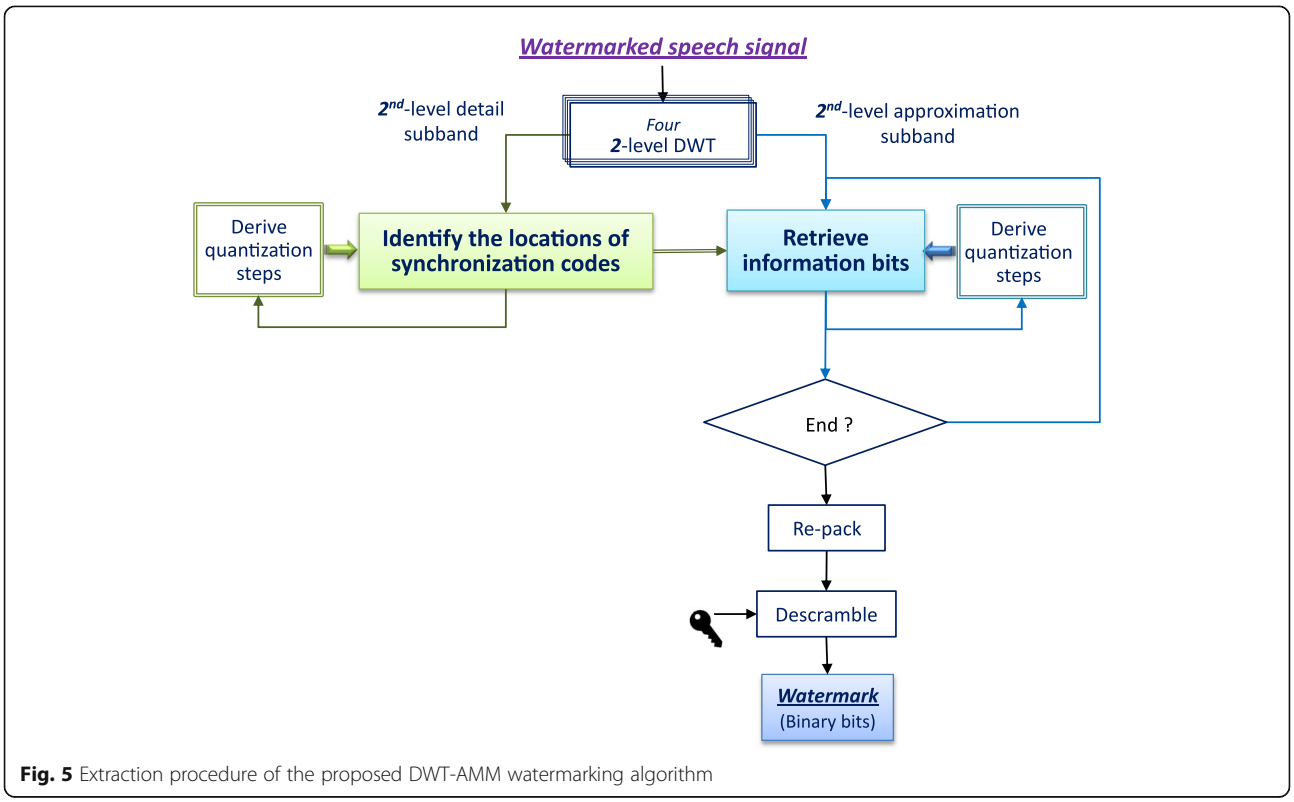

Fig. 5 Extraction procedure of the proposed DWT-AMM watermarking algorithm

sequence ranging from 0.3 to 0.975 in increments of 0.025. We measured the variations in signal-to-noise ratio (SNR), mean opinion score of listening quality objective (MOS-LQO), and bit error rate (BER) under changes in α. Among the three abovementioned measures, SNR and MOS-LQO reflect the impairment of quality due to watermarking, while BER indicates the robustness of the embedded watermark against possible attacks. The definition of SNR is given as follows:

$$\mathrm{SNR} = 10\log_{10}\left(\frac{\sum_n s^2(n)}{\sum_n \left(\hat{s}(n)-s(n)\right)^2}\right), \tag{20}$$

where $s(n)$ and $\hat{s}(n)$ denote the original and watermarked speech signals, respectively. MOS-LQO is the consequence of the perceptual evaluation of speech quality (PESQ) metric [42], which was developed to model subjective tests commonly used in telecommunications. The PESQ assesses speech quality on a −0.5 to 4.5 scale. A mapping function to MOS-LQO is described under ITU-T Recommendation P.862.1, covering a range from 1 (bad) to 5 (excellent). Table 1 specifies the MOS-LQO scale. In this study, we adopted the implementation released from ITU-T website [43].

