Research Article

Optimization of a Blind Speech Watermarking Technique against Amplitude Scaling

Mohammad Ali Nematollahi,1 Chalee Vorakulpipat,2 and Hamurabi Gamboa Rosales3

1Department of Computer Engineering, Islamic Azad University, Safadasht Branch, Tehran, Iran
2Cybersecurity Laboratory, National Electronics and Computer Technology Center, 112 Phahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand
3Department of Electronics Engineering, Universidad Autónoma de Zacatecas, 98000 Zacatecas, DF, Mexico

Correspondence should be addressed to Chalee Vorakulpipat; chalee.vorakulpipat@nectec.or.th

Received 31 August 2016; Revised 26 November 2016; Accepted 18 December 2016; Published 20 February 2017

1. Introduction

Hiding a secret message in an object has a long history, possibly dating back thousands of years. The rapid growth of computer and communication transmissions has inspired the idea of digital data hiding. Digital watermarking, as a major branch of data hiding, has attracted many researchers [1]. The importance of speech watermarking is gradually increasing because of significant speech transmission through insecure communication channels. There are many approaches for speech watermarking, including spread spectrum (SS), auditory masking, patchwork, transformation, and parametric modeling [2]. In the SS approach, a pseudorandom sequence is used to spread the spectrum of the watermark data and add it to the frequency spectrum of the host signal. However, auditory masking uses unimportant perceptual components of the signal to embed the watermark bits. By contrast, the patchwork approach embeds the watermark data by manipulating two sets of the signal to determine the difference between them. The transformation approach embeds the watermark data into the transformation domains, for example, discrete cosine transform, discrete wavelet transform (DWT), and discrete Fourier transform (DFT). Finally, in the parametric modeling approach, the watermark is embedded by modifying the coefficients of the autoregressive (AR) model.

In addition to speech watermarking approaches, four main embedding strategies are widely applied for watermarking: least significant bit (LSB) replacement, quantization, addition, and multiplication. Among these strategies, quantization has attracted much attention because of blindness, robustness, controlled distortion, and payload. For this purpose, a set of quantizers that are associated with various watermark data are used. However, the quantization strategy suffers from amplitude scaling. To rectify this problem, rational dither modulation (RDM) [3] was proposed to enhance the robustness of quantization index modulation (QIM) [4, 5]; however, it degraded the imperceptibility of the watermarked signal. Hence, hyperbolic RDM [6] was proposed to improve the robustness against power law and gain attacks. Another attempt was made by embedding a watermark into the angle of the signal, known as angle QIM.
This paper attempts to mitigate the limitations of previous research by quantizing the ratio between the Lp-norms of even and odd indices. After quantization, the Lagrange optimization method is applied to compute the best watermarked sample that minimizes the embedding distortion and improves imperceptibility. By assuming Laplacian and Gaussian distributions for the speech and noise signals, respectively, the embedding distortion and error probability are driven analytically and validated by performing a simulation. Moreover, experimental results show that the proposed speech watermarking technique outperforms state-of-the-art watermarking techniques.

Generally, speech watermarking should preserve the identity of the speaker, which is important for certain security applications [12, 13]. To preserve speaker-specific information, some investigations have been conducted to embed the watermark into special frequency subbands that have less energy. Two subsets $X$ and $Y$ are formed with respect to even and odd indices of $S$, respectively, based on (1) and (2), each of length $N/2$, with size $N$.

To solve this optimization problem, the Lagrange method should estimate the optimized values of the equation system as follows:

$$\min J(X) = \sum_{j=1}^{N/2} (X_j^Q - X_j) \times L_X,$$

$$\text{subject to: } C(Y) = \sqrt{\sum_{j=1}^{N/2} (X_j^Q - Z_j^Q) \times L_Y} = 0.$$  

These optimized values are simply computed by solving the following:

$$X_j^{Q,\text{opt}} = \lambda_{\text{opt}} \times X_j,$$

$$\lambda_{\text{opt}} = \frac{Z_j^Q \times L_Y}{L_X}.$$  

2. Proposed Speech Watermarking Technique

In this section, a blind speech watermarking technique is developed based on quantization of the Lp-norm ratio between two blocks of even and odd indices. Assume that $S$ represents an original speech signal that consists of $N$ samples. Two subsets $X$ and $Y$ are formed with respect to even and odd indexed terms, respectively, so that both $X$ and $Y$ have approximately the same energy that causes less embedding distortion. Moreover, synchronization between the transmitter and receiver is most efficient in this case. Figure 1 shows the formation of the subsequences of $X$ and $Y$ from the odd and even indices of the original signal, respectively.

Then, the Lp-norm of both subsequences $X$ and $Y$ are computed, respectively, as follows:

$$L_X = \sqrt{\frac{2}{N} \times \sum_{i=1}^{N} |S_{2i}|^p},$$

$$L_Y = \sqrt{\frac{2}{N} \times \sum_{i=1}^{N} |S_{2i-1}|^p}.$$  

The ratio ($Z$) between $L_X$ and $L_Y$, given as

$$Z = \frac{L_X}{L_Y},$$  

is quantized to embed the watermark bit. Although embedding the watermark into the ratio of Lp-norms can provide high robustness against various attacks, imperceptibility can be seriously degraded.

