IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 20, 2020, accepted December 11, 2020, date of publication January 6, 2021, date of current version January 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3049525

Robust Blind Speech Watermarking via FFT-Based Perceptual Vector Norm Modulation With Frame Self-Synchronization

HWAI-TSU HU, (Member, IEEE), HSIEN-HSIN CHOU, AND TUNG-TSUN LEE
Department of Electronic Engineering, National I-Lan University, Yilan 260, Taiwan, R.O.C.

Corresponding author: Hwai-Tsu Hu (hthu@niu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, R.O.C., under Grant MOST 108-2221-E-197-013 and Grant 109-2221-E-197-024.

ABSTRACT Watermarking is an important measure for protecting proprietary digital multimedia data. This paper presents a novel approach to achieving robust and imperceptible blind speech watermarking on a frame-by-frame basis. The proposed method employs two modules operating in the fast Fourier transform (FFT) domain. The first module is referred to as downward progressive quantization index modulation. It modulates the vector norms drawn from FFT coefficients according to a guideline deduced from human auditory masking properties. The second module is referred to as boundary-constrained iterative adjustment. It provides a smooth transition across frames in the resulting speech waveform. Experiment results confirm the imperceptibility of the proposed modulation scheme in terms of the mean opinion score – listening quality objective (MOS–LQO) based on the perceptual evaluation of speech quality (PESQ) metric. The proposed watermarking method matched and exceeded the performance of five state-of-the-art methods in terms of robustness against common speech processing attacks.

INDEX TERMS Blind speech watermarking, FFT-based perceptual vector norm modulation, downward progressive quantization index modulation, boundary constrained iterative adjustment.

I. INTRODUCTION

Rapid advances in computing and communication technology have made it easier than ever to access multimedia data via the internet. The infringement of copyrighted digital multimedia data via illicit duplication, distribution, and/or modification can result in significant losses for content creators and service providers [1]–[3]. The most common approach to protecting online multimedia data is digital watermarking, which involves embedding confidential information pertaining to ownership within the media file.

The effectiveness of any given watermarking technology is measured in terms of imperceptibility, robustness, and capacity [1], [2]. Ideally, a watermark should hide all necessary information and remain robust to attempts at removal or modification without altering the appearance or sound of the original file. The term 'blind watermarking' refers to techniques that require neither the original source nor other side-information for watermark extraction. Considerable research has gone into watermarking images, audio tracks, and videos; however, there has been relatively little research specific to speech. Speech is a type of audio signal; however, it differs from music signals in terms of temporal continuity, spectral intensity distribution, and production modeling [3]–[5]. The techniques developed for watermarking typical audio files are not always applicable to speech [6].

A common approach to watermarking speech files involves the partitioning of the host signal into frames inside which the watermark is embedded (i.e., frame by frame). For speech watermarking, the embedded target can be the waveform itself, a transformed representation of the speech signal, or the modeling parameters of a source-filtering model. A few watermarking methods have been developed based on source-excitation. Hofbauer *et al.* [7] focused on unvoiced segments of the host speech signal. Coumou and Sharma [8] embedded data through the modification of pitch in voice segments. They also introduced a concatenated coding scheme to safeguard synchronization and error recovery. Chen and Liu [9] developed a watermarking method based on codebook-excited linear prediction (CELP), which involves modifying the position indices of selected excitation pulses.

Speech can be regarded as a signal with time-dependent spectral content; therefore, the spectral envelope of the speech signal seems a sensible choice for embedding a watermark.

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

Faundez-Zanuy *et al.* [5] hid watermark information below the formant peaks deduced via linear prediction (LP) analysis. Chen and Zhu [10] inserted watermark bits inside codebook indices corresponding to the LP coefficients obtained from multistage vector quantization (MSVQ). Yan and Guo [11] converted LP coefficients into inverse sine (IS) parameters and then manipulated the IS parameters via odd-even modulation [12].

The discrete wavelet transform (DWT), which captures both frequency and location information, has long been a standard approach in speech/audio watermarking [4], [13], [14]. Lei *et al.* [3] proposed a sophisticated wavelet-based watermarking scheme specifically for breath sounds. Their scheme employs the lifting wavelet transform (LWT), discrete cosine transform (DCT), and singular value decomposition (SVD) for watermarking, in conjunction with particle swarm optimization (PSO) to optimize performance. Nematollahi *et al.* [4] embedded watermark bits by quantizing the eigenvalue derived from the singular value decomposition of the approximation coefficients in the DWT domain. Saadi *et al.* [15] performed norm-space watermarking in a hybrid domain formed by the DWT and DCT in tandem. Their method provides a good tradeoff between imperceptibility and robustness. Hu *et al.* [14] developed a synchronous package scheme in which packaged information bits are embedded within selected frames in various DWT sub-bands. An improved version with satisfactory robustness against common signal processing attacks was proposed in a follow-up study [16].

Note however that DWT-based watermarking methods suffer from degraded speech quality when the payload capacity exceeds 200 bps. In [17], the fast Fourier transform (FFT) proved highly effective for blind-watermarking audio files. In this study, we developed an FFT-based speech watermarking method, which exploits auditory masking to improve the balance between robustness and imperceptibility.

The remaining of this paper is arranged as follows. Section II discusses the packing protocol by which the watermark bits, address tags, and synchronization codes are arranged for watermarking. Section III describes the technical details of embedding and extracting the watermark in the FFT domain. Section IV presents experiment results evaluating the efficacy of the proposed scheme in terms of quality and robustness against commonly-encountered attacks. Conclusions are drawn in Section V.

II. PROTOCOL FOR ARRANGEMENT OF INFORMATION BITS

As mentioned in the previous section, speech utterances differ from typical audio in the manner of production. Speech signals can generally be categorized as voiced, unvoiced, and silent. The human voice is produced by directing airflow through the vocal tract, including the oral, nasal, and pharyngeal resonant cavities. In voiced speech, the oscillation of vocal folds converts the airflow into a train of flow pulses, resulting in a quasi-periodic signal. In unvoiced speech, the vocal folds are held apart to allow the airflow to pass through the glottis until it is partially or totally obstructed somewhere in the vocal tract. Because of the intrinsic characteristics of the glottal flow, the energy of the voiced speech is normally concentrated below approximately 3 kHz, whereas the energy of the unvoiced speech is located mainly in higher frequencies. Silence occurs when no sound is phonated. Silent segments present only a minuscule amount of energy in the time or frequency domain.

