Journal of Information Hiding and Multimedia Signal Processing ©2015 ISSN 2073-4212
Ubiquitous International Volume **6**, Number **6**, November **2015**

Tampering Detection Scheme for Speech Signals using Formant Enhancement Based Watermarking

Shengbei Wang, Ryota Miyauchi, and Masashi Unoki

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan
{wangshengbei, ryota, unoki}@jaist.ac.jp

Nam Soo Kim

School of Electrical Engineering and INMC
Seoul National University, Republic of Korea
nkim@snu.ac.kr

Received November, 2014; revised May, 2015

Abstract. *Unauthorized tampering in speech signals has brought serious problems when verifying their originality and integrity. Digital watermarking can effectively check if the original speech signals have been tampered by embedding digital data into them. To be effective, watermarking method should be implemented according to four requirements of (1) inaudibility to human auditory system, (2) blindness to extract watermarks without referring to the host signal, (3) robustness against speech processing, and (4) fragility against tampering. This paper proposes a tampering detection scheme for speech signals based on formant enhancement-based watermarking. Watermarks are embedded as slight enhancement of formant by symmetrically controlling a pair of linear spectral frequencies (LSFs) of corresponding formant. Such embedding concept not only enables the proposed scheme to be inaudible but also provides the possibilities of robustness against speech processing and fragility against tampering. The proposed scheme was evaluated with respect to inaudibility, robustness, and fragility compared with other two typical methods. The evaluation results showed that the proposed scheme could provide satisfactory performance for all the requirements, and had the ability to detect tampering in speech signals.*
Keywords: Tampering detection, Speech watermarking, Formant enhancement, Inaudibility, Robustness, Fragility

1. **Introduction.** Rapid development in digital technologies has greatly facilitated speech signals to be reduplicated and edited at high fidelity. Although many applications benefit from these advances, new social issues related to malicious attacks and unauthorized tampering to speech have accordingly arisen. For example, by using free editing software, ordinary people are allowed to alter speech without leaving perceptual clues. Some specialized speech analysis/ synthesis tools such as STRAIGHT [1], voice conversion [2], and speech morphing [3], are professional to produce high naturalness and intelligibility of tampered speech, although important information has been changed. As these progresses enable speech to be tampered in a more realistic and credible way, it is becoming difficult to identify the tampering and confirm the originality and integrity of speech.

Since tampering may be motived by malicious intend to mislead the listener, problem becomes urgent when the possibly tampered speech is used as digital evidence in digital

forensics [4, 5]. To confirm the used speech is best suited to the unique acquisition environment and the truth, investigation about whether the speech has been tampered since its creation should be carried out. Such investigation, therefore, is aiming to validate the originality and integrity of speech [6, 7].

There are generally two categories to authenticate speech, i.e., active method and passive method. The cryptography [8], as an active method, can prevent speech from tampering by setting up a secure delivery of speech from the sender side to the receiver side. Legal recipient at the receiver side will be provided with a key to decrypt the speech. Cryptography provides a useful way for transmission, however, it does not examine the original speech data that being protected [9]. Therefore, there exists a defect that once the decrypted speech is edited or distributed, or if the decryption key is captured by illegal user, cryptography cannot provide any information to track the speech for its originality and integrity. As a complement technique to cryptography, speech watermarking has been proposed as a passive method [10, 11, 12, 13] by means of which speech can be automatically authenticated. Compared with cryptography, speech watermarking [14] does not prevent a user from listening to and using the speech. Moreover, it does not suffer from the drawbacks in cryptography, since information (referred as watermarks) is directly embedded within the speech, and the embedded information can permanently exist and is difficult to be removed [15]. Therefore, watermarking enables speech to be authenticated in a more suitable and durable way, and tampering can be detected with the embedded watermarks [16].

Speech watermarking benefits that although tampering may leave no perceptual clue, it has possibly changed the underlying characteristics of speech which would destroy the embedded watermarks. Thus, tampering can be well indicated with destroyed watermarks. To effectively detect tampering, speech watermarking should generally satisfy four requirements [9, 11, 14, 17, 18]: (1) inaudibility, (2) blindness, (3) robustness, and (4) fragility. Inaudibility indicates that the embedding of watermarks should not degrade the sound quality of host speech. Blindness indicates watermarks which will be later used for detecting tampering, should be extracted without referring the host signal. Robustness means allowable speech processing (e.g., re-sampling and re-quantization) to the watermarked signal should not destroy the embedded watermarks and thereby nullify them for tampering detection. Fragility means watermarks are sensitive to tampering and easy to be destroyed once tampering has been made to the watermarked signal. Although robustness and fragility conflict with each other, their collaboration guarantees that watermarks can only be destroyed by tampering, which enables tampering to be reliably identified. To take the advantages of both robustness and fragility, watermarking methods should be explored to tolerate speech processing, and meanwhile, detect tampering [19, 20].

In the past, watermarking-based tampering detection schemes for image [21, 22, 23, 24, 25] and video [20, 26] have been intensively studied. Tampering detection for audio/speech, however, is still unresolved since human auditory system (HAS) is considered more sensitive than human vision system (HVS) [27]. Basically, tampering detection schemes for speech come down to two main categories: i) schemes just verify the originality of speech without localizing the tampering and ii) schemes that can localize the tampering in time domain. The second category is more preferred in practical applications. In the literature, limited tampering detection schemes concerning the above two categories have been found. For example, in [28], Park *et al.* investigated a scheme with watermarking and pattern recovery to detect tampering. A watermark pattern was attached to speech so that when tampering occurred, destroyed watermark pattern could be used to identify the tampering. In this scheme, tampering was only detected after MP3 (at 16 kbps) and code-excited linear prediction (CELP) (11.5 kbps) compression, and only

three tampering, i.e., substitution, insertion, and removal were considered. As the HAS is sensitive, some schemes exploited the properties of HAS and embedded watermarks to the perceptually inaudible components for inaudibility [29]. Celik *et al.* proposed a watermarking method by introducing small changes to pitch (fundamental frequency) [30] with quantization index modulation (QIM) [31]. Insensitivity of human perception to the natural variability of pitch enabled the method to be inaudible. The stability of pitch under low data rate compression (e.g. Global System for Mobile communications coder (GSM) 6.10 and Adaptive Multi-Rate coder (AMR)) also made the method effective for semi-fragile authentication. Nonetheless, the method had not been designed to be robust against attacks that aimed to obstruct detection of watermarks. For example, a systematic modification of pitch such as re-embedding, would typically disable the watermarks. Wu *et al.* implemented a fragile speech watermarking for tampering detection based on odd/even modulation with exponential scale quantization [32]. Watermarks were embedded as pseudo-random noise in the discrete Fourier transform (DFT) domain by roughly approximating the MPEG psychoacoustic model. The time resolution for tampering could be set at 0.5 second or even shorter. However, its compatibility with CELP speech codecs still needed to be improved. In [33, 34, 35], Unoki and Hamada introduced a watermarking method by employing the characteristics of cochlear delay (CD). Watermarks were embedded by enhancing the phase of the host speech with respect to two kinds of group delays. Based on this concept, a tampering detection scheme was presented in [36]. The performance of this scheme was evaluated at variant embedding bit rate (bps). It was found that the scheme could successfully detect tampering (e.g., reverberation), and the detection precision (in second) could be increased with higher bps. Nonetheless, this scheme did not show strong robustness when subjected to speech codecs of G.726 and G.729 [37].

Since robustness and fragility are conflicting, and fragility may occasionally make the watermarking methods [32, 36] not robust, performance of these methods will be much degraded when used for tampering detection. Our work aims to effectively detect tampering with speech watermarking that can satisfy all the requirements (inaudibility, blindness, robustness, and fragility). To achieve this, a formant enhancement-based watermarking is designed to realize inaudibility by taking advantage that humans are not sensitive to slight enhancement of formant. Watermarks are embedded into the host speech as slight enhancement of formant by symmetrically controlling a pair of linear spectral frequencies (LSFs) [38]. The properties of LSFs enable the method to be robust against allowable meaning processing. This watermarking method is then employed in the tampering detection scheme to detect tampering.

The rest of this paper is organized as follows. Section 2 introduces the overall tampering detection scheme. Section 3 details the formant enhancement-based watermarking employed in the proposed tampering detection scheme. Section 4 talks about how to apply this watermarking method to tampering detection scheme. Section 5 evaluates the proposed tampering detection scheme with respect to three requirements of inaudibility, robustness, and fragility by comparing with other typical methods. A short discussion is also given out in this section. Section 6 gives a summary of this work.

2. Overall tampering detection scheme for speech.

Speech can be used for a variety ways, and criminal investigation is a major one. As a kind of digital evidence, speech can record, e.g., what happened in a certain place and time, and the information provided by either the victims or the suspects. However, in most cases, speech is not immediately used after being recorded. They have to be passed from people to people, and delivered

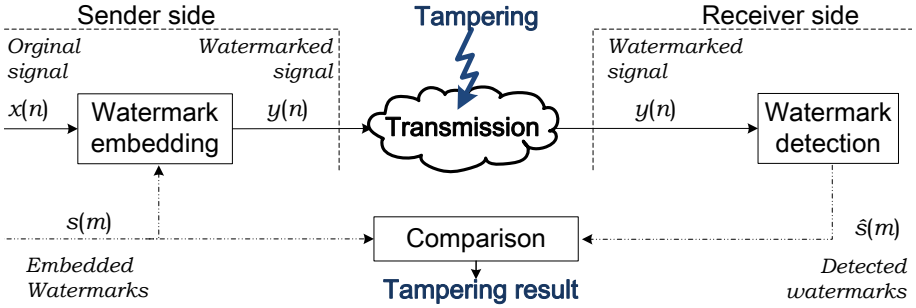

FIGURE 1. Proposed scheme for tampering detection.

over the Internet or through remote distance. Tampering are possibly happened when speech is transmitted to the destined recipient.

