Journal of Information Hiding and Multimedia Signal Processing ©2019 ISSN 2073-4212
Ubiquitous International Volume **10**, Number **4**, December **2019**

Speech Watermarking Based on Source-filter Model of Speech Production

Shengbei Wang, Weitao Yuan *, Jianming Wang

Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems
Tianjin Polytechnic University, Binshuixi Road, Tianjin, China
wangshengbei@tjpu.edu.cn; weitaoyuan@hotmail.com; wangjianming@tjpu.edu.cn

Masashi Unoki

School of Information Science
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
unoki@jaist.ac.jp

Received June 2019; revised September 2019

Abstract. *Illegal use of advanced techniques has enabled the speech to be easily reduplicated and edited. Watermarking can effectively prevent the speech from unauthorized operations. A watermarking method for speech signals is proposed in this paper by taking advantages of speech production and the mechanism of speech codecs. Speech signal is first separated into two components, i.e., sound source information (residue) and vocal tract information (characterized by formants), according to the source-filter model. The sound source and vocal tract information are separately embedded with watermarks using quantization index modulation (QIM) based and the formant enhancement (FE) based watermarking. Evaluations related to inaudibility and robustness were carried out on the proposed method. The results revealed that the proposed method could satisfy inaudibility. Moreover, its robustness could be increased in comparison with single method. These results verified the effectiveness of the proposed method.*
Keywords: Speech watermarking, Source-filter model, Formant enhancement, Quantization index modulation, Inaudibility, Robustness

1. **Introduction.** Digital speech can be easily distributed over the Internet and transmission system. Digital speech offers several advantages over analog speech and the modifications to digital speech can be made at the exact location of the whole signal. However, these advances also enable ordinary people to perform unauthorized operations, such as reduplicating, editing, or tampering to the speech signals. These unauthorized operations have resulted in serious problems in the protection of speech signals.

Speech watermarking is a promising technique to protect the speech. General watermarking methods are required to satisfy three basic requirements: inaudibility, blindness and robustness. Some additional requirements may also need to be satisfied for particular purposes. Typical methods, e.g., least significant bit-replacement (LSB) [1] and direct spread spectrum (DSS) [2] methods can just partially satisfy the requirements. In recent works, Kazemi et al. [3] proposed a watermarking for cellular networks based on spread spectrum. Sarreshtedari et al. [4] proposed to embed the compressed version of speech into original signal. Hofbauer et al. [5] presented a watermarking to the phase of non-voiced speech. Wu et al. [6] implemented a method based on odd/even modulations, however,

*Corresponding author: Weitao Yuan, weitaoyuan@hotmail.com

this method was not robust against code-excited linear prediction (CELP) speech codec. Likewise, such problem was also found in [7]. It should be noted that the trade-off between inaudibility and robustness is difficult to accomplish and most existing methods are not completely robust, and they are especially not robust against different speech codecs.

Recent watermarking scheme which can combine two watermarking methods together has been proposed to achieve stronger robustness. This kind of methods takes advantage of the fact that the watermarks embedded with one method can assist or refine the watermark extraction of the other method. In literature, several methods have been explored for images [8], [9] and audio [10], [11]. In [10], spread spectrum (SS) and singular value decomposition (SVD) were combined for copyright protection. In this method, the destroyed watermarks in SS or SVD were likely to be recovered from the other domain, leading to stronger robustness.

Since watermarks embedded with two methods can mutually complement each other, these methods have superior performance in robustness. However, there are generally three challenging issues in designing speech watermarking with this concept: (i) two methods are combined together as a whole method, it is important to guarantee that the watermarks embedded with one method will not destroy the watermarks embedded with the other method, i.e., one method should not affect the other method; (ii) one speech signal will be applied by two watermarking methods, i.e., sound distortion will be doubly introduced, thus it is difficult to maintain the sound quality of watermarked speech; (iii) speech signals need to be encoded/decoded by various codecs, it is difficult to realize a watermarking that is robust against all kinds of codecs.

This paper proposes a watermarking method based on the source-filter model of speech production. Speech signal is separated into two components, i.e., sound source and vocal tract information, which are separately embedded with watermarks with quantization index modulation (QIM) based and the formant enhancement (FE) based watermarking methods. It can overcome the above three challenging issues from the following aspects. (i) According to the source-filter model, the interactions between source and vocal tract information are not strong and this can effectively prevent one watermarking method from being affected by the other. (ii) Sound distortion can be minimized by carefully controlling the parameters in QIM and FE. (iii) The source-filter model and line spectral frequencies (LSFs) are widely adopted by speech codecs, implementing the watermarking based on this model is helpful in achieving the robustness against different speech codecs.

This paper is organized as follows. Section 2 introduces the main concept of the proposed method. Sections 3 and 4 talk about the implementation of two watermarking methods. Section 5 explains the whole scheme of the proposed method. A frame synchronization scheme is also designed in this section. Section 6 evaluates the proposed method with respect to inaudibility and robustness. Section 7 compares the proposed method with other typical methods. The last section gives a summary of our work.

2. **Concept of watermarking.**

Speech signals usually need to be encoded/decoded with speech codecs. Most speech codecs utilize the mechanism of speech production, i.e., the source-filter model of speech production, to improve the efficiency of speech compression while maintaining good speech quality. Therefore, the source-filter model should be considered to attain the robustness of watermarking against speech codecs.

This paper considers the source-filter model to design speech watermarking. The source-filter model assumes the glottal pulse is the sound source and the vocal tract is a filter. Human beings can independently control glottal pulse and vocal tract. Therefore, the sound source and vocal tract filter are assumed to be independent of each other in most

speech production based research [15]. This inspired us to investigate if speech watermarking could be implemented by separately applying watermarking to the sound source and the vocal tract without affecting each other.

The linear prediction (LP) [16] can be used to separate the speech into sound source and vocal tract information:

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i), \tag{1}$$

where p is the LP order, a_i are the LP coefficients which provide an accurate estimation of the formants to characterize the vocal tract. The $\hat{x}(n)$ is the prediction of $x(n)$ and $x(n-i)$ is the i-th previous sample of $x(n)$. The prediction error $e(n)$ between $x(n)$ and $\hat{x}(n)$ is sound source, i.e., residue:

$$e(n) = x(n) - \hat{x}(n). \tag{2}$$

2.1. **Watermarking for sound source (residue).** Previous studies have proven that the human auditory system is not very sensitive to slight phase modifications [17], [18]. Accordingly, the phase of residue is suitable for watermark embedding. As both sound distortion and robustness can be caused in a manner that is proportional to the magnitude of frequency components, phase should be modified according to the magnitude of frequency components, to balance inaudibility and robustness. In our method, the phase of frequency components which have high magnitude is slightly modified to reduce sound distortion, and also maintain robustness; the phase of frequency components which have low magnitude is sufficiently modified to maintain robustness, where speech quality will not be greatly distorted. These modifications are achieved with the QIM [19] [20].

2.2. **Watermarking for vocal tract.** Vocal tract information can be represented with formants and formants can be enhanced to improve the sound quality [21], [22], [23], [24]. The method of reshaping formants to make them sharper is commonly referred to as formant enhancement (FE). In general, the LP coefficients can be used to estimate formants, however, LP coefficients are sensitive to noise. The LSFs [27], as substitute parameters of LP coefficients, are less sensitive to noise and able to directly control formants [23], [25], [26]. In addition, the LSFs are employed in many source-filter model based speech codecs, embedding watermarks into vocal tract based on FE by controlling LSFs would enable the watermarking method to be robust against different speech codecs.

2.3. **Perturbation analysis.** The interaction between sound source and vocal tract is first checked to ensure the feasibility of the proposed method. Here, we investigate when there is perturbation in sound source (or vocal tract), how vocal tract (or sound source) will be affected. 30 speech samples (0.25 s, 20 kHz, and 16 bits) from the Advanced Telecommunications Research (ATR) database (B set) were used [28]. Each sample was separated into sound source and vocal tract information using LP analysis of order 16. Perturbations were separately added to the phase spectrum $\angle R(W)$ of residue and the LSFs $\Phi=\{\phi_i, i = 1, ..., 16\}$ (converted from LP coefficients), where W indicated frequency bins. The experimental results were calculated on the average of 30 speech samples.

Experiment 1: Random perturbations were added to $\angle R(W)$. The perturbated phase $\angle \hat{R}(W)$ was calculated in Eq. (3),

$$\angle \hat{R}(W) = \angle R(W) + w_r \times \Gamma_r \times \|\angle R(W)\|_\infty \tag{3}$$

where w_r was an additive white Gaussian noise (AWGN) with standard deviation $\sigma=1.0$, Γ_r set as $\{0.25, 0.50, 0.75, 1.00\}$ controlled the perturbation strength, and the infinite norm of $\|\angle R(W)\|$, i.e., $\|\angle R(W)\|_\infty$, adjusted the perturbation to match the range of $\angle R(W)$. The perturbated residue and original LP coefficients was synthesized to obtain

TABLE 1. Statistical analysis under perturbated residue.

