Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 7285624, 10 pages
https://doi.org/10.1155/2019/7285624

Research Article

Speech Watermarking for Tampering Detection Based on Modifications to LSFs

Xiangning Chen,[1] Weitao Yuan ,[1] Shengbei Wang ,[1] Chao Wang,[1] and Lin Wang[2]

[1]*Tianjin Key Laboratory of Autonomous Intelligence Technology and Systems, Tianjin Polytechnic University, Tianjin 300387, China*
[2]*Techfantasy. Co. Ltd., Tianjin 300387, China*

Correspondence should be addressed to Weitao Yuan; weitaoyuan@tjpu.edu.cn and Shengbei Wang; wangshengbei@tjpu.edu.cn

Received 25 June 2019; Revised 20 August 2019; Accepted 5 September 2019; Published 17 December 2019

Academic Editor: Krzysztof Puszynski

There have been serious issues concerning the protection of speech signals from malicious tampering. Digital watermarking has been paid much attention in solving this problem. This paper proposes a tampering detection approach based on speech watermarking by modifying the line spectral frequencies (LSFs). Watermarks are embedded into LSFs that derived from linear prediction (LP) analysis with dither modulation-quantization index modulation (DM-QIM). Minor modifications to LSFs introduced by quantization not only enable the watermarks to be inaudible to human auditory system but also provide the possibility of robustness against meaningful processing and fragility against tampering. We evaluated the proposed approach with objective evaluations with respect to inaudibility, robustness, and fragility. The results indicated that the proposed approach for tampering detection not only satisfied inaudibility but also provided good robustness against meaningful processing and fragility against malicious tampering.

1. Introduction

The development of digital technologies has greatly stimulated the widespread use of multimedia information. These technologies, however, also enable the digital signals to be delivered in a detached manner crossing time and space, which facilitates unforeseen operations (tampering) to be performed. In particular, advanced speech analysis/synthesis methods (e.g., STRAIGHT [1]) and their applications such as voice conversion [2, 3] and speech morphing [4] are capable of producing high fidelity of tampered speech. Since tampering the speech may cover up the fact and mislead the listeners, problems become particularly serious in forensic investigations [5, 6], where the evidence sometimes needs to be recovered from digital media and served as the basis for judicial proceedings. Correspondingly, it is quite necessary to investigate whether there is tampering happened to the speech, to ensure the integrity and originality of the speech.

Digital watermarking [7, 8] has drawn more and more attention in the past few years in speech protection. It can hide digital data at low energy level to the host signals while keeping the perceptual quality undistorted. Watermarking methods should satisfy four main requirements: (1) inaudibility, (2) blindness, (3) robustness, and (4) confidentiality. The importance of a particular requirement may vary upon the applications [9, 10]. For tampering detection, an additional requirement is vitally important, i.e., fragility. Fragility indicates that the embedded watermarks will easily be destroyed once a slight modification has been made to the watermarked signals [11]. Correspondingly, fragile watermarking has the ability to identify where tempering has occurred. However, speech signals usually need to be processed by speech codecs or other meaningful processing, and it seems unable for fragile watermarking to survive from these processing due to its fragility [12]. In practice, effective watermarking methods for tampering detection should satisfy two conflicting requirements: robustness against meaningful processing and fragility against malicious tampering. Only then can the watermarking methods provide reliable and effective protection of speech signals.

Digital watermarking for speech signals is more challenging compared with image watermarking, due to the extreme sensitivity of the human auditory system. Nonetheless, many successful watermarking algorithms for speech signals have been proposed. Sarreshtedari et al. [13] proposed to embed the compressed version of speech into original signal for tampering detection. Celik et al. [14] proposed a robust speech watermarking method by introducing small changes to pitch (fundamental frequency). Stability of such features under low data rate compression makes the method effective for semifragile authentication. Karnjana et al. [15, 16] proposed a scheme based on singular-spectrum analysis to detect the acoustic feature-based tampering. Wu and Jay Kuo [17] implemented a fragile speech watermarking based on odd/even modulation and exponential scale quantization. The pseudorandom noise was embedded as the watermarks in the discrete Fourier transform (DFT) magnitude domain by roughly approximating the MPEG audio psychoacoustic model. Another method proposed by Narimannejad and Mohammad [18] was based on phase quantization of the sinusoidal model, in which the watermarks were hidden via phase quantization of sine wave. Unoki and Hamada [19] proposed a digital audio watermarking method based on the characteristic of human cochlear delay. This approach was also successfully applied to speech signals for tampering detection [20].

We previously proposed a speech watermarking approach based on LSFs [21] and DM-QIM [22]. The quantization step of DM-QIM [23] was reasonably controlled to achieve a good balance between inaudibility and robustness. The evaluation results also suggested that the proposed approach could satisfy inaudibility and robustness. We also found that the proposed approach was very sensitive to processing that could change the shapes of waveform or the values of watermarked signal. This characteristic inspired us to investigate if the proposed approach could be used as fragile watermarking for tampering detection. In this paper, we developed the proposed approach for speech tampering detection.

The rest of this paper is organized as follows. Section 2 talks about the proposed tampering detection scheme based on the proposed watermarking approach, including watermark embedding process, detection process, and the identification of tampering. Subsequently, in Section 3, evaluations concerning inaudibility, robustness, and fragility are carried out. In the last section, we give a summary of this paper.

2. Scheme of Tampering Detection

The overall scheme for tampering detection consists of three main parts: embedding, detection, and tampering identification. Figure 1 illustrates a block diagram of this scheme. The original signal $x(n)$ and watermarks $s(m)$ are used to construct the watermarked signal $y(n)$. Whether the received signal has been tampered with or not can be inferred from the detected watermarks, $\hat{s}(m)$, with nonblind or blind detection.