To resolve this limitation, the variation between the original ratio ($Z$) and quantized ratio ($Z^Q$) should be minimized. Therefore, the Lagrange optimization method is used to minimize this variation; that is, the Lagrange optimization method decreases the embedding distortion after quantization to improve the imperceptibility of the watermarked speech signal. As a result, the Lagrange optimization problems can be formulated as follows:

$$\min J(X) = \sum_{j=1}^{N/2} (X_j^Q - X_j) \times L_X,$$

$$\text{subject to: } C(Y) = \sqrt{\sum_{j=1}^{N/2} (X_j^Q - Z_j^Q) \times L_Y} = 0.$$  

These optimized values are simply computed by solving the following:

$$X_j^{Q,\text{opt}} = \lambda_{\text{opt}} \times X_j,$$

$$\lambda_{\text{opt}} = \frac{Z_j^Q \times L_Y}{L_X}.$$  

2.1. Speech Watermarking Algorithm. The details of the proposed embedding and extraction processes are described in the following algorithms:

**Embedding Process**

(a) Segment the input speech signal ($S$) into different frames ($S_i$) with size $N$.

(b) Form two subsequences $X$ and $Y$, each of length $N/2$, based on the even and odd indices of $S_i$, respectively.

(c) Compute the Lp-norms $L_X$ and $L_Y$ of both the $X$ and $Y$ subsequences, respectively, based on (1) and (2), respectively.

(d) Apply the QIM technique to embed the watermark bit into the ratio between the Lp-norms of $X$ and $Y$ ($Z = L_X/L_Y$) as follows:

$$Z_j^Q = \left[ \frac{Z_j + W_j \times \Delta}{2\Delta} \right] \times 2\Delta + W_j \times \Delta, \quad W_j \in \{0, 1\},$$
Figure 1: Formation of two odd and even subsets from the original speech signal.

where $\Delta$ represents the quantization steps, $W_i$ is the watermark bit, and $Z^Q$ is the modified ratio of the Lp-norms between $X$ and $Y$. Choosing large quantization steps increases the robustness but results in less imperceptibility and vice versa.

(e) Apply the Lagrange method to optimize the values of $X^Q_{2i}$.

(f) Reposition the even and odd subsequences based on $X^Q_{2i}$ and $Y$, respectively.

(g) Rearrange the watermarked speech signal based on the modified frames ($\hat{S}$).

Figure 2 shows the block diagram of the proposed embedding process.

**Extraction Process**

(a) Segment the input watermarked speech signal ($\hat{S}$) into different frames ($\hat{S}_i$) with size $N$.

(b) Form two subsequences $\hat{X}$ and $\hat{Y}$, each of length $N/2$, based on the even and odd indices of $S_i$, respectively.

(c) Compute the Lp-norms $\hat{L}_X$ and $\hat{L}_Y$ of both $\hat{X}$ and $\hat{Y}$ subsequences, respectively, based on (1) and (2), respectively.

(d) Extract the $k$th binary watermark data from the $k$th frame of the watermarked speech signal by selecting the minimum Euclidean distance (nearest quantization step) from the ratio of $\hat{Z}_k = \hat{L}_X/\hat{L}_Y$ as follows:

$$\bar{w}_k = \min \left( \sqrt{\hat{Z}_k^2 + Q_{b_k} \left( \frac{\hat{Z}_k}{\mu_X} \right)^2}, \sqrt{\hat{Z}_k^2 + Q_{b_k} \left( \frac{\hat{Z}_k}{\mu_Y} \right)^2} \right).$$

(9)

where $Q_{b_k}$ is the quantization function while meeting the requirements of watermark bits $b_k = \{0, 1\}$.

Figure 3 shows the block diagram of the proposed extraction process.

3. Statistical Analysis of the Proposed Technique

Generally, Laplacian distribution is the best distribution approach for modeling speech signals within the frame range of 5–50 ms [17, 18]. Laplacian distribution is expressed as

$$f(x) = \frac{b}{2} e^{-b|x-\mu|}, \quad b = \sum_{i=1}^{L} \left| \frac{X_i}{\hat{X}_i} - \mu \right|,$$

(10)

where $L$ is the sample size and $\mu$ is the mean of the random variables. If the subsequences of $X$ and $Y$ are considered as independent, identically distributed (i.i.d) variables, then the distribution of each of them can be assumed to be Laplacian distributions $X = \mathcal{L} (\mu_X, 2\lambda_X^2)$ and $Y = \mathcal{L} (\mu_Y, 2\lambda_Y^2)$, respectively. Based on (3), the ratio ($Z$) between $X$ and $Y$ should be computed. However, the ratio between two Laplacian distributions cannot be computed exactly because the mean and variance are not actually finite in either the Gaussian or Laplacian case. The problem arises because the denominator has nonzero density in the neighborhood of zero. If the denominator is bounded away from zero (immediately it no longer has the ratio of two Laplacian distributions or two normals), then a Taylor expansion should converge to estimate the ratio between two Laplacian distributions. According to Appendix A, the parameters of the ratio can be derived as follows:

$$\mu_Z = \frac{\mu_X}{\mu_Y} \left( 1 + \frac{2\lambda_X^2}{\mu_Y^2} \right),$$

(11)

$$\sigma_Z^2 = \frac{\mu_X^2}{\mu_Y^2} \left( \frac{2\lambda_Y^2}{\mu_X^2} - \frac{4\lambda_Y^4}{\mu_Y^4} + \frac{2\lambda_X^2}{\mu_Y^2} \right).$$

(12)

To estimate the embedding distortion, quantization noise ($\Delta$) should be considered between the original and watermarked speech signals as follows:

$$S_i - \bar{S}_i = (X_{2i} - Y_{2i-1}) - (X_{2i} - Y_{2i-1}).$$

(13)

As in (4) to (6), $\bar{Y}_{2i-1} = Y_{2i-1}$; thus, (12) can be expressed as

$$S_i - \bar{S}_i = \lambda_{opt} \times (X_{2i} - X_{2i}) = (X_{2i} \times (\lambda_{opt} - 1)).$$

(14)