Strength (Γ_r)	Perturbation in phase		Differences between original speech and perturbated speech							
	Pha. (π)	Prop. (%)	LP Env.	Prop. (%)	Mag.	Prop. (%)	LP coeff.	Prop. (%)	LSFs (π)	Prop. (%)
0.25	0.3926	26.44	0.1166	3.01	1.4061	8.35	0.2030	18.07	0.0050	0.34
0.50	0.7884	53.14	0.2518	6.51	2.0847	12.38	0.4316	38.23	0.0107	0.72
0.75	1.1779	79.18	0.2854	7.37	2.4243	14.42	0.5438	48.32	0.0137	0.93
1.00	1.5732	105.9	0.2648	6.85	2.3986	14.28	0.4651	41.43	0.0124	0.84

TABLE 2. Statistical analysis under perturbated LSFs.

Strength (Γ_l)	Averaged perturbation in LSFs				Information in residue					
	LSFs (π)	Prop. (%)	LP coeff.	Prop. (%)	Res.	Prop. (%)	Mag.	Prop. (%)	Pha. (π)	Prop. (%)
0.15	0.0272	15.27	0.0173	1.81	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.25	0.0487	27.31	0.0318	3.32	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.50	0.0886	49.72	0.0567	5.94	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.75	0.1331	74.24	0.0792	8.31	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

the perturbated speech. LP coefficients and LSFs re-calculated from the perturbated speech showed how they were affected by the perturbated phase.

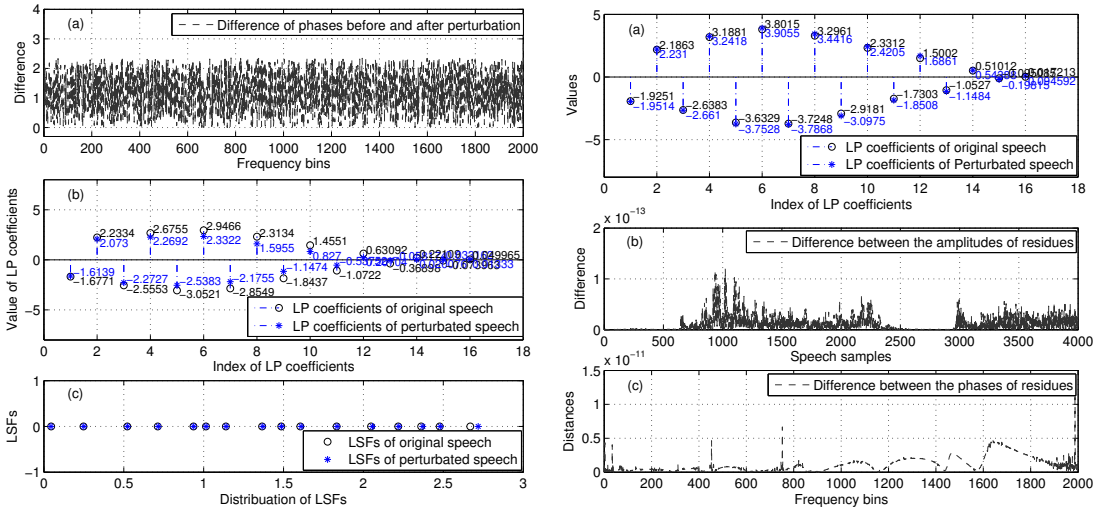
Table. 1 (Column 2-3) lists the perturbation in phase spectrum (in π) and the perturbation proportion (Prop.) under different strengths, where Prop. calculates the perturbation caused changes in proportional to the original phase spectrum. These perturbations caused obvious differences (Column 4-9) between original speech and perturbated speech in LP envelope (LP env.), spectral magnitude (Mag.), and LP coefficients (LP coeff.). In contrast, the differences in LSFs were quite trivial (Column 10-11). For Γ_r=1.00, the perturbation proportion in phase reached to 105.9% and there were obvious differences in LP coefficients, however, LSFs only slightly changed. An example under Γ_r=0.75 is shown in Fig. 1 (left panel). The strong perturbations in phase (see left panel (a)) caused obvious differences in LP coefficients (see left panel (b)) but did not disturb the LSFs (see left panel (c)). These results verified that (1) LSFs were more stable than LP coefficients; (2) the perturbations in phase had slight influence on LSFs; (3) Current perturbation strength in phase was much stronger than those of watermarks and LSFs could keep stable, i.e., watermark in LSFs are not easily affected by the watermarks in phase.

Experiment 2: Random perturbations were added to K LSFs Φ_K=$\{\phi_k, k = 1, ..., K\}$, which was a subset randomly selected from Φ. The perturbated $\hat{\Phi}_K$ was in Eq. (4),

$$\hat{\Phi}_K = \{\phi_k + w_l \times \Gamma_l \times \|\phi_k\|_\infty, k = 1, ..., K\} \tag{4}$$

where w_l was AWGN with standard deviation $\sigma = 1.0$, Γ_l set as $\{0.15, 0.25, 0.50, 0.75\}$ controlled the perturbation strength, and the infinite norm of $\|\phi_k\|$, i.e., $\|\phi_k\|_\infty$, adjusted the perturbation to match the range of ϕ_k. The perturbated LSFs and the other un-perturbated LSFs were converted to LP coefficients and re-synthesized with original residue to obtain the perturbated speech. The re-calculated residue from the perturbated speech showed how it was affected by the perturbated LSFs.

Table. 2 (Column 2-3) lists the perturbation in LSFs and perturbation proportion under different strengths, where three LSFs (K=3) were perturbated. These perturbations in LSFs also disturbed the LP coefficients. Nevertheless, the amplitude of residue (Column 6), the magnitude spectrum (Column 8), and the phase spectrum (Column 10) almost unchanged even under the strong LSFs perturbations. This phenomenon can be found from the right panel in Fig. 1, where Γ_l was set as 0.75. These results suggested that the perturbations in LSFs had slight influence on the phase of residue, indicating that the watermark in the phase of residue will not be affected by the watermark in LSFs.



(a) (b)

FIGURE 1. Left panel: perturbation in residue (sound source) and right eft panel: perturbation in LSFs (vocal tract).

It is worthwhile to note that the strength of perturbations added in phase and LSFs are far exceeds that of watermarks. Therefore, watermarks embedded in sound source/vocal tract will not be destoryed by the other.

3. **Implementation of watermarking for sound source.**

3.1. **Principles of QIM.** Figure 2 illustrates the QIM based watermark embedding process. Suppose s is the signal needed to be quantized. The s lies somewhere in one quantization step Δ. Two functions $Q_0(s,0)$ and $Q_1(s,1)$ in Eqs. (5) and (6) can uniquely map s to s_0 or s_1 for embedding "0" or "1", where [.] stands for the rounding function. After this, s_0 and s_1 can carry the watermark "0" and "1", respectively.

$$s_0 = Q_0(s,0) = \Delta\left[\frac{s}{\Delta} + \frac{1}{2}\right] \tag{5}$$

$$s_1 = Q_1(s,1) = \Delta\left[\frac{s}{\Delta}\right] + \frac{\Delta}{2} \tag{6}$$

In the extraction process, the received $\hat{s}$ is re-quantized with both functions in Eqs. (5) and (6). As outlined in Figs. 2(b) and 2(c), $\hat{s}_0$ calculated from Eq. (5) and $\hat{s}_1$ calculated from Eq. (6) are obtained. The embedded bit w can be determined by comparing the distance between $\hat{s}_0$ and $\hat{s}$, and the distance between $\hat{s}_1$ and $\hat{s}$:

$$d_0 = |\hat{s} - Q_0(\hat{s},0)| = \left|\hat{s} - \left(\Delta\left[\frac{\hat{s}}{\Delta} + \frac{1}{2}\right]\right)\right|, \tag{7}$$

$$d_1 = |\hat{s} - Q_1(\hat{s},1)| = \left|\hat{s} - \left(\Delta\left[\frac{\hat{s}}{\Delta}\right] + \frac{\Delta}{2}\right)\right|, \tag{8}$$

$$w = \begin{cases} 0, & d_0 < d_1 \\ 1, & \text{otherwise.} \end{cases} \tag{9}$$

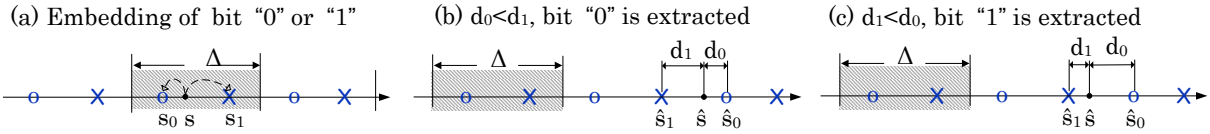


FIGURE 2. QIM based watermark embedding and extraction.

