LETTER *Special Section on Enriched Multimedia—Application of Multimedia Technology and Its Security—*

Non-Blind Speech Watermarking Method Based on Spread-Spectrum Using Linear Prediction Residue

Reiya NAMIKAWA[†], ***Nonmember and*** **Masashi UNOKI**[†a)], ***Senior Member***

SUMMARY We propose a method of non-blind speech watermarking based on direct spread spectrum (DSS) using a linear prediction scheme to solve sound distortion due to spread spectrum. Results of evaluation simulations revealed that the proposed method had much lower sound-quality distortion than the DSS method while having almost the same bit error ratios (BERs) against various attacks as the DSS method.

key words: *speech watermarking, spectrum spreading, linear prediction residue, pseudo-random noise signal*

1. Introduction

The rapid development of advanced information communication technology (ICT) has positively impacted digital speech processing. However, misuse of ICT causes problems, such as speech tampering and speech spoofing in digital speech communication. Therefore, audio information hiding (AIH) techniques have recently been focused on as state-of-the-art techniques enabling secure and safety communication [1], [2]. To protect digital audio content, AIH techniques aim at embedding codes as watermarks, which are inaudible and inseparable by users, and at robustly detecting embedded codes from watermarked signals against any kind of processing and various attacks. Therefore, AIH techniques generally must satisfy two important requirements: inaudibility and robustness.

The direct spread spectrum (DSS) method is a robust AIH method [3]. It spreads the message signal by a pseudo-random noise (PN) signal and adds it to the host signal. This method can robustly detect the embedded messages from the watermarked signal against various signal processing. However, the principle of the DSS method has a critical problem: spreading of the spectrum reduces the sound quality of the watermarked signal without adequately controlling spreading level under masking conditions [2], [3].

This paper proposes a non-blind speech watermarking method that can satisfy inaudibility and robustness simultaneously. A linear prediction (LP) scheme for speech analysis and synthesis is used in the proposed method. On the basis of the principle of the DSS method, the proposed method spectrally spreads a message by using LP residue and embeds the spread spectrum of the message into the host signal. The use of LP residue is a novel idea to solve the above issue.

Manuscript received April 1, 2019.
Manuscript revised August 9, 2019.
Manuscript publicized October 23, 2019.

[†]The authors are with the School of Information Science, Japan Advanced Institue of Science and Technlogy, Nomi-shi, 923–1292 Japan.
a) E-mail: unoki@jaist.ac.jp

DOI: 10.1587/transinf.2019MUL0003

2. Direct Spread Spectrum Method

In the DSS method, it is assumed that the frame-based processing is used to synchronized the message $m(n)$. The watermark signal $m(n)c(n)$ is generated by spread-spectrum modulating the $m(n)$ with the PN signal $c(n)$. A watermarked signal $y(n)$ is generated by adding the $m(n)c(n)$ to the host signal $x(n)$ as follows.

$$y(n) = x(n) + am(n)c(n), \tag{1}$$

where a is a correction parameter for adjusting the embedding strength level of the watermark signal.

The $m(n)$ can be detected in each frame by using the following equation.

$$m(n) = \begin{cases} 0, & E\{y(n)c(n)\} \le 0, \\ 1, & E\{y(n)c(n)\} > 0, \end{cases} \tag{2}$$

where $E\{\cdot\}$ is the expected value of "$\cdot$". Since $x(n)$, $y(n)$, and $c(n)$ are assumed to be ergodicity, $E\{\cdot\}$ can be regarded as the temporal average of "$\cdot$". In addition, note that the PN signal has properties of $E\{c(n)\} = 0$ and $E\{c^2(n)\} = 1$. Hence, a mean of multiplying the watermarked signal $y(n)$ with the same PN signal $c(n)$, $E\{y(n)c(n)\}$, can be used to extract message "0" or "1" from $y(n)$ by a sign or negative of them, as follows.

$$\begin{aligned} E\{y(n)c(n)\} &= E\{[x(n) + am(n)c(n)]c(n)\} \\ &= E\{x(n)c(n)\} + E\{am(n)c^2(n)\} \\ &= am(n). \end{aligned} \tag{3}$$

Since $x(n)$ and $c(n)$ are mutually independent, it was found that the first term in Eq. (3) becomes 0 and the second term in Eq. (3) becomes $E\{am(n)c^2(n)\} = am(n)$ in which $E\{c^2(n)\} = 1$. From these results, $m(n)$ can be correctly detected by using Eq. (3).