2.1. Embedding Process. Figure 1(a) shows the block diagram of embedding process. The main process can be divided into frame segmentation, linear prediction (LP) analysis, parameter (LSFs) extraction, watermark embedding, LP synthesis, and frame connection. We select line spectral frequencies (LSFs) as the carrier of watermarks since LSFs that converted from LP coefficients are less sensitive to noise. We embed watermarks to LSFs as follows. (i) The original signal, $x(n)$, is first segmented into nonoverlapping frames, and the frame number is indexed by "m." (ii) Each frame is analysed by p-th order LP analysis, and then we can extract the LP coefficients, a_{mk} $(k = 1, 2, \ldots, p)$, and LP residue, r_m. (iii) The LP coefficients, a_{mk} $(k = 1, 2, \ldots, p)$, within one frame are converted to LSFs, f_{mk} $(k = 1, 2, \ldots, p)$. The obtained LSFs, f_{mk} $(k = 1, 2, \ldots, p)$, are expressed in the angle domain. All LSFs within one frame satisfy the ordering property from 0 to π as $0 < f_{m1} < f_{m2} < \cdots < f_{mp} < \pi$. (iv) Current watermark $s(m)$ ($s(m) =$ "0" or "1") for frame m is first duplicated to p times for all the LSFs in current frame, and then all LSFs are quantized with one of the DM-QIM quantizers Q_0 and Q_1 in equations (1)–(3), depending on the value of $s(m)$:

$$f_{mkw} = Q_w(f_{mk}), \quad w = 0 \text{ or } 1, k = 1, 2, \ldots, p, \tag{1}$$

$$Q_0(f) = \Delta\left[\frac{f - b_0}{\Delta}\right] + b_0, \quad b_0 = -\frac{\Delta}{4}, \tag{2}$$

$$Q_1(f) = \Delta\left[\frac{f - b_1}{\Delta}\right] + b_1, \quad b_1 = \frac{\Delta}{4}, \tag{3}$$

where f_{mk} in equation (1) is the LSF to be quantized, f_{mkw} is the quantized value of f_{mk} after using Q_w ($w =$ "0" or "1"), the Δ in equations (2) and (3) is the quantization step, "$[\cdot]$" stands for the rounding function, and b_i ($i = 0$ and 1) denotes the dither vectors corresponding to Q_0 and Q_1 to embed "0" and "1." (v) The modified LSFs, f_{mkw}, are converted back to LP coefficients, a_{mkw}. (vi) Current frame is then synthesized with LP coefficients, a_{mkw} (obtained in step (v)) and the residue r_m (obtained in step (ii)). (vii) The whole watermarked signal $y(n)$ is finally reconstructed with all watermarked frames using nonoverlapping and adding function.

Figure 2 demonstrates an example of watermark embedding using quantization step 2.0°. The positions of LSFs on the half unit circle can reflect the formants' information of the speech signal. The LP order, p, was ten; thus, ten LSFs were calculated. In this case, five formants were estimated. Watermark "1" was embedded into each LSF with equations (1) and (3). Since the quantization step of 2.0° was small, the original LSFs as well as positions of formants just slightly shifted from their previous positions. Accordingly, the sound quality was not seriously distorted.

2.2. Detection Process. Figure 1(b) outlines two schemes of detection: (i) nonblind detection (top side) in which both original signal $x(n)$ and watermark signal $y(n)$ are available and (ii) blind detection (bottom side), where watermarks should be detected without using the original signal $x(n)$.

2.2.1. Nonblind Detection. The detailed procedures for (i) nonblind detection in Figure 1(b) involve five steps. (i) First, the

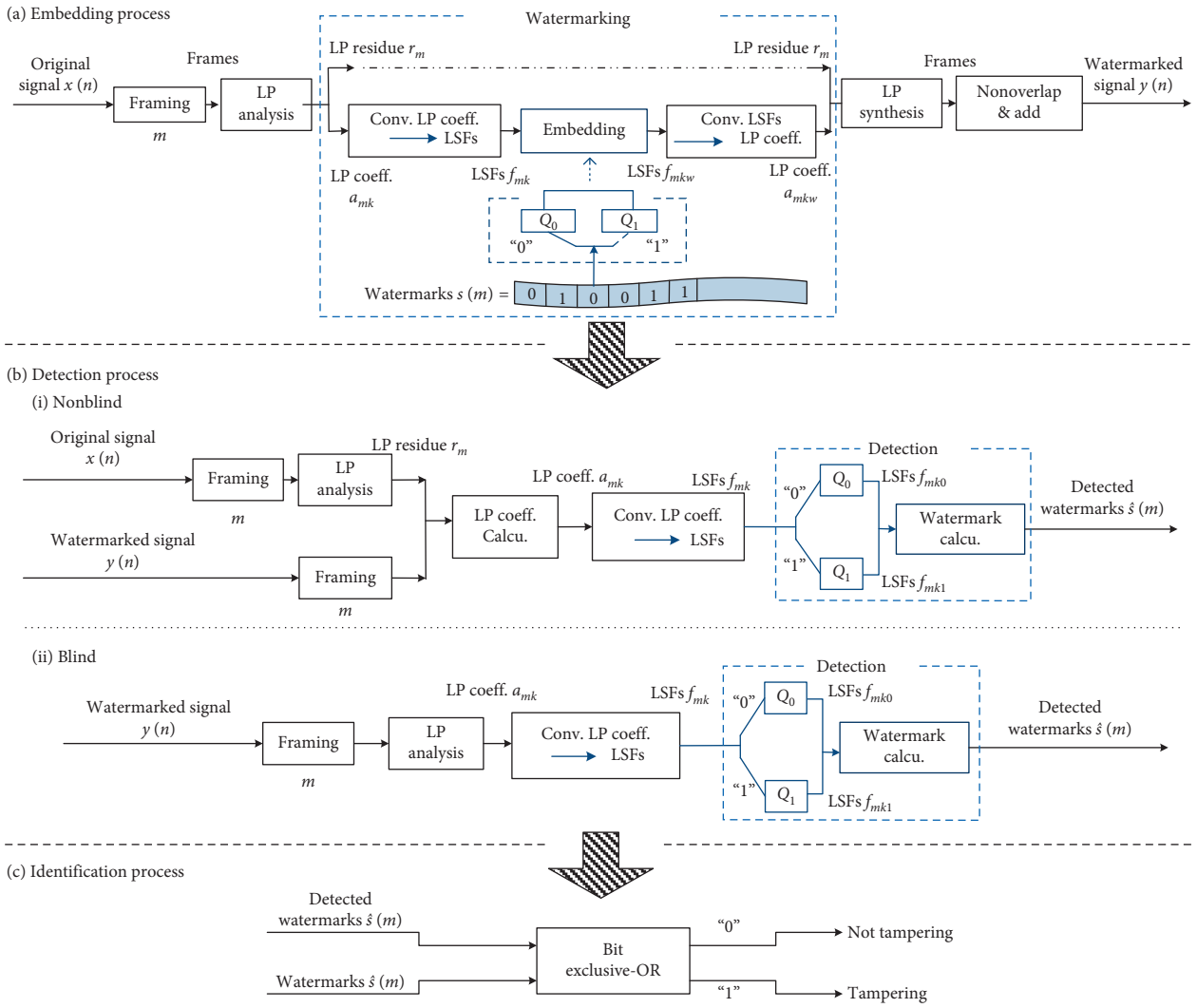

FIGURE 1: Block diagram of the proposed approach: (a) embedding, (b) detection ((i) nonblind and (ii) blind), and (c) tampering identification.