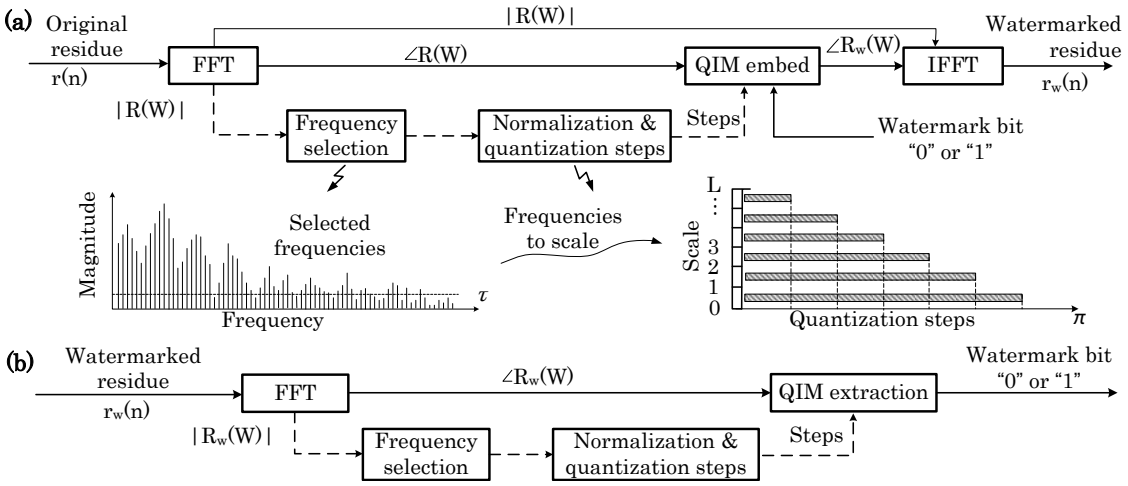


FIGURE 3. Block diagram of QIM based watermarking.

3.2. **QIM based watermark embedding and extraction.** According to Fig. 3(a), watermarks are embedded as below.

Step 1 The magnitude spectrum $|R(W)|$ and phase spectrum $\angle R(W)$ of residue $r(n)$ are first calculated using FFT (fast Fourier Transform). The phase of frequency components whose magnitudes are greater than threshold τ ($|R(W)|>\tau$) are selected to embed watermarks.

Step 2 The selected phase are normalized to the same scale which can be divided into L levels. Each level has its corresponding quantization steps (in π), i.e., higher levels have smaller quantization steps and lower levels have larger quantization steps, to balance inaudibility and robustness.

Step 3 $Q_0(s,0)$ in Eq. (5) and $Q_1(s,1)$ in Eq. (6) are used to embed "0" and "1" (here s is the phase $\angle R(W)$) using the quantization steps determined in **Step 2**.

Step 4 The quantized phase spectrum $\angle R_w(W)$ and the magnitude spectrum $|R(W)|$ are combined into Fourier spectrum and then transformed into time domain to obtain watermarked residue, $r_w(n)$.

According to Fig. 3(b), watermarks are extracted as below.

Step 1 Watermarked residue $r_w(n)$ is transformed into Fourier spectrum $R_w(W)$ with FFT. Magnitude spectrum $|R_w(W)|$ and phase spectrum $\angle R_w(W)$ are calculated.

Step 2 The same threshold, τ, is used to find the frequency components that have been embedded with watermarks. The quantization steps for watermark extraction are determined as those in **Step 2** of the embedding process.

Step 3 The phase of selected frequency components $\angle R_w(W)$ are calculated withEq. (7) to Eq. (9) using their quantization steps (here $\hat{s}$ is phase $\angle R_w(W)$). The final decision on watermark bit w of the residue is determined with a majority decision.

4. Implementation of watermarking for vocal tract.

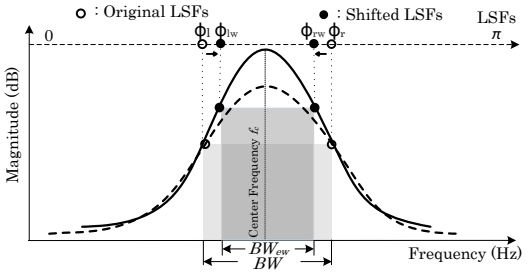


FIGURE 4. Formant enhancement by closing up two LSFs.

4.1. **Principles of formant enhancement.** In general, one formant is controlled by two LSFs, and the closer two LSFs are, the sharper the formant is. Therefore, formants can be enhanced by closing up two LSFs. In Fig. 4, the original formant (dotted curve) is produced by a pair of LSFs, ϕ_l and ϕ_r. Its bandwidth BW can be roughly calculated using Eq. (10), where F_s is the sampling frequency.

$$BW = |\phi_r - \phi_l|/2\pi \times F_s. \tag{10}$$

To enhance this formant, two LSFs, ϕ_l and ϕ_r, are symmetrically shifted more closely, i.e., ϕ_l to ϕ_{lw} and ϕ_r to ϕ_{rw}. This can be expressed with Eq. (11), where ϵ controls the degree of shift, and a larger ϵ indicates a more severe shift of LSFs as well as a much greater enhanced formant.

$$\phi_{lw} = \phi_l + \epsilon \text{ and } \phi_{rw} = \phi_r - \epsilon,\ 0 < \epsilon < |\phi_r - \phi_l|/2. \tag{11}$$

The enhanced formant (solid curve in Fig. 4) becomes much sharper. Its bandwidth BW_{ew} can be calculated as:

$$BW_{ew} = |\phi_{rw} - \phi_{lw}|/2\pi \times F_s. \tag{12}$$

4.2. **FE based watermark embedding and extraction.** According to Fig. 5(a), watermarks are embedded as below.

Step 1 The p LP coefficients a_i $(i = 1, 2, \cdots, p)$ which represent the formants of vocal tract are converted to p LSFs, ϕ_i $(i = 1, 2, \cdots, p)$.

Step 2 The sharpest formant produced by ϕ_a and ϕ_b (labeled "1^{st}") and the second sharpest formant produced by ϕ_c and ϕ_d (labeled "2^{nd}") are extracted for embedding. Their bandwidths are calculated as:

$$BW_{ab} = |\phi_a - \phi_b|/2\pi \times F_s, \tag{13}$$

$$BW_{cd} = |\phi_c - \phi_d|/2\pi \times F_s. \tag{14}$$

In addition, their relationship before embedding is:

$$BW_{cd} > BW_{ab}. \tag{15}$$

Step 3 To embed "0", the sharpest formant will be enhanced. Its bandwidth BW_{ab} will be reduced by Ω (Ω>1.0) times via closing ϕ_a and ϕ_b to ϕ_{aw} and ϕ_{bw}. The obtained bandwidth is BW_{abw}(BW_{abw}=BW_{ab}/Ω,). After this, the bandwidth relationship between two formants is:

$$BW_{cd} > BW_{abw} \times \Omega. \tag{16}$$

To embed"1", the second sharpest formant will be enhanced. Its bandwidth BW_{cd} will be reduced to be the same as BW_{ab}. To achieve this, ϕ_c and ϕ_d are shifted to ϕ_{cw} and ϕ_{dw}. After this, the bandwidth relationship between two formants is:

$$BW_{cdw} = BW_{ab}. \tag{17}$$

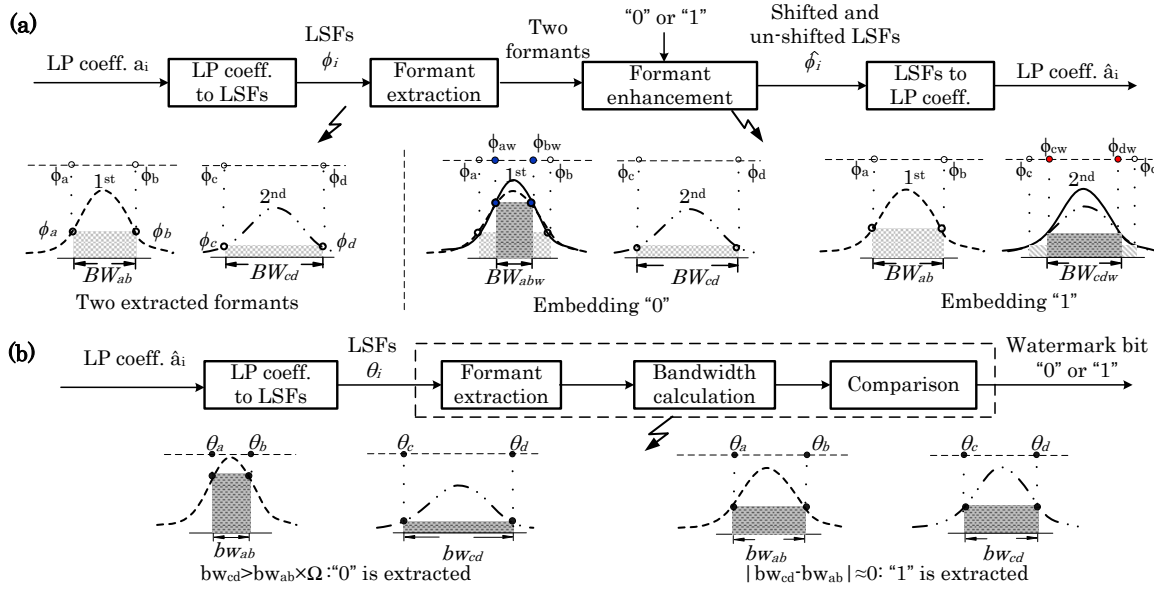


FIGURE 5. Block diagram of FE based watermarking.

