Research Article

Robust Digital Speech Watermarking For Online Speaker Recognition

**Mohammad Ali Nematollahi,¹ Hamurabi Gamboa-Rosales,²
Mohammad Ali Akhaee,³ and S. A. R. Al-Haddad¹**

¹*Department of Computer & Communication Systems Engineering, Faculty of Engineering, University Putra Malaysia (UPM), 43400 Serdang, Selangor Darul Ehsan, Malaysia*

²*Department of Electronics Engineering, Universidad Autónoma de Zacatecas, 98000 Zacatecas, DF, Mexico*

³*Department of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran*

Correspondence should be addressed to Mohammad Ali Nematollahi; greencomputinguae@gmail.com

Received 13 May 2015; Accepted 30 September 2015

Academic Editor: Zhiqiang Ma

Copyright © 2015 Mohammad Ali Nematollahi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A robust and blind digital speech watermarking technique has been proposed for online speaker recognition systems based on Discrete Wavelet Packet Transform (DWPT) and multiplication to embed the watermark in the amplitudes of the wavelet's subbands. In order to minimize the degradation effect of the watermark, these subbands are selected where less speaker-specific information was available (500 Hz–3500 Hz and 6000 Hz–7000 Hz). Experimental results on Texas Instruments Massachusetts Institute of Technology (TIMIT), Massachusetts Institute of Technology (MIT), and Mobile Biometry (MOBIO) show that the degradation for speaker verification and identification is 1.16% and 2.52%, respectively. Furthermore, the proposed watermark technique can provide enough robustness against different signal processing attacks.

1. Introduction

Security and robustness of speaker recognition systems are the main concerns in online environments [1]. Eight potential cracks are available which made online speaker recognition systems vulnerable [2]. Recently, speech watermarking is used to secure the communication channel against intentional and unintentional attacks for speaker verification and identification purpose [3–7]. For this reason, the watermark is embedded to verify the authenticity of the transmitter (i.e., sensor and feature extractors) and the integrity of the entire authentication mechanism. However, applying speech watermarking can seriously degrade the recognition performance. Since the main aim of the speaker recognition technologies is to enhance recognition performance, applying watermark technology in this context is questionable due to its potential degradation on recognition performance. Available speech watermarking techniques [8–11] embed the watermark in the special frequency range or the speech formants. However, these techniques can seriously degrade the speaker

recognition performance. Furthermore, watermarking and speaker recognition systems have opposite goals whenever the Signal-to-Watermark Ratio (SWR) is decreased and the robustness of the watermark is increased. However, the speaker identification and verification performance can be decreased [5, 6, 12, 13]. Therefore, some researchers apply semifragile watermarking to reduce this impact on recognition performance [14, 15]. Although semifragile watermarking techniques can be used for tamper detection, a requirement is still needed for robust watermarking techniques to protect the ownership.

In this paper, a novel digital speech watermarking technique is proposed for online speaker recognition systems by using Discrete Wavelet Packet Transform (DWPT) and multiplication. For this reason, watermark bits are embedded where less speaker-specific subbands are available. Basically, discriminative speaker features are within low and high frequency bands: glottis is between 100 Hz and 400 Hz, piriform fossa is between 4 kHz and 5 kHz, and constriction of the consonants is 7.5 kHz [16–18].

FIGURE 1: The eight selected wavelet subbands (2, 3, 4, 5, 6, 7, 13, and 14) where less speaker-specific information is available for watermarking by applying DWPT decomposition.

The rest of this paper is organized as follows: first, applied methodology is discussed; second, the proposed digital robust speech watermarking algorithm is explained; third, experimental result on the proposed digital speech watermarking is evaluated; the effect of the proposed robust digital speech watermarking technique on speaker recognition performance is given; and finally, conclusion and future trend are drawn.

2. Methodology

Figure 1 shows the critical bands which are chosen to embed the watermark. As seen in Figure 1, the selected bands have less speaker-specific information which has caused less degradation on the recognition performance of online speaker recognition systems. For this reason, the speech signal has decomposed into 16 critical bands by applying

DWPT. Then, 8 critical bands (with numbers 2, 3, 4, 5, 6, 7, 13, and 14), where the amount of Fisher ratio (F -ratio) is not much, were chosen to have minimum degradation on speaker-specific information. F -ratio curve in Figure 1 is captured from previous work [3, 16] specifically.

3. Robust Digital Speech Watermarking Algorithm

In this section, a robust digital speech watermarking technique based on robust multiplicative technique is proposed. In this technique, the watermark is embedded by manipulating the amplitude of the speech signal [19]. For this reason, the speech signal is segmented into nonoverlapping frames with the length of N . Then, all the sampling of the frame is manipulated based on the following equation:

$$\begin{aligned} r_i &= \alpha \times s_i \quad \text{if } m_i = 1, \\ r_i &= \frac{1}{\alpha} \times s_i \quad \text{if } m_i = 0, \end{aligned} \quad (1)$$

where α is the intensity of the watermark which must be slightly greater than 1, m_i is watermark bit, s_i is the original speech samples, and r_i is watermarked speech samples. Whenever α is increased, the robustness of the watermark is increased, but the imperceptibility is decreased. s_i corresponds to i th samples of the frame. r_i is the i th watermarked sample of the frame.