original signal, $x(n)$, and watermarked signal, $y(n)$, are segmented into nonoverlapping frames, where the frame number is indexed by "m." (ii) A p-th LP analysis is applied to the frames of $x(n)$ to obtain LP residue, r_m. (iii) The LP coefficients, a_{mk} ($k = 1, 2, \ldots, p$), can be calculated using LP residue r_m of original frame and the current watermarked frame. (iv) The LP coefficients, a_{mk} ($k = 1, 2, \ldots, p$), are converted to LSFs, f_{mk} ($k = 1, 2, \ldots, p$). Since we embed the same bits to all LSFs of one frame in the embedding process, there exists a possibility that not all the LSFs can be correctly detected. Thus, we use majority decision to decide the embedded bit. According to the block diagram in Figure 3, each LSF within one frame is requantized with both quantizers in equation (4). We calculate the distances, d_{mkw} ($k = 1, 2, \ldots, p$, $w =$ "0" and "1"), between two quantized results, f_{mkw} ($w =$ "0" and "1"), and the obtained LSF, f_{mk} ($k = 1, 2, \ldots, p$), using equation (5). Each LSF can indicate one embedded bit ("0" or "1") using the quantizer that provides a shorter distance using equation (6). We sum up the value of all detected bits to L with equation (7), and the final decision on the embedded bit of current frame is obtained by comparing the value of L and $p/2$ with equation (8):

$$f_{mkw} = Q_w(f_{mk}), \quad w = 0 \text{ and } 1, k = 1, 2, \ldots, p, \tag{4}$$

$$\Delta d_{mkw} = f_{mkw} - f_{mk}, \quad w = 0 \text{ and } 1, k = 1, 2, \ldots, p, \tag{5}$$

$$d_{mk} = \begin{cases} 1, & \Delta d_{mk1} < \Delta d_{mk0}, \\ 0, & \text{otherwise}, \end{cases} \tag{6}$$

$$L = \sum_{k=1}^{p} d_{mk}, \quad k = 1, 2, \ldots, p, \tag{7}$$

$$\hat{s}(m) = \begin{cases} 0, & L < \dfrac{p}{2}, \\ 1, & \text{otherwise}. \end{cases} \tag{8}$$

2.2.2. Blind Detection. Figure 1(b) has the diagram of (ii) blind detection. Most detection procedures of the blind method are similar to those of the nonblind detection. One

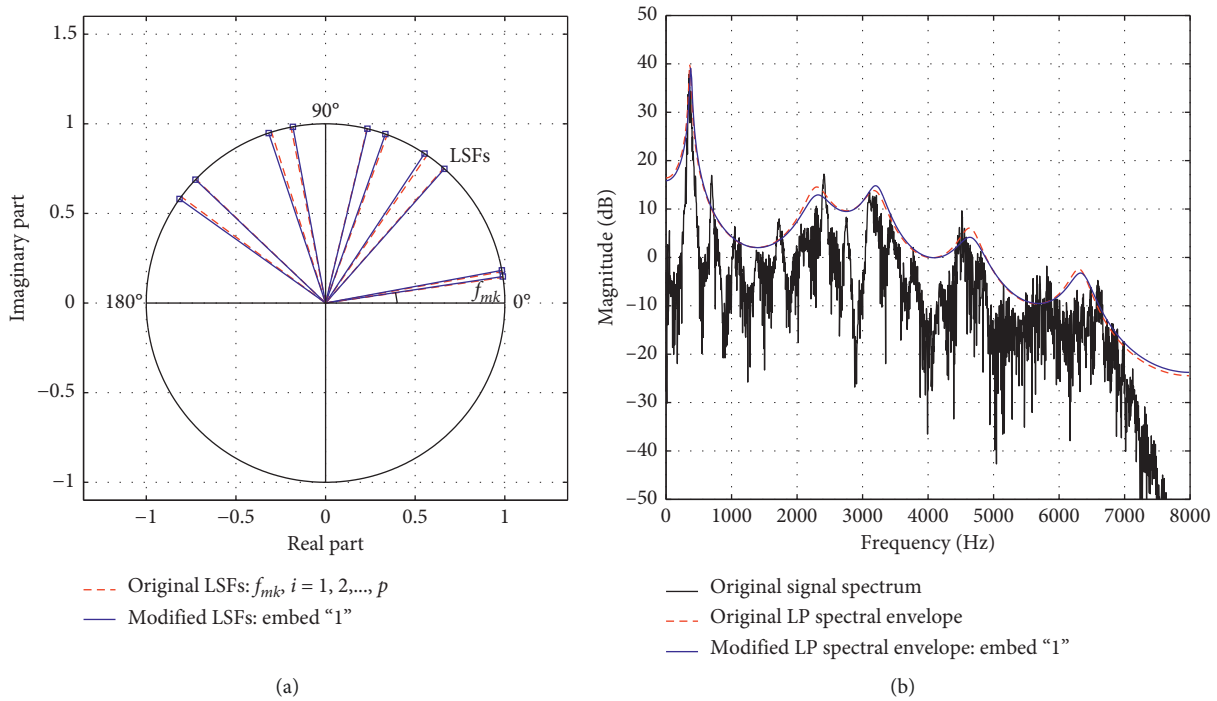

Figure 2: (a) Distribution of LSFs on half unit circle before and after DM-QIM modifications and (b) LP spectral envelope before and after modifications (frame size: 250 ms, quantization step: 2.0°, and original LSFs: f_{mk}, $k = 1, 2, \ldots, p$).