Step 3 The shifted LSFs (ϕ_{aw} and ϕ_{bw} for embedding "0" or ϕ_{cw} and ϕ_{dw} for embedding "1") and the other un-shifted LSFs are converted back to LP coefficients $\hat{a}_i$ to represent the watermarked vocal tract information.

According to Fig. 5(b) , watermarks are extracted as below.

Step 1 The p LSFs, θ_i $(i = 1, 2, \cdots, p)$ are calculated from the watermarked vocal tract information and the sharpest and the second sharpest formant are extracted.

Step 2 Assume the sharpest formant is produced by θ_a and θ_b and the second sharpest formant is produced by θ_c and θ_d. Their bandwidths are:

$$bw_{ab} = |\theta_a - \theta_b|/2\pi \times F_s, \tag{18}$$

$$bw_{cd} = |\theta_c - \theta_d|/2\pi \times F_s. \tag{19}$$

Step 4 If "0" has been embedded, we have bw_{cd}>bw_{ab}×Ω, an equivalent expression is in Eq. (20). If "1" has been embedded, bw_{cd} should be equal to bw_{ab}, as expressed in Eq. (21). Thus, $bw_{ab} \times (\Omega - 1)/2$ in Eq. (22), is set as the threshold to discriminate the watermarks, where w stands for the extracted watermark bit.

$$bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega - 1) \tag{20}$$

$$bw_{cd} - bw_{ab} = 0, \tag{21}$$

$$w = \begin{cases} 0, & bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega - 1)/2 \\ 1, & \text{otherwise} \end{cases} \tag{22}$$

5. **Scheme for the proposed watermarking.** The whole watermarking scheme can be found in Fig. 6. Each one-bit watermark is duplicated for multi-frames' embedding to enhance the robustness. In extraction process, watermarks $\hat{s}_r(m)$ and $\hat{s}_v(m)$ are separately extracted from the sound source and the vocal tract. The watermarks $\hat{s}(m)$ for the proposed method are calculated with $\hat{s}_r(m)$ and $\hat{s}_v(m)$ using majority decision[2].

In addition, a random 0-1 sequence of length T is embedded into the first T samples of each frame for frame synchronization. The last several bits of the T samples (expressed

[2]This process can be implemented using a variety of strategies, such as majority voting, Bayesian methods, or even neural networks. Majority voting is by far the simplest of these methods, because it does not require prior knowledge, and yet has been found to be just as effective as more complicated schemes [29].

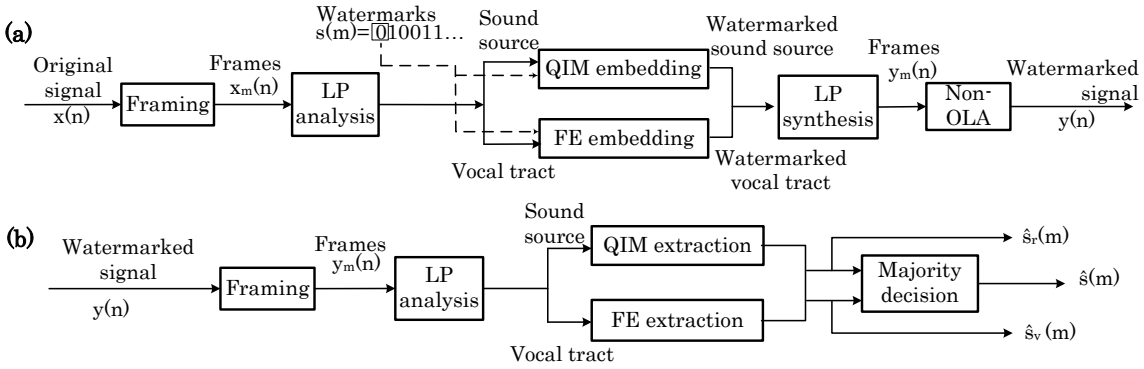


FIGURE 6. Scheme for the proposed watermarking: (a) embedding and (b) extraction.

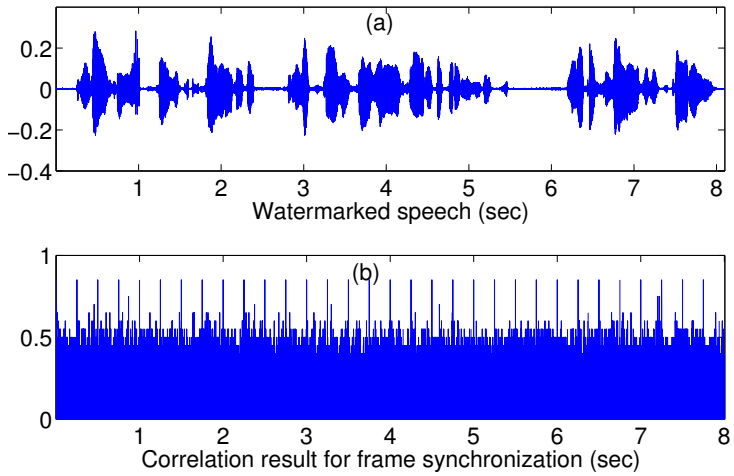


FIGURE 7. Frame synchronization results.

in16-bit binary codes) are sequentially replaced by this 0-1 sequence. In watermark extraction process, the beginning of each frame can be found by applying the correlation technique between the watermarked signal and the 0-1 random sequence. The frame synchronization result is shown in Fig. 7.

6. **Evaluations of proposed method.** Experiments were conducted to evaluate the inaudibility and robustness of the proposed method (this method satisfies blindness). All 12 speech stimuli (8.1 s, 20 kHz, and 16 bits) in the ATR database (B set) were used. The embedded watermark was a random binary code. The LP order was 10-*th*. The FFT size in QIM based watermarking was equal to frame size. The threshold τ for frequency selection was 10^{-6} based on experimental analysis. The selected frequency components were normalized to five levels ($L = 5$), i.e., $(\frac{\pi}{2}, \frac{\pi}{4}, \frac{\pi}{6}, \frac{\pi}{8}, \frac{\pi}{10})$. Ω for embedding "0" in FE based watermarking was fixed as 2.0 to ensure the inaudibility. Each one-bit watermark was duplicated for four frames in the QIM and FE embedding to increase the robustness of the whole scheme. The embedding bit rates were 1, 4, 8, 16, 32, 64, 128, and 256 bps.

Inaudibility was measured by log spectrum distortion (LSD) [30] and the perceptual evaluation of speech quality (PESQ) [31]. The LSD measured the spectral distance between the original speech and watermarked speech. The PESQ evaluated the speech quality with Objective Difference Grades (ODGs), where ODGs were graded from −0.5 (very annoying) to 4.5 (imperceptible), corresponding to Mean Opinion Score (MOS) of 1.0 to 5.0. The criterion for LSD and PESQ are LSD>1.0 dB and PESQ>3.0 ODG. Robustness was measured by Bit Detection Rate (BDR) and BDR of 90% was the criterion.

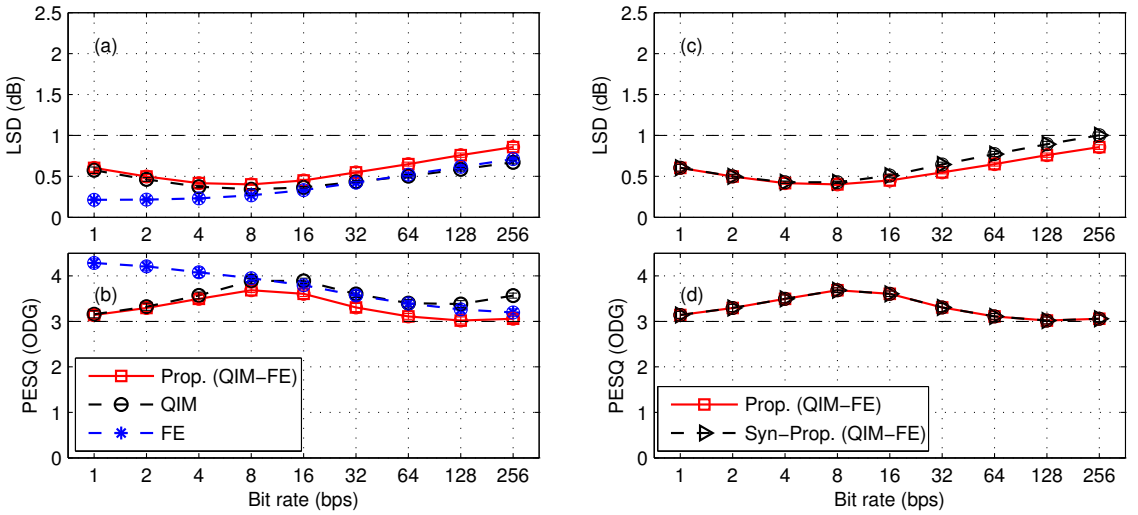


FIGURE 8. Evaluations of Inaudibility.