It is demonstrated [19, 20] that, by knowing the watermark's strength α , variance of the noise, and variance of the original signal, it is possible to extract the watermark bit from the energy of the signal by using a predefined threshold. The detection for watermark bit is based on the following equation:

$$\sum_{i=1}^N r_i^2 \gtrless_1^0 T, \quad (2)$$

where T is the amount of threshold which depends on the variance of the noise and signal. This detection function works well except for gaining attack. If all the samples are multiplied by a constant, the watermark bits cannot be detected at the receiver. In this paper, a rational watermark detection technique has been applied to solve this problem. For this reason, the speech frame is divided into two sets A and B which should have equal length and energy. If their energy is not equal, then their energy can be equalized by using a distortion signal. Next, the watermark bit is embedded into A set based on (1).

For the extraction of the watermark bit from the watermarked frame, (3) has been applied

$$R = \frac{\sum_A r_i^{\text{Order}}}{\sum_B r_i^{\text{Order}}} \gtrless_0^1 T, \quad (3)$$

where Order is an even number and Order = 4 is assumed to provide a tradeoff between robustness and imperceptibility.

Due to the application of DWPT, the distribution of the speech subbands is considered as a Generalized Gaussian

Distribution (GGD) which can be assumed as Weibull distribution when DFT is applied [21]. If GGD is assumed to be $\mu_s^2 = 0$ and σ_s^2 , then it can be expressed as follows:

$$\begin{aligned} f_s(s; \mu, \sigma_s, \nu) \\ = \frac{1}{2\Gamma(1 + 1/\nu) A(\sigma_s, \nu)} \exp \left\{ - \left| \frac{s - \mu}{A(\sigma_s, \nu)} \right|^\nu \right\}, \end{aligned} \quad (4)$$

where $\Gamma(\cdot)$ is gamma function which is represented by $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \cong \sqrt{2\pi} x^{x-1/2} e^{-x}$, ν is the shape of the distribution and can be estimated based on statistical moment of the signal which is discussed briefly in Appendix A.

The amount of threshold for the detection of the watermark bit is estimated for Additive White Gaussian Noise (AWGN) channel. Therefore, the received watermark signal can be expressed based on the following equation:

$$\begin{aligned} r_i &= \alpha \times s_i + n_i \quad \text{if } m_i = 1, \\ r_i &= \frac{1}{\alpha} \times s_i + n_i \quad \text{if } m_i = 0, \end{aligned} \quad (5)$$

where n_i is the noise which is added to the watermarked speech signal. Equation (6) estimates the probability of the watermark bit as follows:

$$\begin{aligned} R | 1 &= \frac{\sum_A (\alpha \times s_i + n_i)^4}{\sum_B (s_i + n_i)^4} \\ \implies R | 1 &= \frac{\alpha^4 \sum_A s_i^4 + 4\alpha^3 \sum_A s_i^3 n_i + 6\alpha^2 \sum_A s_i^2 n_i^2 + 4\alpha \sum_A s_i n_i^3 + \sum_A n_i^4}{\sum_B s_i^4 + 4 \sum_B s_i^3 n_i + 6 \sum_B s_i^2 n_i^2 + 4 \sum_B s_i n_i^3 + \sum_B n_i^4}. \end{aligned} \quad (6)$$

As seen in (6), the amount of the detection threshold depends on the summation of the different parameters. Therefore, different series (which are considered as normal distribution) in the nominator and denominator can be computed based on Central Limit Theorem (CLT). Although some parameters, like $\sum_A n_i^4$, are always positive and cannot be modeled by Gaussian distribution which may be negative, the probability of a negative number which is generated by this Gaussian distribution is very low due to the long length of the speech frames and big amount for μ . As a result, the mean and variance of each parameter of the nominator and denominator are estimated based on (7) and (8), respectively. Consider the following:

$$E \{ \sum s_i^4 \} = \sum E \{ s_i^4 \} = M\mu_4, \quad (7)$$

$$\begin{aligned} \text{var}(\sum s_i^4) &= E \{ (\sum (s_i^4 - M\mu_4))^2 \} \\ &= E \{ (\sum (s_i^4 - \mu_4))^2 \} \\ &= \sum E \{ ((s_i^4 - \mu_4))^2 \} = \sum (E \{ s_i^8 - \mu_4^2 \}) \\ &= M\mu_8 - M\mu_4^2, \end{aligned} \quad (8)$$

where M is the length of each set of A and B . By assuming that $r = 4$ and $r = 8$ and based on the moment of GGD which is computed as in Appendix A, (9) are estimated as follows:

$$\begin{aligned}\mu_4 &= \frac{\sigma_s^4 \Gamma(1/\nu) \Gamma(5/\nu)}{\Gamma^2(3/\nu)}, \\ \mu_8 &= \frac{\sigma_s^8 \Gamma^3(1/\nu) \Gamma(9/\nu)}{\Gamma^4(3/\nu)}.\end{aligned}\quad (9)$$

Therefore, (10) is estimated as follows:

$$\sum s_i^4 \sim \mathcal{N}(M\mu_4, M\mu_8 - M\mu_4^2). \quad (10)$$

By assuming Gaussian signal with zero mean, (11) can be formulated. Consider

$$\begin{aligned}n_i &\sim \mathcal{N}(0, \sigma_n^2) \\ \Rightarrow E\{n_i^m\} \\ &= \begin{cases} 0 & \text{for } m = 2k + 1 \\ (m-1)(m-3)\cdots \times 1 \times \sigma_n^m & \text{for } m = 2k. \end{cases}\end{aligned}\quad (11)$$

