Journal of Signal Processing, Vol.28, No.6, pp.309-313, November 2024

RESEARCH NOTE

Study on Inaudible Speech Watermarking Method Based on Spread-Spectrum Using Linear Prediction Residue

Aulia Adila, Candy Olivia Mawalim, Takuto Isoyama and Masashi Unoki

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology
1–1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {adila, candylim, takuto-i, unoki}@jaist.ac.jp

Abstract **A reliable speech watermarking technique must balance satisfying four requirements: inaudibility, robustness, blind detectability, and confidentiality. A previous study proposed a speech watermarking technique based on direct spread spectrum (DSS) using a linear prediction (LP) scheme, i.e., LP-DSS, that could simultaneously satisfy these four requirements. However, an inaudibility issue was found due to the incorporation of a blind detection scheme with frame synchronization. In this paper, we investigate the feasibility of utilizing a psychoacoustical model, which simulates auditory masking, to control the suitable embedding level of the watermark signal to resolve the inaudibility issue in the LP-DSS scheme. Evaluation results confirmed that controlling the embedding level with the psychoacoustical model, with a constant scaling factor setting, could balance the trade-off between inaudibility and detection ability with a payload up to 64 bps.**

Keywords: speech watermarking, LP-DSS, psychoacoustical model, auditory masking, embedding level

1. Introduction

Digital audio watermarking, commonly known as speech watermarking, has established itself as a reliable technology for secure communication [1] to prevent the risk of illegal distribution and misuse through non-authentic media. Direct spread spectrum (DSS) is one of the most widely used digital watermarking methods, renowned for its high robustness and security [2]. However, it has an inaudibility issue due to its principle of spreading messages using a pseudo-random noise (PN) signal across the host signal's spectrum. To address this, the LP-DSS scheme was proposed, which uses the linear prediction (LP) residue obtained from speech analysis and synthesis techniques to spread messages [3]. Additionally, another study proposed enhancements to the LP-DSS method to meet the requirements of blind detectability and confidentiality by incorporating two new forms of data embedding [4]. However, the blind scheme with frame synchronization causes inaudibility issue.

Inaudibility is closely related to psychoacoustics, which is the science of sound perception, i.e., the study of the statistical relationship between acoustic stimuli and hearing sensations [5]. A natural phenomenon occurs when the perception of one sound is likely to be obscured by the presence of another, which is called auditory masking [2]. Several studies have introduced the masking concept in speech watermarking to improve inaudibility [6, 7, 8].

In this paper, we aim to investigate the feasibility of using the auditory masking concept to resolve the inaudibility issue in the LP-DSS scheme proposed by previous studies [3, 4]. We use the masking threshold (i.e., masking curve) of the host signal, derived from a psychoacoustical model, to adjust a message's embedding strength so that it cannot be perceived by human ears. Moreover, we analyze how different embedding level settings could affect the inaudibility of a watermarked signal.

2. Watermarking Based on LP-DSS Scheme

LP-DSS is an advanced version of DSS that adopts the most basic speech coding method, linear predictive coding (LPC). The sound source of a speech signal is represented by the LP residue, and the spectral envelope is represented by the LP coefficient, which are provided by LPC. To create a watermark signal $\boldsymbol{m(n)r(n)}$, a message $\boldsymbol{m(n)}$ is modulated by the LP residue $\boldsymbol{r(n)}$ and then subsequently added to the host signal $\boldsymbol{x(n)}$ per frame to create a watermarked signal

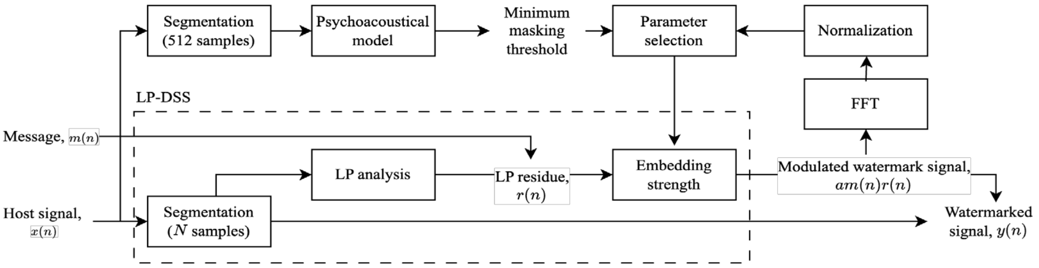

Fig. 1 Embedding process using Psychoacoustical LP-DSS embedding scheme

$\boldsymbol{y(n)}$:

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

$$a = 10^{L_{\mathrm{all}}/20} \tag{2}$$

$$L_{\mathrm{all}} = L_{\mathrm{PHS}} - L_{\mathrm{PWS}} + L_{\mathrm{SSL}} \tag{3}$$

where $\boldsymbol{a}$ is the scaling factor used to control the amplitude of the watermark signal $\boldsymbol{m(n)r(n)}$ to keep the signal inaudible, $\boldsymbol{L_{\mathrm{PHS}}}$ is the power level of the host signal, $\boldsymbol{L_{\mathrm{PWS}}}$ is the power level of the LP residue signal, and $\boldsymbol{L_{\mathrm{SSL}}}$ is the embedding-strength level in dB. The message $\boldsymbol{m(n)}$ can be defined as

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

where $\boldsymbol{x(n)}$, $\boldsymbol{y(n)}$, and $\boldsymbol{r(n)}$ are assumed to be ergodic. The message $\boldsymbol{m(n)}$ is identified by multiplying the watermarked signal $\boldsymbol{y(n)}$ with the LP residue $\boldsymbol{r(n)}$, resulting in the expected value $\boldsymbol{E\{y(n)r(n)\}}$, which is calculated using the Fourier transform:

$$\begin{aligned} E\{y(n)r(n)\} &= E\{[x(n) + am(n)r(n)]r(n)\} \\ &= E\{x(n)r(n)\} + E\{am(n)r^2(n)\} \\ &= am(n) \end{aligned} \tag{5}$$

The LP residue has the statistical properties $\boldsymbol{E\{r(n)\} = 0}$ and $\boldsymbol{E\{r^2(n)\} = 1}$. Additionally, since $\boldsymbol{x(n)}$ and $\boldsymbol{r(n)}$ are mutually orthogonal, we can derive the first term in Eq. (5) as $\boldsymbol{0}$ and the second term as $\boldsymbol{am(n)}$.

3. Proposed Method

The psychoacoustical model is a quantitative model that mimics the human hearing mechanism. Among the many phenomena in the hearing process, one crucial task for this model is simultaneous frequency masking [2, 7]. The model aims to analyze which frequency components contribute more to the masking threshold and calculate the amount of noise signal that can be added without being perceived. The masking condition is achieved when the first tone, known as the "maskee," is barely audible in the presence of the "masker" as the second tone. The difference in sound pressure level between the "masker" and "maskee" is defined as the "masking level" [7].

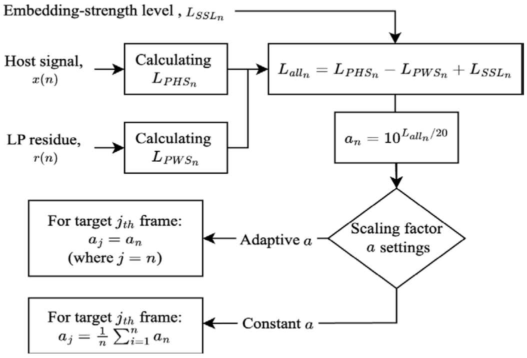
Embedding-strength level , L_{SSL_n}
Host signal,
$x(n)$
Calculating
L_{PHS_n}
$L_{all_n} = L_{PHS_n} - L_{PWS_n} + L_{SSL_n}$
LP residue,
$r(n)$
Calculating
L_{PWS_n}
$a_n = 10^{L_{all_n}/20}$
For target j_{th} frame:
$a_j = a_n$
(where $j = n$)
Adaptive a
Scaling factor
a settings
For target j_{th} frame:
$a_j = \frac{1}{n}\sum_{i=1}^{n} a_n$
Constant a

Fig. 2 Two methods for determining $\boldsymbol{a}$ for $\boldsymbol{n_{th}}$ frame signal: (1) adaptive $\boldsymbol{a}$ and (2) constant $\boldsymbol{a}$

The psychoacoustical model processes the audio information to derive the final masking threshold, i.e., the minimum masking threshold (MMT).