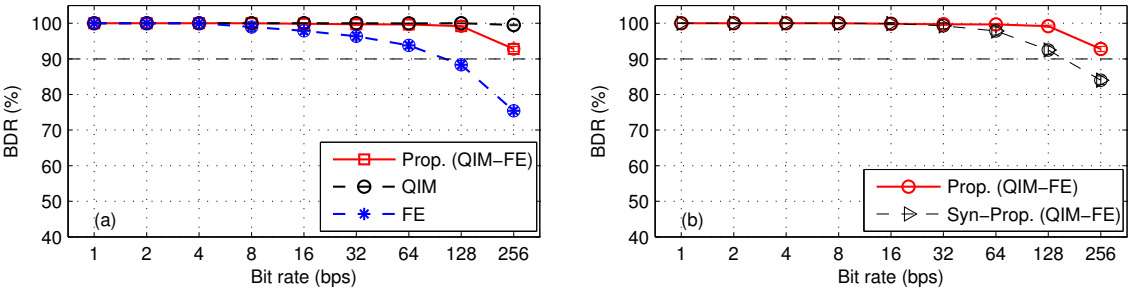


FIGURE 9. BDR results for normal extraction.

6.1. **Evaluations for inaudibility and normal extraction.**

6.1.1. *Inaudibility.* The proposed method embeds information into three channels: sound source, vocal tract, and the beginning of each frame (for synchronization). Therefore, inaudibility was checked when informaiton was embedded in (*i*) the sound source ("QIM"), (*ii*) the vocal tract ("FE"), (*iii*) both the sound source and vocal tract ("Prop. (QIM-FE)"), and (*iv*) sound source, vocal tract, and each frame ("Syn-Prop. (QIM-FE)").

The results concerning (*i*) to (*iii*) are plotted in Figs. 8(a) and 8(b). The inaudibility of FE based watermarking was better than QIM based watermarking. Nevertheless, both the two single watermarking could satisfy the criteria for LSD and PESQ. The inaudibility of the proposed method was a little worse compared with single watermarking but could satisfy inaudibility. We especially found that the speech quality was better around 8 and 16 bps. According to speech synthesis [32], the frame lengths at these bit rates were more suitable for speech analysis/synthesis. The comparative results between (*iii*) and (*iv*) are plotted in Figs. 8(c) and 8(d). The two curves for LSD were close to each other and the two curves for PESQ overlapped. These results suggested the distortion introduced by synchronization was almost negligible.

6.1.2. *Normal extraction.* The proposed method embeds information to the sound source, vocal tract, and the beginning of each frame. It is necessary to check whether one channel will interfere with the other two channels and obstruct the information extraction.

Figure 9(a) plots the BDR results in the condition of no synchronization. The curves labelled "QIM" and "FE" are the BDR results calculated from the sound source and vocal tract, respectively. The curve labelled "Prop. (QIM-FE) " is the BDR of the proposed method calculated based on the "QIM" and "FE" results. These results indicated that the

watermarks could be successfully extracted from the sound source and vocal tract, i.e., one method did not severely affect the other. Furthermore, the BDR results for the proposed method were almost as good as the better ones for "QIM" and "FE". This implied that the successful extraction with either method would provide the proposed method with successful extraction. From Fig. 9(b), one can find that the watermarks could be correctly extracted when the bit rate was lower than 128 bps, i.e., the synchronization information did not greatly affect the extraction performance of the proposed method.

According to these results, the embedded information in one channel did not affect the others, i.e., three channels were nearly independent of one another. This attribute guaranteed the feasibility of the proposed method.

6.2. **Evaluations for robustness.**

6.2.1. *Robustness against speech codecs.* Several speech codecs were applied to the watermarked speech, i.e., G.711, G.723.1, G.726, and G.729. Figure 10 plots the results. Both QIM and FE methods were robust against G.711. Therefore, the proposed method was robust against G.711. The QIM method failed to extract the watermarks after G.723.1, G.726 and G.729. These results were consistent with our prediction that as phase information is not employed in speech codecs, watermarks in phase could not survive from speech codecs. In contrast, the FE method provided higher BDR results. The BDR results for the proposed method were almost as good as the better ones for "QIM" and "FE". This verified that the disadvantage of one watermarking method can be concealed by incorporating it with another method. Therefore, the robustness of the proposed method could be increased in comparison with singe watermarking method.

6.2.2. *Robustness against speech processing.* We evaluated the robustness of the proposed method against several speech processing, including re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits, speech analysis/synthesis using gammatone filterbank (GTFB) and short-time Fourier transform (STFT), bandpass filtering (BPF) with passband [0.1, 6] kHz and stopband attenuation of -12 dB/octave, and signal scaling by two times. It can be found from Fig. 11 that (1) the FE method was robust against most processing except for re-quantization with 8 bits and BPF, (2) the QIM method was robust against some of these processing, and (3) the performances of proposed method was obviously better than QIM and FE methods since watermarks could be extracted even if one method (FE or QIM) failed.

6.2.3. *Robustness against common attacks.* We evaluated the proposed method against several attacks. These included signal flanger, signal flipping, signal jitter, signal sample repetition, Gaussian noise addition (signal-to-noise ratio (SNR) of 36 dB), and echo addition (ECHO,100 ms echo of -6 dB). Signal flanger was an operation to create a signal by mixing a slightly delayed copy of itself. The delay time was decided by frame size. Around one third of frame size was delayed for each frame in our evaluation. The values of two randomly chosen samples in each frame were exchanged in signal flipping. According to the embedding rules, the exchanged samples in one second increased from 8 to 2048 from 1 to 256 bps. The randomly chosen samples of each frame was set to be 0 in signal jitter. One randomly chosen sample in each frame was repeated in sample repetition, and the repeated samples in one second were increased when the bit rate was increased. The duration of attacked signals was also increased in this case. In noise addition, Gaussian noise with an overall average signal-to-noise ratio (SNR) of 36 dB was added to watermarked speech. A single 100 ms echo of -6 dB was added to watermarked speech in echo attacks. The first four attacks were referred to by Steinebach et al. [33], and the last two attacks were recommended by the Information Hiding and its Criteria (IHC)

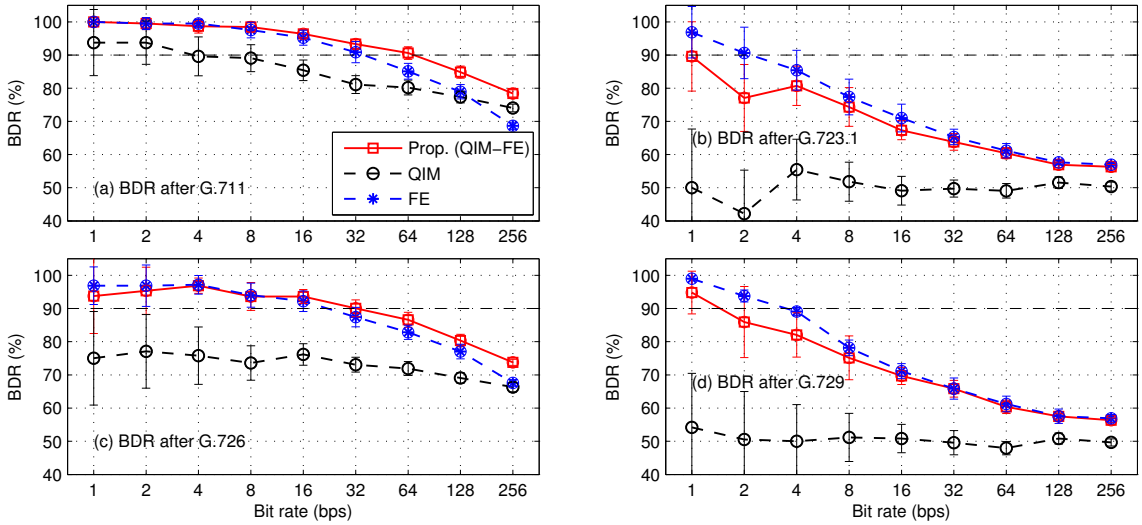


FIGURE 10. Robustness of proposed watermarking against different speech codecs.