If $Z_i^Q = (L_X + \epsilon)/L_Y$, then $\lambda_{opt}$ can be expressed as

$$\lambda_{opt} = \left( \frac{L_X + \epsilon}{L_Y} \right) \times \frac{L_Y}{L_X} = \left( 1 + \frac{\epsilon}{L_X} \right).$$

(15)
Thus, (13) can be approximately estimated by
\[ S_i - \hat{S}_i \approx \left( 1 + \frac{e}{L_X} \right) X_{2i}, \] (16)
Therefore, the expected values of (13) can be estimated as
\[ E\left[ \| S - \hat{S} \|_2^2 \right] = E \left( \left( 1 + \frac{e}{L_X} \right)^2 E\left[ X_{2i}^2 \right] \right) \]
\[ = \left[ E \left( \frac{1}{L_X} \right) E\left[ (e)^2 \right] + E\left( e \right) E \left( \frac{2}{L_X} \right) + 1 \right] E\left[ X_{2i}^2 \right]. \] (17)
If quantization noise \( e \) is considered as a uniform distribution in \([-\Delta/2, \Delta/2]\) then \( E\left( e \right) = 0 \) and \( E\left( (e)^2 \right) = \Delta^2/48 \). Additionally, as the mean value of the speech signal is considered to be zero, then the zero mean Laplacian distribution is used to model the speech signal as \( E\left[ X_{2i}^2 \right] = 0 \). As a result, \( (X_{2i}^2) = 2b_X^2 \). To model \( E\left( 1/L_X^2 \right) \), the absolute moment of the Laplacian distribution should be estimated using Appendix B as follows:
\[ E\left[ |X|^P \right] = \left( e^{\mu/b} n! \right) \left( \frac{\mu^2}{2} \right) [(-1)^n \cdot I_n + n!], \] (18)
where \( I_n = n I_{n-1} \) and \( I_n = \int_0^\infty t^n e^{-t} dt \). Thus, we can derive the mean and variance for the Pth absolute moment of the Laplacian distribution as
\[ L_X^P = \sum_{j=1}^{N/2} |X_j|^P \sim \mathcal{L} \left( \frac{2 \left( \mu_X(2P) - \mu_X^2 \right)}{N}, \frac{\mu_X(P)}{N} \right). \] (19)

Now, based on (1) and (19), we can compute \( E\left( 1/L_X^2 \right) = E\left( 1/(\sqrt{L_X^2}) \right) = \mu_{X(2P)} \). Therefore, the signal-to-watermark ratio (SWR) can be estimated as
\[ \text{SWR} = \frac{E\left[ \| S \|_2^2 \right]}{E\left[ \| S - \hat{S} \|_2^2 \right]}, \]
\[ \approx \frac{2b_X^2 + 2b_Y^2}{((\Delta^2/48) \times \mu_{X(2P)} + 1) \times (2b_X^2)}. \] (20)
Because both X and Y sets have been selected from the neighboring samples, it can be assumed that \( 2b_X^2 \equiv 2b_Y^2 \). As a result, (20) can be expressed based on the quantization step as
\[ \Delta = \sqrt{-2 \left( 1 + 12 \times \text{SWR} \right) / \text{SWR} \times \mu_{X(2P)}}. \] (21)
To model the error probability, it is assumed that the watermarked speech signal passes through an AWGN channel with zero mean Gaussian noise \( \mathcal{N}(0, \sigma_n^2) \). Therefore, (3) must be rewritten as
\[ \bar{Z} = \sum_{j=1}^{N/2} X_j + N_{X_j} \]
\[ \sum_{j=1}^{N/2} Y_j + N_{Y_j}, \] (22)
where \( N_{X_j} \) and \( N_{Y_j} \) correspond to the odd and even components of the AWGN, respectively. Because the term
$\sum_{j=1}^{N/2} |X_j|^p$ is a known parameter, it is not possible to estimate $Z$ using a chi-square with $N$ degrees of freedom, $\chi^2(N)$. To compute the distribution of $\tilde{Z}$, it should be decomposed and estimated as

$$\tilde{Z} = \frac{\sum_{j=1}^{N/2} \left( |X_j|^p + P |X_j|^{P-1} N_{X_j} + (P (P-1)/2) |X_j|^{P-2} N_{X_j}^2 + \cdots + N_{X_j}^P \right)}{\sum_{j=1}^{N/2} \left( |Y_j|^p + P |Y_j|^{P-1} N_{Y_j} + (P (P-1)/2) |Y_j|^{P-2} N_{Y_j}^2 + \cdots + N_{Y_j}^P \right)}.$$  \hfill (23)

Equation (23) can be expressed as

$$\tilde{Z} = \text{Original } Z + \frac{\sum_{j=1}^{N/2} |X_j|^p}{\sum_{j=1}^{N/2} |Y_j|^p},$$  \hfill (24)

where each part of $\tilde{Z}$ is estimated as follows:

Original $Z = \frac{\sum_{j=1}^{N/2} |X_j|^p}{\sum_{j=1}^{N/2} |Y_j|^p},$

$$y_1 = \frac{\sum_{j=1}^{N/2} P |X_j|^{P-1} N_{X_j}}{\sum_{j=1}^{N/2} |Y_j|^p},$$

$$y_2 = \frac{\sum_{j=1}^{N/2} P |X_j|^{P-1} N_{X_j}}{\sum_{j=1}^{N/2} |Y_j|^p} \times \frac{\sum_{j=1}^{N/2} P |Y_j|^{P-1} N_{Y_j}}{\sum_{j=1}^{N/2} |Y_j|^p},$$

$$y_3 = \frac{\sum_{j=1}^{N/2} P |X_j|^{P-1} N_{X_j}}{\sum_{j=1}^{N/2} |Y_j|^p} \times \frac{\sum_{j=1}^{N/2} P |Y_j|^{P-1} N_{Y_j}}{\sum_{j=1}^{N/2} |Y_j|^p} \times \frac{\sum_{j=1}^{N/2} P |Y_j|^{P-1} N_{Y_j}}{\sum_{j=1}^{N/2} |Y_j|^p}.$$  \hfill (25)