The distribution of the noise component with the moment of 4 can be estimated based on the following equation:

$$\sum n_i^4 \sim \mathcal{N}(3M\sigma_n^4, 96M\sigma_n^8). \quad (12)$$

The rest of the components of (6) are simply expressed as follows:

$$\begin{aligned}\sum s_i^3 n_i &\sim \mathcal{N}(0, M\mu_6 \sigma_n^2), \\ \mu_6 &= \frac{\sigma_s^6 \Gamma^2(1/\nu) \Gamma(7/\nu)}{\Gamma^3(3/\nu)}, \\ \sum s_i^2 n_i^2 &\sim \mathcal{N}(M\sigma_s^2 \sigma_n^2, 3M\mu_4 \sigma_n^4 - M\sigma_s^4 \sigma_n^4), \\ \sum s_i n_i^3 &\sim \mathcal{N}(0, 15M\sigma_s^2 \sigma_n^6).\end{aligned}\quad (13)$$

Therefore, by using two free auxiliary parameters p and q which are stated in (14), $R \mid 1, p, q$ is expressed by (15). Consider the following:

$$p = \sum_B s_i^4, \quad (14)$$

$$q = \frac{\sum_A s_i^4}{\sum_B s_i^4},$$

$$\begin{aligned}R \mid 1, p, q \\ &= \frac{\alpha^4 pq + 4\alpha^3 \sum_A s_i^3 n_i + 6\alpha^2 \sum_A s_i^2 n_i^2 + 4\alpha \sum_A s_i n_i^3 + \sum_A n_i^4}{p + 4 \sum_B s_i^3 n_i + 6 \sum_B s_i^2 n_i^2 + 4 \sum_B s_i n_i^3 + \sum_B n_i^4} \\ &= \frac{u}{w},\end{aligned}\quad (15)$$

where u and w are estimated based on the following equations:

$$\begin{aligned}f_U(u) &\sim \mathcal{N}(\alpha^4 pq + 6\alpha^2 M\sigma_s^2 \sigma_n^2 \\ &\quad + 3M\sigma_n^4, 16\alpha^6 M\mu_6 \sigma_n^2 + 36\alpha^4 (3M\mu_4 \sigma_n^4 - M\sigma_s^4 \sigma_n^4) \\ &\quad + 16\alpha^2 \times 15M\sigma_s^2 \sigma_n^6 + 96M\sigma_n^8),\end{aligned}\quad (16)$$

$$\begin{aligned}f_W(w) &\sim \mathcal{N}(p + 6M\sigma_s^2 \sigma_n^2 + 3M\sigma_n^4, 16M\mu_6 \sigma_n^2 \\ &\quad + 36(3M\mu_4 \sigma_n^4 - M\sigma_s^4 \sigma_n^4) + 16 \times 15M\sigma_s^2 \sigma_n^6 \\ &\quad + 96M\sigma_n^8).\end{aligned}\quad (17)$$

For estimating the PDF of $R \mid 1, p, q$, computing the density of u/w is required. By assuming u and w as normal distribution and that they are independent, (18) can be expressed (more details in Appendix B) as follows:

$$f_{R|1,p,q}(r) = \int_{-\infty}^{\infty} |w| f_{U,W}(wr, w) dw. \quad (18)$$

By the assumption of independent and normal distribution of U and W , $f_{U,W}(u, w)$ can be expressed as follows:

$$f_{U,W}(u, w) = f_U(u) \times f_W(w). \quad (19)$$

It should be mentioned that the closed-form solution for (17) is available which is fully discussed in the literature and formulated as in the following equation:

$$\begin{aligned}D(r) &= \frac{b(r)c(r)}{a^3(r)} \frac{1}{\sqrt{2\pi}\sigma_u\sigma_w} \left[2\Phi\left(\frac{b(r)}{a(r)}\right) - 1 \right] \\ &\quad + \frac{1}{a^3(r)\pi\sigma_u\sigma_w} e^{-(1/2)(\mu_u^2/\sigma_u^2 + \mu_w^2/\sigma_w^2)},\end{aligned}\quad (20)$$

where each parameter is expressed as follows:

$$\begin{aligned}a(r) &= \sqrt{\frac{r^2}{\sigma_u^2} + \frac{1}{\sigma_w^2}}, \\ b(r) &= \frac{\mu_u}{\sigma_u^2} r + \frac{\mu_w}{\sigma_w^2}, \\ c(r) &= \exp\left\{\frac{1}{2} \frac{b^2(r)}{a^2(r)} - \frac{1}{2} \left(\frac{\mu_u^2}{\sigma_u^2} + \frac{\mu_w^2}{\sigma_w^2}\right)\right\}, \\ \Phi(r) &= \int_{-\infty}^r \frac{1}{\sqrt{2\pi}} e^{-(1/2)u^2} du.\end{aligned}\quad (21)$$