To determine the effect on robustness, we examined the BER between the recovered watermark $\tilde{W} = \{\tilde{w}_n\}$ and the original watermark $W = \{w_n\}$:

$$\mathrm{BER}\left(W, \tilde{W}\right) = \frac{\sum_{n=1}^{N_w} w_n \oplus \tilde{w}_n}{N_w}, \tag{21}$$

where N_w denotes the number of watermark bits.

Figure 6 presents the average BER, SNR, and MOS-LQO obtained from the test set with the parameters $l_a^{(2)} = 20$ and $\eta_a^{(2)} = 20$. In this experiment, non-embeddable frames were not excluded from average calculation. The obtained MOS-LQOs were therefore slightly lower than that attained by the actual watermarking scheme, and the resulting BERs were somewhat higher than the outcomes involving merely the embeddable frames. As shown in Fig. 6, the average BER, SNR, and MOS-LQO remain roughly steady when $\alpha < 0.5$ and gradually descend with an increase in α. The increasing tendency become increasingly obvious once α exceeds 0.8. The lower SNR values at $\alpha > 0.9$ can be attributed to the fact that the computation of $\overline{\rho}_k$ refers more to previous data than recent data. This often results from large quantization steps at the end of a speech segment where the volume drops abruptly. A lower SNR also implies a more pronounced modification to the speech signal; therefore, the MOS-LQO presents a downward trend. In subsequent experiments, we eventually set α to 0.8 for embedding watermark bits, as this achieves suitable BER and SNR values without deviating MOS-LQO too far from a desirable score.

Table 1 Speech quality characterized by MOS-LQO scores

Numerical indication	Perceived quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

4.2 Detection rate of synchronization codes

In Section 2.2, we discuss how to embed and detect the synchronization codes in the second-level detail subband. All the theoretical analysis in that section is deduced from a probability aspect. Here, we present the experiment results with respect to the test materials. In accordance to the rule given in Section 3, there were 781 speech segments selected for embedding synchronization codes over a length of 9,229,695 samples in total (or equivalently, 576.86 s). The embedding of synchronization codes as per the specifications in Section 2.2 led to a SNR of 27.73 dB and a MOS-LQO score of 4.35. The competence of the proposed method was verified by inspecting the frequency counts of miss and false alarm in the presence of various attacks. The attack types in this study involved resampling, requantization, amplitude scaling, noise corruption, low-pass and high-pass filtering, DA/AD conversion, echo addition, jittering, and compression. Table 2 lists the details of these attacks. The time-shifting attack considered in case N is intended to find out the consequence if frames are slightly misaligned. This particular attack is not designed for the synchronization test but will be examined in the evaluation of watermarking performance. For the other attack types ranging from A to M, the test results are tabulated in Table 3.

As revealed by the results in Table 3, the false alarm events seldom occurred because of the choice of a relative high detection threshold, i.e., $T = 0.45L_{\mathrm{code}} = 54$. The proposed synchronization technique survived most attacks except for high-pass filtering above 1 kHz. The reason can be attributed to the fact that the synchronization codes are inserted into the second-level detail subband, of which the spectrum is primarily distributed from 2 to 4 kHz. Consequently, obliterating the frequency components above 1 kHz will ruin the synchronization. Apart from the high-pass filtering, the noise corruption with SNR = 20 dB was another attack that caused obvious damage. Nonetheless, the miss rate 66/781 is still considered acceptable since over 91.5% of the package locations are recoverable.