In this paper, our approach is to adopt the psychoacoustical model into the LP-DSS scheme by using the calculated MMT of the host signal $\boldsymbol{x(n)}$ to control the shape of the watermark signal $\boldsymbol{m(n)r(n)}$. The scaling factor $\boldsymbol{a}$, which corresponds to the embedding strength, is the selected parameter that is adjusted accordingly to ensure it remains below the masking threshold and is therefore imperceptible. We call our proposed method Psychoacoustical LP-DSS.

As shown in Fig. 1, the watermarking embedding process consists of two parts. The first part, marked with the dotted line box, is the watermarking embedding process using the LP-DSS scheme [3]. The second part involves the selection of the embedding strength parameter based on a psychoacoustical model. In this work, we use Psychoacoustical Model 1 (ISO/IEC MPEG-1 Standard) [9] to derive the host signal's MMT, which is then used as a criterion for selecting the scaling factor $\boldsymbol{a}$.

To obtain the masking curve, the host signal is divided into $\boldsymbol{N}$ frames, each with a fixed length of 512 samples. An FFT is performed on the segmented signal for accurate analysis of frequency components. The power spectral density (PSD) is then calculated and normalized to a sound pressure level (SPL) of 96 dB. The normalized PSD is used to discern fre-

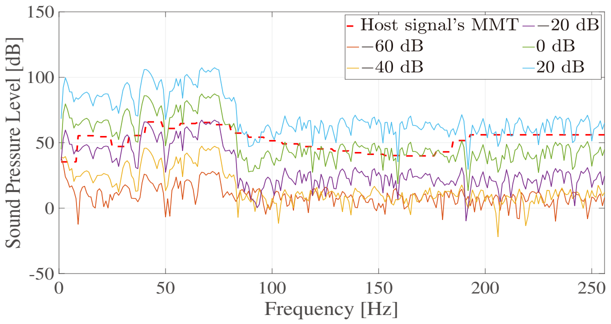

Fig. 3 Watermark signal comparison with host signal

quency components as tonal (more sinusoid-like) and nontonal (more noise-like) maskers. Invalid tonal and nontonal maskers are removed, i.e., maskers below the threshold in quiet (human hearing threshold) and maskers with a lower SPL compared with other maskers within the distance of 0.5 Bark. An individual masking threshold is computed for each remaining tonal and nontonal masker. A global masking threshold is calculated as the combination of individual masking thresholds and the threshold in quiet. Finally, the MMT of the host signal for each frame is derived from the global masking threshold obtained.

To keep the watermark signal $m(n)r(n)$ inaudible, we adjust the scaling factor a accordingly to be below the host signal's MMT. We considered two different approaches in setting the scaling factor a: adaptive a and constant a, as shown in Fig. 2. Initially, we determine an a value for each n_{th} frame obtained from the original LP-DSS scheme in Eq. (2), using the predefined embedding-strength level $L_{\mathbf{SSL}}$. This value is used to control the message's energy spread level throughout the host signal, according to the power of the host signal and the LP residue in each signal frame. Thus, this method of determining the a value is referred to as the adaptive a setting. Moreover, we calculate the constant a setting by averaging the a values from all signal frames obtained from the previous adaptive a setting.

Considering both approaches for determining the scaling factor a, the adaptive a setting is estimated to yield a higher bit-detection error rate in the lower power and silent parts of the signal. Since the a value in the adaptive setting is adjusted frame-by-frame according to the power of the host signal and the LP residue, it results in a small a value in the "quiet" regions of the audio, leading to a weak level of message embedding. This configuration offers better inaudibility but reduced robustness. Thus, we suggest the constant a setting as our proposed method for determining the scaling factor, which provides a stable a value in every part of the signal, regardless of its power. Moreover, this constant value is still sufficient

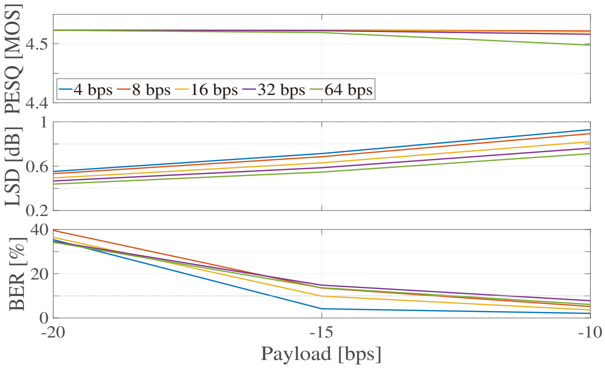

Fig. 4 Determining suitable embedding-strength level $L_{\mathbf{SSL}}$ for constant scaling factor a

and represents the host signal condition, as it is derived from the mean a value of the overall signal.