As a result, the density of $R \mid 1$ can be formulated as follows:

$$\begin{aligned}f_{R|1}(r \mid 1) \\ &= \int_L^U \int_{-\infty}^{\infty} f_{R|1,p,q}(r \mid 1, p, q) f_P(p) f_Q(q),\end{aligned}\quad (22)$$

where L and U are the lowest bound and the highest bound of the energy ratio between two sets of A and B , respectively. As discussed, these two sets should be selected and somehow have equal energy approximately. This situation can be stated as in the following equation:

$$L < \frac{\sum_A r_i^4}{\sum_B r_i^4} < U. \quad (23)$$

The density of parameter P is expressed as in (10). However, the density of parameter q is formulated as in (23) which is estimated from the ratio between normal and independent distributions

$$f_Q(q) = \frac{D(q)}{\int_L^U D(q) dq}. \quad (24)$$

FIGURE 2: Block diagram of embedding process in the proposed robust digital speech watermarking technique.

The probability of $r | 0$ can be computed by using the same manner. Then, the probability of detection error can be estimated based on the following equation:

$$P_e = \frac{1}{2} \int_{-\infty}^{\infty} f(r | 0) dr + \frac{1}{2} \int_{-\infty}^{\infty} f(r | 1) dr. \quad (25)$$

As the main aim of the watermark detector is the minimization of the error, the threshold is calculated as follows:

$$\begin{aligned} \frac{\partial P_e}{\partial T} &= 0 \\ \Rightarrow f_r(T | 0) &= f_r(T | 1). \end{aligned} \quad (26)$$

The amount of the threshold is experimentally computed by using simulation. In the following, the robust digital speech watermarking technique has been developed based on the statistical model which is fully described in this section.

As discussed, the watermark bits are embedded into the specific frequency subbands of DWPT. Details of the embedding and extraction process are presented in the following algorithms.

Embedding process is as follows:

- Segment the original speech signal into frame F_i with lengths of N .
- Apply DWPT on each frame with L levels to compute the different sub-bands.
- Select specific frequency subbands in the last level and arrange them into a data sequence.
- Divide the data sequence into two sets of A and B with equal length of $N/2$ for each set. If these two sets have different energy, their energy is equalized by using a distortion.
- Apply a channel coding technique to improve the robustness of the watermark bits.
- Embed the coded watermark into A set based on multiplication which is expressed in (1).
- Apply inverse DWPT to reconstruct the watermarked signal.

Figure 2 shows the block diagram of the embedding process in the proposed robust speech watermarking technique.

Extraction process is as follows:

- Segment the watermarked speech signal into frame F_i with lengths of N (which can be considered as a public key between the transmitter and receiver).
- Apply DWPT on each frame with L levels to compute the different subbands (which can be considered as a public key between the transmitter and receiver).
- Select specific frequency subbands in the last level and arrange them into a data sequence.
- Divide the data sequence into two sets of A and B with equal length of $N/2$ for each set.
- Extract the watermark bits based on (3).
- Decode the watermark bits which are extracted from all embedding frames.

4. Experimental Setup

In this part, the simulation was done to evaluate the performance of the developed robust digital speech watermarking technique. For this reason, this technique was applied separately to evaluate its performance. The simulation results fully confirmed the mathematical models which were developed in previous section. It must be mentioned that 6300 speech signals of TIMIT database were used in this experiment. The simulation parameters were assumed as follows:

- The frame size was assumed to be 32 ms which was equal to $0.032 \times F_s = 512$ samples. A watermark bit was embedded into each frame. It is clear that whenever the size of the original speech signal is increased, the watermark capacity is increased.
- The required level for DWPT was assumed to be 4. The watermarked subbands were considered as in Figure 1. Daubechies' wavelet function was applied for DWPT.

- (c) Although the watermark's intensity (α) was changed for simulation purpose, the overall assumption was $\alpha = 1.15$.
- (d) For channel coding, Hamming method was used with its parameters being assumed to be $n = 15, k = 11$.
- (e) The threshold was assumed to be 0.95. However, the proper amount for the threshold was expected to be a number near to 1 due to the equalization of energy blocks in the developed algorithm.

In the following, simulations were done to study robustness, imperceptibility, and payload of the developed robust digital speech watermarking technique by using MATLAB. The developed robust digital speech watermarking technique was compared to state-of-the-art digital watermarking techniques including DWT-SVD [22], LWT-SVD [23], and SVD-QIM [24].

4.1. Robustness. Various intentional and unintentional attacks were used for this experiment to evaluate the robustness of the developed technique. The most common method using BER was applied to evaluate the robustness of watermarking which is defined as follows:

$$\begin{aligned} \text{BER}(W, \widehat{W}) &= \frac{\text{Number_of_error_bits}}{\text{Number_of_total_bits}} \\ &= \frac{1}{N} \sum_{i=1}^N W(i) \otimes \widehat{W}(i), \end{aligned} \quad (27)$$

where $\otimes$ is the exclusive OR (XOR) operator, N is the length of the watermark, W is the original watermark, and $\widehat{W}$ is the extracted watermark.

Whenever BER is close to zero (due to less errors between the original and extracted watermarks), the watermark technique provides good robustness.

For better, valid, and fair comparisons of the advantages and disadvantages of each digital watermarking technique, different speech attacks were designed to evaluate the robustness of the different watermarking techniques. The following attacks are common during the speech transmission through telephony channel.