In [14], [16], Hu *et al.* described the process of frame-synchronous speech watermarking. The central idea is to pack information bits into a frame-wise format. The speech signal is initially partitioned into frames of equal size. Next, the frame size suitable for watermarking is identified according to signal intensity. The subsequent steps involve embedding watermark bits into the designated frames. In a similar manner, the scheme proposed here begins with the division of the host speech into non-overlapping frames of length N_f. We selected $N_f = 1024$ to enable the use of FFT to explore the spectral features of the host speech signal. The choice of $N_f = 1024$ reflects an updating rate of 15.625 frames per second, and it renders an adequate frequency resolution within each frame. Let $X_w[l; m]$ denote the l^{th} FFT coefficient derived from a windowed speech signal $x_w(n; m) = w(n)x(n; m)$:

$$X_w(l; m) = \sum_{n=0}^{N_f-1} x_w(n; m) e^{-i\frac{2\pi}{N_f}ln}, \tag{1}$$

where $(n; m)$ denotes the n^{th} element in the m^{th} frame, and $w(n)$ represents a windowing function defined as follows:

$$w(n) = \begin{cases} 0.5\left(1 - \cos\left(2\pi \dfrac{n + 1/2}{N_f/8}\right)\right), \\ n = 0, 1, \cdots, \dfrac{N_f}{16} - 1; \\ 1, \\ \qquad n = \dfrac{N_f}{16}, \cdots, \dfrac{15N_f}{16}; \\ 0.5\left(1 - \cos\left(2\pi \dfrac{n - \frac{7N_f}{8} + \frac{1}{2}}{N_f/8}\right)\right), \\ \qquad n = \dfrac{15N_f}{16}, \cdots, N_f - 1. \end{cases} \tag{2}$$

Function $w(n)$ given above is formed by inserting a boxcar function into the middle of a Hanning window of length $N_f/8$. The advantage of using this special window is twofold. First, it avoids the spectral leakage caused by the rectangular window. Second, this type of window plays a pivotal role in constrained spectral modification, as discussed below.

To enhance robustness against malicious attacks, we tactically embed the watermark in a region of higher spectral intensity. This region generally coincides with the distributive range of the first formant of voiced speech. A frame is classified as embeddable if the spectral energy in the low-frequency range exceeds a given threshold, ψ. "Non-embeddable" frames can also be used to embed information bits; however, they cannot provide sufficient robustness to

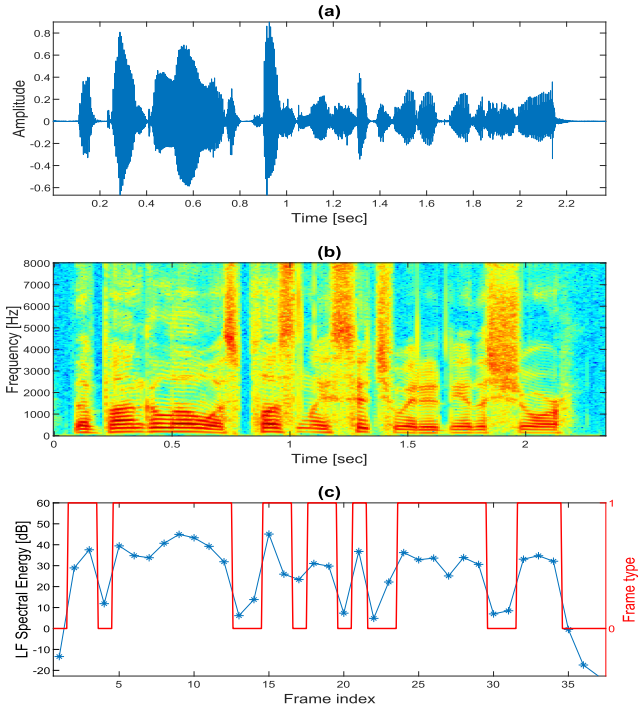

FIGURE 1. Determination of embeddable frames. (a) speech signal, (b) spectrogram, (c) low-frequency spectral energy (marked as "∗") in each frame. The embeddable frames are signified by a high level "1" along a red solid line.

guarantee the retrieval of embedded bits. Mathematically, the frame type $\Lambda(m)$ is defined as

$$\Lambda(m) = \begin{cases} \text{"embeddable"}, & \text{If } \sigma(m) \geq \psi \ ; \\ \text{"non-embeddable"}, & \text{otherwise}, \end{cases} \tag{3}$$

with

$$\sigma(m) = \sum_{n=i_b}^{i_e} X_w(n;m) X_w^*(n;m); \tag{4}$$

$$\psi = 0.01 \max_{m\in\{0,\cdots,M-1\}} \{\sigma(m)\}. \tag{5}$$

where i_b and i_e denote the beginning and ending indexes, respectively. M is the total number of frames. In the current study, these were respectively assigned values of 9 and 57 (roughly corresponding to 150 and 900 Hz for speech sampled at 16 kHz). The superscript symbol "$*$" signifies a complex conjugate operation. Figure 1 presents a typical example of frame classification.

The proposed watermarking scheme was designed for the embedding of ten binary bits into each frame. Figure 2 illustrates the bit arrangement protocol used for speech watermarking, in which 2 out of the 10 embedded bits are reserved for frame-type classification and the remaining 8 bits (or equivalently one byte) conveys watermark information. As shown in Fig. 2, [0 0] indicates non-embeddable frames, and [1 1] indicates embeddable frames containing information bits. Note that [0 1] signifies synchronization frames, which are used to synchronize frames during watermark extraction. Essentially, the synchronization code is a 10-bit sequence (with [0 1] on the top) periodically embedded into the leading frame (i.e., every 16 frames). Subsequent frames can be embeddable or non-embeddable. The first embeddable frame that is encountered is used as an address tag, which points to the starting position of the recovered bytes within the group of 16 frames. We therefore designate this as an "address" frame, designated as [1 0].

III. FFT-BASED SPEECH WATERMARKING

We developed a two-phase speech watermarking method in the FFT domain, referred to as perceptual vector norm modulation (PVNM). Figure 3 illustrates the watermark embedding procedure. The first phase involves data preprocessing, while the second phase deals with watermark embedding. In the first phase, the host speech signal is divided into non-overlapping frames to which the FFT is respectively applied. At the same time, the embeddable frames are identified and the ceiling reference level of the FFT derivatives is determined. In the second phase, downward progressive quantization index modulation (DPQIM) and boundary constrained iterative adjustment (BCIA) are jointly applied to modify the speech signal in accordance with the intended watermark bits.

A. DOWNWARD PROGRESSIVE QUANTIZATION INDEX MODULATION (DPQIM)

To make the embedded watermark resistant to adversary attacks, we deliberately select speech segments of relatively high intensity for watermarking. Most of the energy in a speech signal is concentrated in the low-frequency region; therefore, we embed the watermark into low-frequency FFT coefficients. The embedding procedure involves grouping a set of low-frequency FFT coefficients (termed $\mathbf{I}_k$) into vectors, as follows:

$$\mathbf{I}_k = \{i_b + (k-1)n_b + 1, \cdots, i_b + kn_b\} \tag{6}$$

where $\mathbf{I}_k$ denotes the k^{th} index set of the FFT coefficients in a given vector, i_b is the beginning index, and n_b is the number of coefficients in each vector.

After packing n_b FFT coefficients into a vector, we compute the vector norm as

$$\rho(k;m) = \left(\sum_{l\in\mathbf{I}_k} X_w(l;m) X_w^*(l;m) \right)^{1/2}, \tag{7}$$

where $X_w(l;m)$ is the result of the FFT of windowed speech $x_w(n;m)$. The norm value obtained in (7) is subsequently converted into a decibel value as an indication of the perceived sound level:

$$\rho_{dB}(k;m) = 20\log_{10}(\rho(k;m)). \tag{8}$$

Before we launch the DPQIM, the FFT coefficients and vector norms (in dB) must be ready for all of the frames.