Figure 1 shows the block diagram of the DSS method. Embedding and detection processes in the DSS method are shown in Figs. 1 (a) and 1 (b). In Fig. 1 (a), the $x(n)$ is segmented out N-frames by framing with the rectangular window function, where N is the payload in bps. Then $m(n)$ is watermarked by spectral-spreading with $c(n)$ with the embedding strength level a. In Fig. 1 (b), the $y(n)$ is also segmented out to N-frames by the same framing processing. The fast Fourier transform (FFT) is used to detect the sign of

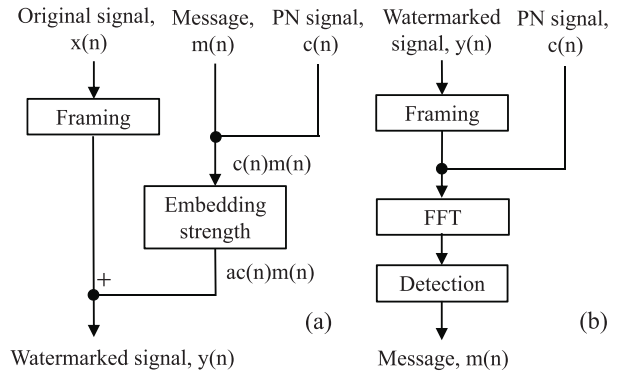

Fig. 1 Block diagram of DSS method: (a) embedding and (b) detection.

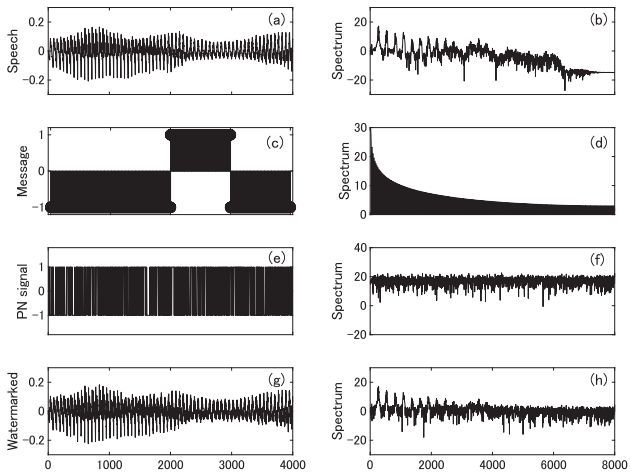

Fig. 2 Example of speech watermarking by DSS method: (a) host signal, (b) amplitude spectrum of host signal, (c) message, (d) amplitude spectrum of message, (e) PN signal, (f) amplitude spectrum of PN signal, (g) watermarked signal, and (h) amplitude spectrum of watermarked signal.

$E\{y(n)c(n)\}$ in each frame and then $m(n)$ is calculated from the sign by using Eqs. (2) and (3).

Figure 2 shows examples of speech watermarking by using the DSS method. Figure 2 (a) shows the host signal (one of speech signals in the ATR database (B set) [5]) where the watermark signal will be embedded. The watermark signal is obtained by multiplying the message of "0010" in Fig. 2 (c) and the PN signal in Fig. 2 (e). In this method, frame-based processing is used to segment out N-frames. The sampling frequency is 16 kHz and the payload is 16 bps. Thus, the frame length is 1000 samples (62.5 ms) and there are the four frames in Fig. 2 (c). The watermarked signal in Fig. 2 (g) is obtained by adding the watermark signal addition into the host signal $x(n)$. In this case, the embedding strength level is −35 dB so that parameter a is $10^{-35/10}$. The right panels in Fig. 2 show the respective amplitude spectra of the signals at the left panels.

The PN signal has a white spectrum as shown in Fig. 2 (f), so the watermarked signal $y(n)$ has been spread over a wide band as shown in Fig. 2 (h). Since this feature distorts the $y(n)$, the DSS method has a problem of inaudibility due to the spread spectrum.

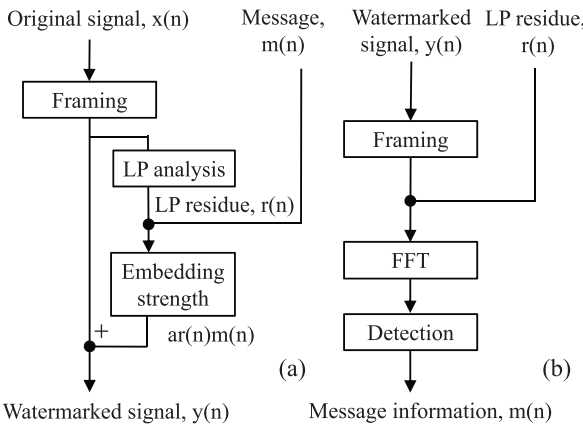

Fig. 3 Block diagram of proposed method: (a) embedding and (b) detection.