Speech Attacks

Additive White Gaussian Noise (AWGN). A 5 dB noise was added to the signal for simulating ambient distortion.

Low Pass Filter (LPF). An elliptic LPF with 4 KHz cutoff frequency was performed on the watermarked signal.

Band Pass Filter (BPF). An elliptic BPF (from 300 Hz to 3400 Hz for simulation of the narrowband telephony channel) was performed on the watermarked signal.

A-Law. The watermarked signal was compressed and then expanded by A-law with the prevailing parameter of $A = 87.6$.

μ -Law. The watermarked signal was compressed and then expanded by μ -law with the prevailing parameter of $\mu = 255$.

CELP9600. Apply 9.6 kbps CELP codecs on the watermarked speech signal.

CELP 16 K. Apply 16 kbps CELP codecs on the watermarked speech signal.

Amplitude Variation. Increase or decrease the amplitude of the watermarked signal up to 400% (multiplied by 4) or down to 25% (divided by 4).

Resample 8 KHz. Convert the sampling rate of the watermarked signal to 8 KHz and then convert it back to the original one.

Requantization. 16 bits of the watermarked samples were quantized to 8 bits and then were requantized to 16 bits.

Table 1 summarizes the average BER of different digital watermarking techniques. As reported, the average BER for robust DWPT-Multiplication seriously outperformed the other techniques. Based on the results in Table 1, it appeared that embedding the watermarks in the transform domains (DWPT and DWT) was superior to the time domain (SVD-QIM) as the maximum BER in the robust watermarking technique was for SVD-QIM.

4.2. Imperceptibility. Table 2 presents the average of the objective and subjective imperceptibility comparisons among different watermarking techniques (see International Telecommunications Union (ITU-T) method for the subjective measurement of speech quality [25]). For fair comparison, similar frame lengths, watermark bits, and assumptions were used for these watermarking techniques. As seen in Table 2, approximately less MOS music was reported for the speech signals due to the short durations and less energy of the speech signals. Despite the signals' duration and energy, the most important issue for the listeners was reporting the dissimilarities, that is, how much of the quality of speech signals was understood and how much of the quality of the speech signals was enjoyable. Although SNR values for music were expected to be more than those for speech, the listeners still expected more quality from music than speech. Therefore, they reported more MOS speech for this experiment. It can be concluded that the listeners' expectations from the speech signals are different from those for the music signals.

4.3. Capacity. In this experiment, a memoryless binary symmetric channel C_{BSC} as in (28) was applied to compute the capacity of the data channel for error-free watermark transmission through telephony channel, where bitrate (R) for the channel was assumed to be 64 kbps [26]. Consider

$$\begin{aligned} C_{\text{BSC}}(R, \text{BER}) \\ = R \left[1 + \text{BER} \log_2^{\text{BER}} + (1 - \text{BER}) \log_2^{(1-\text{BER})} \right]. \end{aligned} \quad (28)$$

Figure 4 shows C_{BSC} based on the average BER (see Table 1) for different watermarking techniques. It can be seen that the

TABLE 1: Comparison of robustness (BER %) among different watermarking techniques.

Attack	DWT-SVD	Robust DWPT-Multiplication	LWT-SVD	SVD-QIM
No attack	0	0	0	0.02
AWGN (5 dB)	0.41	0.22	0.45	0.53
LPF (4 KHz)	0.43	0.04	0.12	0.41
BPF (300 Hz–3300 Hz)	0.39	0.05	0.25	0.48
A-law	0	0	0	0.08
μ -law	0	0	0	0.14
CELP9600	0.30	0.09	0.32	0.49
CELP16K	0.0817	0.11	0.41	0.42
Amplitude variation	0.49	0	0.42	0.36
Resample 8 KHz	0.51	0.03	0.50	0.54
Requantization	0	0	0	0.44
Average	0.24	0.05	0.22	0.36

TABLE 2: Comparison among different watermarking techniques in terms of objective (SNR), subjective (MOS) measurements, and capacity (bps).

Watermark techniques	MOS speech**	Speech SNR (dB)	MOS music*	Capacity (bps)
Robust DWPT-Multiplication	4.11	28.08	3.22	31.25–125
DWT-SVD	4.88	31.36	3.45	31.25
LWT-SVD	4.51	29.30	3.56	170.67
SVD-QIM	5	42.23	4.14	196

* Efforts to understand the meaning scale were applied.

** Quality scale was applied.

robust DWPT-Multiplication watermarking technique had more capacity than LWT-SVD, QIM-SVD, and DWT-SVD techniques.

4.4. Discussion on the Developed Robust Digital Speech Watermarking. This section discusses the various aspects of the DWPT-Multiplication robust watermarking technique. Firstly, the effects of the threshold are discussed. Figure 5 shows BER in respect to different thresholds (T) for the various AWGN channels. Whenever the threshold became larger, robustness of the watermark improved. Due to the equalization of energy in each block, it was expected that the proper value for the threshold was a number near to 1.