Speech watermarking in each frame is performed by modulating $\rho_{dB}(k;m)$ in accordance with the watermark bit $b_{wm}(k;m) \in \{0,1\}$. DPQIM stemmed from the technique

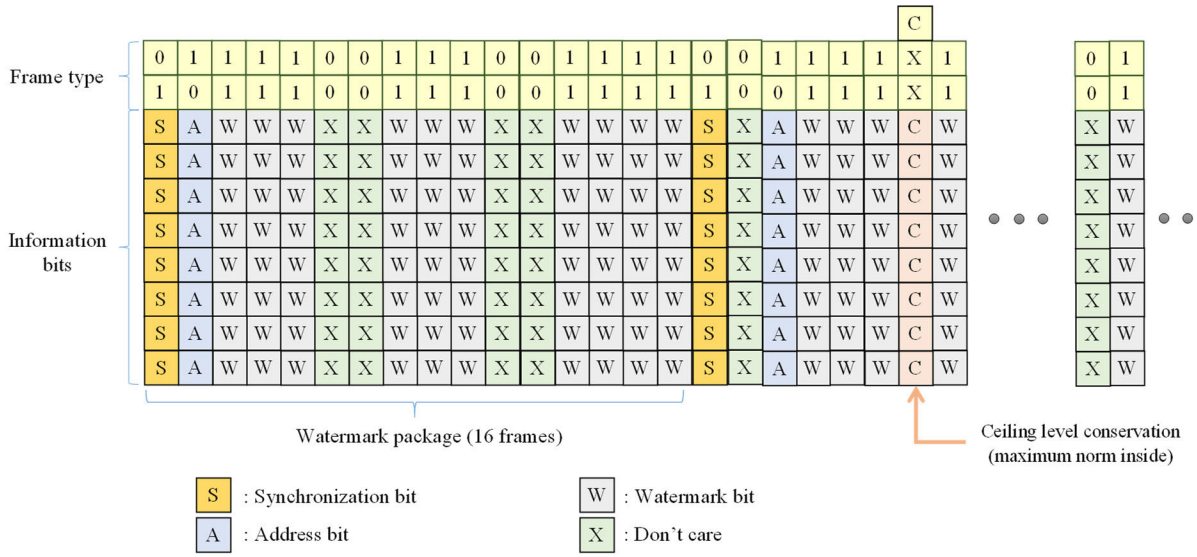

FIGURE 2. Bit arrangement protocal.

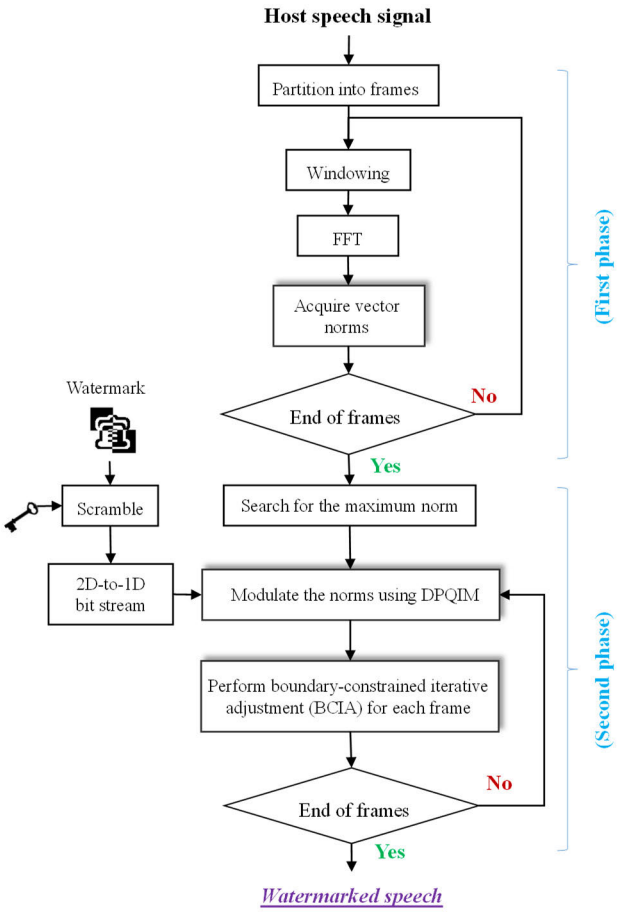

FIGURE 3. Embedding procedure for the proposed FFT-based watermarking method.

developed in [18], in which an incremental quantization step is used to perform the QIM. We observed in experiments that a norm of large magnitude is generally less susceptible to noise perturbation than are norms of smaller magnitude. Thus, it is preferable to use a rather large quantization step for norms of low intensity to preserve the robustness of the watermark. The use of a larger quantization step certainly introduces more distortion; nonetheless, the alteration for a low-level $\rho_{dB}(k;m)$ is relatively minor on a linear scale. It is very likely that the alteration introduced by the watermark can still be perceptually masked by neighboring spectral components of higher intensity. This is why we selected $\rho_{dB}(k;m)$ instead of $\rho(k;m)$ as the target for watermarking.

DPQIM begins with a search for the maximum value among the vector norms across all frames:

$$\alpha = \max_{\substack{0\le l\le L-1\\ 0\le m\le M-1}} \{\rho_{dB}(l;m)\} - \eta = \rho_{dB}(\widehat{l};\widehat{m}) - \eta, \tag{9}$$

where index combination $(\widehat{l};\widehat{m})$ implies that the maximum norm is the $\widehat{l}^{\,th}$ norm in the $\widehat{m}^{\,th}$ frame, η is a small margin used to preserve $\rho_{dB}(\widehat{l};\widehat{m})$, and α serves as a reference level. Throughout the watermarking process, the maximum norm $\rho_{dB}(\widehat{l};\widehat{m})$ must remain intact, while all the other norms are recast to quantized levels not exceeding α. The maximum norm cannot be used in watermark embedding; therefore, we allocate an extra norm to embed the intended bit. The protruding square block depicted in Fig. 2 exemplifies this situation. Based on the formulation in Eq. (9), $\rho_{dB}(\widehat{l};\widehat{m})$ is retrievable as long as the altered norm, $\hat{\rho}_{dB}(l;m)$, does not exceed α.

The gap between reference level α and selected norm $\rho_{dB}(l;m)$ is the target for modulation. Let λ represent the gap, which is defined as follows:

$$\lambda = \alpha - \rho_{dB}(l;m). \tag{10}$$

DPQIM converts λ to a value indicating the number of quantization steps:

$$s_\lambda(l) = \text{sgn}(\lambda) \cdot \frac{-\Delta_l + \sqrt{\Delta_l^2 + 2\delta_l |\lambda|}}{\delta_l}, \tag{11}$$

where Δ_l and δ_l respectively denote the basic and incremental step sizes. We impose a link between the watermark alteration and perceived auditory signal by relating Δ_l and δ_l to the auditory masking threshold $\xi(l)$ using two scaling factors, γ_Δ and γ_δ:

$$\begin{cases} \Delta_l = \gamma_\Delta \xi(l); \\ \delta_l = \gamma_\delta \xi(l), \end{cases} \tag{12}$$

where $\xi(l)$ can be derived from $\rho_{dB}(l;m)$, as outlined in Section III-B.