3. Proposed Method

Linear predictive coding (LPC) is the most basic speech coding method using linear prediction (LP). LPC provides an LP coefficient corresponding to the spectral envelope and an LP residue corresponding to the sound source of the speech signal. LP residue has properties of $E\{c(n)\} = 0$ and $E\{c^2(n)\} = 1$. By using these properties, a DSS method with the LP residue was examined to propose a speech watermarking method.

In the proposed method, the message $m(n)$ is spectrally spread by the LP residue $r(n)$ instead of the PN signal $c(n)$, because the LP residue has similar properties to the PN signal. Figure 3 shows a block diagram of the proposed method. Embedding and detection processes are shown in Figs. 3 (a) and 3 (b).

Embedding process: In Fig. 3 (a), the $x(n)$ is segmented out to N-frames by framing with the rectangular window function, where N is the payload in bps. In each frame, the watermark signal $m(n)r(n)$ is obtained by multiplying the $m(n)$ with the LP residue $r(n)$ which is the spread signal by the $m(n)$. The watermarked signal $y(n)$ is obtained by adding the $m(n)r(n)$ into the host signal $x(n)$ in each frame.

Detection process: In Fig. 3 (b), the $y(n)$ is also segmented out to N-segments by the same frame processing. Detection properties in the proposed method are the same as those in the DSS method as explained in Eqs. (2) and (3). Thus, FFT is used to detect the sign of $E\{y(n)r(n)\}$ in each frame and then the message $m(n)$ can be obtained by the following equation.

$$m(n) = \begin{cases} 0, & E\{y(n)r(n)\} \le 0, \\ 1, & E\{y(n)r(n)\} > 0. \end{cases} \tag{4}$$

Figure 4 shows examples of speech watermarking using by the proposed method. Figure 4 (a) shows the host signal as the same as Fig. 2 (a) where the watermark signal will embed. The watermark signal is obtained by multiplying the message of "0010" in Fig. 4 (c) with the LP residue

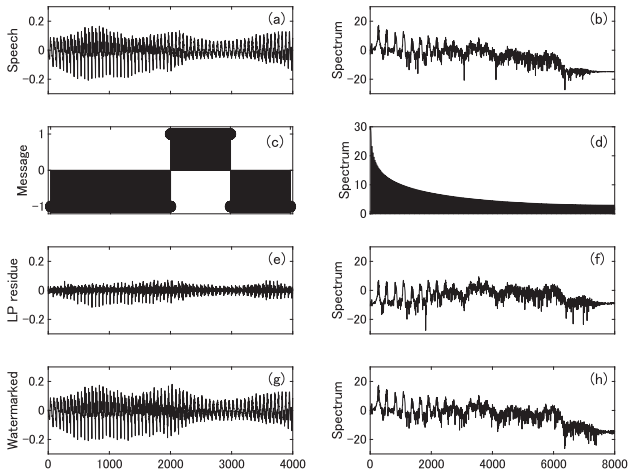

Fig. 4 Example of speech watermarking by proposed method: (a) host signal, (b) amplitude spectrum of host signal, (c) message, (d) amplitude spectrum of message, (e) LP residue, (f) amplitude spectrum of LP residue, (g) watermarked signal, and (h) amplitude spectrum of watermarked signal.

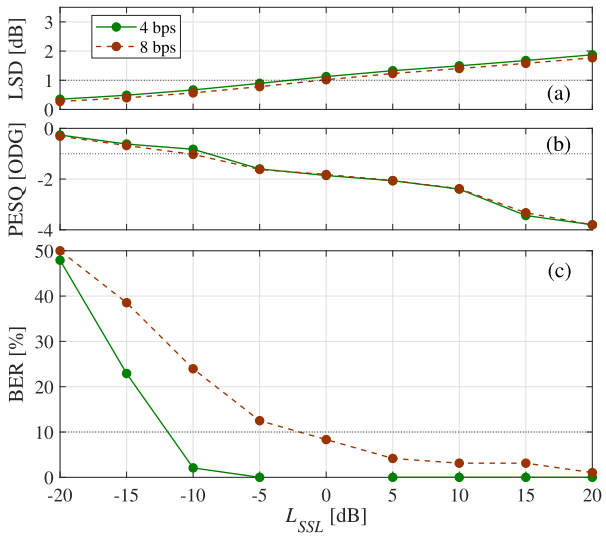

Fig. 5 Relationship between strength setting level L_{SSL}, (a) LSD, (b) PESQ, and (c) BER.