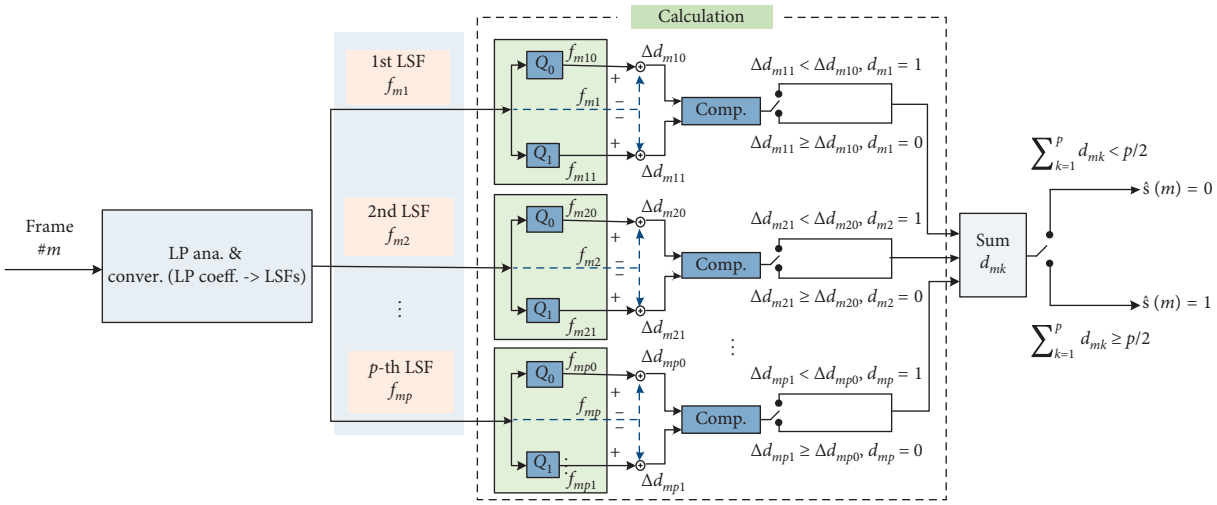

Figure 3: Block diagram of nonblind detection with majority decision.

difference between them is on how the LSFs of one frame are obtained. In blind detection, the LP coefficients, a_{mk} ($k = 1, 2, \ldots, p$), are directly calculated from each watermarked frame. The LSFs, f_{mk} ($k = 1, 2, \ldots, p$), are converted from these LP coefficients. Embedded bit is calculated using the same equations (4)–(8) as in the blind detection.

2.3. Tampering Identification. Using the above nonblind and blind detection, we can detect watermarks, $\hat{s}(m)$, from watermarked signal, $y(n)$. As to verify if there is tampering happens to the watermarked signal, $y(n)$, before it is received at the receiver side, we should compare the original watermarks, $s(m)$, and the detected watermark, $\hat{s}(m)$, and find the mismatches. According to Figure 1(c), this can be simply figured out by bit exclusive-OR. If there are no mismatches, the received signal, $y(n)$, is the original signal, $x(n)$, with no tampering occurred; otherwise, each mismatch indicates the possible tampering of the corresponding frame.

3. Evaluations

We carried out three experiments with respect to (1) inaudibility, (2) robustness against codecs and meaningful processing, and (3) fragility against tampering, to evaluate

the performance of the proposed approach [8]. The ATR dataset (B set) consisting of 12 speech stimuli (Japanese sentences, 20 kHz, 16 bits) [24] was used to evaluate the proposed method. This dataset is also widely used to investigate the speaker properties, e.g., speaker individuality and the acoustic/phonetic features. Therefore, it is quite suitable to evaluate the tampering detection performance of the proposed method. Each stimulus was clipped to 8.1 sec duration and embedded with watermarks at different bit rates. The bit rates in our experiments were set to 4, 8, 16, 32, 64, 128, 256, 512, and 1024. The embedded watermark was a 122×77 bitmap image in Figure 4.

The LP order is important for the performance of the proposed method. High LP order is beneficial to follow the details of the spectral contour while low LP order can provide global frequency information only. Under low-order LP analysis, each LSF carries more information compared with those under high-order LP analysis. As a result, the sound distortion brought by quantizing LSFs of low-order LP analysis will be severe. On the other hand, most processing will bring distortions to the watermarked signal; if LP order is so high to follow all the spectral details, any distortion will disturb the LSF deviation, which obstructs the watermark detection. In this case, LP order should be low to achieve robustness. According to the above analysis, we selected suitable LP order for the proposed method based on preliminary experiments. The LP order was finalized as 10 to balance inaudibility and robustness performance.

The quantization step in QIM also affects the performance of the proposed method, which results in a trade-off among the conflicting requirements of inaudibility and robustness. A small quantization step provides better sound quality of the watermarked signal; however, the robustness will be degraded. In this work, we chose 1.0° as a suitable quantization step to achieve good balance between inaudibility and robustness.

3.1. Evaluations for Inaudibility. The log-spectrum distortion (LSD) [25] and perceptual evaluation of speech quality (PESQ) [26] were adopted to check the inaudibility of the proposed methods. The LSD is distance measure (in decibel (dB)) of the two spectra between the original signal and watermarked signal. LSD of 1 dB was usually chosen as the criterion, and a lower value indicated less distortion. The PESQ recommended by ITU-T recommendation P.862 is a family of standards for automated assessment of the speech quality. The results of PESQ are Objective Difference Grades (ODGs), which are graded from −0.5 (very annoying) to 4.5 (imperceptible), corresponding to Mean Opinion Score (MOS) values. The ODG of 3.0 (slightly annoying) was set as the criterion, and a higher value indicated better quality.

Figure 5 shows an example of embedding watermark "1" into one frame of the original signal using quantization step of 1.0°. The waveforms and the spectra of the original signal and the watermarked signal are shown in the top two panels, and the differences between them are shown in the bottom panel. One can see that the differences between the original signal and the watermarked signal in both the time domain and frequency domain were negligible, which indicated that the proposed method introduced almost imperceptible distortion to the human auditory system.

Figure 4: Original message of 122×77 bitmap image.

The objective evaluation results are provided by LSD (Figure 6(a)) and PESQ (Figure 6(b)). The straight blue dashed lines in each subfigure indicated the criteria for LSD (≤ 1 dB) and PESQ (≥ 3.0 ODG). Since the embedding processes of the nonblind and blind detection were the same, we got the same evaluation results for LSD and PESQ. As we can see, sound quality got worse when bit rate increased. Nevertheless, for all bit rates from 4 bps to 1024 bps, the watermarked signals could satisfy the criteria for both LSD and PESQ. These results indicated that for the quantization steps of 1.0°, the proposed approach (with nonblind and blind detection) could satisfy inaudibility for all bit rates.