Note that the $s_\lambda(l)$ obtained in a silent frame tends to be large, due to the slight amount of energy contained in the vector norm. The embedded information in the silent frames is thus susceptible to intentional attacks or unintentional modifications. To secure the frame-type tag "00" residing in the non-embeddable frames, we lay a virtual demarcation, termed I_{bound}, on $s_\lambda(l)$ while carrying out binary embedding. The $s_\lambda(l)$ with a value larger than I_{bound} is assumed to represent a binary "0". Since we use odd and even numbers to signify binary "1" and "0", respectively, I_{bound} is bound to be an even integer, which is 16 in this study. Depending on whether $s_\lambda(l) \geq I_{bound}$, we adopt two different strategies to embed the watermark bit, termed $b_{wm}(l;m)$. In case $s_\lambda(l) \geq I_{bound}$, the watermarked version of $s_\lambda(l)$ will be

$$s_w(l) = \begin{cases} I_{bound} - 1, & \text{if } b_{wm}(l;m) = 1; \\ s_\lambda(l), & \text{if } b_{wm}(l;m) = 0. \end{cases} \tag{13}$$

Note also that there is no need to modify $s_\lambda(l)$ if $s_\lambda(l) \geq I_{bound}$. When $s_\lambda(l) < I_{bound}$, we quantize $s_\lambda(l)$ using

$$s_w(l) = \begin{cases} 2\left\lfloor \dfrac{s_\lambda(l)+1}{2} \right\rfloor + b_{wm}(l;m), \\ \quad \text{if } s_\lambda(l) > 2\left\lfloor \dfrac{s_\lambda(l)+1}{2} \right\rfloor \\ 2\left\lfloor \dfrac{s_\lambda(l)+1}{2} \right\rfloor - b_{wm}(l;m), \\ \quad \text{otherwise.} \end{cases} \tag{14}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. In the above equation, the resulting $s_w(l)$ is designated as the nearest odd integer, if $b_{wm}(l;m) = 1$. Similarly, it is designated as an even integer, if $b_{wm}(l;m) = 0$.

Because of the insignificant energy level of the silent frames, the $s_\lambda(l)'s$ obtained within a silence segment are generally larger than I_{bound}. Accordingly, the retrieved watermark bits are mostly identified as "0's". The two frame-type bits "00" coincide with the tag of a non-embeddable frame. A minor perturbation to $s_\lambda(l)$ will not affect the interpretation of the watermark bits unless a rather large variation occurs. Consequently, the above arrangement conduces to the robustness of the frame-type bits in the non-embeddable frames.

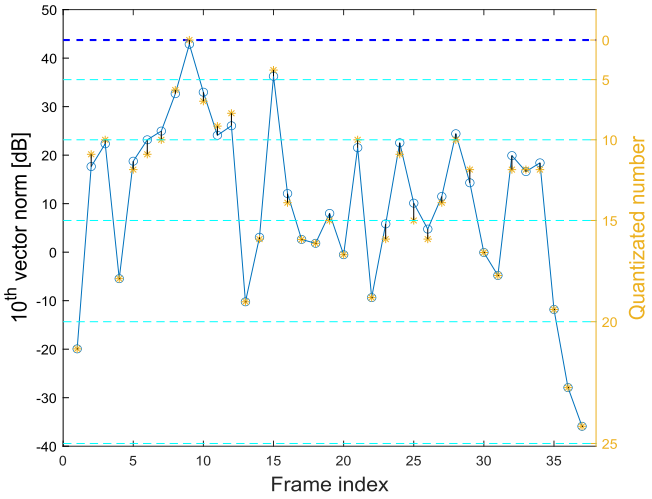

FIGURE 4. Illustration of the DPQIM applied to the 10th FFT vector norms across frames. In each frame, the value of the original norm in decibel scale (as symbolized as circle markers) is moved to a quantized integer (as symbolized as star markers) according to the intended watermark bit.

Once $s_w(l)$ is settled, we recalculate the corresponding gap distance $\hat{\lambda}$ to derive the modified norm $\hat{\rho}(l;m)$ from $s_w(l)$ using the following three equations:

$$\hat{\lambda} = \text{sgn}(s_w(l)) \left(\Delta_l s_w(l) + \delta_l \frac{s_w^2(l)}{2} \right), \tag{15}$$

$$\hat{\rho}_{dB}(l;m) = \alpha - \hat{\lambda}, \tag{16}$$

$$\hat{\rho}(l;m) = 10^{\hat{\rho}_{dB}(l;m)/20}. \tag{17}$$

Equations (11)-(17) illustrate the watermarking process in which $\rho(l;m)$ is changed into $\hat{\rho}(l;m)$ based on binary bit $b_{wm}(l;m)$. Figure 4 illustrates the alteration of the vector norm due to DPQIM. In this example, the original norm (marked as a circle) is moved to a level (marked as an asterisk) equal to a quantized integer. After obtaining $\hat{\rho}(l;m)$, we still need to find a way to realize the necessary changes in the FFT spectrum to ensure accurate retrieval of the vector norms.

B. PERCEPTUAL CONSIDERATIONS IN DETERMINING QUANTIZATION STEP SIZE

The selection of Δ_l and δ_l especially influences watermarking performance. The use of large values for Δ_l and δ_l is conducive to robustness but detrimental to imperceptibility. A common strategy for determining Δ_l and δ_l is to increase the parameters as much as possible without sacrificing imperceptibility. Based on the same principle, we exploited the effect of auditory masking to determine appropriate values for these two parameters. Auditory masking in the frequency domain is a phenomenon in which faint sounds (e.g., watermarking noise) can be rendered inaudible in the presence of loud sounds (e.g., speech). The auditory masking threshold can be modeled as follows:

$$\begin{aligned} a(z) &= \lambda a_{tmn}(z) + (1-\lambda) a_{nmn}(z) \\ &\geq a_{tmn}(z) = -0.275z - 15.025, \end{aligned} \tag{18}$$

where $a_{tmn}(z)$ and $a_{nmn}(z)$ respectively refer to tone-masking-noise and noise-masking-noise functions [19], [20]

and z corresponds to a psychoacoustic measure of frequency on the Bark scale. Conversion from frequency f_{rep} (in Hz) to the Bark scale is performed as follows:

$$z = 13\tan^{-1}\left(0.00076 f_{rep}(l)\right) + 3.5\tan^{-1}\left(\left(f_{rep}(l)/7500\right)^2\right), \quad (19)$$

where $f_{rep}(l)$ denotes the representative frequency for FFT coefficients in the l^{th} vector. Specifically,

$$f_{rep}(l) = \left(i_b + \frac{1}{2}\left((l-1)n_b + 1 + ln_b\right)\right)\frac{F_s}{N_f}, \quad (20)$$

where F_s denotes the sampling frequency.

Suppose that watermark embedding produces a change in the l^{th} vector norm from $\rho_{dB}(l;m)$ to $\rho_{dB}(l;m)+\xi(l)$. Making the watermarking noise inaudible requires that

$$10\log_{10}\left(\frac{10^{\frac{\rho_{dB}(l;m)+\xi(l)}{20}} - 10^{\frac{\rho_{dB}(l;m)}{20}}}{10^{\frac{\rho_{dB}(l;m)}{20}}}\right)^2 \le a\left(z(l)\right) = -0.275z(l) - 15.025. \quad (21)$$

Consequently,

$$10^{\frac{\xi(l)}{20}} - 1 \le 10^{\frac{a(z(l))}{20}} \quad (22)$$

or

$$\xi(l) \le 20\log_{10}\left(1 + 10^{\frac{a(z(l))}{20}}\right). \quad (23)$$

The above inequality suggests that watermark embedding will not cause perceptual degradation as long as $\xi(l)$ is less than the value specified on the right side of Eq. (23). Consequently, we set $\xi(l)$ as the upper bound value.