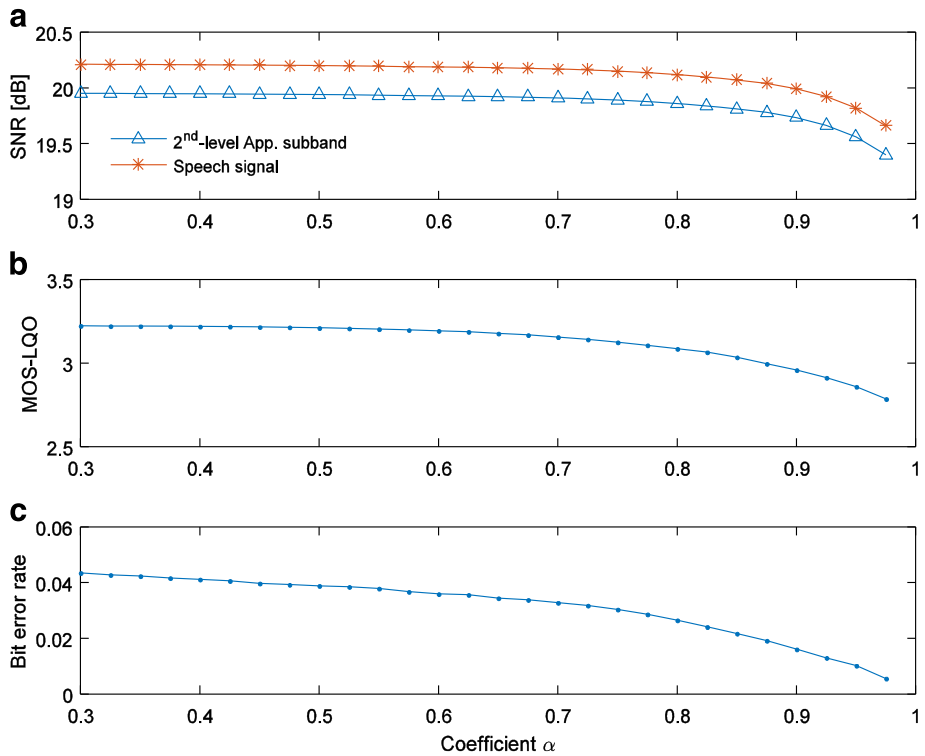

Fig. 6 The effect due to the use of various αs. Subplot **a** presents the SNRs observed in the second-level approximation subband and time-domain signal. Subplot **b** delineates the MOS-LQO scores. Subplot **c** is the result when the watermarked speech is corrupted by white Gaussian noise with SNR = 20

4.3 Comparison with other WT-based watermarking methods

This study compared the performance of three wavelet transform (WT)-based speech watermarking methods, namely, DWT-SVD [4], LWT-DCT-SVD, [3] and the proposed DWT-AMM. For the sake of a fair comparison, the watermark bits were embedded in the second-level approximation subband using an identical payload capacity of 200 bps for all three methods. It should also be noted that the idea of embedding the watermark bits and synchronization codes within different subbands is applicable to any wavelet-based method. We assumed that the second-level detail subband was reserved for the embedding of synchronization codes in all cases to ensure that each method was equally capable of resisting cropping and/or time-shifting attacks. Only frames satisfying the conditions in (17) were used to embed binary information. Furthermore, in order to provide more insights into the proposed DWT-AMM approach, we also implemented watermark embedding at a rate of 100 bps with respect to the third-level approximation and detail subbands, both of which were obtained by splitting the second-level approximation subband. The parametric settings followed those in the second-level approximation subband. That is, $l_a^{(3)} = l_d^{(3)} = 20$ and $\eta_a^{(3)} = \eta_d^{(3)} = 20$.

The quality of the watermarked speech signal obtained using the abovementioned methods was evaluated based on SNR and PESQ. We intentionally adjusted the parameters of the three methods to permit SNR values nearby 22 dB, which is above the level (20 dB) recommended by International Federation of the Phonographic Industry (IFPI) [28]. The commensurate SNR values also imply the use of comparable embedding strengths for all three methods. As shown in Table 4, the MOS-LQO values for DWT-SVD, LWT-DCT-SVD, and DWT-AMM were distributed over a range just above 3.2. These outcomes suggest that these three methods render comparable quality. Nonetheless, the score of 3.2 merely reflects a fair auditory perception. The cause is conceivably connected with the embedding strength and payload capacity. For the DWT-AMM implemented in the third-level approximation subband with a payload capacity of 100 bps, the average MOS-LQO value has been raised above 4.0. The MOS-LQO score could be further lifted beyond 4.2 when the DWT-AMM was applied to the third-level detail subband with the same capacity.