For example in criminal investigation, the recorded speech has to pass through a series of judicial procedures in which different people may be involved. Since improper actions taken to handle, examine, and store the speech are possible to destroy the integrity of it (intentionally or unintentionally), and not everyone involved is trustful, it is difficult to ensure the originality and integrity of speech. Especially those who are responsible for examining the speech have malicious intent to conceal important information or cover up the reality. For example, by using voice conversion, speech content (what is the speaker saying) can be tampered, e.g., a word replacement from "YES" to "NO"; by using speech morphing, the individuality of speaker (who is saying) can be deliberately transformed to that of another speaker. As the speaker identity plays a key role in the criminal investigation, any single word change or forged speaker will result in serious problem for judgement.

To check whether tampering has occurred to speech signals during the transmission, an overall block diagram for tampering detection is given out in this section, where speech watermarking is employed. As shown in Fig. 1, before sending the original signal $x(n)$ to the recipient, watermarks $s(m)$ will be embedded into it to construct the watermarked signal $y(n)$. Then the watermark signal $y(n)$ will be transmitted. After receiving $y(n)$ at the receiver side, watermarks will be extracted from received $y(n)$. The extracted watermarks, named as $\hat{s}(m)$, will be compared with $s(m)$ to check whether tampering has occurred. If the speech watermarking method can satisfy both robustness and fragility, tampering could be detected by the mismatched bits between $s(m)$ and $\hat{s}(m)$.

3. **Formant enhancement-based speech watermarking.**

3.1. **Related studies on formant enhancement.** Formant corresponds to a concentration of frequencies that are close to the resonance frequency of the vocal tract. As a crucial acoustic feature for speech perception, formant needs to be enhanced when the speech quality is impaired by noise or other reasons. The method of re-shaping formant to make it sharper is generally referred as formant enhancement. This kind of methods has been widely used to improve the speech quality for speech recognition system where speech quality is reduced by noise [39], and the hidden Markov model (HMM) based speech synthesis [40, 41] where speech is muffled by the over-smoothed spectral envelope. Most of these methods try to obtain a more prominent formant structure by enhancing formant without shifting the center frequency of formant.

Since formant can be enhanced to improve speech quality [42], and such modifications do not cause perceptual distortion to the original speech, watermarking based on formant enhancement is possible to be imperceptible to human. Therefore, we employ this concept

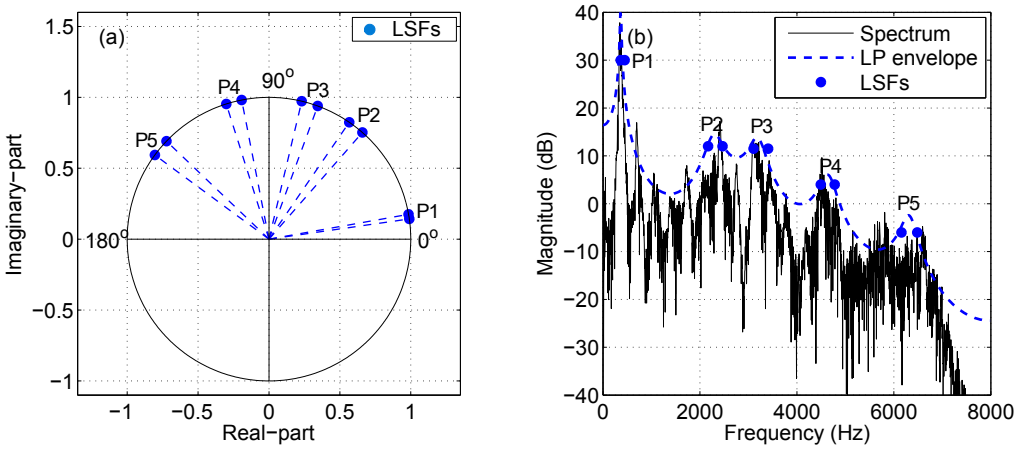

FIGURE 2. (a) LSFs distribution on half unit circle and (b) relationship between LSFs and formants.

to achieve inaudibility for watermarking. Watermarks will be embedded through formant enhancement. In most speech synthesis methods [40, 41], formants are enhanced with complicated methods so that the dynamics between formant peaks and spectral valleys can be increased. However, such complicated methods are not suitable for watermark embedding and robust watermark extraction. As to inherited formant enhancement for watermarking, we have investigated a simple but effective formant enhancement method in our previous work [43, 44]. This work has showed how the formants could be estimated, enhanced, and applied for watermarking. The following subsections give a quick review of the formant enhancement-based watermarking.

3.2. **Formant enhancement by controlling LSFs.** To enhance formant for watermarking, formant should be first estimated. Based on the theory of source-filter model, the set of linear prediction (LP) coefficients as an all-pole model, can provide accurate estimate of formants. In practice, LP coefficients are usually substituted with LSFs to ensure the stability of predictor. Moreover, LSFs have other excellent properties: (1) they are less sensitive to noise; (2) the influences caused by the deviation of LSFs can be limited to local spectral, which suggests that if LSFs are used to enhance formant for watermark embedding, the distortion in both spectral and sound quality could be minimized; (3) LSFs are universal features in different speech codecs, thus watermarks in LSFs can be preserved even after the coding/decoding processes. Hence we employ LSFs to enhance formant. The LSFs converted from LP coefficients satisfy the ordering property from 0 to π as follows [45]:

$$0^o < \phi_1 < \phi_2 < \phi_3 < \cdots < \phi_p < 180^o, \tag{1}$$

where p is LP order, $\phi_i, 1 \leq i \leq p$, are LSFs. The positions of LSFs on half unit circle can reflect the formants of speech. Figure 2 shows an example of the distribution of LSFs on half unit circle and the relationship between LSFs and formants of a speech segment (length: 250 ms, sampling frequency: 16 kHz). The LP order for formant estimation is ten, so there are ten LSFs obtained in Fig. 2(a). In theory, two adjacent LSFs (a pair of LSFs) can produce a formant, and the closer two LSFs are, the sharper formant is. As shown in Fig. 2(b), five formants are estimated in the LP spectral envelope. These formants correspond to the "P1" to "P5" labelled LSFs pairs in Fig. 2(a).

Since formant can be produced by a pair of LSFs, and the closer two LSFs are, the sharper the formant is, formant can be effectively enhanced by directly closing up two LSFs. Figure 3 illustrates how this idea can be implemented. In Fig. 3, original formant

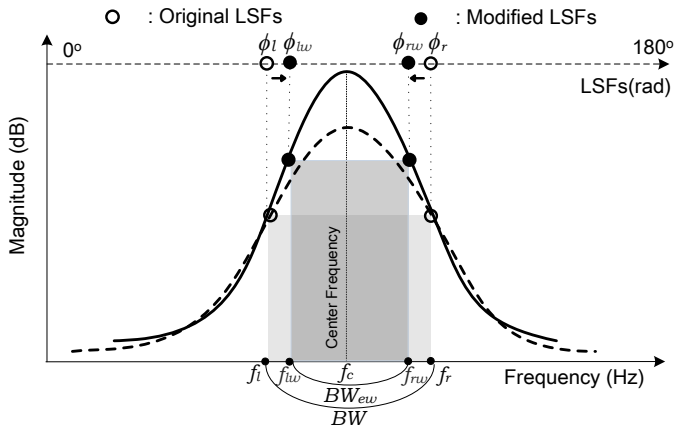

FIGURE 3. Formant enhancement by controlling a pair of LSFs.

(dotted curve) is produced by a pair of LSFs, ϕ_l and ϕ_r. Its sharpness can be mathematically measured by the tuning level, that is Q-value defined in Eq. (2), where f_c is the center frequency of formant, BW is the bandwidth between f_l and f_r that converted from ϕ_l and ϕ_r with Eq. (3), in which F_s is the sampling frequency of signal.

$$Q = \frac{f_c}{BW} = \frac{f_c}{f_r - f_l} \tag{2}$$

$$f_r = \frac{\phi_r}{2\pi} \times F_s \quad \text{and} \quad f_l = \frac{\phi_l}{2\pi} \times F_s \tag{3}$$

To enhance this formant, as seen in Fig. 3, two LSFs ϕ_l and ϕ_r are symmetrically shifted to close to each other, that is ϕ_l to ϕ_{lw} and ϕ_r to ϕ_{rw}. This process can be expressed with Eq. (4), where Δ is used to control the degree of shift, a bigger Δ indicates a more severe shift of LSFs as well as a much enhanced formant.

$$\phi_{lw} = \phi_l + \Delta \quad \text{and} \quad \phi_{rw} = \phi_r - \Delta, \quad 0 < \Delta < (\phi_r - \phi_l)/2 \tag{4}$$

After obtaining two shifted LSFs ϕ_{lw} and ϕ_{rw}, a narrower bandwidth BW_{ew} is produced. According to Eq. (5), the tuning level of original formant has been increased to Q_{ew}, and the enhanced formant (solid curve in Fig. 3) has become much sharper.

$$Q_{ew} = \frac{f_c}{BW_{ew}} = \frac{f_c}{f_{rw} - f_{lw}} \tag{5}$$

where f_{lw} and f_{rw} are calculated as follows:

$$f_{rw} = \frac{\phi_{rw}}{2\pi} \times F_s \quad \text{and} \quad f_{lw} = \frac{\phi_{lw}}{2\pi} \times F_s \tag{6}$$

Note that in the above manipulation, two LSFs are symmetrically shifted, so there is no deviation between the center frequency of the original formant and the enhanced formant which furthest maintains the sound quality of the host signal.