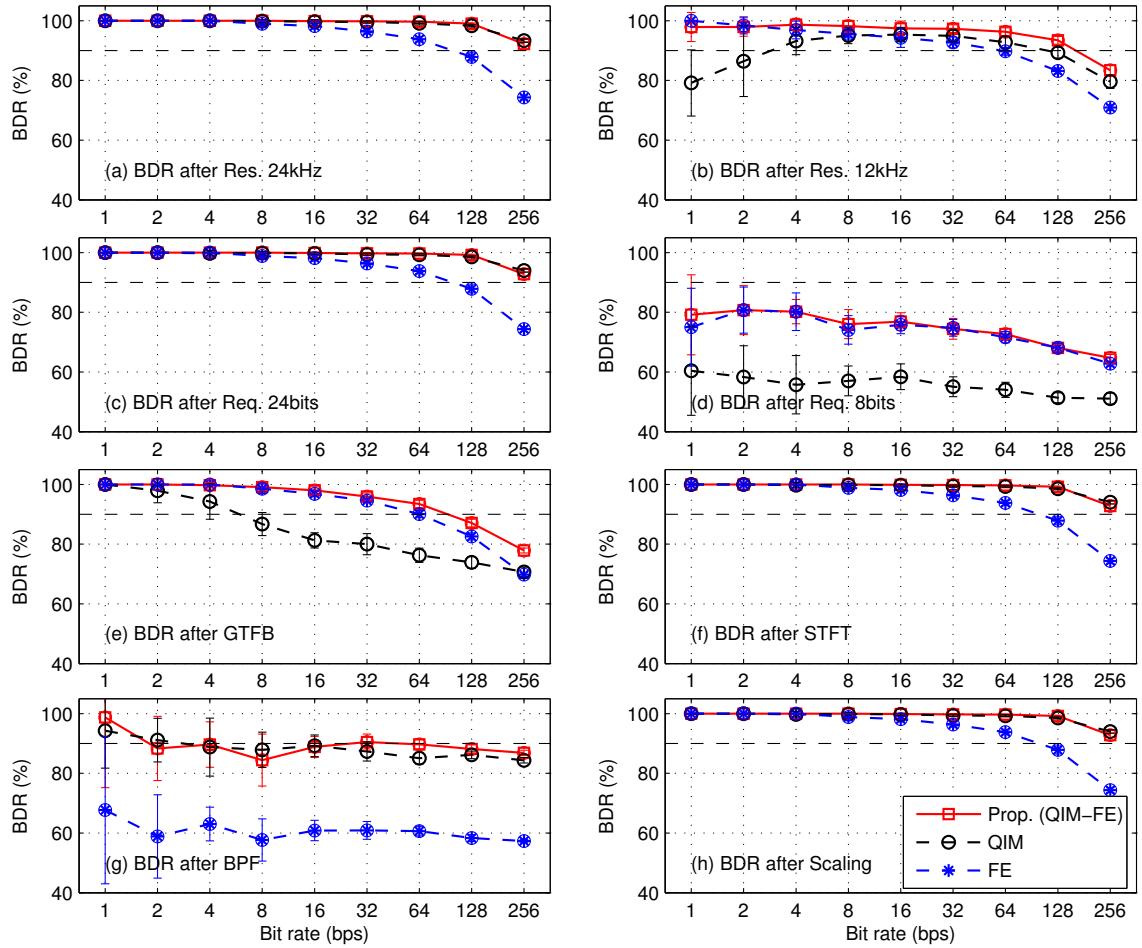


FIGURE 11. Robustness of proposed watermarking against different speech processing.

committee [34]. According to Fig. 12, the QIM method was only robust against signal flipping, signal jitter, and Gaussian noise addition and the FE method was robust against all attacks. Thus, the proposed method was robust against all these attacks. Therefore, the proposed method demonstrated stronger robustness than single watermarking and this was its obvious advantage.

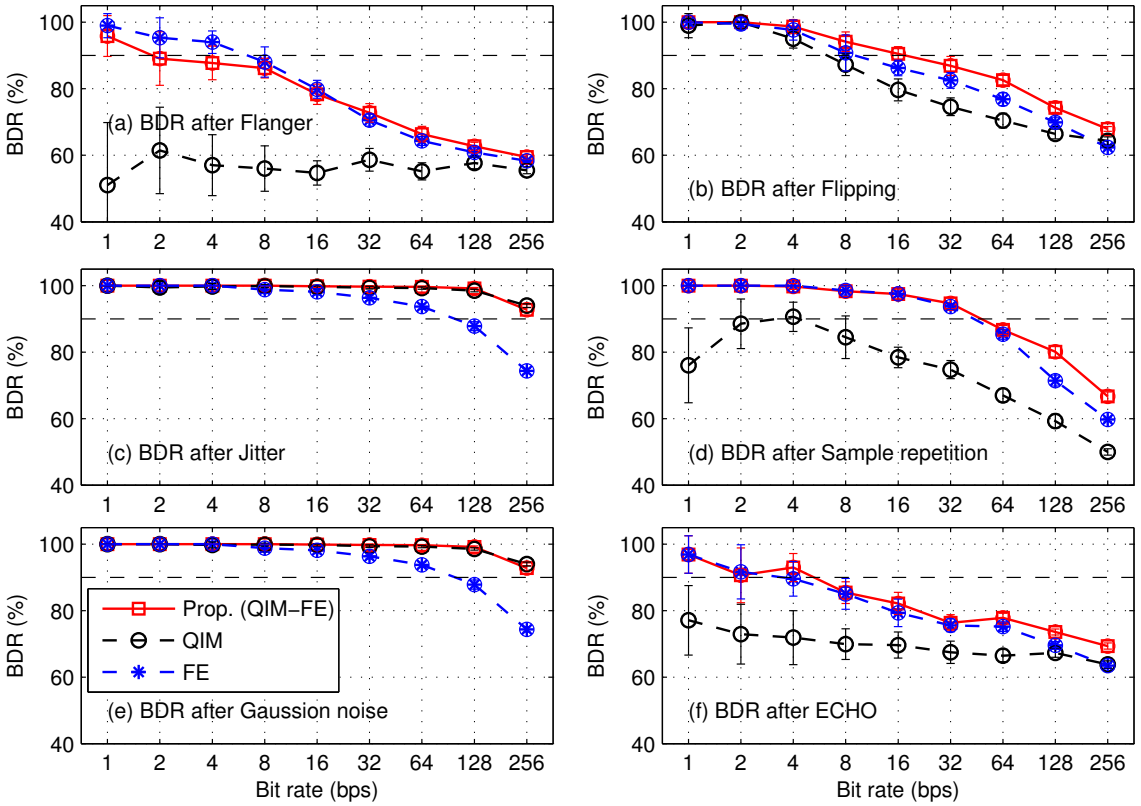


FIGURE 12. Robustness of proposed watermarking against common attacks.

6.3. **Discussion.** This section evaluates the proposed method with respect to inaudibility and robustness. The proposed method can satisfy inaudibility. The sound distortion introduced by synchronization is trivial and the synchronization does not obstruct watermark extraction. In robustness evaluations, the QIM method is not sufficiently robust. The FE method shows good robustness. This phenomenon can be attributed to its embedding and extraction mechanism that identifying the bandwidth relationship for watermark extraction can tolerate small modifications caused by speech codecs, processing, and attacks. We also find that the QIM and FE method can complement each other. Therefore, the proposed method demonstrates excellent robustness.

7. **Comparative evaluations.** Many watermarking methods have been proposed in recent years. We chose three typical methods to make comparative evaluations. These were LSB [1], DSS [2], and Cochlear delay (CD) methods [7]. The main reason these three methods were chosen is because they separately exhibited excellent performance in inaudibility, robustness, and both inaudibility and robustness.

A quick review of these methods is provided in what follows: LSB replaces the least significant bits with watermarks at the quantization level so that the replacement does not cause severe distortion, DSS spreads watermarks over many (possibly all) frequency bands so that the watermarks cannot easily be destroyed, and CD embeds watermarks by enhancing the phase information of speech signals with respect to two kinds of cochlear delays (one is for bit "0" and the other is for bit "1").

7.1. **Comparison for inaudibility and normal detection.** The inaudibility results are plotted in Figs. 13(a) and 13(b). LSB had the best performance of all the four methods. The proposed method could satisfy the criteria for both LSD and PESQ. CD could satisfy inaudibility when the bit rate was no more than 16 bps. DSS could not

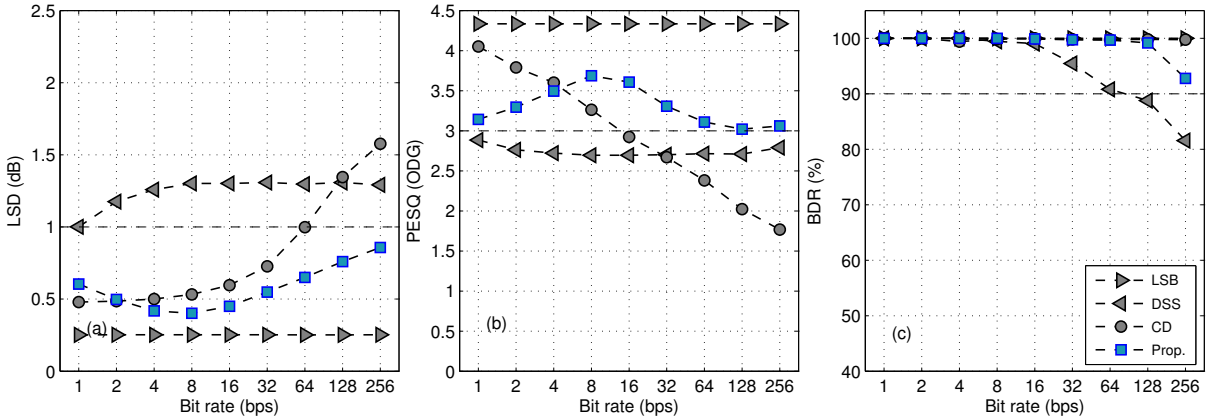


FIGURE 13. Comparison of inaudibility and normal detection performance for the proposed method, LSB, DSS, and CD.