in Fig. 4 (e). In this method, frame-based LP analysis was used to segment out N-frames and then derive LP residue in each frame where LP order is 12. The sampling frequency is 16 kHz and the payload is 16. Thus, the frame length is 1000 samples (62.5 ms) and there are the four frames in Fig. 4 (c). The watermarked signal in Fig. 4 (g) is obtained by adding the watermark signal $m(n)r(n)$ into the host signal $x(n)$ in each frame. In this case, the embedding strength level is −35 dB so that parameter a is $10^{-35/10}$. The right panels in Fig. 4 show the respective amplitude spectra of the signals at the left panels.

The LP residue was used to spread the message in the host signal. The LP residue has a similar spectral shape to the host signal as shown in Fig. 4 (f), so the message was spectrally spread into a similar spectrum to the host signal. Therefore, the sound distortion due to the spread spectrum will be resolved by using the DSS with LP residue. In addition, since an advantage of the proposed method is that it has the same properties as DSS, the proposed method will have the same high robustness for speech watermarking as DSS while having inaudibility by using DSS with LP residue.

In the proposed method, parameter a can be used to control the embedding strength for speech watermarking in Eq. (1). AIH techniques generally have a trade-off between inaudibility and robustness. Thus, the embedding strength should be controlled by using parameter a to satisfy inaudibility and robustness simultaneously. In the proposed method, parameter a is determined by minimizing sound quality distortion (SQD) and the bit error rate (BER) simultaneously.

The embedding strength level, L_{all}, was derived by

$$L_a = L_{\mathrm{PHS}} - L_{\mathrm{PWS}} + L_{\mathrm{ESL}}, \tag{5}$$

where L_{PHS} is the power level of the host signal, L_{PWS} is the power level of the watermark signal, and L_{ESL} is the embedding strength level in dB. Parameter a can be determined as $a = 10^{L_a}/20$. Thus, L_a was determined by minimizing sound distortion and BER under various L_{ESL} conditions from −20 to 20 dB with 5-dB steps.

The BER and SQD tests were carried out to determine the optimal L_{SSL}. All 12 speech signals in the ATR database (B set) [5] were used in these tests. In the BER tests, BER of 10% was set as the criterion. In the SQD tests, log-spectrum distortion (LSD) was used as an objective measure [8]. Moreover, perceptual evaluation of speech quality (PESQ) was also used to objectively evaluate sound quality by outputting the objective difference grades (ODGs) [7], [8]. The ODGs were graded on a five-point scale as 0 (imperceptible), −1 (perceptible), −2 (slightly annoying), −3 (annoying), and −4 (very annoying). In these tests, LSD of 1 dB or less and PESQ of −1 ODG or higher were used as criteria for sound distortion.

Figure 5 shows evaluation results for BER and SQD (LSD and PESQ). The results show that the BER decreases as L_{ESL} increases and distortions increase as L_{ESL} increases. The L_{ESL} was determined to be −10 dB when the payload was 4 bps. At this time, the mean L_a was −20.6 dB. When the payload was 8 bps, there was no optimal L_{all}. Thus, the L_{SSL} was determined to be 0 dB, in which BER < 10%, LSD < 1 dB, and PESQ > −2 ODGs, in this payload. Hence, the mean L_a was −14.0 dB.

4. Evaluation

Two kinds of evaluations were carried out to investigate whether or not the proposed method can meet the requirements of inaudibility and robustness in comparison with the DSS method. The same speech signals in Fig. 5 [5] were used in these evaluations. Only the utterance sections in 12 speech signals were used as the stimulus. The sampling frequency is 16 kHz and the quantization of 16-bits.