3.3. **Formant enhancement for watermarking.** We apply the above concept for watermarking. Different watermarks "0" and "1" are embedded into the LSFs of the host signal by enhancing different formants.

3.3.1. *Embedding concept.* The proposed method is a frame-based method. For each frame, one bit watermark will be embedded. "0" is embedded by enhancing the sharpest formant and "1" is embedded by enhancing the second sharpest formant. Since the closer two LSFs are, the sharper the formant is, these two formants can be easily extracted from the speech frame by checking the bandwidths of each formant and selected two smallest

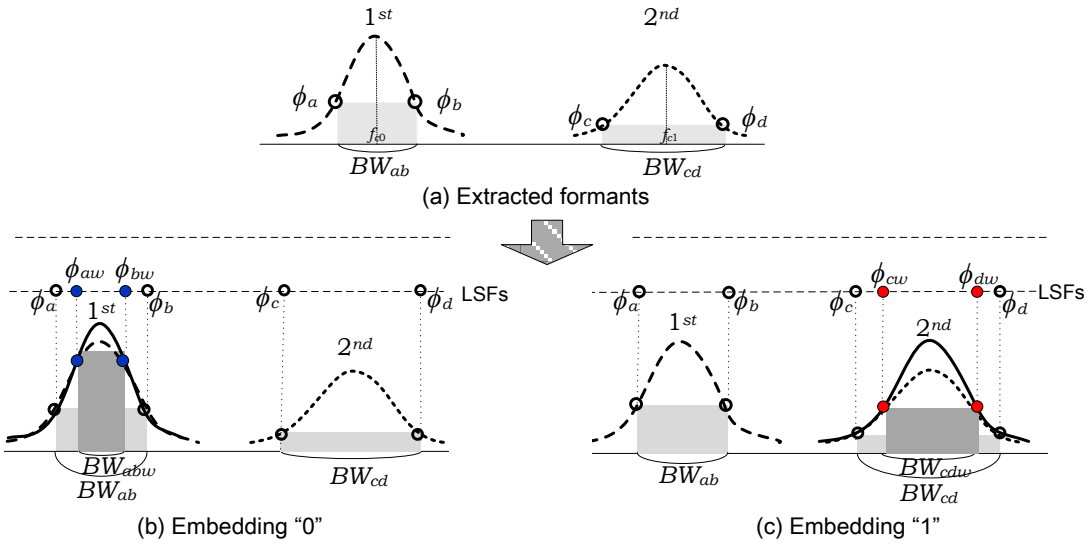

FIGURE 4. Concept of watermark embedding: (a) extracted two formants (b) embedding "0", and (c) embedding "1".

ones. Note that the first formant in low frequency and the last formant in high frequency are not involved in the selection process, since enhancing them will drastically distort the sound quality of the host signal. Figure 4(a) illustrates the extracted two formants, where the sharpest formant (labelled as 1^{st}) is produced by ϕ_a and ϕ_b and the second sharpest formant (labelled as 2^{nd}) is produced by ϕ_c and ϕ_d. According to Sec. 3.2, the sharpest formant has the smallest bandwidth BW_{ab} and its tuning level is $Q_0 = f_{c0}/BW_{ab}$, the second sharpest formant has the second smallest bandwidth BW_{cd}, and its tuning level is $Q_1 = f_{c1}/BW_{cd}$. The rules for embedding are as follow:

A. Rule of embedding "0": To embed "0", as seen in Fig. 4(b), the sharpest formant will be enhanced. An enhancing factor Ω_{e0} ($\Omega_{e0} > 1$) in Eq. (7) is used to control how much the formant is enhanced. According to Eq. (7), BW_{ab} has to be reduced to its $1/\Omega_{e0}$ for the enhancement, that is the newly obtained bandwidth BW_{abw} equals BW_{ab}/Ω_{e0}. To achieve this, original LSFs ϕ_a and ϕ_b will be shifted to ϕ_{aw} and ϕ_{bw} with the modification degree Δ_{e0} in Eq. (8), where Δ_{e0} is calculated by ϕ_a, ϕ_b, and Ω_{e0} with Eq. (9). Since BW_{cd} is originally bigger than BW_{ab}, after enhancing the sharpest formant, an updated relationship, $BW_{cd} > BW_{abw} \times \Omega_{e0}$, has been established in the current frame.

$$Q_0 \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}} \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}/\Omega_{e0}} = \frac{f_{c0}}{BW_{abw}}, \Omega_{e0} > 1 \quad (7)$$

$$\phi_{aw} = \phi_a + \Delta_{e0} \quad \text{and} \quad \phi_{bw} = \phi_b - \Delta_{e0} \quad (8)$$

$$\Delta_{e0} = \frac{1}{2}[(\phi_b - \phi_a) \times (1 - \frac{1}{\Omega_{e0}})] \quad (9)$$

B. Rule of embedding "1": To embed "1", as seen in Fig. 4(c), the second sharpest formant will be enhanced. An enhancing factor Ω_{e1} ($\Omega_{e1} = \frac{BW_{cd}}{BW_{ab}}$) in Eq. (10) is used for the enhancement. With this factor, BW_{cd} will be reduced to the same as BW_{ab}. This is achieved by shifting ϕ_c and ϕ_d to ϕ_{cw} and ϕ_{dw} with Eq. (11), where Δ_{e1} is calculated by ϕ_c, ϕ_d and Ω_{e1} with (12). Therefore, after embedding "1", the bandwidth relationship

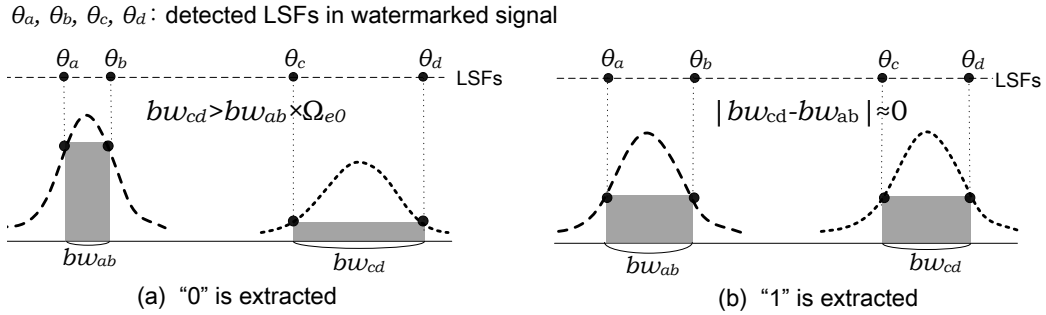

FIGURE 5. Concept of watermark extraction: (a) "0" is extracted and (b) "1" is extracted.

$BW_{cdw} = BW_{ab}$ has been established in the current frame.

$$Q_1 \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}} \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}/\Omega_{e1}} = \frac{f_{c1}}{BW_{ab}} = \frac{f_{c1}}{BW_{cdw}}, \Omega_{e1} = \frac{BW_{cd}}{BW_{ab}} \quad (10)$$

$$\phi_{cw} = \phi_c + \Delta_{e1} \quad \text{and} \quad \phi_{dw} = \phi_d - \Delta_{e1} \quad (11)$$

$$\Delta_{e1} = \frac{1}{2}[(\phi_d - \phi_c) \times (1 - \frac{1}{\Omega_{e1}})] \quad (12)$$

In summary, different watermarks are embedded by establishing different bandwidth relationships between the sharpest and the second sharpest formants via formant enhancement. The different bandwidth relationships enable watermarks to be blindly extracted. Note that this watermarking method can be applied for both voiced/unvoiced speech frames, while the formants detected from unvoiced speech segment are just pseudo-formants.

3.3.2. *Extraction concept.* According to the embedding rules, bandwidth relationships always exist in the sharpest and the second sharpest formants no matter embed "0" or "1". Therefore, in extraction process, for each frame of watermarked signal, we extract these two formants respectively. As seen in Fig. 5, the sharpest formant should have the smallest bandwidth, we name it as bw_{ab} (produced by θ_a and θ_b). The second sharpest formant should have the second smallest bandwidth, we name it as bw_{cd} (produced by θ_c and θ_d). If "0" has been embedded, according to Fig. 5(a), the relationship between bw_{ab} and bw_{cd} should be $bw_{cd} > bw_{ab} \times \Omega_{e0}$, an equivalent representation is given in Eq. (13); if "1" has been embedded, bw_{cd} in Fig. 5(b) should be similar to bw_{ab}, as expressed in Eq. (14). Since LP analysis calculates LP coefficients (or LSFs) with the criterion that the mean-squared error is always minimized, the LP coefficients (or LSFs) that are derived from watermarked frame are not exactly the same as those after embedding process even there is no modifications. Therefore, as shown in Eq. (15), we set a threshold (half of the difference between two extracted bandwidths) to discriminate two cases of embedding "0" or "1", and enable the method to be error-tolerant.

$$\text{embedding "0": } \quad bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1) \quad (13)$$

$$\text{embedding "1": } \quad bw_{cd} \approx bw_{ab} \quad (14)$$

$$\hat{s}(m) = \begin{cases} 0, & bw_{cd} - bw_{ab} > bw_{ab} \times \frac{\Omega_{e0}-1}{2} \\ 1, & \text{otherwise} \end{cases} \quad (15)$$