C. BOUNDARY-CONSTRAINED ITERATIVE ADJUSTMENT (BCIA)

Below, we present an iterative approach to modifying the FFT coefficients without yielding an obvious discontinuity in the waveform at frame boundaries. Figure 5 illustrates the process of BCIA in each iteration. Suppose that we have the following windowed speech signal:

$$x_w^{(t)}(n) = w(n)x^{(t)}(n), \quad (24)$$

where superscript (t) signifies the t^{th} iteration. For the sake of simplicity, we drop the frame index in the mathematical expressions. Following application of the FFT to $x_w^{(t)}(n)$ to derive the FFT sequence $X_w^{(t)}(k)$, we alter the FFT coefficients in each respective vector to achieve norm modulation, as follows:

$$\hat{X}_w^{(t)}(l) = X_w^{(t)}(l)\frac{\hat{\rho}(l)}{\rho(l)+\varepsilon}, l \in \mathbf{I}_k,\ k = 0, 1, \cdots, K-1. \quad (25)$$

where K denotes the number of bits to be embedded. The FFT sequence of a real signal maintains Hermitian symmetry; therefore, the modification must also reflect the mirrored components, as follows:

$$\hat{X}_w^{(t)}(L_f - l) = \left(\hat{X}_w^{(t)}(l)\right)^*. \quad (26)$$

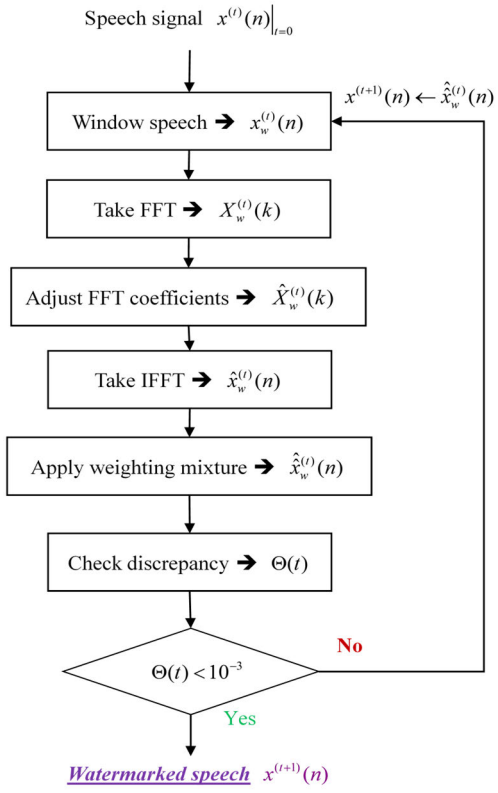

FIGURE 5. Processing flow of the boundary-constrained iterative adjustment.

Taking the inverse FFT with respect to the modified FFT sequence renders a windowed speech signal $\hat{x}_w^{(t)}(n)$; i.e.,

$$\hat{x}_w^{(t)}(n) = \frac{1}{N_f}\sum_{l=0}^{N_f-1}\hat{X}_w^{(t)}(l)e^{i\frac{2\pi}{N_f}ln}. \quad (27)$$

The resulting $\hat{x}_w^{(t)}(n)$ may occasionally exhibit noticeable discontinuities at the frame boundaries; therefore, we adopt a weighting strategy to restrain the artifact as follows:

$$\hat{\hat{x}}_w^{(t)}(n) = w(n)\hat{x}_w^{(t)}(n) + (1 - w(n))\, x_w^{(t)}(n). \quad (28)$$

The windowing function in Eq. (2) is a handy tool to perform this task. The use of Eq. (28) achieves a smooth transition between $\hat{x}_w^{(t)}(n)$ and $x_w^{(t)}(n)$ at frame boundaries. Using $\hat{\hat{x}}_w^{(t)}(n)$, we can further reconstruct the watermarked speech signal as follows:

$$x^{(t+1)}(n) = (1 - w(n))x^{(t)}(n) + \hat{\hat{x}}_w^{(t)}(n). \quad (29)$$

Here we have increased the iteration number by one for $x^{(t+1)}(n)$. The appropriateness of $x^{(t+1)}(n)$ can be evaluated by examining whether there is a noticeable difference between the intermediate result $\hat{\hat{x}}_w^{(t)}(n)$ and $x_w^{(t)}(n)$. As long as the criterion $\Theta(t) = \sum_{n=0}^{N_f-1}\left|\hat{\hat{x}}_w^{(t)}(n) - x_w^{(t)}(n)\right| < 10^{-3}$ is satisfied, we can assume that the FFT sequences of $\hat{\hat{x}}_w^{(t)}(n)$ and $x_w^{(t)}(n)$ resemble each other, and that they hold similar norms for the vectors drawn from each respective FFT sequence.

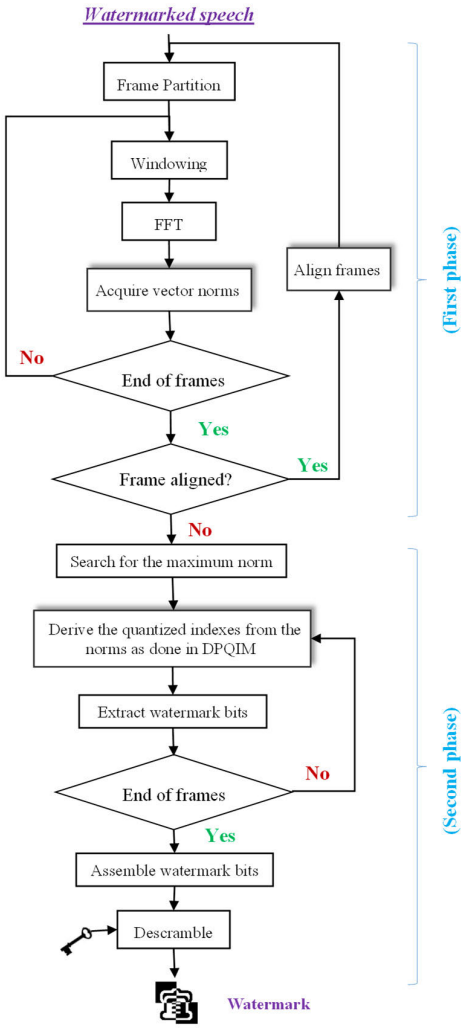

FIGURE 6. Extraction procedure for the proposed FFT-based watermarking method.

Thus, if the criterion is met, then $x^{(t+1)}(n)$ is the final result of the watermarked speech. If the criterion is not met, then the process is repeated for another iteration.