The inaudibility was evaluated by using SQD (LSD and PESQ) of the watermarked signal. Robustness was

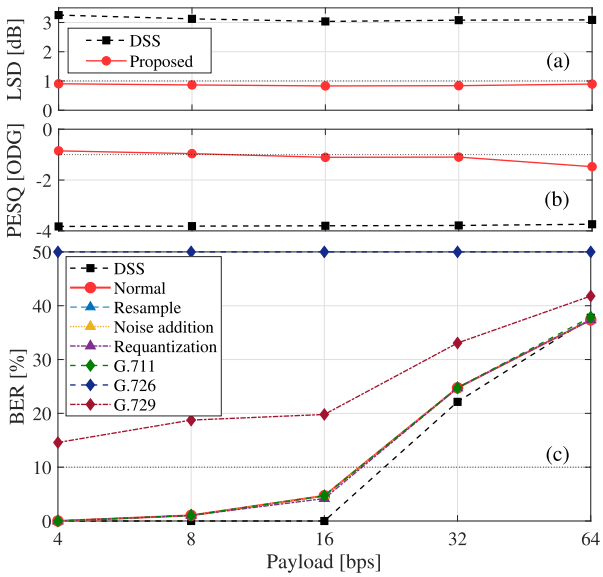

Fig. 6 Evaluation results: (a) LSD, (b) PESQ, and (c) BER (Normal, down-sampling (16 to 8 kHz), WGN (SNR 36 dB), re-quantization (16 to 8 bits), and speech coding (G.711, G.726, and G.729)).

evaluated by using BERs against various attacks such as down-sampling (16 to 8 kHz), addition of white Gaussian noise (WGN, signal-to-noise ratio (SNR) of 36 dB), re-quantization (RQZ, 16 bit quantization to 8 bit quantization), and speech coding (G.711, G.726, and G.729).

The payloads in these evaluations were 4, 8, 16, 32, and 64 bps. Messages were random bit streams. No error correction schemes were used in the previous (DSS) or proposed methods.

Figure 6 (a) shows the averaged LSD for the watermarked signals. The dotted line indicates the evaluation criteria (1 dB or lower). Under various payload conditions, the LSDs for the proposed method were under 1 dB, whereas the LSDs in the DSS method were at or above 3 dB. The proposed method thus improved LSD by over 2 dB compared with the DSS method.

Figure 6 (b) shows the averaged ODGs of the PESQ for the watermarked signals. The dotted line indicates the evaluation criteria (-1 ODG or higher). The PESQs in the proposed method exceeded -1 ODG under 4 to 8 bps conditions, whereas the LSDs in the DSS method was always about -4 ODG. The proposed method thus improved PESQ by about 3 ODGs compared with the DSS method.

Figure 6 (c) shows the BERs of the watermarked signals against various attacks. The dotted line indicates the evaluation criteria (10% or lower). The proposed method was robust against some attacks except those with G.726 and G.729. The proposed method has similar robustness to the DSS method.

5. Conclusions

This paper proposed a robust speech watermarking method based on direct spectrum spreading (DSS) with linear prediction (LP) residue to solve DSS's problem in terms of inaudibility. Two kinds of evaluations were carried out to investigate whether or not the proposed method can satisfy the two requirements of inaudibility and robustness in comparison with the DSS method. The evaluations verified that the proposed method can embed watermarks with low sound quality distortion. Moreover, the proposed method has similar robustness to the DSS method. These results show that the proposed method can satisfy the two important requirements for audio information hiding (AIH) techniques: inaudibility and robustness. Therefore, the proposed method can be regarded as an inaudible and robust speech watermarking. However, problems of the frame synchronization and blind detection remain to be solved in future work.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (B) (No. 17H01761) and I-O DATA foundation.

References

[1] N. Cvejic and T. Seppänen, Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks, IGI Global, Hershey, PA 2007.
[2] G. Hua, J. Huang, Y.Q. Shi, J. Goh, and V.L.L. Thing, "Twenty years of digital audio watermarking—A comprehensive review," Signal Processing, vol.128, pp.222–242, 2016.
[3] L. Boney, A.H. Tewfik, and K.N. Hamdy, "Digital watermarks for audio signals," Proc. ICMCS, pp.473–480, 1996.
[4] A. Bregman, Auditory Scene Analysis: The Perceptual Organization of Sound, MIT Press, 1990.
[5] K. Takeda, et al., "Speech database user's manual," ATR Technical Report, TR-I-0028, 2010.
[6] M. Unoki, J. Karnjana, S. Wang, N.M. Ngo, and R. Miyauchi, "Comparative evaluations of inaudible and robust watermarking for digital audio signals," Proc. ICSV21, g1–8, 2014.
[7] Information Hiding and its Criteria for Evaluation, https://www.ieice.org/iss/emm/ihc/, Retrieved 2018 Dec.
[8] M. Unoki and R. Miyauchi, "Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay characteristics," IEICE Trans. Inf. & Syst., vol.E98-D, no.1, pp.38–48, Jan. 2015.