To estimate the probability of error, the noise term can be analyzed because it makes the original $Z$ into a wrong region. Therefore, the distribution of each term of (24) can be estimated by the central limit theorem (CLT) because of the large number of samples in each block. Regardless of the type of original speech signal distribution and because of the independence between the signal and noise samples, the mean and variance of the noise can be computed as

$$\mu_{\text{Noise}} = \mu_{y_1} + \mu_{y_2} + \mu_{y_3},$$

$$\sigma^2_{\text{Noise}} = \sigma^2_{y_1} + \sigma^2_{y_2} + \sigma^2_{y_3}.$$  \hfill (26)

By assuming equal probabilities for both zero and one bit of the watermark data, the probability of error for a fixed quantization step ($\Delta$) can be estimated as

$$P_e = \sum_{i=1}^{\infty} \Pr \left\{ T_{(i-1)/2} < Z^p < T_{(i+1)/2} \right\} \times \sum_{j=-[i/2]}^{\infty} \Pr \left\{ V_{2j+1} < 2^p < V_{2j+1} \right\}. $$  \hfill (27)

A close-form solution for (27) is computed as

$$P_e = \sum_{i=1}^{\infty} \left( Q \left( \frac{T_{(i-1)/2} - \mu_{y_1}}{\sigma_{y_1}} \right) - Q \left( \frac{T_{(i+1)/2} - \mu_{y_1}}{\sigma_{y_1}} \right) \right) \times \sum_{j=-[i/2]}^{\infty} \left( Q \left( \frac{V_{2j+1} - \mu_{y_2}}{\sigma_{y_2}} \right) - Q \left( \frac{V_{2j+1} - \mu_{y_2}}{\sigma_{y_2}} \right) \right),$$  \hfill (28)

where $Q(\cdot)$ is the complementary error function defined as $Q(x) = (1/\sqrt{2\pi}) \int_{x}^{\infty} e^{-u^2/2} du.$

4. Discussion on the Experimental Results

To validate the performance of the developed watermarking technique, a simulation was performed on the TIMIT database to verify the robustness, imperceptibility, and capacity of the technique. The TIMIT database included 630 speakers (438 males and 192 females) with sampling frequency 16 KHz [19]. Each speaker pronounced 10 sentences, which are contained in 6,300 sentences. For the experimental results, the average results of 630 speech signals with duration 1 s to 3 s from 630 speakers were used.

Figure 4 shows the bit error rate (BER) with respect to different $P$ for various frame lengths under Watermark to Noise Ratio (WNR) = 40 dB. In this figure, each curve is plotted separately in order to appear the changes. As can be observed, the frame size was positively correlated with the BER. Whenever the frame size decreased, the BER increased. Additionally, it seems that $P$ was not highly correlated with the BER for $P$ values greater than two. Only a small fluctuation can be observed for the BER when $P$ changed.

Figure 5 shows the BER with respect to different $P$ for various quantization steps. As expected, whenever the quantization step increased, the BER decreased. Furthermore, the variation of $P$ did not seriously change the BER. It must be mentioned that because of perfect watermark detection under clean conditions, a small AWGN was induced on the watermarked signals for the experiments shown in Figures 4 and 5.

Figure 6 shows the variation of the signal-to-noise ratio (SNR) with respect to different $P$ values for different frame lengths. There was not a significant difference in the SNR when the frame size increased. As can be observed, whenever
the frame size increased, the energy level between the two sets of $L_X$ and $L_Y$ increased. Consequently, the ratio between them increased, which caused a lower SNR. Additionally, it seems that changing $P$ was not highly correlated with the SNR for different frame lengths.

Figure 7 illustrates different SNRs with respect to different $P$ for various quantization steps. As observed, $P$ did not highly affect the SNR. However, the quantization step highly affected the SNR. As the quantization step increased, the SNR decreased.
To compute the payload of the proposed watermark, a memoryless binary symmetric channel (BSC) ($C_{\text{BSC}}$) defined as

$$C_{\text{BSC}} = R \times \left[1 + H(P_e)\right]$$

where

$$H(P_e) = P_e \times \log_2(P_e) + (1 - P_e) \times \log_2(1 - P_e)$$

was applied to estimate the capacity of the channel with bitrate ($R$) for error-free watermark transmission [20].

Because the sampling rate of the TIMIT was 16 KHz, $R$ was assumed to be 64 Kbps (8 KHz for speech bandwidth × 8 bits per sample = 64 Kbps) for a telephony channel and $P_e$ was assumed to be equal to the BER in the watermark detection process. Figure 8 shows the amount of the BSC for different WNRs for various quantization steps. As observed, the capacity increased whenever the WNR increased. This is because the watermark was extracted with a minimum BER when the WNR increased. Moreover, it can be inferred that the amount of the BSC increased while the quantization step increased because the watermark was embedded with high intensity when the quantization step increased. As observed, the BSC capacity for fewer quantization steps ($\Delta \leq 0.25$) was approximately zero under a high noisy channel.

Figure 9 shows the variation of the BSC capacity with respect to different WNRs for different frame lengths. As observed, it seems that, under serious noise, the frame size was not a significant factor for the BSC capacity. Despite this, the frame size was likely to be important whenever the WNR increased. Thus, for a large WNR, it is obvious that whenever the frame size increased, the BER in the watermark detection process decreased, which caused an improvement in the BSC capacity.