3.3.3. *Analysis.* In the proposed watermarking method, the enhanced formant is selected according to the frequency characteristics of each frame and the watermark "0" or "1", therefore, the enhanced formant is possible to exist in any frequency range, which enables

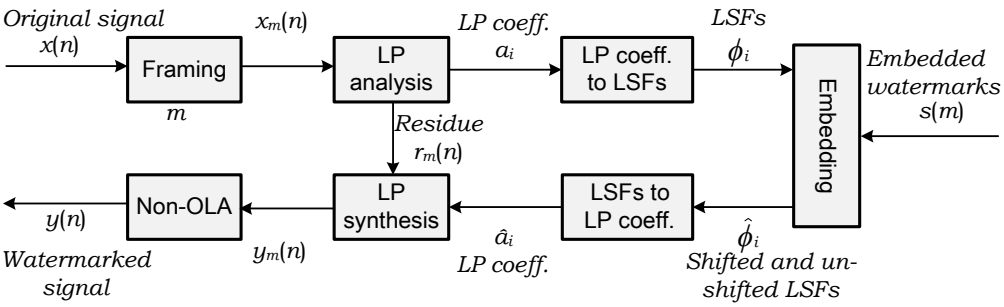

FIGURE 6. Block diagram of watermark embedding at the sender side.

the watermarks to be well hidden. Moreover, since watermarks are embedded into the intrinsically irregular formant structures, it is difficult for the attackers or the third party to confirm whether the formant structure was formed by artificial manipulation as the embedded bandwidth relationship is also possible in a rough speech. Especially when the LP order for estimating formants is unknown, bandwidth relationship is unable to discover. For more detailed reasons about why the sharpest and the second sharpest formants are selected to carry the watermarks, please refer to our previous study [44].

4. **Tampering detection based on watermarking.** This section talks about how to apply the watermarking method for detect tampering. Watermarks will be embedded at the sender side and then will be extracted at the receiver side for detecting tampering.

4.1. **Process at the sender side.** At the sender side, we have the host speech signal $x(n)$ and the watermarks $s(m)$. To construct the watermarked signal $y(n)$, $s(m)$ will be embedded into $x(n)$. Figure 6 has a block diagram of embedding process. Watermarks are embedded as follow. First, $x(n)$ is segmented into non-overlapping frames (indexed by $x_m(n)$). For each frame, a p-th order LP analysis is applied to obtain LP coefficients $a_i, i = 1, 2, ..., p$ and LP residue $r_m(n)$. LP coefficients a_i are then converted to LSFs ϕ_i to represent the formants in each frame. Each frame will be embedded with one bit watermark "0" or "1" (according to $s(m)$) according to the rules introduced in Sec. 3.3, after which, a pair of shifted LSFs (ϕ_{aw} and ϕ_{bw} for embedding "0", or ϕ_{cw} and ϕ_{dw} for embedding "1") are generated. All LSFs $\hat{\phi}_i$ including the shifted LSFs and the other unshifted LSFs will be converted back to LP coefficients $\hat{a}_i$. The current frame $y_m(n)$ is then synthesized by the LP residue $r_m(n)$ and LP coefficients $\hat{a}_i$. The watermarked signal $y(n)$ is reconstructed with all watermarked frames using non-overlapping and adding function. Finally, watermark signal $y(n)$ will be transmitted.

4.2. **Process at the receiver side.**

4.2.1. *Watermark extraction.* After receiving $y(n)$ at the receiver side, watermarks will be extracted to confirm whether tampering has occurred to watermarked signal before receiving it. The block diagram of watermark extraction process is shown in Fig. 7. We apply the same procedures as those in the embedding process to watermarked signal $y(n)$ to obtain the LSFs θ_i of each frame. Each frame will be extracted with one bit by using the method in Sec. 3.3. All extracted bits can construct the whole watermark signal $\hat{s}(m)$.

4.2.2. *Verification of tampering.* To check whether tampering has occurred, extracted watermarks $\hat{s}(m)$ will be compared with embedded watermarks $s(m)$. Ideally, if the watermarking method could satisfy fragility, once tampering occurred, watermarks in tampered segment will be destroyed. Therefore, tampering could be detected by the

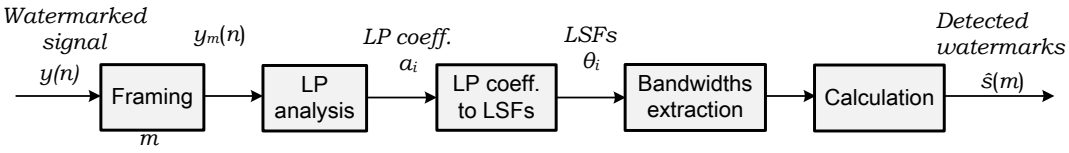

FIGURE 7. Block diagram of watermark extraction at the receiver side.

mismatched bits between $s(m)$ and $\hat{s}(m)$. If there is no mismatch, it means the received signal is the original signal and no tampering occurred; otherwise, each mismatch indicates that the corresponding frame in received signal has been possibly tampered. For instance, if $s(m)$=01001101... while the detected $\hat{s}(m)$=01101101..., this indicates the third frame may have been tampered.

5. **Evaluations.**

5.1. **Database and conditions.** In this section, we evaluated the proposed scheme with respect to inaudibility, robustness, and fragility (the proposed scheme is a blind method). All 12 speech stimuli in the ATR database (B set) [47] (Japanese sentences uttered by six males and six females, 8.1-sec, 20 kHz, 16 bits) were used as the host signals. Since our scheme is based on speech analysis/synthesis, the frame size was fixed at 25 ms (40 frames in 1.0 second) to attain better sound quality. Every 10 frames were embedded with the same watermark and detected the watermark with a majority decision. Thus, the bit rate for embedding was 4 bps. The order of LP analysis was chosen as 10 based on our previous analysis in [44]. Ω_{e0} for embedding '0' was adopted as 2.0 to balance the conflicting requirements of inaudibility, robustness, and fragility. Ω_{e1} for '1' was automatically fixed based on bandwidth characteristics of each frame. Embedded watermarks was a single word "GOOD". Evaluations were also done to two other methods: the least significant bit-replacement (LSB) method [48] and the cochlear delay (CD) method [36]. A quick review of these methods is given as follows: LSB replaces the least significant bits with watermarks at the quantization level so that the replacement does not cause severe distortion to the host signal; CD embeds watermarks "0" and "1" as two kinds of group delays by exploring the characteristics of human cochlear delay. All of the evaluations were conducted on Linux operating system with kernel 3.4.87-2vl6. The CPU is Intel (R) Core (TM) i7-4771 with frequency of 3.50 GHz, and the memory is 15.6 GiB. The time consumptions of watermark embedding and detection for one 8.1-sec speech signal are 1.02 sec and 1.06 sec with the proposed method, 1.45 sec and 1.12 sec with the LSB method, and 1.06 sec and 1.24 sec with the CD method.

5.2. **Evaluations for inaudibility.** Inaudibility was checked by objective experiments. The log spectrum distortion (LSD) [49] and the perceptual evaluation of speech quality (PESQ) [50] are objective measures. They can estimate the degradation between the host signal and the watermarked signal.

LSD defined in Eq. (16) can measure the spectral distance between the host signal and the watermarked signal, where m indicates the frame index, M is the total numbers of frames, $X(\omega, m)$ and $Y(\omega, m)$ are the spectra of m-th frame in the host signal and the watermarked signal, respectively. LSD of 1.0 dB is chose as the criterion, and a lower value indicates a less distortion.

$$\text{LSD} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(10\log_{10}\frac{|Y(\omega,m)|^2}{|X(\omega,m)|^2}\right)^2} \quad \text{(dB)} \tag{16}$$

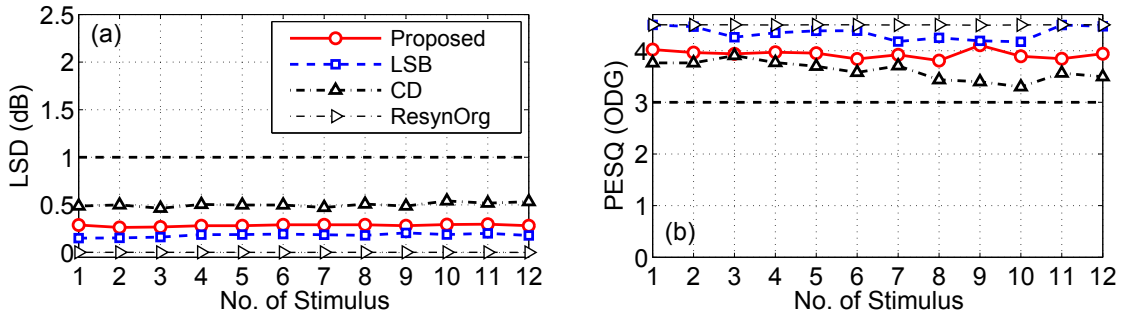

FIGURE 8. Objective evaluation results of sound quality: (a) LSD and (b) PESQ.

PESQ evaluates the speech quality with Objective Difference Grades (ODG) that range from -0.5 (very annoying) to 4.5 (imperceptible). ODG of 3.0 (slightly annoying) was set as the criterion, and a higher value indicates a better speech quality.