Figure 6 presents BER versus the frame lengths for the developed digital watermarking techniques in various AWGN channels. For Figures 6 and 7, it can be seen that the decreasing length of the frames could increase BER and decrease SNR due to the increase in watermark distortion which was induced by more watermark bits. Increasing SNR in respect to the frame length (when longer frame length was applied to watermarking) caused lesser distortion in the watermarked signal. Choosing a shorter frame length could increase BER because the shorter frame length was easily affected by noise and had small energy. It is clear that increasing the length of the frame means increasing the duration of the host speech signal which can decrease

BER and increase SNR, as demonstrated in Figures 6 and 7. There is a limitation for frame length as increasing the frame length does not improve imperceptibility and robustness anymore. This length is determined by the signal when it appears as quasi-stationary. Furthermore, choosing a longer frame length can be effective to reduce BER in serious noisy conditions. As seen in Figure 6, whenever the size of frame increased, the watermark was detected with less BER for serious noise (SNR = 0 dB). It shows that the frame length is directly affected by robustness and imperceptibility.

Figure 8 shows how BER was changed by the variation of the strength of the watermark (α). The robustness of the developed watermarking technique was improved when the strength of the watermark (α) was increased. This situation is because the watermark is embedded with more strength in the speech signal. As seen in Figure 8, increasing the strength of the watermark (α) was effective only for serious noisy channels (SNR = 0 dB or SNR = 20 dB). It seems that increasing the strength of the watermark (α) for clean conditions is unnecessary.

Figure 9 shows how SNR was changed by the strength of the watermark (α). It is clear that increasing the strength of the watermark (α) can decrease SNR which also decreased the signal quality. Furthermore, more distortion was injected into the host signal by choosing a high value for α . Although $\alpha = 1.5$, SNR was more than 22 dB. There was a difference of 10 dB when α was decreased to 1.01.

FIGURE 3: Block diagram of extraction process in the proposed robust digital speech watermarking technique.

5. Effects of Robust Watermarking on the Performance of Speaker Recognition Systems

In this section, the effects of robust DWPT-Multiplication digital speech watermarking on the performance of speaker

verification were evaluated in terms of two speaker verification systems, that is, i-vector and GMM-UBM systems.

Table 3 shows the effect of robust digital speech watermarking on the performance of different speaker verification systems for different speech databases

$$\frac{|0.71 - 4.52| + |1.45 - 1.59| + |0.79 - 2.39| + |1.9 - 5.24| + |15.04 - 15.34| + |46 - 46.02| + |49.43 - 49.66| + |45.96 - 46.20| + |46 - 47.8| + |46.66 - 47.33| + |47 - 47.59|}{11} = 1.16\%. \quad (29)$$

As seen, the best results were reported for TIMIT speech database which was a clean speech database.

Due to the mismatch in the channel, microphone, and environment, other databases had a lesser performance than TIMIT. As shown in Table 3, it appeared that i-vector with MFCC outperformed other speaker verification systems. Furthermore, robust speech watermarking affected MFCC more than LPRC. From Table 3, the total effect of robust digital speech watermarking was calculated to be 1.16%. This amount is not small and shows that robust digital speech watermarking has some degradation on the performance of online speaker recognition systems which can justify the application of frame selection before robust digital speech watermarking.

Table 4 presents the effect of the developed robust digital speech watermarking system on the performance of GMM speaker identification system in terms of recognition rate. As seen, the best recognition rates were reported for TIMIT speech database which was a clean speech database. Due to the mismatch in the channel, microphone, and environment, other databases had less recognition rate than TIMIT. Furthermore, it seems that MFCC outperformed LPRC. From Table 4, the total degradation of robust digital speech watermarking was calculated to be 2.52%. Therefore, the degradation effect of robust digital speech watermarking can be considerable and can be decreased by applying the frame selection technique

$$\frac{|94.36 - 90.33| + |54.42 - 49.53| + |51.32 - 50.30| + |88.80 - 85.81| + |47.56 - 45.64| + |45.87 - 45.61|}{6} = 2.52\%. \quad (30)$$

As seen in Tables 3 and 4, robust watermarking and the speaker recognition system have opposite goals. Whenever SWR decreased, the robustness of the watermark increased.

However, the performance of speaker recognition decreased as confirmed in previous studies [5, 6, 12, 13]. Therefore, selecting the frames which have less speaker-specific

FIGURE 4: Binary symmetric channel capacity for different watermarking techniques with average BER.

FIGURE 5: Plot BER in respect to the threshold for different AWGN channels.

information for watermarking can result in minimum degradation on the recognition performance of the online speaker recognition systems.

6. Conclusion and Future Works

In this paper, new robust digital speech watermarking technique was developed by applying DWPT and multiplication. This watermarking technique is very robust against different attacks such as filtering, additive noise, resampling, gain, and compression. By embedding the watermark in less speaker specific of the speech subbands, the degradation effect on the recognition performance for this watermarking technique is the minimum.

FIGURE 6: Figure 3 shows the block diagram of the extraction process in the proposed robust speech watermarking technique. Plot BER in respect to frame size for different AWGN channels.

FIGURE 7: Plot SNR in respect to frame length.

It will be the future work to study new adaptive multiplication technique. Also, proposing synchronization technique for this approach could be an improved watermark extraction process.

Appendices

A. Shape of the Distribution Based on Statistical Moment

The estimation for GGD shape parameters has been calculated by using moments of a signal. If $\mu = 0$ and $|s| = y$ is

FIGURE 8: Plot BER versus strength of watermark (α) for different AWGN channels.