To demonstrate the efficiency and performance of the proposed speech watermarking technique, the robustness, capacity, and inaudibility of the proposed technique must be compared with other state-of-the-art speech watermarking techniques.

Table 1 describes the benchmark for simulating the results for the robustness test. Many of these attacks are based on the StirMark Benchmark for Audio (SMBA) [24].
Table 1: Benchmark for speech watermarking.

<table>
<thead>
<tr>
<th>Attack type</th>
<th>Attack name</th>
<th>Description</th>
<th>Parameter(s)</th>
<th>Default value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AddBrumm</td>
<td>AddBrumm</td>
<td>It adds buzz or low frequency sinus tone to the watermarked signal to simulate the impact of a power supply</td>
<td>⟨STRENGTH⟩ ⟨FREQUENCY⟩</td>
<td>2500 : 55 to 3000 : 75 A</td>
</tr>
<tr>
<td>AddDynNoise</td>
<td>AddDynNoise</td>
<td>It adds a dynamic white noise to the watermarked signal</td>
<td>⟨STRENGTH⟩</td>
<td>20 to 40 B</td>
</tr>
<tr>
<td>AddFFTNoise</td>
<td>AddFFTNoise</td>
<td>It adds white noise to the watermarked signal in the frequency domain</td>
<td>⟨FFTSIZE⟩ ⟨STRENGTH⟩</td>
<td>256 : 1000 to 1024 : 3000 C</td>
</tr>
<tr>
<td>AddNoise</td>
<td>AddNoise</td>
<td>A white Gaussian noise is contaminated the watermarked signal to simulate ambient distortion</td>
<td>⟨STRENGTH⟩</td>
<td>35 dB level to 5 dB D</td>
</tr>
<tr>
<td>AddSinus</td>
<td>AddSinus</td>
<td>It adds a sinus signal to the watermarked signal</td>
<td>⟨AMPLITUDE⟩ ⟨FREQUENCY⟩</td>
<td>120 : 3000 to 150 : 3500 E</td>
</tr>
<tr>
<td>Resampling</td>
<td>Resampling</td>
<td>The sampling rate of the watermarked signal is converted to ⟨SAMPLERATE1⟩ and then is reconverted to ⟨SAMPLERATE2⟩</td>
<td>⟨SAMPLERATE1⟩ ⟨SAMPLERATE2⟩</td>
<td>4 KHz : 16 KHz to 8 KHz : 16 KHz F</td>
</tr>
<tr>
<td>Conversion</td>
<td>Requantization</td>
<td>The sample of the watermarked signal is quantized to ⟨QUANTIZATION1⟩ and then is requantized to ⟨QUANTIZATION2⟩</td>
<td>⟨QUANTIZATION1⟩ ⟨QUANTIZATION2⟩</td>
<td>8 bits and 16 bits G</td>
</tr>
<tr>
<td>Invert</td>
<td>Invert</td>
<td>It inverts all samples in the watermarked signal, like a 180 degree phase shift</td>
<td></td>
<td>NO PARAMETER REQUIRED None H</td>
</tr>
<tr>
<td>Ambience</td>
<td>Echo</td>
<td>An echo with a delay ⟨DELAY⟩ and decay ⟨DECAY⟩ is added to the watermarked signal</td>
<td>⟨DELAY⟩ ⟨DECAY⟩</td>
<td>20 ms and 10% to 100 ms and 50% I</td>
</tr>
<tr>
<td>Cut samples</td>
<td>⟨REMOVENUMBER⟩</td>
<td>⟨REMOVEDIST⟩ samples are removed from the watermarked signal from every ⟨REMOVEDIST⟩ period</td>
<td>⟨REMOVENUMBER⟩ ⟨REMOVEDIST⟩</td>
<td>1 and 1000 to 7 and 1000 J</td>
</tr>
<tr>
<td>Copy samples</td>
<td>⟨PERIOD⟩ ⟨COPYDIST⟩ ⟨COPYCOUNT⟩</td>
<td>Some of the samples of the watermarked signal are copied between the samples values</td>
<td>⟨PERIOD⟩ ⟨COPYDIST⟩ ⟨COPYCOUNT⟩</td>
<td>1000 : 100 : 30 to 1000 : 200 : 60 K</td>
</tr>
<tr>
<td>LSB Zero</td>
<td>⟨PERIOD⟩ ⟨COPYDIST⟩ ⟨COPYCOUNT⟩</td>
<td>Set all samples of the watermarked signal to zero</td>
<td>NO PARAMETER REQUIRED None</td>
<td>L</td>
</tr>
<tr>
<td>Smooth</td>
<td>⟨PERIOD⟩ ⟨COPYDIST⟩ ⟨COPYCOUNT⟩</td>
<td>The new sample value depends on the samples before and after the modifying point</td>
<td>NO PARAMETER REQUIRED None</td>
<td>M</td>
</tr>
<tr>
<td>Stat1</td>
<td>⟨PERIOD⟩ ⟨COPYDIST⟩ ⟨COPYCOUNT⟩</td>
<td>It averages the sample with its next neighbors</td>
<td>NO PARAMETER REQUIRED None</td>
<td>N</td>
</tr>
<tr>
<td>Dynamics</td>
<td>Amplify</td>
<td>The amplitude of the watermarked signal is increased up to ⟨FACTOR1⟩ and is decreased down to ⟨FACTOR2⟩, respectively</td>
<td>⟨FACTOR1⟩ ⟨FACTOR2⟩</td>
<td>150% and 75% and 200% and 50% O</td>
</tr>
<tr>
<td>Denoising</td>
<td>⟨FACTOR⟩</td>
<td>The watermarked signal is denoised by ⟨FACTOR⟩.</td>
<td>⟨FACTOR⟩</td>
<td>−80 dB to −60 dB P</td>
</tr>
<tr>
<td>Filters</td>
<td>Low Pass Filter (LPF)</td>
<td>The watermarked signal is filtered by an elliptic LPF with cutoff frequency of ⟨FREQUENCY⟩</td>
<td>⟨FREQUENCY⟩</td>
<td>5 KHz to 4 KHz Q</td>
</tr>
<tr>
<td></td>
<td>Band Pass Filter (BPF)</td>
<td>The watermarked signal is filtered by an elliptic filter with bandwidth from ⟨FREQUENCY1⟩ to ⟨FREQUENCY2⟩ to simulate a narrowband telephony channel</td>
<td>⟨FREQUENCY1⟩ ⟨FREQUENCY2⟩</td>
<td>500 Hz &amp; 4000 Hz to 300 Hz &amp; 3400 Hz R</td>
</tr>
<tr>
<td></td>
<td>High Pass Filter (HPF)</td>
<td>The watermarked signal is filtered by an elliptic HPF with cutoff frequency of ⟨FREQUENCY⟩</td>
<td>⟨FREQUENCY⟩</td>
<td>500 Hz to 800 Hz S</td>
</tr>
</tbody>
</table>
Table 1: Continued.