The evaluation results of LSD and PESQ for three watermarking methods are plotted in Fig. 8, where the straight dashed-lines in each sub-figure indicate the criteria for LSD (≤ 1.0 dB) and PESQ (≥ 3.0 OGD). As we can see, all three methods could satisfy the criteria for LSD and PESQ. The LSB method performed the best and the proposed method was a little better than CD method. These results indicated that these methods could objectively satisfy the inaudibility requirement. Especially, in this figure, an additional result labelled as "ResynOrg" was also given out. This result was calculated between the host signal and the resynthesized host signal (LP analysis/synthesis of host signal without watermarking) for checking whether sound distortion could be caused by speech analysis/synthesis in spite of the embedding of watermarks in the proposed method. Based on the obtained result, the resynthesized host signal had almost the same sound quality as the host signal, which suggested sound distortion caused by speech analysis/synthesis was imperceptible.

5.3. **Evaluations for robustness.** Watermarking method should be robust against allowable speech processing (e.g., speech codecs, re-sampling, re-quantization) to guarantee the effectiveness of the embedded watermarks. In this section, the robustness of proposed method is evaluated in comparison with LSB and CD.

Robustness can be indicated by Bit Detection Rate (BDR), i.e., the ratio between correctly extracted watermarks and all embedded watermarks. The BDR can be calculated with Eq. (17), where $s(m)$ represents embedded watermarks, $\hat{s}(m)$ is detected watermarks, and M is the total length of $s(m)$. The symbol "$\oplus$" denotes the operation of "exclusive-OR", that is, if the bit values of $s(m)$ and $\hat{s}(m)$ are different, "$s(m) \oplus \hat{s}(m)$" equals 1; otherwise, "$s(m) \oplus \hat{s}(m)$" equals 0. We chose BDR of 90% as the criterion, and a higher BDR indicates a stronger robustness.

$$\mathrm{BDR} = \frac{M - \sum_{m=0}^{M} s(m) \oplus \hat{s}(m)}{M} \times 100 \quad (\%) \tag{17}$$

5.3.1. *Robustness against speech codecs.* Speech codecs is a kind of necessary processing for speech transmission over the Internet and telecommunication systems. Speech codecs can generally be classified into waveform-based and parameter-based schemes. Therefore, we separately applied these typical speech codecs of G.711 (pulse code modulation (PCM)), G.726 (adaptive differential PCM (ADPCM)), and G.729 (Code-excited linear prediction (CELP)) to the watermarked signals obtained from three watermarking methods. The BDR results calculated after speech codecs are presented in Fig. 9, where

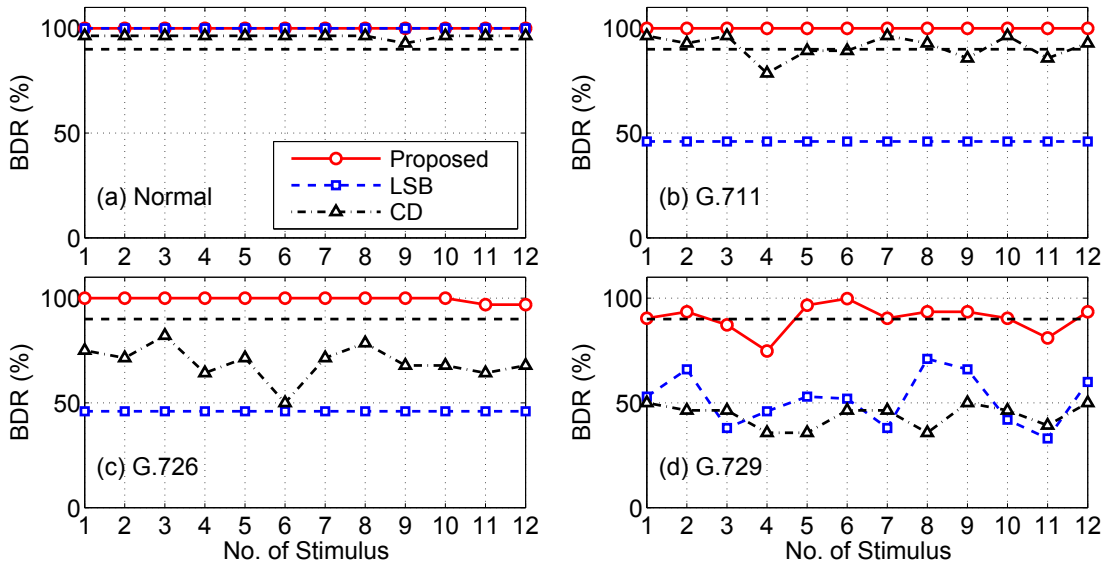

FIGURE 9. Evaluation of robustness against speech codecs: (a) normal extraction, (b) G.711, (c) G.726, and (d) G.729.

normal extraction (without codec) is also given out in Fig. 9(a). The straight dashed line in each sub-figure indicated the criteria for BDR ($\geq 90\%$). It is clear that LSB was not robust against any speech codec except for normal extraction, CD was only robust against normal extraction and G.711. In contrast, the proposed method could survive from normal extraction and all speech codecs (100% for G.711 and G.726, around 90% for G.729). These implied the proposed method was more robust against these speech codecs compared with LSB and CD.

5.3.2. *Robustness against practical speech processing.* We also evaluated the proposed method against several practical speech processing [33]. These included re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits, signal amplifying by 2.0 times, a single 100 ms echo addition of -6 dB (recommended by the Information Hiding and its Criteria (IHC) committee [51]), speech analysis/synthesis by short-time Fourier transform (STFT), and gammatone filterbank (GTFB). The BDR results after each processing have been plotted in Fig. 10. LSB was only robust against re-sampling at 24 kHz, re-quantization with 24 bits, and STFT; CD was robust against most processing except for re-quantization with 8 bits, echo addition, and GTFB. In comparison, the proposed method could correctly detect watermarks after these processing, which meant it was more robust than LSB and CD.

5.4. **Evaluations for fragility.** Many previous works, e.g., [28] and [30], have confirmed the fragility of their methods by carrying out various types of tampering. However, there is no consistent definition for tampering among these works. In general, tampering are performed based on the motivation of the attackers. In this case, any operation that can be used to tamper a speech should be evaluated for watermarking method when verifying its fragility and ability for tampering detection. Therefore, we evaluated the fragility of the proposed method against several possible tampering in this section. Since LSB and CD are not completely robust, even they are fragile against tampering, they are unable to tell whether the failed extraction of watermarks is caused by speech processing or tampering. That is to say, they cannot successfully detect tampering unless robustness being improved. Therefore, fragility evaluation was only conducted to the proposed method.

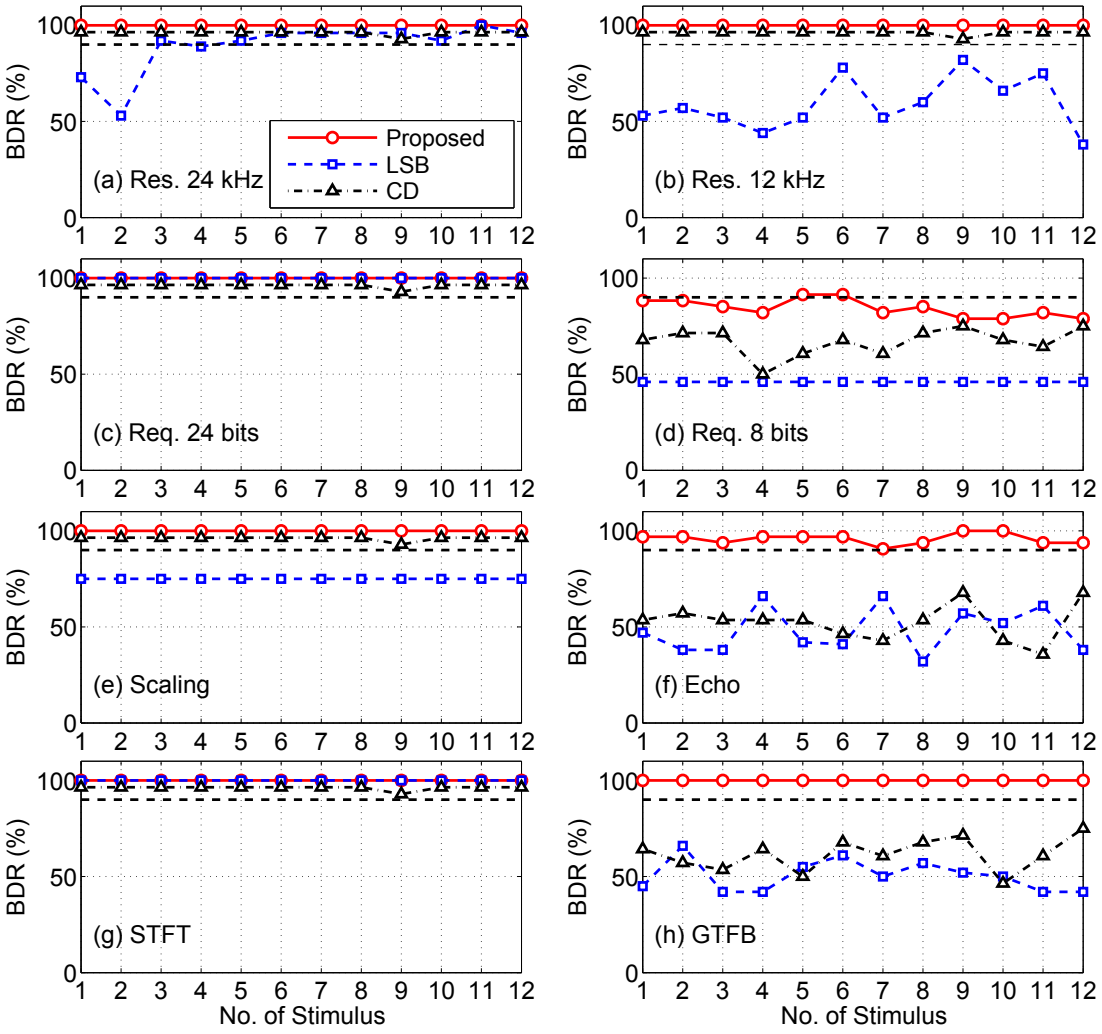

FIGURE 10. Evaluations of robustness against various practical speech processing.