FIGURE 9: Plot SNR versus the strength of watermark (α).

assumed for a signal with GGD distribution, then (A.1) can be expressed as follows:

$$f_Y(y) = \frac{1}{\Gamma(1+1/\nu) A(\sigma_s, \nu)} \exp \left\{ -\frac{y^\nu}{(A(\sigma_s, \nu))^\nu} \right\}. \quad (\text{A.1})$$

Therefore, moments of this variable are calculated as follows:

$$\begin{aligned} E\{|s|^r\} &= E\{y^r\} \\ &= \frac{1}{\Gamma(1+1/\nu) A(\sigma_s, \nu)} \int_0^\infty y^r e^{-(y^\nu/(A(\sigma_s, \nu))^\nu)} dy, \quad (\text{A.2}) \\ &\quad r > 0. \end{aligned}$$

TABLE 3: The effect of robust watermarking on the performance of speaker verification for different speech databases.

Database	System	EER (%) without robust WM	EER (%) with robust WM
TIMIT	i-vector + MFCC	0.71	4.52
	i-vector + LPRC	1.45	1.59
	GMM-UBM + MFCC	0.79	2.39
	GMM-UBM + LPRC	1.90	5.24
MIT	i-vector + MFCC	15.04	15.34
	i-vector + LPRC	24.20	24.20
	GMM-UBM + MFCC	46	46.02
	GMM-UBM + LPRC	49.43	49.66
MOBIO	i-vector + MFCC	45.96	46.20
	i-vector + LPRC	46	47.80
	GMM-UBM + MFCC	46.66	47.33
	GMM-UBM + LPRC	47	47.59

TABLE 4: The effect of robust watermarking on the performance of speaker identification for different speech databases.

Feature	Speech database	Recognition rate (%) without robust WM	Recognition rate (%) with robust WM
MFCC	TIMIT	94.36	90.33
	MIT	54.42	49.53
	MOBIO	51.32	50.30
LPRC	TIMIT	88.80	85.81
	MIT	47.56	45.64
	MOBIO	45.87	45.61

By changing the variable $w = y^\nu / (A(\sigma_s, \nu))^\nu$, (A.3) is formulated:

$$\begin{aligned} E\{y^r\} &= \frac{1}{\Gamma(1+1/\nu) A(\sigma_s, \nu)} \\ &\cdot \int_0^\infty [A(\sigma_s, \nu)]^r w^{r/\nu} e^{-w} A(\sigma_s, \nu) \frac{1}{\nu} w^{1/\nu-1} dw \\ &= \frac{[A(\sigma_s, \nu)]^r}{\nu \Gamma(1+1/\nu)} \int_0^\infty w^{(r+1)/\nu-1} e^{-w} dw \\ &= \frac{[A(\sigma_s, \nu)]^r}{\nu \Gamma(1+1/\nu)} \Gamma\left(\frac{r+1}{\nu}\right). \end{aligned} \quad (\text{A.3})$$

Therefore, (A.4) is expressed as follows:

$$E\{y^r\} = \left[\frac{\sigma_s^2 \Gamma(1/\nu)}{\Gamma(3/\nu)} \right]^{r/2} \frac{\Gamma((r+1)/\nu)}{\Gamma(1/\nu)}. \quad (\text{A.4})$$

For $r = 1$, (A.5) is expressed as follows:

$$E\{y\} = \frac{\sigma_s \Gamma(2/\nu)}{\sqrt{\Gamma(3/\nu) \Gamma(1/\nu)}}. \quad (\text{A.5})$$

Finally, the estimation for GGD shape parameters is estimated as follows:

$$\frac{E|s|}{\sqrt{E\{s^2\}}} = \frac{\Gamma(2/\nu)}{\sqrt{\Gamma(3/\nu)\Gamma(1/\nu)}}. \quad (\text{A.6})$$

B. The Computation of Statistical Density of a Ratio between Two Independent Normal Variables

The computation of statistical density of a ratio between two independent normal variables is formulated as follows:

$$\begin{aligned} Z &= \frac{U}{W} \\ \Rightarrow f_Z(z) &=? \\ F_Z(z) &= P(Z \leq z) = P\left(\frac{U}{W} \leq z\right) \stackrel{w>0}{=} P(U \leq Wz) \\ &= E_w \left\{ \frac{P(U \leq wz | w)}{F_{U|W}(wz|w)} \right\} \\ \Rightarrow f_Z(z) &= \frac{\partial}{\partial z} F_Z(z) \\ &= \frac{\partial}{\partial z} E_w \{F_{U|W}(wz | w)\} \\ &= E_w \left\{ \frac{\partial}{\partial z} F_{U|W}(wz | w) \right\} \\ &= E_w \{f_{U|W}(wz | w)\} \\ &= \int w f_{U|W}(wz | w) f_W(w) dw \\ &= \int w f_{U,W}(wz | w) dw. \end{aligned} \quad (\text{B.1})$$