<table>
<thead>
<tr>
<th>Attack type</th>
<th>Attack name</th>
<th>Description</th>
<th>Parameter(s)</th>
<th>Default value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time stretch and pitch shift</td>
<td>Pitch scale</td>
<td>The pitch of the watermarked signal is nonlinearly scaled without changing the time</td>
<td>SCALEFACTOR</td>
<td>1.05 to 1:10</td>
</tr>
<tr>
<td>Time stretch</td>
<td>Time stretch</td>
<td>The time of the watermarked signal is nonlinearly stretched</td>
<td>TEMPOFACTOR</td>
<td>1.05 to 1.10</td>
</tr>
<tr>
<td>Compression</td>
<td>CELP coding</td>
<td>The watermarked signal is coded with rate of (BITRATE) by CELP codecs and then is decoded to original one</td>
<td>BITRATE</td>
<td>16 kbps to 9.6 kbps</td>
</tr>
<tr>
<td>Compression</td>
<td>MP3 compression</td>
<td>The watermarked signal is compressed by MP3 with different rate (BITRATE)</td>
<td>BITRATE</td>
<td>128 to 32</td>
</tr>
<tr>
<td></td>
<td>G.711</td>
<td>The watermarked signal is coded by standard 64 kbps, A/µ-law PCM</td>
<td>NO PARAMETER REQUIRED</td>
<td>None</td>
</tr>
</tbody>
</table>

Figure 8: Variation of the BSC capacity with respect to different WNRs for different quantization steps.

Table 2 compares the BER with state-of-the-art speech watermarking techniques. We implemented all the techniques and tested them for the entire TIMIT corpus under different attacks. As can be observed, the proposed speech watermarking technique has a lower BER overall compared with other techniques.

The perceptual quality of the watermarked signal is critical for the evaluation of the proposed watermarking technique, which can be measured based on the mean opinion score (MOS) (as proposed by the International Telecommunications Union (ITU-T) [23]) and SNR. The MOS uses a subjective evaluation technique to score the watermarked signal, which is presented in Table 3. In the MOS evaluation method, 10 people were asked to listen blindly to the original and watermarked signals. Then they reported the dissimilarities between the quality of the original and watermarked speech signals. The average of these reports were computed for MOS music and MOS speech and presented in Table 4.

An objective evaluation technique, such as SWR and SNR, attempts to quantify this amount based on the following formula:

\[
\text{SNR} = 10 \times \log_{10} \frac{\sum_{n} s^2}{\sum_{n} (\hat{s} - s)^2}, \tag{31}
\]

where \(S\) and \(\hat{S}\) are the original and watermarked signals, respectively.