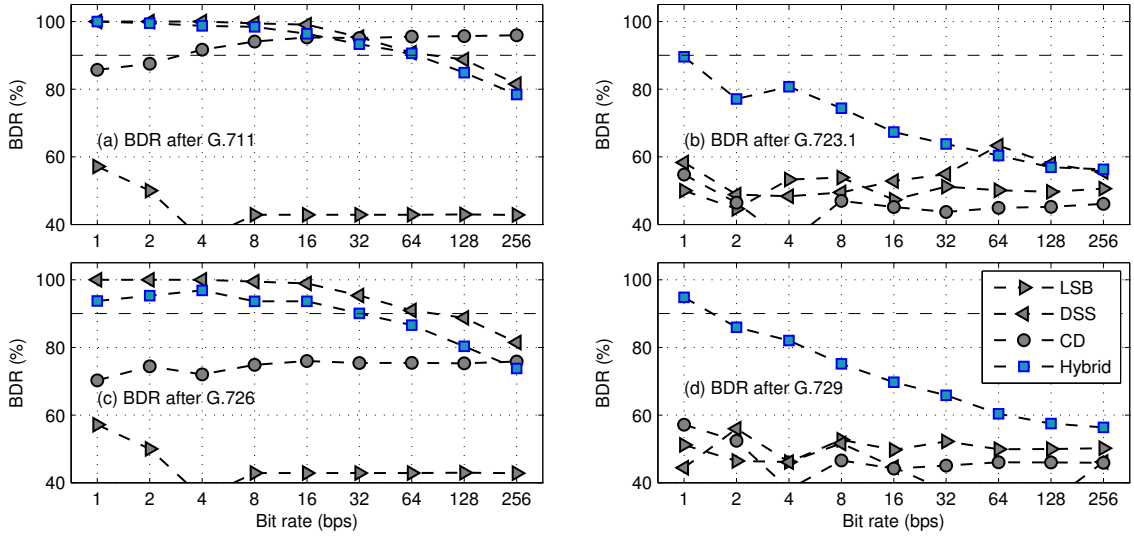


FIGURE 14. Comparison of robustness against different speech codecs.

satisfy the criteria for either LSD or PESQ. In summary, the LSB method had the best performance in inaudibility and the proposed method was better than DSS and CD. The normal detection results are plotted in Fig. 13(c). LSB and CD could correctly detect the watermarks for all the embedding bit rates and the proposed method could correctly detect watermarks when the bit rate was less than 256 bps, while the BDR of DSS started dropping from 64 bps.

7.2. **Comparison for robustness.** The results against speech codecs are plotted in Fig. 14. We found that LSB was not robust against any speech codecs, CD was only robust against G.711, DSS was robust against G.711 and G.726, and the proposed method was basically robust against all kinds of speech codecs although its performance against G.723.1 and and G.729 still needs to be improved. These results implied that the proposed method had better robustness than the other methods.

The results against speech processing are plotted in Fig. 15. DSS obviously performed the best. LSB was robust against re-sampling at 24kHz, re-quantization with 24 bits, and STFT. CD provided provided satisfactory BDR results except for re-quantization

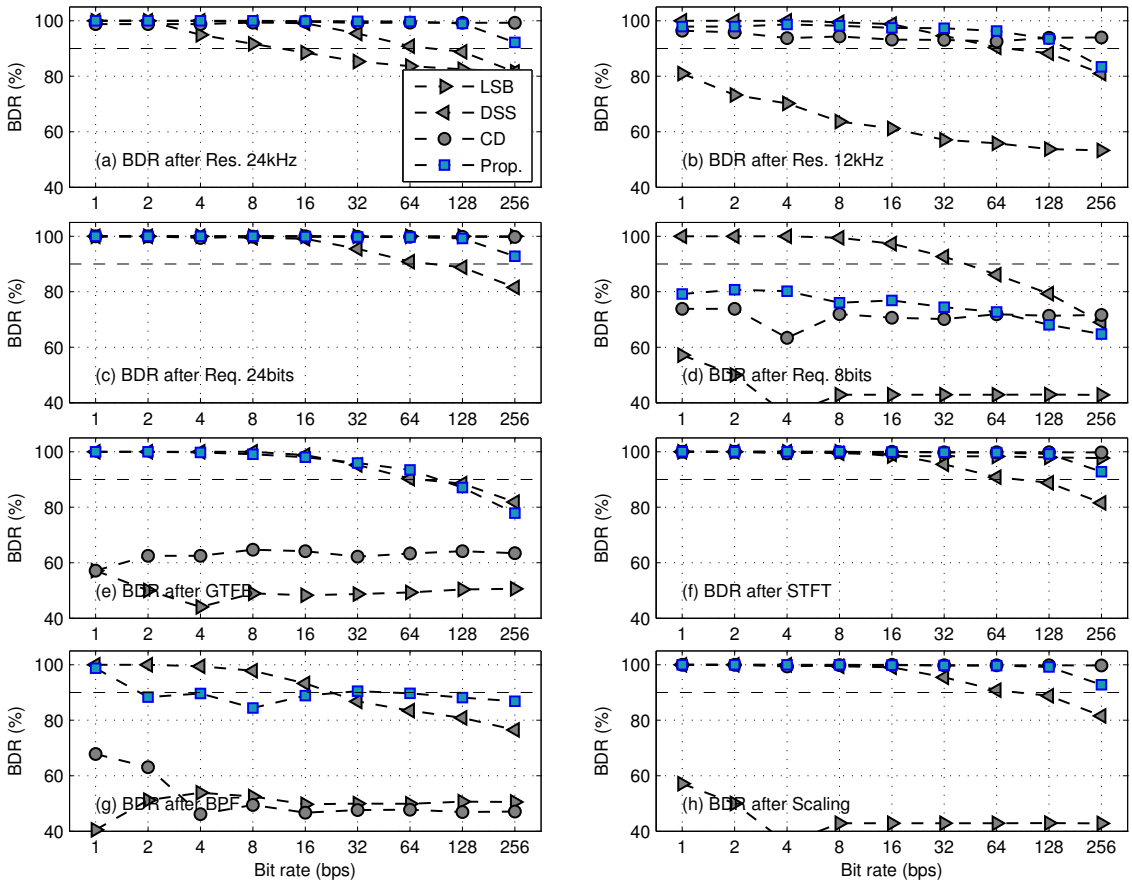


FIGURE 15. Comparison of robustness against different speech processing.

with 8 bits, GTFB, and BPF. The proposed method was basically robust except for requantization with 8 bits. We could conclude from these results that DSS and the proposed method were more robust against these processing than LSB and CD.

The results against common attacks are plotted in Fig. 16. DSS failed to detect watermarks after sample repetition. LSB and CD were only robust against some of the attacks. Overall, the proposed method performed better than the others.

7.3. **Discussion.** This section compared the proposed method with other typical methods. LSB was inaudible but not robust, DSS was not inaudible and not completely robust, CD could conditionally satisfy inaudiblity and robustness. In comparison, the proposed method was better than these methods, and it could basically satisfy both inaudibility and robustness. To explore the reasons behind this, we analyzed all the methods we evaluated. As the LSB method embedded watermarks in the least significant bits, the distortion to the original signal was negligible and this thus enabled LSB to be perfectly inaudible. However, watermarks in the least significant bits could easily be reset by operations such as amplitude modifications and lossy processing, which deteriorated the robustness of LSB method. DSS was relatively more robust than LSB since watermarks were spread over a wide frequency range, and the watermarks could only be eliminated when all possible frequencies were destroyed with considerable strength. Therefore, DSS exhibited excellent robustness for most processing. However, watermarks over a wide range of frequencies made them perceptually significant. Watermarks in CD were embedded as phase modifications by modelling cochlear delay. Watermarks detection strongly depended on the cues in low-frequency phase, according to the characteristics of cochlear delay. Correspondingly, once phase information in the low frequency is destroyed or erased