D. FRAME SYNCHRONOUS WATERMARK EXTRACTION

The process of watermark extraction is similar to the process of watermark embedding, as shown in Fig. 6. In the first phase, we acquire the vector norms from the FFT sequence of the watermarked speech signal in each frame. After the maximum norm is identified, we compute gap $\tilde{\lambda}$ between the norm and referential level, as in Eq. (10). Here the variable with a tilde indicates an acquisition from a speech signal that is possibly under attack. Substituting $\tilde{\lambda}$ into Eq. (11) results in a quantized index $\tilde{s}_\lambda(l)$. The watermark bit $\tilde{b}_{wm}(l)$ is "1" if $\tilde{s}_\lambda(l)$ is close to an odd integer, and "0" if $\tilde{s}_\lambda(l)$ is near to an even integer. That is,

$$\tilde{b}_{wm}(l) = \begin{cases} \mathrm{mod}\left(\left\lfloor \tilde{s}_\lambda(l) + \frac{1}{2} \right\rfloor, 2\right), & \text{if } \tilde{s}_\lambda(l) \le I_{bound} = 16; \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

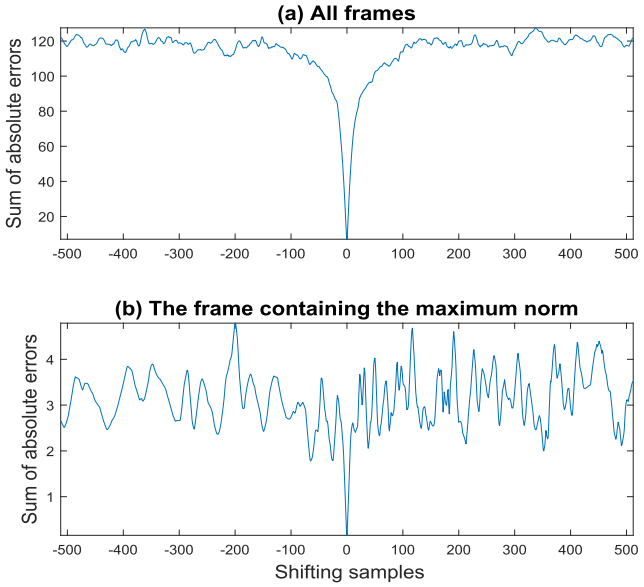

FIGURE 7. Frame synchronization using the sum of absolute quantization errors involving (a) all frames, and (b) the frame with the maximum norm.

where $\mathrm{mod}(\cdot, 2)$ refers to the modulus of 2. The second branch of Eq. (30) implies that $\tilde{b}_{wm}(l)$ will always be regarded as "0" in case the retrieved $\tilde{s}_\lambda(l)$ exceeds a threshold of 16.

Time-shifting and cropping are common desynchronizing attacks leading to the displacement of watermarks. Extracting watermark bits from erroneous locations seldom produces the anticipated results. To ensure the correctness of the extracted watermark, we insert a frame-aligning unit at the end of the first phase to synchronize the speech frames. The aligning unit first selects a speech segment that covers the frame with the maximum norm, and then examines the sum of the absolute quantization errors, termed $\Omega(k)$, for all the vector norms gathered from available frames:

$$\Omega(k) = \sum_{m} \sum_{l=0}^{L-1} \left| \tilde{s}_\lambda(l; m) - \left\lfloor \tilde{s}_\lambda(l; m) + \frac{1}{2} \right\rfloor \right|_{\tilde{x}[\vec{n}+k]}, \tag{31}$$

where $\vec{n}$ is the initial onset of the segment, and $\tilde{s}_\lambda(l; m)$ denotes the quantized outcome resulting from $\tilde{\rho}_{dB}(l; m)$ using the formula given in Eqs. (10) and (11). Theoretically (i.e., according to DPQIM), $\tilde{s}_\lambda(l)$ should be an integer. Hence the function $\Omega(k)$ defined in Eq. (31) is small if the speech frames are perfectly aligned. Figure 7 depicts the outcome of $\Omega(k)$ when one or all of the frames are involved in the alignment process. Frame synchronization can be established by setting the index $\vec{n} + \arg\min_k \{\Omega(k)\}$ as the boundary demarcation.

IV. PERFORMANCE EVALUATION

The test materials comprised 192 sentences uttered by 24 speakers (16 males and 8 females) drawn from the core set of the TIMIT database [21]. Speech utterances were recorded at 16 kHz with 16-bit resolution. For the sake of convenience, all utterances with the same dialect region were cascaded to form a longer file, resulting in a total of 8 speech files for testing. The watermark for the test was a binary image logo measuring 64 × 64 pixels. To enhance the security

of the watermark, the image logo was scrambled using the Arnold transform [22] and converted into a one-dimensional (1-D) bit sequence. The 1-D bit sequence recurred continuously if multiple watermarks could be inserted. In accordance with the arrangement protocol specified in Figure 2, the proposed watermarking method requires at least 32 ($= 64 \times 64/128$) speech segments (each containing one synchronization frame, one address frame, and 16 embeddable frames) to cover the entire watermark. This is equivalent to saying that the watermark requires at least 36.864 ($=32\times(1+1+16)\times(1024/16000)$) seconds in order to embed the entire watermark. In practice, the duration of embedding a full-size watermark may be much longer, depending on the proportion of the embedding frames to the non-embedding frames.

The parameters involved in the FFT-PVNM were as follows: $N_f = 1024$, $i_b = 9$, $i_e = 57$, $n_b = 3$, $\gamma_\Delta = 1$, and $\gamma_\delta = 0.14$. We conducted a comparative performance evaluation (in imperceptibility and robustness) of the proposed FFT-PVNM and five state-of-the-art watermarking methods, namely DWT-SVD [4], LWT-DCT-SVD [3], DWT-DCT-norm [15], DWT-AMM [14], and DWT-IAMM [16] (all expressed in abbreviated form). One similarity among the investigated methods is that they all apply a 1-D transformation to the host speech and perform watermarking on transformed coefficients frame by frame. For those involving the SVD, conversion between a 1-D coefficient sequence and a 2-D matrix is also needed. The frame length and matrix size used in each method follow the specifications in the literature. To ensure a fair comparison, the watermark capacity was set at 200 bits per second (bps). For the proposed FFT-PVNM, we considered two different scenarios: (1) Embedding 10 bits ($K = 10$) in each frame resulted in a payload capacity of 156.25 bps, and (2) Embedding 13 bits ($K = 13$) resulted in a payload capacity of 203.125 bps, which is just above the baseline (i.e., 200 bps). Note that $K = 10$ conforms to the package arrangement discussed in Section II and $K = 13$ is used to illustrate how the setups can affect the FFT sequence. Figure 8 presents the frequency ranges for $K = 10$ and $K = 13$ spanning an illustrative spectrum.

The quality of the watermarked audio signal obtained using the abovementioned methods was evaluated in terms of signal-to-noise ratio (SNR) and the mean opinion score of listening quality objective (MOS-LQO) rendered using the perceptual evaluation of speech quality (PESQ) metric [23]. The definition of SNR is given as follows:

$$SNR = 10\log_{10}\left(\frac{\sum_n \hat{s}^2(n)}{\sum_n \left(\hat{s}(n) - s(n)\right)^2}\right), \tag{32}$$

where $s(n)$ and $\hat{s}(n)$ respectively denote the original and watermarked speech signals. PESQ assesses the speech quality on a scale between -0.5 to 4.5 in terms of MOS-LQO scores. ITU-T Recommendation P.862.1 provides a mapping function to convert the MOS-LQO to a range from 1 (bad) to 5 (excellent) [24].