Table 4 presents a comparison of the proposed technique and other techniques in terms of imperceptibility and capacity. Based on the results, it seems that the proposed speech watermarking technique outperformed the other techniques in terms of capacity and imperceptibility. Although the SNR for formant tuning [21] is higher than the proposed technique, the capacity and robustness of the proposed technique are greater than those for formant tuning [21] and Analysis-by-Synthesis [22].
Table 2: Comparison with the robustness of different speech watermarking techniques in terms of BER (%).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No attack</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>A</td>
<td>1.91–4.23</td>
<td>2.09–5.43</td>
<td>3.65–6.45</td>
<td>7.96–9.65</td>
</tr>
<tr>
<td>E</td>
<td>0.32–2.02</td>
<td>0.763–1.14</td>
<td>1.23–2.32</td>
<td>2.98–4.32</td>
</tr>
<tr>
<td>F</td>
<td>13.54–17.23</td>
<td>14.32–17.65</td>
<td>26.23–37.83</td>
<td>29.45–33.06</td>
</tr>
<tr>
<td>G</td>
<td>3.23</td>
<td>2.65</td>
<td>19.32</td>
<td>23.87</td>
</tr>
<tr>
<td>H</td>
<td>0.23</td>
<td>0.00</td>
<td>9.43</td>
<td>12.45</td>
</tr>
<tr>
<td>I</td>
<td>1.34–4.65</td>
<td>2.34–5.11</td>
<td>4.65–10.43</td>
<td>8.23–16.43</td>
</tr>
<tr>
<td>J</td>
<td>1.23–2.54</td>
<td>1.32–4.67</td>
<td>6.54–10.54</td>
<td>11.54–18.87</td>
</tr>
<tr>
<td>K</td>
<td>1.32–3.16</td>
<td>1.78–4.23</td>
<td>7.51–10.34</td>
<td>11.49–19.43</td>
</tr>
<tr>
<td>L</td>
<td>0.92</td>
<td>1.98</td>
<td>1.50</td>
<td>4.04</td>
</tr>
<tr>
<td>M</td>
<td>3.12</td>
<td>5.76</td>
<td>10.34</td>
<td>21.68</td>
</tr>
<tr>
<td>N</td>
<td>4.10</td>
<td>4.23</td>
<td>6.65</td>
<td>9.54</td>
</tr>
<tr>
<td>O</td>
<td>1.21–2.54</td>
<td>0.00–1.43</td>
<td>5.97–8.76</td>
<td>8.98–15.54</td>
</tr>
<tr>
<td>P</td>
<td>1.00–3.54</td>
<td>2.43–5.43</td>
<td>9.65–14.56</td>
<td>19.65–26.45</td>
</tr>
<tr>
<td>Q</td>
<td>21.43–29.43</td>
<td>24.54–31.43</td>
<td>40.54–44.43</td>
<td>50.09–50.32</td>
</tr>
<tr>
<td>R</td>
<td>4.84–9.54</td>
<td>5.32–10.32</td>
<td>16.65–29.44</td>
<td>20.54–36.98</td>
</tr>
<tr>
<td>S</td>
<td>13.32–18.54</td>
<td>15.00–19.43</td>
<td>20.43–29.23</td>
<td>28.54–30.76</td>
</tr>
<tr>
<td>T</td>
<td>1.32–2.32</td>
<td>2.01–3.13</td>
<td>7.43–10.43</td>
<td>9.65–15.32</td>
</tr>
<tr>
<td>U</td>
<td>0.15–0.23</td>
<td>0.18–0.43</td>
<td>1.45–3.21</td>
<td>4.32–5.43</td>
</tr>
<tr>
<td>V</td>
<td>6.54–9.54</td>
<td>11.43–14.54</td>
<td>1.32–4.21</td>
<td>2.32–4.32</td>
</tr>
<tr>
<td>W</td>
<td>10.43–20.34</td>
<td>11.43–25.34</td>
<td>36.32–45.65</td>
<td>33.43–50.32</td>
</tr>
<tr>
<td>X</td>
<td>23.11</td>
<td>24.17</td>
<td>48.32</td>
<td>50.65</td>
</tr>
</tbody>
</table>

Table 3: MOS grades [23].

<table>
<thead>
<tr>
<th>MOS (5)</th>
<th>Quality</th>
<th>Quality scale</th>
<th>Effort required to understand meaning scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5)</td>
<td>Excellent</td>
<td>Imperceptible</td>
<td>No effort required</td>
</tr>
<tr>
<td>(4)</td>
<td>Good</td>
<td>Perceptible, but not annoying</td>
<td>No appreciable effort required</td>
</tr>
<tr>
<td>(3)</td>
<td>Fair</td>
<td>Slightly annoying</td>
<td>Moderate effort required</td>
</tr>
<tr>
<td>(2)</td>
<td>Poor</td>
<td>Annoying</td>
<td>Considerable effort required</td>
</tr>
<tr>
<td>(1)</td>
<td>Bad</td>
<td>Very annoying</td>
<td>No meaning was understood</td>
</tr>
</tbody>
</table>

As observed in Table 4, each entity was bounded between two values that related a particular value of imperceptibility (SNR and MOS) to a particular capacity. Consequently, when the capacity increased, imperceptibility decreased. The trade-off value is completely application dependent and should be determined by the user.

5. Performance Analysis

Generally, two types of errors, false positive probability (FPP) and false negative probability (FNP), must always be analyzed to validate the security of a watermarking system [25]. FPP is defined when an unwatermarked speech signal is declared as a watermarked speech signal by the watermark extractor. Similarly, FNP is defined when the watermarked speech signal is declared as an unwatermarked speech signal by the watermark extractor. By assuming that the watermark bits are independent random variables, both the FPP and FNP can be formulated based on Bernoulli trials, which is expressed as follows:

$$ P_e = \sum_{i=0}^{T-1} \binom{N}{i} P_{FN}^i (1 - P_{FN})^{(N-i)} + \sum_{i=T}^{N} \binom{N}{i} P_{FP}^i (1 - P_{FP})^{(N-i)} $$

(32)

where $N$ is the total number of watermark bits; $i$ is the number of matching bits; $\binom{N}{i}$ is a binomial coefficient; $P_{FP}$ is the probability of a false positive, which is assumed to be 0.5;
Table 4: Comparison of various watermarking techniques in terms of payload and imperceptibility.

<table>
<thead>
<tr>
<th>Technique</th>
<th>Quality scale</th>
<th>Effort required to understand meaning scale</th>
<th>SNR (dB)</th>
<th>Theoretical payload (bps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formant tuning [21]</td>
<td>4.98–4.32</td>
<td>5.00–4.55</td>
<td>30.32–27.54</td>
<td>33.33–50</td>
</tr>
<tr>
<td>The proposed method</td>
<td>4.87–3.65</td>
<td>5.00–4.05</td>
<td>42.11–20.71</td>
<td>40–400</td>
</tr>
</tbody>
</table>

Figure 10: FPP with respect to various total number of watermark bits for different BER.

Figure 11: FNP with respect to various total number of watermark bits for different BER.