We examined the BER defined in Eq. (21) to evaluate the robustness of the algorithms against various attacks previously specified in Table 2. Table 5 presents the average BERs obtained by each of the methods in the presence of various attacks. All three methods successfully

Table 2 Attack types and specifications

Item	Type	Description
A	Resampling	Conducting downsampling to 8 kHz and then upsampling back to 16 kHz.
B	Requantization	Quantizing the watermarked signal to 8 bits/sample and then back to 16 bits/sample.
C	Low-pass filtering (I)	Applying a low-pass filter with a cutoff frequency of 4 kHz.
D	Low-pass filtering (II)	Applying a low-pass filter with a cutoff frequency of 1 kHz.
E	High-pass filtering	Applying a high-pass filter with a cutoff frequency of 1 kHz.
F	Amplitude scaling	Scaling the amplitude of the watermarked signal by 0.85.
G	Noise corruption (I)	Adding zero-mean white Gaussian noise to the watermarked audio signal with SNR = 30 dB.
H	Noise corruption (II)	Adding zero-mean white Gaussian noise to the watermarked audio signal with SNR = 20 dB.
I	DA/AD conversion	Converting the digital audio file to an analog signal and then resampling the analog signal at 16 kHz. The DA/AD conversion is performed through an onboard Realtek ALC892 audio codec, of which the line-out is linked with the line-in using a cable line during playback and recording.
J	Echo addition	Adding an echo signal with a delay of 50 ms and a decay to 5% to the watermarked audio signal.
K	Jittering	Randomly deleting or adding one sample for every 100 samples within each frame.
L	G.722 speech coding	Encoding and decoding the watermarked speech signal with a G.722 wideband audio codec at 64 kbps.
M	G.726 speech coding	Encoding and decoding the watermarked speech signal with a G.726 wideband audio codec at 32 kbps.
N	Time-shift by 1 sample	Purposely shifting the watermarked audio signal by one sample.

retrieved the watermark when no attack was present. All of them demonstrated comparable satisfactory resistances against G.722 and G.726 codecs. They also survived low-pass filtering (I) and resampling, due to the fact that these two attacks do not have a severe effect on coefficients in the second-level approximation subband. For the same reason, these three methods did not pass the high-pass filtering attack, by which the low-frequency components below 1 kHz were destroyed. The low-pass filtering with a cutoff frequency of 1 kHz inflicted obvious damage on DWT-SVD and LWT-DCT-SVD; however, only minor damage was observed in the results of DWT-AMM. This can be ascribed to the use of the statistical mean for watermarking.

In cases involving echo addition (Attack J) and slight time-shift (Attack N), DWT-AMM outperformed DWT-SVD and LWT-DCT-SVD, due primarily to its adaptability to signal intensity. Adaptively adjusting the quantization steps also enables DWT-AMM to withstand amplitude scaling attacks. By contrast, both DWT-SVD and LWT-DCT-SVD failed in the case of amplitude scaling, due to the use of a fixed quantization step.

Table 3 Detection of the embedded synchronization codes under various attacks. Overall, the synchronization code has been embedded in 781 speech segments over a length of 9,229,695 samples in total

Attack type	Number of misses	Number of false alarms
0. none	0	1
A	2	3
B	0	3
C	3	2
D	756	0
E	0	1
F	0	1
G	0	1
H	66	3
I	1	1
J	0	2
K	0	2
L	0	2
M	0	0