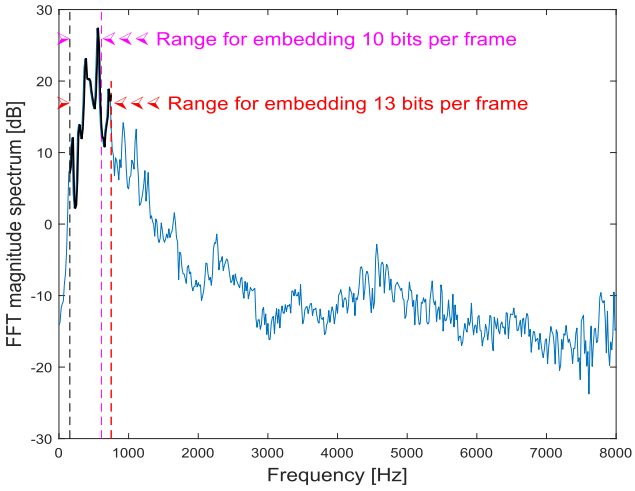

FIGURE 8. Frequency ranges for two settings of embedding watermark bits.

TABLE 1. 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-DCT-norm	15.804 [0.799]	3.078 [0.198]	200
DWT-AMM	21.932 [0.210]	3.254 [0.212]	200
DWT-IAMM	21.354 [0.235]	3.208 [0.195]	200
FFT-PVNM$^{(1)}$	19.224 [0.824]	3.713 [0.058]	156.25
FFT-PVNM$^{(2)}$	17.762 [0.637]	3.633 [0.055]	203.125

As shown in Table 1, the MOS-LQO scores for DWT- SVD, LWT-DCT-SVD, DWT-AMM, and DWT-IAMM were distributed over a narrow range slightly above 3.2, which indicates that these four methods render speech files of comparable perceptual quality. DWT-DCT-norm produced the worst SNR and MOS-LQO values, due to the fact that this method requires a longer duration (or an equivalently higher frequency resolution) for subsampling. The proposed FFT-PVNM achieved the highest score in the PESQ metric despite its moderate performance in terms of SNR. The average MOS-LQO score was 3.633 when the capacity was set at 203.125 bps and 3.713 when the capacity was reduced to 156.25 bps. The higher MOS-LQO scores achieved by the FFT-PVNM can be attributed to the exploitation of auditory masking effects in determining embedding strength.

The robustness of the watermarking methods against various attacks was assessed in terms of BER between the original watermark $\mathbf{B}_{wm} = \{b_{wm}(n)\}$ and recovered watermark $\tilde{\mathbf{B}}_{wm} = \left\{\tilde{b}_{wm}(n)\right\}$:

$$BER\left(\mathbf{B}_{wm}, \tilde{\mathbf{B}}_{wm}\right) = \frac{\sum_{n=0}^{N_w-1} b_{wm}(n) \oplus \tilde{b}_{wm}(n)}{N_w}, \tag{33}$$

TABLE 2. Attack types and specifications.

Item	Type	Description
A	Resampling	Conducting down-sampling to 8 kHz and then up-sampling back to 16 kHz.
B	Requantization	Quantizing the watermarked signal to 8 bits/sample and then back to 16 bits/sample.
C	Lowpass filtering (I)	Applying a lowpass filter with a cutoff frequency of 4 kHz.
D	Lowpass filtering (II)	Applying a lowpass filter with a cutoff frequency of 2 kHz.
E	Amplitude scaling	Scaling the amplitude of the watermarked speech signal by 0.85.
F	Noise corruption (I)	Adding zero-mean white Gaussian noise to the watermarked speech signal with SNR = 30 dB.
G	Noise corruption (II)	Adding zero-mean white Gaussian noise to the watermarked speech signal with SNR = 20dB.
H	DA/AD conversion	Converting the digital speech 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 during playback and recording.
I	Echo addition	Adding an echo signal with a delay of 50 ms and a decay to 5% to the watermarked speech signal.
J	Time shift by 1 sample	Deliberately shifting the watermarked speech signal by one sample.
K	Time shift by 3 samples	Deliberately shifting the watermarked speech signal by three samples.
L	G.722 speech coding	Encoding and decoding the watermarked speech signal with a G.722 wideband speech codec at 64 kbps.
M	G.726 speech coding	Encoding and decoding the watermarked speech signal with a G.726 wideband speech codec at 32 kbps.
N	G.729 speech coding	Encoding and decoding the watermarked speech signal with a G.729 narrowband speech codec at 8 kbps.
O	G.723.1 speech coding	Encoding and decoding the watermarked speech signal with a G.723.1 narrowband speech codec at 6.7 kbps.

TABLE 3. Average bit error rates (in percentage) obtained from the compared watermarking methods under various attacks.

Attack type	DWT-SVD	LWT-DCT-SVD	DWT-DCT-norm	DWT-AMM	DWT-IAMM	FFT-PVNM$^{(1)}$	FFT-PVNM$^{(2)}$
0. none	0	0	0	0	0	0	0
A	0	0	1.7	0	0	0	0
B	0	0	0.32	0.76	0.23	0	0
C	0	0	1.5	0	0	0	0
D	1.19	10.09	14.96	0.39	0.24	0	0
E	46.33	31.82	0	0	0	0	0
F	0	0	0.04	0.73	0	0	0
G	0	0	5.06	3.46	1.79	0.99	0.79
H	53.14	57.48	0.89	1.69	0.31	0.07	0.05
I	1.22	1.92	6.87	0.02	0.13	0.49	0.6
J	5.35	14.47	32.45	1.64	0.21	0	0
K	18.41	29.54	44.82	17.65	14.67	0.18	0.17
L	1.17	0.86	1.02	0.9	0.34	0.7	0.56
M	0.58	0.33	0.81	0.81	0.39	0.37	0.28
N	41.51	49.43	43.99	49.84	50.28	28.57	28.4
O	34.79	43.91	39.11	33.67	33.69	28.85	29.36

where N_w denotes the total number of watermark bits in both $\{b_{wm}(n)\}$ and $\left\{\tilde{b}_{wm}(n)\right\}$. The symbol $\oplus$ stands for the exclusive-OR operation. Table 2 outlines the attacks considered in this study, and Table 3 lists the corresponding average BERs. All of the methods succeeded in retrieving the watermark in the absence of attack and they all performed well in speech compression using the G.722 and G.726 codecs. All of the methods except for the DWT-DCT-norm also survived low-pass filtering and resampling, due to the fact that these attack methods do not have a severe impact on the low-frequency region in which the watermark resides. The inferior performance of DWT-DCT-norm can be attributed to the use of a rather small gap in the norm comparison. Note that a larger gap would tend to have a more pronounced effect on perceptual quality. In the case of amplitude scaling, DWT-SVD and LWT-DCT-SVD failed because they use a fixed quantization step. Noise corruption with SNR = 30 and 20 dB appears not to have caused any serious problems for DWT-SVD and LWT-DCT-SVD. FFT-PVNM merely suffered minor damage in the case where SNR = 20 dB. The effects of re-quantization, echo addition, and DA/AD conversion can be regarded as noise-like attacks; therefore, FFT-PVNM can be said to provide adequate resistance to this type of attack. Note that FFT-PVNM was relatively insensitive to short-range temporal shifting, wherein a shift of 3 samples produced a BER of only 0.17%. Short-range temporal shifting had a profoundly negative effect on the BER results of all the other methods (>14%). The proposed FFT-PVNM passed the tests of both waveform codecs (i.e., G.722 and G.726); however, it provided only limited resistance to CELP-based speech coding. Despite the resultant BERs were nearly 30% in the attacks using the G.723.1 and G.729 standard codecs, the FFT-PVNM still performed better than the others.