By assuming Laplacian distribution for the speech signal and Gaussian distribution for the noise signal, the probability of error and watermarking distortion were modeled based on a statistical analysis of the proposed technique. Additionally, experimental results not only proved that the developed watermarking technique was highly robust against different attacks, such as compression, AWGN, filtering, and resampling, but also demonstrated the validity of the analytical model. For future work, an investigation on synchronization and adaptive quantization techniques might contribute to the proposed watermarking technique.

6. Conclusion and Future Work

In this paper, a gain invariant speech watermarking technique was developed using the Lagrange optimization method. For this purpose, samples of the signal were separated based on odd and even indices. Then the ratio between the $L_p$-norms was quantized using the QIM method. Finally, the Lagrange method was used to estimate the optimized values. In a similar manner, the extraction process detected the watermark data blindly by finding the nearest quantization step.

Appendix

A. Estimation of the Mean and Variance of the Ratio of Two Laplacian Variables Based on Taylor Series

In [26], the bivariate second-order Taylor expansion for $f(x, y)$ around $\theta = (E(x), E(y))$ is expressed as follows:

$$f(x, y) = f(\theta) + f_x(\theta)(x - \theta_x) + f_y(\theta)(y - \theta_y) + \frac{1}{2} \left\{ f_{xx}(\theta)(x - \theta_x)^2 + 2f_{xy}(\theta)(x - \theta_x)(y - \theta_y) + f_{yy}(\theta)(y - \theta_y)^2 \right\} + \text{remainder}. \quad (A.1)$$
Therefore, \( E[f(X, Y)] \) can be expanded about \( \theta = (E(X), E(Y)) \) to compute the approximate values as follows:

\[
E(f(X, Y)) = f(\theta) + \frac{1}{2} \left\{ f''_{XX}(\theta) \text{ var}(X) + 2f''_{XY}(\theta) \text{ cov}(X, Y) + f''_{YY}(\theta) \text{ var}(Y) \right\} + O\left(n^{-1}\right).
\]  

(A.2)

For \( f = R/S \), \( f''_{RR} = 0 \), \( f''_{RS} = -S^{-2} \), and \( f''_{SS} = 2R/S^3 \). Then, the mean and variance of the ratio between \( R \) and \( S \) \( E(R/S) \), respectively, can be estimated as follows:

\[
E\left(\frac{R}{S}\right) = E\left(f(R, S)\right) = \frac{E(R)}{E(S)} - \frac{\text{cov}(R, S)}{E(S)^2} + \frac{\text{var}(S)}{E(S)^3} \mu_R \left(1 + \frac{\sigma_R^2}{\mu_R^2}\right),
\]

\[
\text{var}\left(\frac{R}{S}\right) = \frac{1}{E^2 S} \frac{\text{var}(R)}{E^3} + 2 \frac{-E R}{E^3 S} \text{cov}(R, S)
\]

(A.3)

B. Compute the Absolute Moment of the Laplacian Distribution

The moment of Laplacian distribution expressed as follows:

\[
E(|X|^n) = \int_{-\infty}^{\infty} |X|^n \cdot e^{-((X-\mu)/b)} \, dx \quad \text{for} \quad n \geq 1,
\]

(B.1)

\[
= \frac{1}{2b} \int_{-\infty}^{\infty} |X|^n \cdot e^{-(X-\mu)/b} \, dx.
\]

There are two cases, \( X \geq \mu \) and \( X < \mu \):

\[
E(|X|^n) = \begin{cases} 
\frac{1}{2b} \int_{\infty}^{\infty} |X|^n \cdot e^{-(X-\mu)/b} \, dx & \text{if } X \geq \mu \\
\frac{1}{2b} \int_{-\infty}^{\infty} |X|^n \cdot e^{-(X-\mu)/b} \, dx & \text{if } X < \mu
\end{cases}
\]

(B.2)

For first case, when \( X \geq \mu \),

\[
E(|X|^n) = \frac{1}{2b} \left[ \int_{\infty}^{\infty} X^n \cdot e^{-(X-\mu)/b} \, dx + \int_{0}^{\infty} X^n \cdot e^{-(X-\mu)/b} \, dx \right].
\]

(B.3)

If \( t = -X/b \), then \( I \) can be expressed as

\[
I = b^{n+1} \int_{-\infty}^{0} t^n e^{-t} \, dt = b^{n+1} \cdot n! = n!.
\]

(B.4)

\( I_n \) can also be expressed as

\[
I_n = \int_{-\infty}^{0} X^n e^{X/b} \, dx = \int_{-\infty}^{0} (b \cdot t)^n e^{-t} \cdot b \cdot dt
\]

\[
= b^{n+1} \int_{-\infty}^{0} t^n e^{-t} \, dt
\]

\[
= b^{n+1} \int_{-\infty}^{0} \sum_{k=0}^{n} \frac{(-1)^{k} t^k}{k!} \cdot \frac{d}{dt} \left[ e^{-t} \right] \, dt
\]

\[
= \sum_{k=0}^{n} \frac{(-1)^{k} b^{n+1}}{k!} \cdot \frac{d}{dt} \left[ e^{-t} \right] \bigg|_{t=-\infty}^{0}
\]

\[
= \sum_{k=0}^{n} \frac{(-1)^{k} b^{n+1}}{k!} \cdot \frac{d}{dt} \left[ e^{-t} \right] \bigg|_{t=-\infty}^{0} = 0 + n I_{n-1}.
\]

Substituting (B.4) and (B.5) into (B.3), the absolute moment of the Laplacian distribution can be computed based on

\[
E(|X|^n) = \left(\frac{e^{\mu b} b^n}{2}\right) \left[(-1)^n \cdot I_n + n! \right].
\]

(B.6)

Competing Interests

The authors declare that they have no competing interests.

References


Submit your manuscripts at
https://www.hindawi.com