As to intuitively reflect fragility, a 32×32 bitmap image in Fig. 11(a) was used as watermarks. Since bit rate was 4 bps, as to embed the complete image, 12 speech tracks are repeatedly connected to construct a long host signal (256 second). After embedding the image to the host signal, the middle segment of watermarked signal was separately tampered with the tampering listed in Tab. 1 (Line 2 to Line 9). These evaluations referred to [36] and our previous work [46]. Adding white noise and reverberation are channel distortion, tampering speech with these operations can be considered as disturbing the speech. Concatenating the watermarked signal with un-watermarked speech can be considered as content replacement. Filtering with low-pass and high-pass filters is regarded as removing specific frequency information of speech. Speed change (speech up and speed down) can modify the duration and tempo of speech without affecting its pitch. Pitch shift is to proportionally shift frequency components while preserving the duration of speech, which can be regarded as manipulating the individualities of the speaker.

The extracted image from un-tampered watermarked signal is shown in Figs. 11(b). where watermarks could be correctly extracted. The extracted images from other tampered watermarked signals are separately shown in Figs. 11(c) to (j). It is noticeable that watermarks in the tampered segment were destroyed. Tab. 1 gives out the accurate BDR results. Since the BDR calculated from the tampered segment were quite low compared with no tampering (normal extraction), we can conclude that the proposed method

TABLE 1. Bit detection rates in fragility evaluations.

No.	Tampering type	Description	BDR (%)
(b)	No tampering	$--$	100.0
(c)	Add white noise	normal distribution, $N(0.01, 1)$	45.19
(d)	Reverberation	real impulse response of 0.3 s	68.80
(e)	Concatenation	concatenate with un-watermarked speech	42.86
(f)	Low-pass filtering	order: 32-th, normalized cut-off frequency: 0.99	41.98
(g)	High-pass filtering	order: 32-th, normalized cut-off frequency: 0.01	49.85
(h)	Speed up	speed up the whole speech by +4%	71.56
(i)	Speed down	speed down the whole speech by −4%	79.51
(j)	Pitch shift	change the pitch of speech −4% in real time	68.12

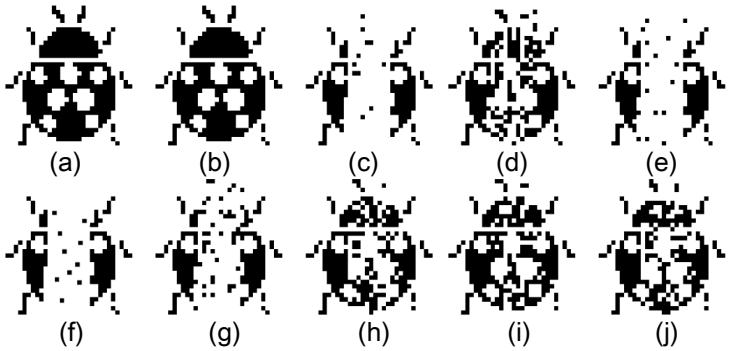

FIGURE 11. Evaluations of fragility against tampering.

was fragile against the evaluated tampering. Figure 12 illustrates an example of how detection errors happened, for example, after tampering by adding white noise. In Fig. 12(b), detection errors (shown in red cross) densely appeared in the tampered segment. In contrast, watermarks in the un-tampered segment could be correctly detected due to the robustness the proposed method. Therefore, tampering could be indicated with the destroyed watermarks. Figure 12(c) examined the bandwidth relationships before and after tampering of one tampered frame, where "1" has been embedded. Before tampering, bandwidth relationship, $BW_{ab} = BW_{cd}$, could be easily observed to correctly extracted the watermark. After tampering, BW_{ab} has been much narrowed to bw_{ab}, that is, bw_{ab} is much narrower than bw_{cd}. Therefore, it would be easily taken as that "0" has been embedded.

These obtained results suggested that the proposed method was fragile against tampering, and the destroyed watermarks could provide a strong evidence that signal has been tampered. As the embedding bit rate of watermarks is 4 bps, each embedded bit was able to account for 0.25 s speech segment when locating the tampering, although 0.25 s is too short to make a meaningful tampering of speech content. Therefore, as shown in Fig. 11, even some correct bits could still be intermittently extracted from the tampered speech segment, tampering could also be detected by checking several adjacent bits where detection errors densely appeared. Additionally, the detection precision is possible to be improved by increasing the embedding bit rate.

5.5. **Ability for tampering detection.** The above two subsections evaluated the robustness and fragility of the proposed method. Evaluation results indicated the proposed method was enough robust and fragile. To investigate the tampering detection ability of proposed method in more realistic situation, we considered the following evaluations.

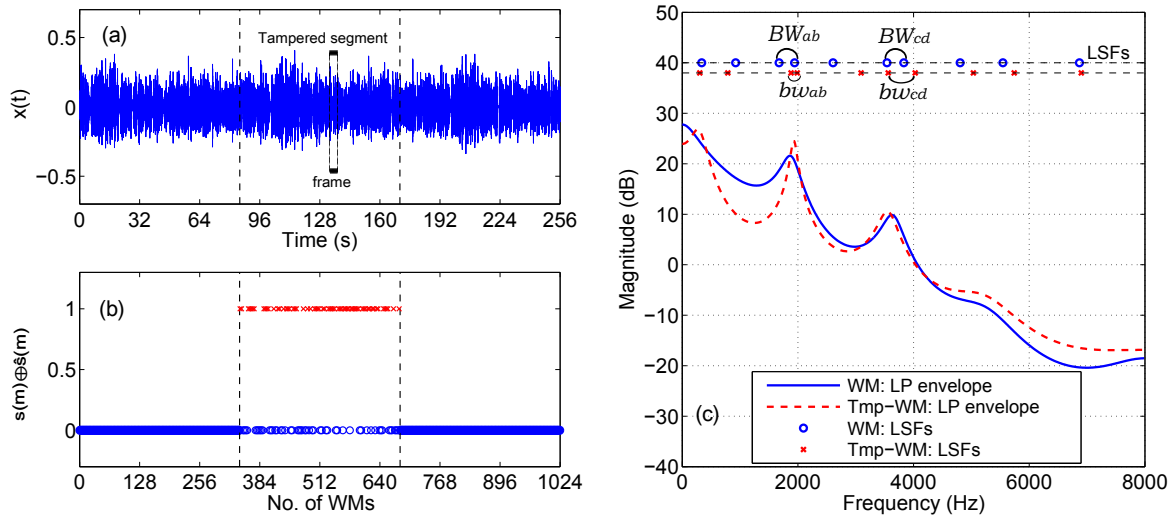

FIGURE 12. Example on fragility analysis: (a) host signal with middle segment tampered by adding white noise, (b) detection errors densely appear in the tampered segment, and (c) one frame analysis: bandwidth relationship for watermark "1" has been destroyed due to tampering.

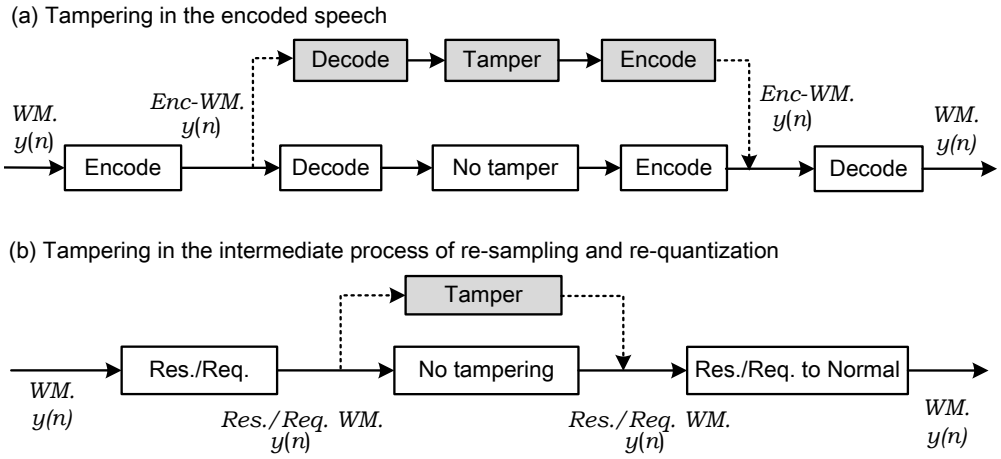

FIGURE 13. Flowchart of tampering in more realistic situation.

In realistic situation, encoding process is generally performed for watermarked signal at the sender side, and the decoding process is performed before watermark detection at the receiver side. To tamper the transmitted speech, as seen in Fig. 13(a), attackers should firstly decode the speech to raw data, make tampering, and then encoded back with the original coder. Likely, tampering also possibly happens to watermarked signals which are in the intermediate process of re-sampling and re-quantization. To investigate whether the proposed method could identify tampering under the situations that speech processing (speech codecs, re-sampling, and re-quantization) also exist, we followed the tampering process in Fig. 13 and then extract watermarks. Note that, to make a fair comparison, encoded watermarked signal in Fig. 13(a) was decoded and encoded even no tampering occurred. This process was made to compensate the speech codecs caused extraction error in the tampering case. In these evaluations, 12 speech tracks were embedded with the watermarks "GOOD". The types of tampering were the same as those in Sec. 5.4. All evaluation results were calculated on the average of 12 host signals.