3.2. Evaluations for Robustness. Robustness of the proposed approach was evaluated from two aspects: (a) robustness against different speech codecs and (b) robustness against general processing. We adopted bit detection rate to measure the robustness, and a higher bit detection rate suggests a better robustness. The calculation of bit detection rate is defined in equation (9), where $s(m)$ represents embedded watermarks, $\hat{s}(m)$ represents detected watermarks, and W is the length of watermarks. The symbol "$\oplus$" denotes the operation of "exclusive-OR," that is, if the bit values of $s(m)$ and $\hat{s}(m)$ are the same ($s(m) = 1$ and $\hat{s}(m) = 1$, or $s(m) = 0$ and $\hat{s}(m) = 0$), "$s(m) \oplus \hat{s}(m)$" equals 0; otherwise, "$s(m) \oplus \hat{s}(m)$" equals 1:

$$R = \frac{\sum_{m=1}^{W} s(m) \oplus \hat{s}(m)}{W} \times 100\%. \quad (9)$$

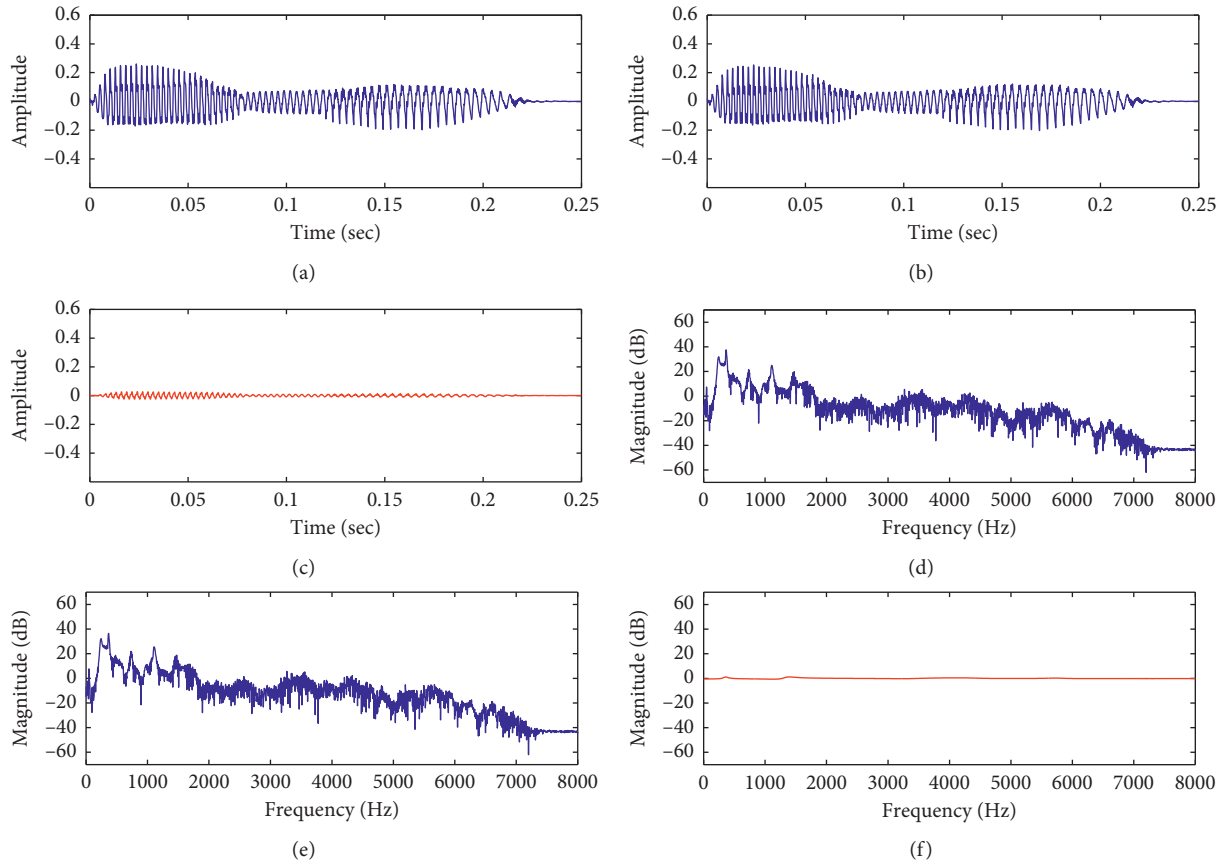

Figure 5: Waveform and spectra differences between the original signal and the watermarked signal (LP order: 10th; quantization step: 1.0°). (a) Waveform of the original signal, (b) waveform of the watermarked signal: embed "1," (c) difference in waveform, (d) spectrum of the original signal, (e) spectrum of the watermarked signal: embed "1," and (f) difference in spectrum.

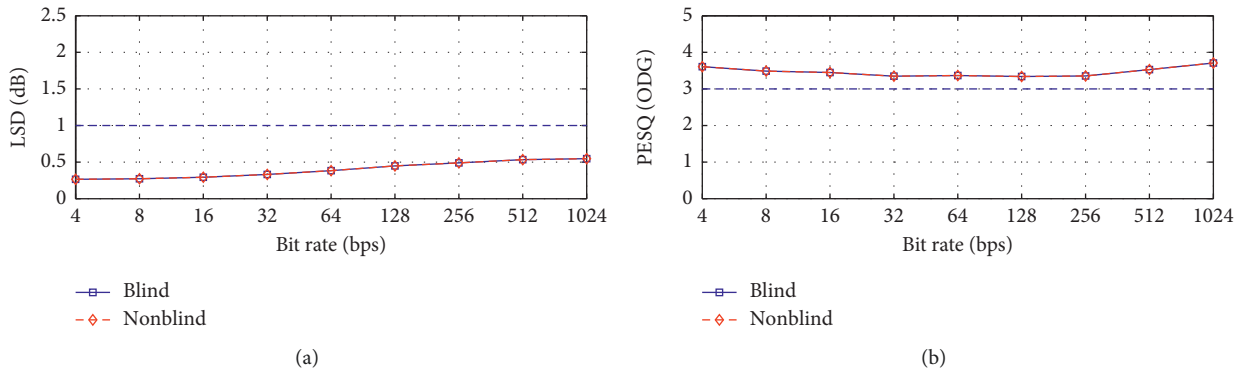

Figure 6: Inaudibility performance of the proposed approach measured by (a) LSD and (b) PESQ.