Furthermore, the modulated watermark signal $am(n)r(n)$ is produced from the watermark embedding process, utilizing the scaling factor a obtained from our proposed constant a setting. The modulated signal is then transformed into a frequency domain signal using FFT, and its PSD is normalized using the same normalization as applied in the psychoacoustical model. The normalized modulated signal, which represents a different predefined $L_{\mathbf{SSL}}$, is subsequently compared with the host signal's MMT by examining the proportion of data samples conditioned under the host signal's MMT, as illustrated in Fig. 3.

As a higher proportion leads to a better inaudibility, we investigate watermarked signals $y(n)$ with $L_{\mathbf{SSL}}$ ranging from $-\mathbf{10}$ dB to $-\mathbf{20}$ dB since the corresponding modulated watermark signals $am(n)r(n)$ yield proportions ranging from $\mathbf{70}\%$ to $\mathbf{90}\%$ under the host signal's MMT. We then evaluate the signals to determine the most suitable $L_{\mathbf{SSL}}$. The final watermarked signal $y(n)$ is obtained by embedding the message into the host signal using the LP-DSS scheme with the selected $L_{\mathbf{SSL}}$ and the constant a setting.

In the detection process, the watermarked signal $y(n)$ is divided into N frames using the same frame processing as in the embedding process. We use the same detection properties as indicated in Eq. (4). Thus, the message $m(n)$ is derived by using the following Eq. (4) after applying FFT to determine the sign of $E\{y(n)r(n)\}$ in each frame.

4. Evaluation

We evaluated our proposed method using two steps. First, we determined the optimal $L_{\mathbf{SSL}}$ by comparing it with the host signal's MMT, followed by evaluation using objective tests. Second, we measured the robustness and inaudibility of the watermarked signal using three objective tests: BER, LSD, and PESQ. To ensure unbiasedness in the evaluation due to the use of random embedded messages and a limited number of test data, we repeated the experiments for a total of five rounds. The final evaluation scores were calculated as the mean scores across all rounds, providing a more stable and reliable assessment of the proposed method.

4.1 Metrics and dataset

The BER test is used to measure signal robustness, with the criterion set at 10%. Additionally, LSD was conducted to determine how well the watermarked signal was perceived compared with the host signal, with the typical criterion for speech watermarking being 1 dB. As for PESQ, which is expressed as the mean opinion score (MOS), it has a scale ranging from 1 (bad) to 5 (excellent), with a standard threshold of 3 (fair or slightly annoying) for speech watermarking [4]. Our aim was to minimize the evaluation scores on the three objective tests simultaneously.

The tests were conducted on 12 utterances from the Advanced Telecommunications Research Institute International (ATR) speech dataset (B set) [10], which is sampled at 44.1 kHz and has an 8.1-sec duration each. The embedded message payloads were 4, 8, 16, 32, and 64 bps, respectively.

4.2 Results

To determine the suitable embedding-strength level, we analyzed the relationship between different $\boldsymbol{L}_{\mathbf{SSL}}$ values ($\mathbf{-20}$ dB, $\mathbf{-15}$ dB, and $\mathbf{-10}$ dB) in the watermarked signal, as shown in Fig. 4. It was observed that the BER decreased as $\boldsymbol{L}_{\mathbf{SSL}}$ increased, while distortions increased correspondingly. As illustrated in the figure, $\boldsymbol{L}_{\mathbf{SSL}}$ was determined to be $\mathbf{-10}$ dB for all payloads, resulting in a BER of less than 10%, an LSD of less than 1 dB, and a MOS score greater than 3. Therefore, we selected $\mathbf{-10}$ dB as the suitable $\boldsymbol{L}_{\mathbf{SSL}}$.

After determining the optimal $\boldsymbol{L}_{\mathbf{SSL}}$, we investigated whether our proposed method could satisfy the requirements of inaudibility and robustness by comparing it with the LP-DSS scheme [3, 4] using three objective tests: PESQ, LSD, and BER. As shown in Fig. 5, the horizontal axis represents the payload in bits per second (bps), and the vertical axis displays the PESQ, LSD, and BER scores, respectively.