Overall, the FFT-PVNM outperformed the other methods in re-sampling, lowpass filtering, amplitude scaling, AD/DA

conversion, few-samples time-shifting, and CELP-based codec compression. The superiority of the FFT-PVNM can be attributed to the exploitation of auditory masking effects in controlling embedding strength and implementation based on DPQIM and BCIA.

V. CONCLUSION

This paper presents a novel FFT-based blind watermarking method operating at 156.25 bps for speech signals. Watermark embedding and extraction are achieved by perceptually manipulating the vector norms drawn from the FFT coefficients of speech frames. The two pivotal modules of the proposed FFT-PVNM method are DPQIM and BCIA. DPQIM exploits auditory masking effects in the assignment of quantization level to the vector norm with the aim of balancing robustness vs. imperceptibility. Based on instructions from DPQIM, the BCIA alters the FFT spectrum to render a smooth transition across frames while simultaneously providing a self-synchronization mechanism. In experiments using the TIMIT core corpus, our watermarked speech files achieved a MOS-LQO score of approximately 3.7. The proposed FFT-PVNM clearly outperformed other watermarking methods in terms of robustness against commonly-encountered attacks. The FFT_PVNM can therefore be considered a dependable tool to secure watermarked speech signals from malicious attacks.

REFERENCES

[1] N. Cvejic and T. Seppanen, *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. Hershey, PA, USA: Information Science Reference, 2008.

[2] X. He, *Watermarking in Audio: Key Techniques and Technologies*. Youngstown, NY, USA: Cambria Press, 2008.

[3] B. Lei, I. Song, and S. A. Rahman, "Robust and secure watermarking scheme for breath sound," *J. Syst. Softw.*, vol. 86, no. 6, pp. 1638–1649, Jun. 2013.

[4] M. A. Nematollahi, S. A. R. Al-Haddad, and F. Zarafshan, "Blind digital speech watermarking based on Eigen-value quantization in DWT," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 27, no. 1, pp. 58–67, Jan. 2015.

[5] M. Faundez-Zanuy, J. J. Lucena-Molina, and M. Hagmüller, "Speech watermarking: An approach for the forensic analysis of digital telephonic recordings," *J. Forensic Sci.*, vol. 55, no. 4, pp. 1080–1087, Mar. 2010.

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

[7] K. Hofbauer, G. Kubin, and W. B. Kleijn, "Speech watermarking for analog flat-fading bandpass channels," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 8, pp. 1624–1637, Nov. 2009.

[8] D. J. Coumou and G. Sharma, "Insertion, deletion codes with feature-based embedding: A new paradigm for watermark synchronization with applications to speech watermarking," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 2, pp. 153–165, Jun. 2008.

[9] O. T.-C. Chen and C.-H. Liu, "Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 5, pp. 1605–1616, Jul. 2007.

[10] N. Chen and J. Zhu, "Multipurpose speech watermarking based on multistage vector quantization of linear prediction coefficients," *J. China Universities Posts Telecommun.*, vol. 14, no. 4, pp. 64–69, Dec. 2007.

[11] B. Yan and Y.-J. Guo, "Speech authentication by semi-fragile speech watermarking utilizing analysis by synthesis and spectral distortion optimization," *Multimedia Tools Appl.*, vol. 67, no. 2, pp. 383–405, Nov. 2013.

[12] D. Kundur, "Multiresolution digital watermarking: Algorithms and implications for multimedia signals," Ph.D. dissertation, Graduate Dept. Elect. Comput. Eng., Univ. Toronto, Toronto, ON, Canada, 1999.

[13] H.-T. Hu and T.-T. Lee, "Hybrid blind audio watermarking for proprietary protection, tamper proofing, and self-recovery," *IEEE Access*, vol. 7, pp. 180395–180408, 2019.

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

[15] S. Saadi, A. Merrad, and A. Benziane, "Novel secured scheme for blind audio/speech norm-space watermarking by Arnold algorithm," *Signal Process.*, vol. 154, pp. 74–86, Jan. 2019.

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

[17] H.-T. Hu and T.-T. Lee, "High-performance self-synchronous blind audio watermarking in a unified FFT framework," *IEEE Access*, vol. 7, pp. 19063–19076, 2019.

[18] H.-T. Hu and L.-Y. Hsu, "Exploring DWT–SVD–DCT feature parameters for robust multiple watermarking against JPEG and JPEG2000 compression," *Comput. Electr. Eng.*, vol. 41, pp. 52–63, Jan. 2015.

[19] X. He and M. S. Scordilis, "An enhanced psychoacoustic model based on the discrete wavelet packet transform," *J. Franklin Inst.*, vol. 343, no. 7, pp. 738–755, Nov. 2006.

[20] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[21] W. Fisher, G. Doddington, and K. Goudie-Marshall, "Ther DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognit.*, 1986, pp. 93–99.

[22] V. I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics*. New York, NY, USA: Benjamin, 1968.

[23] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," Dept. Elect. Comput. Eng., McGill Univ., Montreal, QC, Canada, TSP Lab Tech. Rep., 2002. Accessed: Jan. 2, 2021. [Online]. Available: http://www-mmsp.ece.mcgill.ca/Documents/Reports/

[24] ITU-T Recommendation P.862 Amendment 1. (2003). *Source Code for Reference Implementation and Conformance Tests*. Accessed: Sep. 9, 2020, [Online]. Available: http://www.itu.int/rec/T-REC-P.862-200303-S!Amd1/en

HWAI-TSU HU (Member, IEEE) received the B.S. degree from National Cheng Kung University, Taiwan, in 1985, and the M.S. and Ph.D. degrees from the University of Florida, Gainesville, FL, USA, in 1990 and 1993, respectively, all in electrical engineering. Since 1998, he has been a Professor with the Department of Electronic Engineering, National I-Lan University, Taiwan. His research interests include speech, audio, and image signal processing.

HSIEN-HSIN CHOU received the B.Sc. and M.S. degrees from the Department of Electrical Engineering, Tatung University, in 1982 and 1984, respectively, and the Ph.D. degree from the Department of Electronic Engineering, National Chiao Tung University, in 1989. He is currently a Professor with the Department of Electronic Engineering, National I-Lan University, Taiwan. His research interests include multimedia processing and intelligence system application.

TUNG-TSUN LEE received the B.S. and M.S. degrees in computer science and engineering from National Chiao Tung University, Taiwan, in 1983 and 1985, respectively. Since 1992, he has been a Lecturer with the Department of Electronic Engineering, National I-Lan University, Taiwan. His research interests include software engineering and computer networks.