3.2.1. Robustness against Speech Codecs. In general, speech codes can be classified into waveform-based and parameter-based schemes. Watermarking methods are thus required to satisfy both kinds of speech codecs. We chose two typical speech codecs G.711 (waveform-based) and G.729 (parameter-based) to evaluate the robustness of the proposed approach.

Figure 7 presents bit detection results for normal detection without any modifications (Figure 7(a)), detection after G.711 (Figure 7(b)), and detection after G.729 (Figure 7(c)). The straight blue dashed lines in each subfigure indicated the criteria for bit detection rate of 90%. As we can see from Figure 7(a), the nonblind approach had almost 100% bit detection rates for all bit rates, while for the blind detection, the bit detection rates were a little lower. For bit detection after G.711 and G.729, the nonblind approach in Figures 7(b) and 7(c) provided very good results, indicating that it had very good robustness

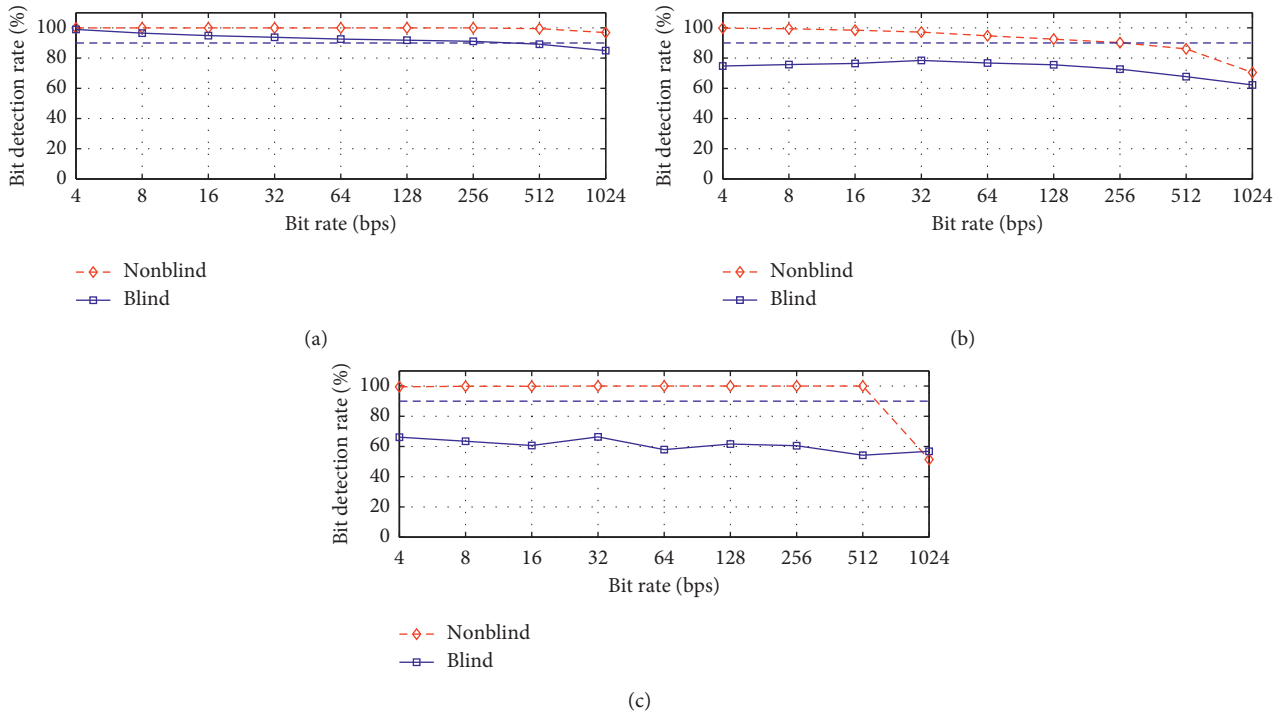

Figure 7: Robustness of the proposed approach under (a) normal, (b) G.711, and (c) G.729.

against these speech codecs. However, the bit detection rates dramatically reduced in blind detection approach (Figure 7).

We investigated the reason why there are such big differences in bit detection between the nonblind approach and the blind approach. In the nonblind approach, the original LP residue can be used to calculate the LSFs in detection process. Thus, the obtained LSFs in this approach were almost the same with those in the embedding process, which facilitates the watermark detection. However, for blind approach, the LSFs are directly derived from the watermarked signal without the assistance of the original LP residue. As we know, the LP analysis calculates LP coefficients (LSFs) based on the squared error criterion, and thus, the LSFs derived directly from the watermarked signal are different from those modified LSFs in the embedding process. As a result, the watermarks cannot be detected accurately (Figure 8).

3.2.2. Robustness against General Processing. We carried out robustness evaluation of the proposed approaches against several meaningful processing, which are listed below:

(1) Scaling by 0.5 and 2.0

(2) Resampling at 12 kHz and 24 kHz

(3) Requantization with 8 bits and 24 bits

(4) Spectrum modification by short-time Fourier transform (STFT)

The 16×16 bitmap image (i.e., watermarks) in Figure 9(a) was embedded to the original signal at 4 bps. In this case, each embedded bit was able to account for 0.25 s speech segment when locating the tampering. In fact, 0.25 s is too short to make a meaningful tampering of speech content. Therefore, embedding bit rate of 4 bps is able to locate the tampering in time domain at sufficient precision in practical. We processed the middle segment of watermarked signals with the above processing and detected the embedded watermarks from watermarked signals. Watermarks should be correctly detected if the watermarking method is robust against these processing.

Figure 9 illustrates all the nonblind and blind detection results. The nonblind detection had better performance than the blind detection. The bit detection rate for each subfigure has been listed in Table 1. For the nonblind detection, the proposed approach had very good robustness against general speech processing, since almost all the bit detection rates were over 90% except for requantization with 8 bits. For the blind detection, the bit detection rates were slightly lower, as shown in Figure 9(b). The bit detection rates under, e.g., Figures 9(c), 9(d), 9(f), 9(h), and 9(i), were still satisfactory. Nevertheless, the proposed blind approach was not robust again resampling at 12 kHz and requantization with 8 bits. The main reason for this was that the resampling or requantization at lower rate introduced distortions to the watermarked signals.