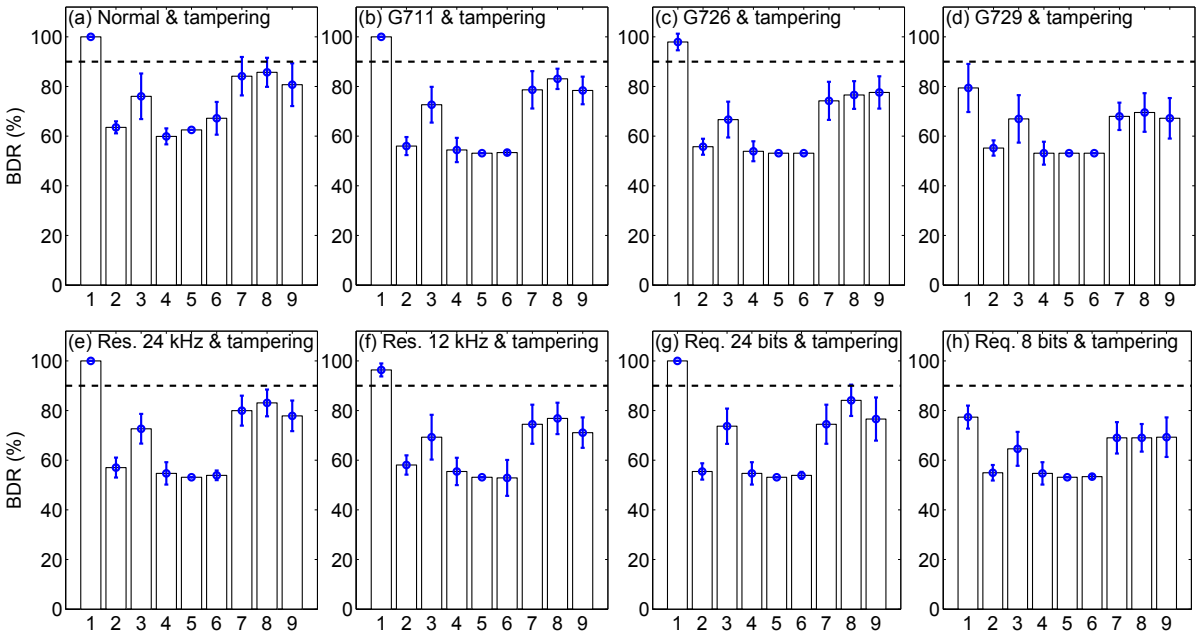

FIGURE 14. Evaluations for the tampering detection ability of the proposed method: (a) BDR comparison between normal extraction and after tampering, (b) to (h) BDR comparison after one kind of speech processing (G.711, or G.726, or G.729, or re-sampling at 24 kHz, or re-sampling at 8 kHz, or re-quantization with 24 bits, or re-quantization with 8 bits) and after both speech processing and tampering.

Figure 14(a) compares the BDR results of normal extraction (1st bar) and those after different tampering (2nd bar: addind white noise, 3rd bar: reverberation, 4th bar: concatenation, 5th bar: low-pass filtering, 6th bar: high-pass filtering, 7th bar: speed up, 8th bar: speed down, 9th bar: pitch shift). We got the similar results as those in Sec. 5.4 that when tampering occurred, BDR drastically reduced which enabled tampering to be easily figured out. Figures 14(b) to (h) compare the BDR results between two cases, one is BDR after one kind of speech processing (1st bar), and the other one is BDR after both the speech processing and different tampering (2nd bar: adding white noise, 3rd bar: reverberation, 4th bar: concatenation, 5th bar: low-pass filtering, 6th bar: high-pass filtering, 7th bar: speed up, 8th bar: speed down, 9th bar: pitch shift). For Figs. 14(b), (e), and (g), BDR was quite high when only speech processing applied, while after tampering, BDR was reduced. In Figs. 14(c) and (f), speech processing slightly introduced some bit extraction errors, while compared with those after tampering, the discrepancy in BDR could be equivalently kept as that in Fig. 14(a) (normal extraction & tampering). This was because speech processing had the same influence to watermarked signal no matter there was tampering or not. These results suggested that speech processing did not affect the detection of tampering and tampering could be detected no matter there is speech processing or not. However, in Fig. 14(d), BDR after G.729 were deteriorated even without tampering, this was because watermarked signal was encoded and decoded twice by G.729 which doubly introduced bit detection errors, thus it would be easily mistaken G.729 as tampering. Similarly, in Fig. 14(h), BDR after re-quantization with 8 bits was also deteriorated and made it difficult to distinguish it from tampering. To overcome these problems, robustness of the proposed method should be continually improved in the next step.

5.6. **Discussion.** The above sections evaluate the performance of the proposed method with respect to inaudibility, robustness, and fragility. In inaudibility evaluations, the proposed method can satisfy the criteria of both LSD and PESQ, which indicates it can objectively satisfy inaudibility. In comparison, LSB method achieves better performance in audibility, since this method directly embeds watermarks in the least significant bits so that the distortion to the host signal can be minimized. The CD method utilizes that human cannot distinguish enhanced-delay from the host signal so that watermarks can also be inaudibly embedded.

In robustness evaluations, performances of the proposed method, LSB, and CD are evaluated against speech codecs and speech processing. LSB method cannot show strong robustness when subjects to various processing. This is because watermarks in the least significant bits can be easily reset by operations related to amplitude modifications or lossy processing. CD method is basically robust. However, according to the characteristics of enhanced cochlear delay, extraction for different watermarks strongly depends on the cue in low-frequency phase. Correspondingly, once phase information in low frequency is destroyed or erased by other processing, such as GTFB and G.729 codec, watermarks cannot be extracted. In the proposed method, watermark extraction by identifying bandwidth relationship is able to tolerate slight distortions of frequency components caused by other processing, especially when the frequency distortions are not in the modified bandwidth range. Therefore, the proposed method exhibits stronger robustness compared with other methods.

Based on the results from robustness evaluations, fragility evaluations are only conduced to the proposed method. In these evaluations, a series of tampering are performed to the watermarked signals, due to which watermarks cannot be correctly extracted. Therefore, the destroyed watermark can function as a sign to indicate that tampering has occurred. Additionally, to check the detection ability of the proposed method under the situation that speech processing also exist, an in-depth evaluation is also carried out. By comparing the BDR results obtained from watermarked signal processed by speech processing, and the results from watermarked signal processed by both speech processing and tampering, tampering can be distinguish from most speech processing. These results further verify the tampering detection ability of the proposed method.

In summary, the proposed scheme has good performance in inaudible, robustness, and fragility. Moreover, it can detect tampering with its fragility. The embedding capacity of the scheme, though relatively low, is sufficient for locating tampering in time domain. Nonetheless, some remaining issues need to be addressed in the current work. E.g., if the attacker tries to add or crop segment to the transmitted signal, since the proposed scheme is a frame-based watermarking, such kind of tampering will disturb the frame segmentation for watermark extraction. As a result, extraction error will appear from the start point where adding or cropping occurred, just like Fig. 15. Although it is easy to judge how long the watermarked signal has been added or cropped by checking the length of watermarked signal, watermarks are quite difficult to be correctly extracted without using a frame synchronization scheme, especially when several speech segments are cropped or added to different positions of the watermarked signal. Therefore, an automatic scheme for frame synchronization should be implemented in the proposed method. Besides, more types of tampering should be investigated to verify the detection ability of the proposed scheme.

6. **Conclusions.** This paper proposes a tampering detection scheme for speech signals based on speech watermarking. The proposed method utilizes the concept associated with formant enhancement to realize inaudibility. Watermarks are embedded as formant

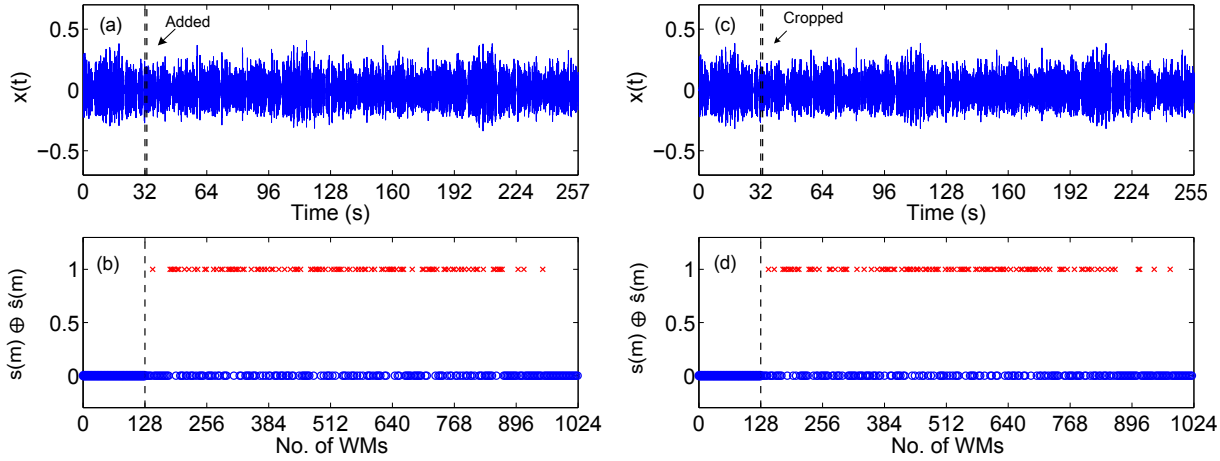

FIGURE 15. Analysis on the defect of proposed method: (a) adding speech segment to watermarked signal and (b) watermark detection errors start to appear at the point where speech segment is added; (c) cropping speech segment from watermarked signal and (d) watermark detection errors start to appear at the point where speech segment is cropped.

enhancement with a straight-forward way by symmetrically controlling a pair of LSFs. We evaluate the proposed scheme with respect to inaudibility, robustness, and fragility. The evaluation results reveal that the proposed scheme cannot only satisfy inaudibility but also provide good robustness. Moreover, the proposed method is capable of locating the tampering in time-domain at sufficient precision with its fragility, and its detection ability will not be degraded even speech processing exist. Therefore, we can conclude that the proposed method can effectively detect tampering in speech signals. However, a frame synchronization scheme should be implemented for the proposed for improved detection ability, and the robustness of the proposed method against G.729 speech codec and re-quantization at lower bits (e.g. 8 bits) also need to be improved in the next step.