For $w < 0$, the expression $-w$ appeared which is finally expressed as in the following equation:

$$f_Z(z) = \int_{-\infty}^{\infty} |w| f_{U,W}(wz | w) dw. \quad (\text{B.2})$$

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgments

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

References

- [1] M. A. Nematollahi and S. A. Al-Haddad, "Distant speaker recognition: an overview," *International Journal of Humanoid Robotics*, vol. 12, no. 3, pp. 1–45, 2015.
- [2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.
- [3] M. A. Nematollahi and S. A. R. A.-H. Shyamala Doraisamy, "Speaker frame selection for digital speech watermarking," *National Academy Science Letters*, In press.
- [4] M. A. Nematollahi, S. A. R. Al-Haddad, S. Doraisamy, and M. Ranjbari, "Digital speech watermarking for anti-spoofing attack in speaker recognition," in *Proceedings of the IEEE Region 10 Symposium*, pp. 476–479, Kuala Lumpur, Malaysia, April 2014.
- [5] M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, "Speaker verification security improvement by means of speech watermarking," *Speech Communication*, vol. 48, no. 12, pp. 1608–1619, 2006.
- [6] M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, "Speaker identification security improvement by means of speech watermarking," *Pattern Recognition*, vol. 40, no. 11, pp. 3027–3034, 2007.
- [7] S. A. R. S. Al-Haddad, M. Iqbal, A. R. Ramli, and M. A. Nematollahi, "A method for speech watermarking in speaker verification," Google Patents, 2015.
- [8] M. A. Nematollahi, S. A. R. Al-Haddad, S. Doraisamy, and M. I. B. Saripan, "Digital audio and speech watermarking based on the multiple discrete wavelets transform and singular value decomposition," in *Proceedings of the 6th Asia Modelling Symposium (AMS '12)*, pp. 109–114, IEEE, Bali, Indonesia, May 2012.
- [9] M. A. Nematollahi and S. A. R. Al-Haddad, "An overview of digital speech watermarking," *International Journal of Speech Technology*, vol. 16, no. 4, pp. 471–488, 2013.
- [10] M. A. Nematollahi, S. A. R. Al-Haddad, S. Doraisamy, F. Zarafshan, and M. Ranjbari, "Interacting video information via speech watermarking for mobile second screen in Android smartphone," in *Proceedings of the 11th IEEE Student Conference on Research and Development (SCORED '13)*, pp. 325–327, IEEE, Putrajaya, Malaysia, December 2013.
- [11] M. A. Nematollahi, S. Al-Haddad, and F. Zarafshan, "Blind digital speech watermarking based on Eigen-value quantization in DWT," *Journal of King Saud University—Computer and Information Sciences*, vol. 27, no. 1, pp. 58–67, 2015.
- [12] W. Al-Nuaimy, M. A. M. El-Bendary, A. Shafik et al., "An SVD audio watermarking approach using chaotic encrypted images," *Digital Signal Processing*, vol. 21, no. 6, pp. 764–779, 2011.
- [13] A. F. Baroughi and S. Craver, "Additive attacks on speaker recognition," in *Media Watermarking, Security, and Forensics*, vol. 9028 of *Proceedings of SPIE*, International Society for Optics and Photonics, February 2014.
- [14] J. Hämmerle-Uhl, K. Raab, and A. Uhl, "Watermarking as a means to enhance biometric systems: a critical survey," in *Information Hiding*, vol. 6958 of *Lecture Notes in Computer Science*, pp. 238–254, Springer, Berlin, Germany, 2011.
- [15] M. A. Nematollahi, M. A. Akhaee, S. A. R. Al-Haddad, and H. Gamboa-Rosales, "Semi-fragile digital speech watermarking for online speaker recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, article 31, 2015.

- [16] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech Communication*, vol. 50, no. 4, pp. 312–322, 2008.
- [17] S. Hyon, "An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution," in *Proceedings of the IEEE Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC '12)*, pp. 1–4, Hollywood, Calif, USA, December 2012.
- [18] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling," *Speech Communication*, vol. 31, no. 2, pp. 89–106, 2000.
- [19] M. A. Akhaee, N. K. Kalantari, and F. Marvasti, "Robust multiplicative audio and speech watermarking using statistical modeling," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, pp. 1–5, IEEE, Dresden, Germany, June 2009.
- [20] N. K. Kalantari, M. A. Akhaee, S. M. Ahadi, and H. Amindavar, "Robust multiplicative patchwork method for audio watermarking," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 6, pp. 1133–1141, 2009.
- [21] M. A. Akhaee, N. Khademi Kalantari, and F. Marvasti, "Robust audio and speech watermarking using Gaussian and Laplacian modeling," *Signal Processing*, vol. 90, no. 8, pp. 2487–2497, 2010.
- [22] V. Bhat K, I. Sengupta, and A. Das, "An adaptive audio watermarking based on the singular value decomposition in the wavelet domain," *Digital Signal Processing*, vol. 20, no. 6, pp. 1547–1558, 2010.
- [23] B. Lei, I. Yann Soon, F. Zhou, Z. Li, and H. Lei, "A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition," *Signal Processing*, vol. 92, no. 9, pp. 1985–2001, 2012.
- [24] V. Bhat K, I. Sengupta, and A. Das, "A new audio watermarking scheme based on singular value decomposition and quantization," *Circuits, Systems, and Signal Processing*, vol. 30, no. 5, pp. 915–927, 2011.
- [25] I. Rec, *P. 800: Methods for Subjective Determination of Transmission Quality*, International Telecommunication Union, Geneva, Switzerland, 1996.
- [26] I. Rec, *G. 711: Pulse Code Modulation (PCM) of Voice Frequencies*, International Telecommunication Union, Geneva, Switzerland, 1988.