3.3. Evaluations for Fragility. Similar to robustness evaluation, we embedded the same 16×16 bitmap image in fragility evaluation. We manually modified the middle segment of the watermarked signals with malicious tampering listed below and then checked whether the embedded watermarks were destroyed:

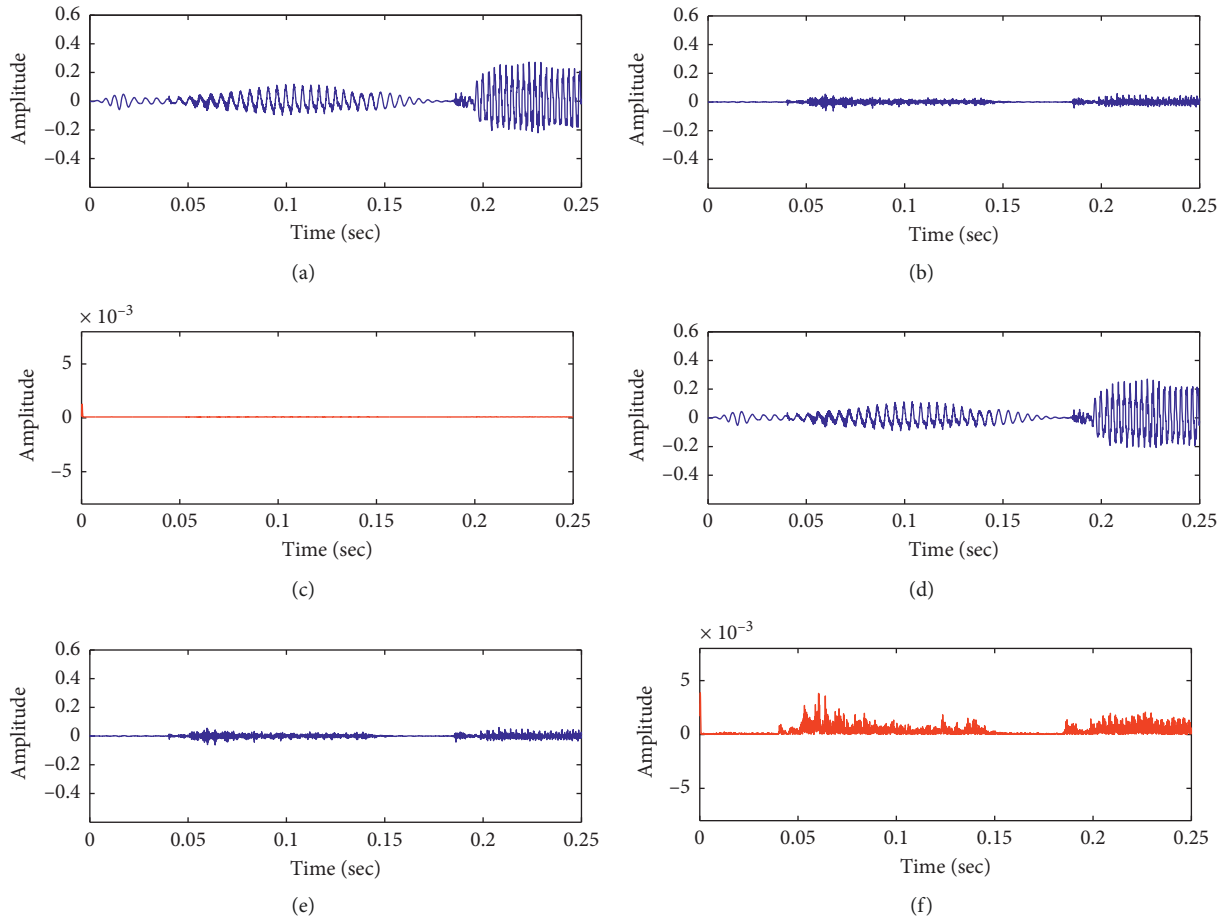

Figure 8: Comparison of LP residues between the nonblind approach and the blind approach (LP order: 10th ; quantization step: 1.0°). (a) Waveform of the original signal, (b) residue of the original signal, (c) residue difference (original vs. nonblind), (d) waveform of the watermarked signal: embed "1," (e) residue of the watermarked signal: embed "1," and (f) residue difference (original vs. blind).

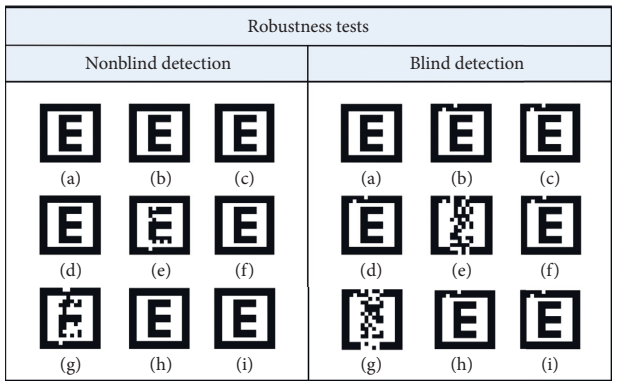

Figure 9: Robustness evaluation of the nonblind approach (left side) and the blind approach (right side): (a) embedded watermarks (a 16 × 16 binary image) and detected watermarks under (b) no modification, (c) scaling by 0.5, (d) scaling by 2.0, (e) resampling at 12 kHz, (f) resampling at 24 kHz, (g) requantization with 8 bits, (h) requantization with 24 bits, and (i) STFT.

Table 1: Bit detection rates for robustness tests.

Processing	Bit detection rates (%)	
	Nonblind detection	Blind detection
No modifications	100.00	98.75
Scaling by 0.5	100.00	98.75
Scaling by 2.0	100.00	98.75
Resampling at 12 kHz	91.25	45.00
Resampling at 24 kHz	100.00	97.50
Requantization with 8 bits	80.00	51.25
Requantization with 24 bits	100.00	98.75
STFT	100.00	98.75

(1) Modifying temporal information by gammatone filterbank (GTFB)

(2) Adding white noise

(3) Reverberation (time: 0.3 sec)

(4) Filtering with low-pass filter (order: 32nd; normalized cutoff frequency: 0.99)

(5) Filtering with high-pass filter (order: 32nd; normalized cutoff frequency: 0.01)

(6) Concatenation with original speech.