Acknowledgment. This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 23300070), an A3 foresight program made available by the Japan Society for the Promotion of Science,the telecommunication advancement foundation, and funding by China Scholarship Council.

REFERENCES

[1] H. Kawahara, I. Masuda-Kasuse, and A. de Cheveigne, Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a reptitive Structure in Sounds, *Speech Communication*, vol. 27, pp. 187–207, 1999.

[2] T. Toda, A. W. Black, and K. Tokuda, Voice conversion based on maximum likelihood estimation of spectral parameter trajectory, *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] H. Kawahara, H. Banno, T. Irino and P. Zolfaghari, ALGORITHM AMALGAM: Morphing waveform based methods, sinusoidal models and STRAIGHT, *Proc. ICASSP*, pp. 13–16, 2004.

[4] *http://en.wikipedia.org/wiki/Digital-forensics*

[5] M. C. Stamm and K. J. R. Liu, Forensic detection of image manipulation using statistical intrinsic fingerprints, *IEEE Trans. Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010.

[6] C. Fei, D. Kundur, and R. H. Kwong, Analysis and design of secure watermark-based authentication systems, *IEEE Trans. Information Forensics and Security*, vol. 1, no. 1, pp. 43–55, 2006.

[7] Q. Y. Zhang, P. F. Xing, Y. B. Huang, R. H. Dong, and Z. P. Yang, An efficient speech perceptual hashing authentication algorithm based on wavelet packet decomposition, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 6, no. 2, pp. 311-322, Mar. 2015.

[8] I. J. Cox, G. Dorr, T. Furon, Watermarking is not cryptography, *Lecture Notes in Computer Science*, vol. 4283, pp. 1–15, Springer, 2006.
[9] C. I. Podilchuk, E. J. Delp, Digital watermarking: algorithms and applications, *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 33–46, 2002.
[10] Y. H. Chen and H. C. Huang, Coevolutionary genetic watermarking for owner identification, *Neural Computing and Applications*, vol. 26, no. 2, pp. 291-298, Feb. 2015.
[11] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, Multimedia data embedding and watermarking technologies, *Proc. IEEE*, vol. 86, no. 6, pp. 1064–1087, 1998.
[12] S. V. Dhavale, R. S. Deodhar, D. Pradhan, and L.M. Patnaik, High payload adaptive audio watermarking based on cepstral feature modification, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 5, no. 4, pp. 586-602, Oct. 2014.
[13] W. Song, X. Sun, C. Liu, and L. Tang, A new watermarking frame based on the genetic algorithms and wavelet packet decomposition, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 6, no. 3, pp. 613–621, 2015.
[14] A. P. F. Petitcolas, R. J. Anderson, M. G. Kuhn, Information hiding-a survey, *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.
[15] H. C. Huang, S.C. Chu, J.S. Pan, C.Y. Huang, and B.Y. Liao, Tabu search based multi-watermarks embedding algorithm with multiple description coding, *Information Sciences*, vol. 181, no. 16, pp. 3379–3396, Aug. 2011.
[16] S. A. N. Thajeel and G. Sulong, A novel approach for detection of copy move forgery using completed robust local binary pattern, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 6, no. 2, pp. 351–364, 2015.
[17] H. C. Huang, F. C. Chang, Robust image watermarking based on compressed sensing techniques, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 5, no. 2, pp. 275–285, 2014.
[18] R. Noriega, M. Nakano, B. Kurkoski, and K. Yamaguchi, High payload audio watermarking: toward channel characterization of MP3 compression, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 2, no. 2, pp. 91–107, 2011.
[19] J. Sang and M. S. Alam, Fragility and robustness of binary phase only filter based fragile/semi-fragile digital image watermarking, *IEEE Trans. Instrumentation and Measurement*, vol. 57, no. 3, pp. 595–606, 2008.
[20] M. Fallahpour, S. Shirmohammadi, M. Semsarzadeh, J. Zhao, Tampering detection in compressed digital video using watermarking, *IEEE Trans. Instrumentation and Measurement*, vol. 63, no. 5, pp. 1057–1072, 2014.
[21] C. C. Chang, Y. S. Hu, and T. C. Lu, A watermarking-based image ownership and tampering authentication scheme, *Pattern Recognition Letters*, vol. 27, no. 5, pp. 439–446, 2006.
[22] D. Kundur, D. Hatzinakos, Digital watermarking for telltale tamper proofing and authentication, *Proc. IEEE*, vol. 87, pp. 1167–1180, 1999.
[23] J. Grim, P. Somol,and P. Pudil, Image forgery detection by local statistical models, *Proc. IIHMSP*, pp. 579–582, 2010.
[24] L. Li, S. Li, and H. Zhu, An efficient scheme for detecting copy-move forged images by local binary patterns, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 4, no. 1, pp. 46–56, 2013.
[25] H. C. Huang and F. C. Chang, Robust image watermarking based on compressed sensing techniques, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 5, no. 2, pp. 275–285, 2014.
[26] P. Bestagini, S. Milani, M. Tagliasacchi, and S. Tubaro, Local tampering detection in video sequences, *Proc. Multimedia Signal Processing (MMSP)*, pp. 488–493, 2013.
[27] T. Painter and A. Spanias, Perceptual coding of digital audio, *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, 2000.
[28] C. M. Park, D. Thapa, G. N. Wang, Speech authentication system using digital watermarking and pattern recovery, *Pattern Recognition Letters*, vol. 28 pp. 931-938, 2008.
[29] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, Robust audio watermarking using perceptual masking, *Signal Processing*, vol. 66, no. 3, pp. 337-355, 1998.
[30] M. Celik, G. Sharma, and A. M. Tekalp, Pitch and duration modification for speech watermarking, *Proc. ICASSP*, vol. II, pp. 17–20, 2005.

[31] 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.
[32] 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.
[33] M. Unoki and D. Hamada, Method of digital-audio watermarking based on cochlear delay characteristics, *J. Inn. Com. Inf., and Cont.*, vol. 6, no.(3(B)), pp. 1325–1346, 2010.
[34] M. Unoki, K. Imabeppu, D. Hamada, A. Haniu, and R. Miyauchi, Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics, *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 2, no. 1, pp. 1-23, 2011.
[35] M. Unoki and R. Miyauchi, Reversible watermarking for digital audio based on cochlear delay characteristics, *Proc. IIHMSP*, pp. 314–317, 2011.
[36] M. Unoki and R. Miyauchi, Detection of tampering in speech signal with inaudible watermarking technique, *Proc. IIHMSP*, pp. 118–121, 2012.
[37] *http://www.itu.int/rec/T-REC/en.*
[38] F. Itakura, Line spectrum representation of linear predictive coefficients of speech signals, *J. Acoust. Soc. Am.*, vol. 57. no. 537(A), pp. 35-35, 1975.
[39] H. Brouckxon1, W. Verhelst1, and B. D. Schuymer, Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments, *Proc. Interspeech*, pp. 557-560, 2008.
[40] T. Raitio, A. Suni, H. Pulakka1, M. Vainio, and P. Alku, Comparison of formant enhancement methods for HMM-based speech synthesis, *Proc. ISCA Speech Synthesis Workshop*, 2010.
[41] HTS, HMM-based speech synthesis system, *http://hts.sp.nitech.ac.jp*, 2009.
[42] Recommendation ITU-T P.800, Methods for subjective determination of transmission quality, *International Telecommunication Union*, 1996.
[43] S. Wang and M. Unoki, Watermarking of speech signals based on formant enhancement, *Proc. EUSIPCO*, pp. 1257–1261, 2014.
[44] S. Wang and M. Unoki, Speech Watermarking Method based on Formant Tuning, *IEICE Trans. Enriched Multimedia*, vol. E98-D, no. 1, pp. 29-37, 2015.
[45] S. Wang and M. Unoki, Watermarking method for speech signals based on modifications to LSFs, *Proc. IIHMSP*, pp. 283–286, 2013.
[46] S. Wang, M. Unoki, and N. S. Kim, Formant enhancement based speech watermarking for tampering detection, *Proc. Interspeech*, pp. 1366-1370, 2014.
[47] K. Takeda et al, Speech database user's manual, *ATR Technical Report*, TR-I-0028, 2010.
[48] P. Bassia and I. P. Pitas, Robust audio watermarking in the time domain, *Proc. EUSIPCO*, pp. 25-28, 1998.
[49] A. Gray, Jr., and J. Markel, Distance measures for speech processing, *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.
[50] Y. Hu and P. C. Loizou, Evaluation of objective quality measures for speech enhancement, *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.
[51] Information hiding and its criteria for evaluation, *http:// www.ieice.org/iss/emm/ihc/en/index.php*