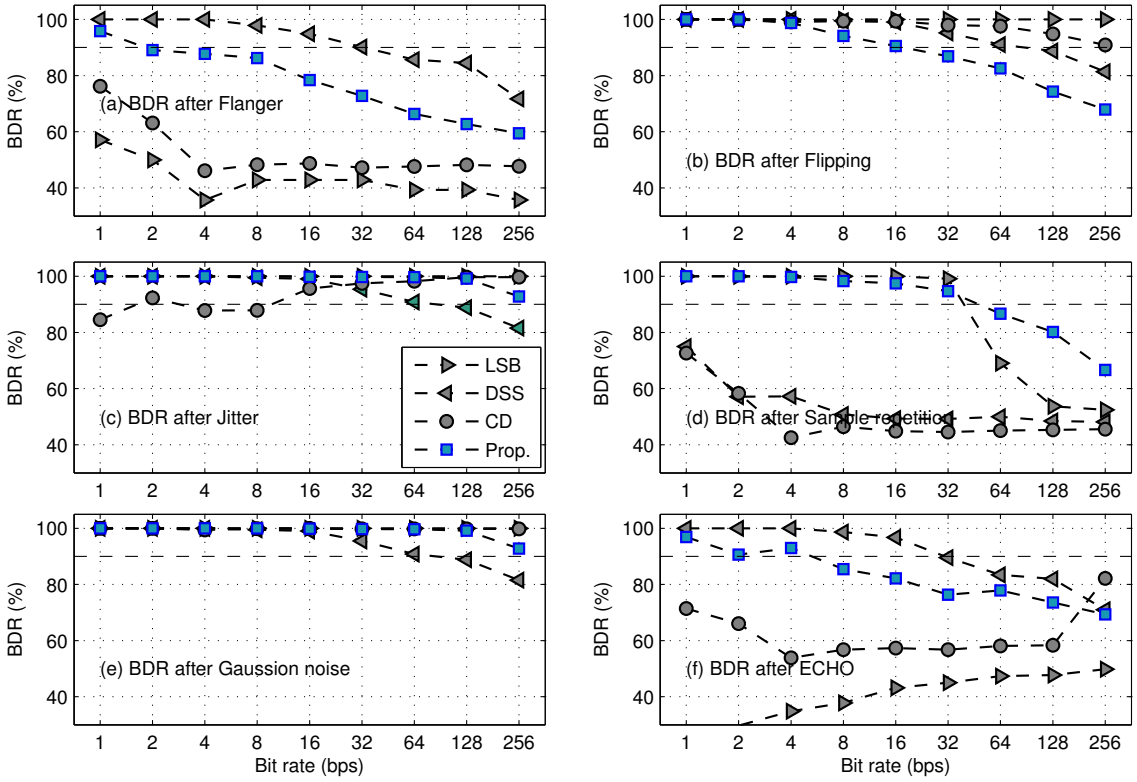


FIGURE 16. Comparison of robustness against common attacks.

by processing, watermarks cannot be detected. The proposed method could take advantage of each single method and gain overall superiority. Therefore, the proposed method exhibited better performance than the others.

8. **Conclusions.** This paper proposed a speech watermarking method for speech signals based on the source-filter model of speech production. LP analysis was used to separate the speech signal into two components, i.e., sound source and vocal tract information. These two components were separately embedded with watermarks using QIM based and FE based watermarking. The independence between the sound source and vocal tract ensured the feasibility of the proposed watermarking. We investigated the inaudibility and robustness of the proposed method. The results revealed that (1) the proposed method could satisfy inaudibility and (2) the combination of FE and QIM methods enabled the proposed method to benefit from both methods to attain stronger robustness. Finally, we compared the proposed method with other typical methods and the results revealed that the proposed method outperformed the other methods. All these results verified the effectiveness of the proposed method.

Acknowledgment. This work was supported by Natural Science Foundation of Tianjin (No. 17JCQNJC00100), the Science&Technology Development Fund of Tianjin Education Commission for Higher Education (No. 2017KJ089 and No. 2018KJ218), a Grant-in-Aid for Scientific Research (B) (No. 17H01761), and I-O DATA foundation. It was also supported by National Natural Science Foundation of China (No. 61373104), the Program for Innovative Research Team in University of Tianjin (No. TD13-5032), and the Program for Science and Technology of Tianjin (No. 19PTZWHZ00020).

REFERENCES

[1] P. Bassia and I. P. Pitas, Robust audio watermarking in the time domain, *Proc. EUSIPCO*, pp. 25–28, 1998.

[2] L. Boney, H. H. Tewfik, and K. H. Hamdy, Digital watermarks for audio signals, *Proc. ICMCS*, pp. 473–480, 1996.

[3] R. Kazemi, F. Pérez-González, M. Ali Akhaee, and F. Behnia, Data hiding robust to mobile communication vocoders, *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2345–2357, 2016.

[4] S. Sarreshtedari, M. A. Akhaee, and Aliazam Abbasfar, A watermarking method for digital speech self-recovery, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1917–1925, 2015.

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

[6] C. Wu and C. Jay Kuo, Fragile speech watermarking based on exponential scale quantization for tamper detection, *Proc. ICASSP*, vol. IV, pp. 3305–3308, 2002.

[7] M. Unoki and R. Miyauchi, Reversible watermarking for digital audio based on cochlear delay characteristics, *Proc. IIHMSP*, pp. 314–317, 2011.

[8] V. Phuoc-Hung, N. Thai-Son, H. Van-Thanh, D. Thanh-Nghi, A robust hybrid watermarking scheme based on DCT and SVD for copyright protection of stereo images, *Proc. NAFOSTED Conference on Information and Computer Science*, pp. 331–335, 2017.

[9] N. Fatima, D. J. Tuptewar, Comparison of hybrid watermarking technique on different color spaces,' *Proc. Advances in Signal Processing (CASP)*, pp.13–17, 2016.

[10] A. Dhawan and S. K. Mitra, Hybrid audio watermarking with spread spectrum and singular value decomposition, *Proc. India Conference*, pp. 11–16, 2008.

[11] B. Y. Lei, K. T. Lo, and H. j. Lei, Hybrid SVD-based audio watermarking scheme, *Proc. Communications, Circuits and Systems (ICCCAS)*, pp. 428–432, 2010.

[12] S. Wang and M. Unoki, Hybrid speech watermarking based on formant enhancement and cochlear delay, *Proc. IIHMSP*, pp. 272–275, 2014.

[13] V. Nikunj Tahilramani, B. Ninad, A hybrid scheme of information hiding incorporating steganography as well as watermarking in the speech signal using Quantization index modulation (QIM), *Proc. CSCITA*, pp. 220–234, 2017.

[14] A. Doraisamy and R. Venkatachalam, Secure robust and hybrid watermarking for speech signal using discrete wavelet transform, discrete cosine Transform, and singular value decomposition, *Journal of Engineering Science and Technology*, vol. 12, no. 6, pp. 1627–1639, 2017.

[15] J. D. Markel and A. H. Gray Jr., Linear prediction of speech, *Springer-Verlag*, 1976.

[16] J. Makhoul, Linear prediction: A tutorial review, *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[17] A. Takahashi, R. Nishimura, and Y. Suzuki, Multiple watermarks for stereo audio signals using phase-modulation techniques, *IEEE Trans. Signal Proc.*, vol. 53, no. 2, pp. 806–815, 2005.

[18] X. Dong, M. F. Bocko, and Z. Ignjatovic, Data hiding via phase manipulation of audio signals, *Proc. ICASSP*, vol. 5, pp. 377-380, 2004.

[19] N. M. Ngo, B. M. Kurkoski, and M. Unoki,, Robust and reliable audio watermarking based on dynamic phase coding and error control coding, *Proc. EUSIPCO*, pp. 2316–2320, 2015.

[20] B. Chen and G. W. Wornel, Quantization index modulation: a class of provably good methods for digital watermarking and information embedding, *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[21] Y. Ueda, S. Hario, and T. Sakata, Formant based speech enhancement for listening speech sound in noisy place, *Proc. ICSV*, pp. 515–522, 2008.

[22] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, Comparison of formant enhancement methods for HMM based speech synthesis, *Proc. ISCA Speech Synthesis Workshop*, 2010.

[23] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, The HMM-based speech synthesis system (HTS) version 2.0, *Proc. ISCA Workshop on Speech Synthesis*, pp. 294–299, 2007.

[24] Recommendation ITU-T P.800, Methods for subjective determination of transmission quality, *International Telecommunication Union*, 1996.

[25] S. Wang and M. Unoki, Watermarking of speech signals based on formant enhancement, *Proc. EUSIPCO*, pp. 1257–1261, 2014.

[26] S. Wang and M. Unoki, Speech Watermarking Method based on Formant Tuning, *IEICE Trans. Inf. & Syst.*, vol. E98–D, no. 1, pp. 29–37, 2015.
[27] F. Itakura, Line spectrum representation of linear predictor coefficients of speech signals, *Journal of the Acoustical Society of America*, vol. 57, no. 537(A), pp. 35–35, 1975.
[28] K. Takeda et al, Speech database user's manual, *ATR Technical Report* TR-I-0028, 2010.
[29] L. Lam and C. Y. Suen, Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance, *IEEE Trans. Systems Man and Cybernetics-Part A: Systems and Humans*, vol. 27, no. 5, pp. 553–568, 1997.
[30] A. Gray, Jr., and J. Markel, Distance measures for speech processing, *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. 24, no. 5, pp. 380–391, 1976.
[31] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 16, no. 1, pp. 229–238, 2008.
[32] S. Ragot, ITU-T G.729.1: An 8–32 kb/s scalable coder interoperable with G.729 for wideband telephony and voice over IP, *Proc. ICASSP*, vol. 4, pp. 529–532, 2007.
[33] M. Steinebach, F. A. P. Petitcolas , F. Raynal, J. Dittmann , C. Fontaine, S. Seibel, N. Fates, and L. C. Ferri, Stirmark benchmark: Audio watermarking attacks, *Proc. Information Technology: Coding and Computing*, pp. 49–54, 2001.
[34] *http:// www.ieice.org/iss/emm/ihc/en/index.php*