The detected images are shown in Figure 10, and we calculated the bit detection rate of each image (Table 2). It is found that the bit detection rates after tampering were

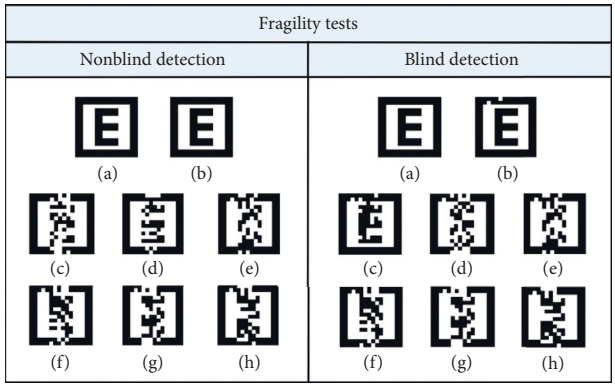

Figure 10: Fragility evaluation of the nonblind approach (left side) and the blind approach (right side): (a) embedded watermarks (a 16×16 binary image) and detected watermarks under (b) no modifications, (c) GTFB, (d) adding white noise, (e) reverberation, (f) filtering with low-pass filter, (g) filtering with high-pass filter, and (h) concatenation.

Table 2: Bit detection rates for fragility tests.

Processing	Bit detection rates (%)	
	Nonblind detection	Blind detection
No modifications	100.00	98.75
GTFT	55.00	86.25
White noise	68.75	47.50
Reverberation	63.73	46.25
Low-pass filter	43.75	58.75
High-pass filter	41.25	43.75
Concatenation	37.50	50.00

dramatically reduced. These results suggested that the proposed approach with nonblind detection and blind detection was fragile against these tampering. Therefore, it was easy to identify tampering with such results.

4. Conclusions

This paper proposed a watermarking-based tampering detection approach for speech signals. Watermarks are embedded by modifying the line spectral frequencies (LSFs) using dither modulation-quantization index modulation (DM-QIM). We evaluated the proposed approach by carrying out three objective evaluations, i.e., inaudibility, robustness, and fragility. The evaluations results suggested that the proposed approach could satisfy inaudibility and provided good robustness. Furthermore, it was also fragile against malicious tampering. Therefore, it is effective for speech tampering detection.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported by the Scientific Research Project of Tianjin Education Commission (no. 2017KJ089). The authors are grateful for the financial support.

References

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, 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 Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[3] Y. Zhao, M. Kuruvilla-Dugdale, and M. Song, "Structured sparse spectral transforms and structural measures for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2267–2276, 2018.

[4] H. Kawahara, H. Banno, T. Irino, and P. Zolfaghari, "Algorithm amalgam: morphing waveform based methods, sinusoidal models and STRAIGHT," in *Proceedings of the 29th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 13–16, Montreal, Canada, May 2004.

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

[6] W. Mazurczyk and S. Wendzel, "Information hiding: challenges for forensic experts," *Communications of the ACM*, vol. 61, no. 1, pp. 86–94, 2018.

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

[8] G. Hua, J. Huang, Y. Q. Shi, J. Goh, V. L. L. Thing, and L. Thing, "Twenty years of digital audio watermarking-a comprehensive review," *Signal Processing*, vol. 128, pp. 222–242, 2016.

[9] S. Wang, M. Unoki, and N. S. Kim, "Formant enhancement based speech watermarking for tampering detection," in *Proceedings of the INTERSPEECH 2014*, pp. 1366–1370, Singapore, September 2014.

[10] S. Wang, W. Yuan, J. Wang, and M. Unoki, "Speech watermarking based on robust principal component analysis and formant manipulations," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2082–2086, Calgary, Canada, April 2018.

[11] J. Li, W. Lu, C. Zhang, J. Wei, X. Cao, and J. Dang, "A study on detection and recovery of speech signal tampering," in *Proceedings of the 2016 IEEE Trustcom/BigDataSE/ISPA*, pp. 678–682, Tianjin, China, August 2016.

[12] Z. Liu, F. Zhang, J. Wang, H. Wang, and J. Huang, "Authentication and recovery algorithm for speech signal based on digital watermarking," *Signal Processing*, vol. 123, pp. 157–166, 2016.

[13] S. Sarreshtedari, M. Ali Akhaee, and A. Abbasfar, "A watermarking method for digital speech self-recovery," *IEEE/*

ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1917–1925, 2015.

[14] M. Celik, G. Sharma, and A. Murat Tekalp, "Pitch and duration modification for speech watermarking," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005(ICASSP '05)*, vol. II, pp. 17–20, Philadelphia, PA, USA, March 2005.

[15] J. Karnjana, K. Galajit, P. Aimmanee, W. Chai, and M. Unoki, "Speech watermarking scheme based on singular-spectrum analysis for tampering detection and identification," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2017)*, pp. 193–202, Kuala Lumpur, Malaysia, December 2017.

[16] K. Galajit, J. Karnjana, M. Unoki, M. Intarauksorn, and P. Aimmanee, "Speech watermarking technique based on singular spectrum analysis and automatic parameter estimation using differential evolution for tampering detection," in *Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, Pattaya, Thailand, November 2018.

[17] C.-P. Wu and C.-C. Jay Kuo, "Fragile speech watermarking based on exponential scale quantization for tamper detection," in *Proceedings of the 27th IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3305–3308, Orlando, FL, USA, May 2002.

[18] M. Narimannejad and A. S. Mohammad, "Watermarking of speech signal through phase quantization of sinusoidal model," in *Proceedings of the 19th Iranian Conference on Electrical Engineering*, pp. 1–4, Tehran, Iran, May 2011.

[19] M. Unoki and D. Hamada, "Method of digital-audio watermarking based on cochlear delay characteristics," *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 3(B), pp. 1325–1346, 2010.

[20] M. Unoki and R. Miyauchi, "Detection of tampering in speech signals with inaudible watermarking technique," in *Proceedings of the 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 118–121, Piraeus-Athens, Greece, July 2012.

[21] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, no. 1, pp. 35–55, 1975.

[22] S. Wang and M. Unoki, "Watermarking method for speech signals based on modifications to LSFs," in *Proceedings of the 9th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2013)*, pp. 283–286, Beijing, China, October 2013.

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

[24] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR Technical Report TR-I-0028, 1988.

[25] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[26] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.