The addition of Gaussian white noise with SNR controlled at 30 and 20 dB did not appear to cause any problems for DWT-SVD or LWT-DCT-SVD; however, DWT-AMM suffered minor deterioration. The reason is conceivably due to the imperfect acquisition of quantization steps from noise-corrupted speech. Requantization can be regarded as a type of noise corruption [44]; therefore, DWT-AMM is also subject to performance degradation. The same explanation applies to the results obtained under DA/AD conversion attacks, which led to composite impairment in time scaling, amplitude scaling, and noise corruption [44]. DWT-AMM was unable to entirely avoid damage under these conditions; however, DWT-SVD

Table 4 Statistics of the measured SNRs and MOS-LQO scores. The data in the second and third columns are interpreted as "mean [±standard deviation]." The payload capacity for each method is listed in the last column

Watermarking method	SNR [dB]	MOS-LQO	Payload (bps)
DWT-SVD	21.844 [±0.888]	3.260 [±0.108]	200
LWT-DCT-SVD	21.985 [±0.910]	3.209 [±0.119]	200
DWT-AMM	21.932 [±0.210]	3.254 [±0.212]	200
$DWT\text{-}AMM_a^{(3)}$	22.872 [±0.350]	4.006 [±0.110]	100
$DWT\text{-}AMM_d^{(3)}$	30.611 [±1.112]	4.238 [±0.045]	100

Table 5 Average bit error rates (in percentage) for three compared watermarking methods under various attacks

Attack type	DWT-SVD	LWT-DCT-SVD	DWT-AMM	DWT-AMM$_a^{(3)}$	DWT-AMM$_d^{(3)}$
0. none	0.00	0.00	0.00	0.00	0.00
A	0.00	0.00	0.00	0.00	0.03
B	0.00	0.00	0.76	0.50	0.95
C	0.00	0.00	0.00	0.00	0.03
D	19.57	25.07	3.66	0.21	49.45
E	50.05	29.06	49.68	46.61	8.96
F	46.33	31.82	0.00	0.00	0.00
G	0.00	0.00	0.73	0.34	1.38
H	0.00	0.00	3.46	1.56	9.32
I	53.14	57.48	1.69	0.82	4.08
J	1.22	1.92	0.02	0.01	0.01
K	2.21	3.43	0.94	0.86	10.92
L	1.07	0.83	0.91	0.70	2.22
M	0.58	0.33	0.81	0.53	2.55
N	5.35	14.47	1.64	0.67	28.80

and LWT-DCT-SVD suffered even more due to a lack of amplitude scaling.

For the two 100-bp versions of DWT-AMM, the one implemented in the third-level approximation subband, termed DWT-AMM$_a^{(3)}$, generally exhibited superior robustness in terms of BER, and yet, the resultant MOS-LQO is above 4.0. The reduction of BER is ascribed to the fact that the watermark embedding is performed over a subband with higher intensity, while the imperceptibility seems improvable at the cost of payload capacity. By contrast, the DWT-AMM implemented in the third-level detail subband, termed DWT-AMM$_d^{(3)}$, offered an average MOS-LQO of 4.238. The associated SNR of 30.611 dB reflects a weaker embedding strength, thus resulting in a worse BER in comparison with the one obtained from DWT-AMM$_a^{(3)}$.

5 Conclusions

This paper proposes a novel DWT-based speech watermarking scheme. In the proposed scheme, information bits and synchronization codes are embedded within the second-level approximation and detail subbands, respectively. The synchronization code serves in frame alignment and indicates the start position of an enciphered bit sequence referred to as a package. The watermarking process is executed on a frame-by-frame basis to facilitate detectability. Binary embedding in the second-level subband is performed by adaptively modifying the mean value of the coefficients gathered in each subgroup. During watermark extraction, all fragments of binary bits are retrieved with the assistance of a synchronization scheme and repacked according to the header content of each package. The robustness of the embedded watermark is reinforced through the selection of frames with sufficient intensity. The proposed formulation makes it possible to specify the embedding strength in terms of the SNR of the intended subband. Specifically, the quantization steps can be acquired from the speech signal by referring to the energy level of the passing coefficients in a recursive manner.

The watermarking scheme outlined in this paper has a maximum rate of 200 bps. PESQ test results indicate that the proposed DWT-AMM renders speech quality comparable to that obtained using two existing wavelet-based methods. With the exception of attacks that compromise the retrieval of quantization steps, the proposed DWT-AMM generally outperforms the compared methods. Overall, the proposed DWT-AMM demonstrates satisfactory performance. The incorporation of the package synchronization scheme allows the splitting of the watermark to cope with the intermittent characteristic of speech signals.

Acknowledgements

This research work was supported by the Ministry of Science and Technology, Taiwan, Republic of China, under grants MOST 104-2221-E-197-023 & MOST 105-2221-E-197-019.

Authors' contributions

In this research work, HTH and LYH jointly developed the algorithms and conducted the experiments. HTH was responsible for drafting the manuscript. SJL provided valuable comments and helped to improve the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Electronic Engineering, National Ilan University, Yi-Lan, Taiwan. [2]Department of Information Management, St. Mary's Junior College of Medicine, Nursing and Management, Yi-Lan, Taiwan.

Received: 16 November 2016 Accepted: 19 April 2017
Published online: 12 May 2017

References

1. N Cvejic, T Seppänen, *Digital audio watermarking techniques and technologies: applications and benchmarks* (Information Science Reference, Hershey, 2008)
2. X He, *Watermarking in audio: key techniques and technologies* (Cambria Press, Youngstown, 2008)
3. B Lei, I Song, SA Rahman, Robust and secure watermarking scheme for breath sound. J Syst Softw **86**(6), 1638–1649 (2013)
4. MA Nematollahi, SAR Al-Haddad, F Zarafshan, Blind digital speech watermarking based on eigen-value quantization in DWT. J King Saud Univ Comp Inf Sci **27**(1), 58–67 (2015)
5. MA Nematollahi, SAR Al-Haddad, An overview of digital speech watermarking. Int J Speech Tech **16**(4), 471–488 (2013)
6. K Hofbauer, G Kubin, WB Kleijn, Speech watermarking for analog flat-fading bandpass channels. IEEE Trans on Audio Speech and Language Processing **17**(8), 1624–1637 (2009)
7. OTC Chen, CH Liu, Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks. IEEE Trans on Audio Speech Language Processing **15**(5), 1605–1616 (2007)
8. 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)
9. N Chen, J Zhu, Multipurpose speech watermarking based on multistage vector quantization of linear prediction coefficients. J China Univ Posts Telecom **14**(4), 64–69 (2007)
10. B Yan, Y-J Guo, B Yan, Y-J Guo, Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization. Multimed Tools Appl **67**(2), 383–405 (2013)
11. D Kundur, (1999). Multiresolution digital watermarking: algorithms and implications for multimedia signals, Ph. D. Thesis, University of Toronto, Ontario, Canada.
12. W Li, X Xue, P Lu, Localized audio watermarking technique robust against time-scale modification. IEEE Trans Multimedia **8**(1), 60–69 (2006)
13. R Tachibana, S Shimizu, S Kobayashi, T Nakamura, An audio watermarking method using a two-dimensional pseudo-random array. Signal Process **82**(10), 1455–1469 (2002)
14. D Megías, J Serra-Ruiz, M Fallahpour, Efficient self-synchronised blind audio watermarking system based on time domain and FFT amplitude modification. Signal Process **90**(12), 3078–3092 (2010)
15. X-Y Wang, H Zhao, A novel synchronization invariant audio watermarking scheme based on DWT and DCT. IEEE Trans Signal Processing **54**(12), 4835–4840 (2006)
16. I-K Yeo, HJ Kim, Modified patchwork algorithm: a novel audio watermarking scheme. IEEE Trans Speech Audio Processing **11**(4), 381–386 (2003)
17. BY Lei, IY Soon, Z Li, Blind and robust audio watermarking scheme based on SVD–DCT. Signal Process **91**(8), 1973–1984 (2011)
18. B Lei, IY Soon, F Zhou, Z Li, H Lei, A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition. Signal Process **92**(9), 1985–2001 (2012)
19. H-T Hu, L-Y Hsu, Robust, transparent and high-capacity audio watermarking in DCT domain. Signal Process **109**, 226–235 (2015)
20. X-Y Wang, P-P Niu, H-Y Yang, A robust digital audio watermarking based on statistics characteristics. Pattern Recogn **42**(11), 3057–3064 (2009)
21. S Wu, J Huang, D Huang, YQ Shi, Efficiently self-synchronized audio watermarking for assured audio data transmission. IEEE Trans Broadcasting **51**(1), 69–76 (2005)
22. X Wang, P Wang, P Zhang, S Xu, H Yang, A norm-space, adaptive, and blind audio watermarking algorithm by discrete wavelet transform. Signal Process **93**(4), 913–922 (2013)
23. H-T Hu, L-Y Hsu, H-H Chou, Variable-dimensional vector modulation for perceptual-based DWT blind audio watermarking with adjustable payload capacity. Digital Signal Processing **31**, 115–123 (2014)
24. A Al-Haj, An imperceptible and robust audio watermarking algorithm, EURASIP J Audio Speech Music Processing **2014**, 37 (2014) doi:10.1186/s13636-014-0037-2
25. X Li, HH Yu, Transparent and robust audio data hiding in cepstrum domain, in IEEE Int Conf Multimedia and Expo (2000), 397–400
26. SC Liu, SD Lin, BCH code-based robust audio watermarking in cepstrum domain. J Inf Sci Eng **22**(3), 535–543 (2006)
27. H-T Hu, W-H Chen, A dual cepstrum-based watermarking scheme with self-synchronization. Signal Process **92**(4), 1109–1116 (2012)
28. S Katzenbeisser, FAP Petitcolas, in *Information hiding techniques for steganography and digital watermarking*, ed. by FAP Petitcolas (Artech House, Boston, 2000)
29. I Daubechies, *Ten lectures on wavelets* (SIAM, Philadelphia, 1992)
30. B Chen, GW Wornell, Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. IEEE Trans Inf Theory **47**(4), 1423–1443 (2001)
31. B Chen, GW Wornell, Quantization index modulation methods for digital watermarking and information embedding of multimedia. J VLSI Signal Processing Systems Signal Image Video Technol **27**(1), 7–33 (2001)
32. P Moulin, R Koetter, Data-hiding codes. Proc IEEE **93**(12), 2083–2126 (2005)
33. H-T Hu, J-R Chang, L-Y Hsu, Windowed and distortion-compensated vector modulation for blind audio watermarking in DWT domain, Multimed Tools Appl 1-21 (2016) doi:10.1007/s11042-016-4202-8
34. H-T Hu, L-Y Hsu, Supplementary schemes to enhance the performance of DWT-RDM-based blind audio watermarking. Circuits Syst Signal Process **36**(5), 1890–1911 (2016)
35. M Fallahpour, D Megias, DWT-based high capacity audio watermarking. IEICE Trans Fundam Electron Commun Comput Sci **E93-A**(1), 331–335 (2010)
36. M Fallahpour, D Megias, High capacity robust audio watermarking scheme based on FFT and linear regression. Int J Innovative Comput Inf Control **8**(4), 2477–2489 (2012)
37. H-T Hu, L-Y Hsu, A DWT-based rational dither modulation scheme for effective blind audio watermarking. Circuits Syst Signal Process **35**(2), 553–572 (2016)
38. H-T Hu, L-Y Hsu, Incorporating spectral shaping filtering into DWT-based vector modulation to improve blind audio watermarking. Wireless Personal Communications **94**(2), 221–240 (2017)
39. G Forney Jr, On decoding BCH codes. IEEE Trans Inf Theory **11**(4), 549–557 (1965)
40. VI Arnold, A Avez, *Ergodic problems of classical mechanics* (Benjamin, New York, 1968)
41. W Fisher, G Doddington, K Goudie-Marshall, The DARPA speech recognition research database: Specifications and status, in Proceedings of DARPA Workshop on Speech Recognition (1986), pp. 93–99
42. P Kabal, An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality, TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University (2002)
43. ITU-T Recommendation P.862 Amendment 1, Source code for reference implementation and conformance tests, [Online]. (2003) Available: http://www.itu.int/rec/T-REC-P.862-200303-S!Amd1/en
44. S Xiang, Audio watermarking robust against D/A and A/D conversions. EURASIP J Adv Signal Process **2011**(1), 1–14 (2011)