Compared with the LP-DSS scheme, our constant $\boldsymbol{a}$ setting resulted in better inaudibility, as evidenced by the higher PESQ and lower LSD scores, due to the suitable selection of $\boldsymbol{L}_{\mathbf{SSL}}$ based on the psychoacoustical model. In terms of robustness, we assessed our proposed method by evaluating the BERs against various attacks, including G.711 speech coding, downsampling to 22.5 kHz, bit compression to 8 bits, conversion to the Ogg format, conversion to MP4, and the addition of white Gaussian noise with a signal-to-noise ratio (SNR) of 36 dB, as shown in Fig. 5. As can be observed, our method, which uses a constant scaling factor $\boldsymbol{a}$, was robust against all of the attacks used in the test. However, compared with the LP-DSS scheme, it demonstrated poorer robustness, as indicated by a higher BER under normal conditions. Despite this, because all BER scores still fell within the acceptable watermarking threshold of 10%, the level of error is considered acceptable.

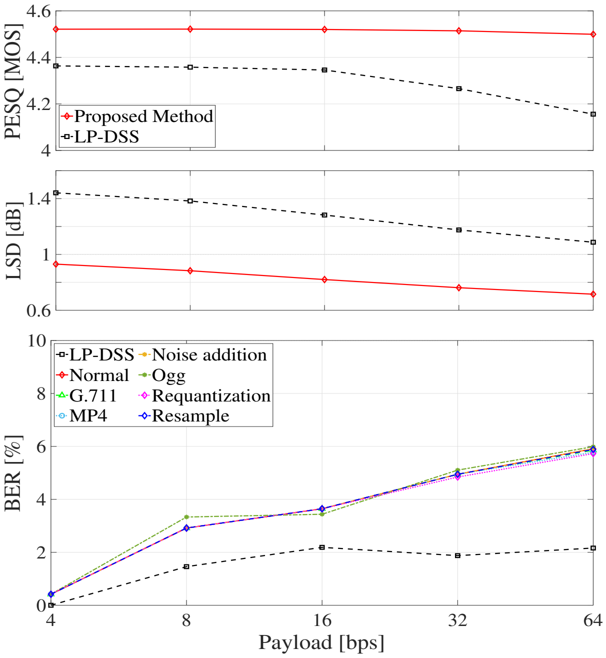

Fig. 5 Objective evaluation results for constant scaling factor $\boldsymbol{a}$ setting, compared with LP-DSS

5. Conclusions

This paper proposed a novel approach to determining a suitable watermark embedding level by utilizing a psychoacoustical model to resolve the inaudibility issue in the LP-DSS scheme. Our evaluation confirmed that the selected $\boldsymbol{L}_{\mathbf{SSL}}$ of $\mathbf{-10}$ dB results in an inaudible and robust watermarked signal that exhibits low sound distortion and an acceptable bit detection rate under 10% for payloads of 4, 8, 16, 32, and 64 bps under normal and attack conditions. Moreover, a constant setting was considered for determining the scaling factor $\boldsymbol{a}$ due to its ability to maintain robustness regardless of the signal's power level. As a future direction, we will incorporate the auditory masking concept using a psychoacoustical model into the blind LP-DSS speech watermarking method.

Acknowledgment

This work was supported by JSPS KAKENHI (23K18491) and the SCAT Foundation.

References

[1] G. Hua, J. Huang, Y.Q. Shi, J. Goh and V. Thing: Twenty years of digital audio watermarking – A comprehensive review, Signal Processing, Vol. 128, No. C, pp. 222-242, 2016.

[2] Y. Lin and W.H. Abdulla: Audio watermark: A comprehensive foundation using MATLAB, Springer Cham, 2015.

[3] R. Namikawa and M. Unoki: Non-blind speech watermarking method based on spread-spectrum using linear prediction residue, IEICE Trans. Inf. & Sys., Vol. E103.D, No. 1, pp. 63–66, 2020.

[4] T. Isoyama, S. Kidani and M. Unoki: Blind speech watermarking method with fame self-synchronization based on sread-spectrum using linear prediction residue, Entropy, Vol. 24, No. 5, p. 677, 2022. DOI: 10.3390/e24050677.

[5] M. Bosi and R.E. Goldberg: Introduction to Digital Audio Coding and Standards, Springer, New York, 2003.

[6] 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.

[7] R.A. Gracia: Digital Watermarking of Audio Signals Using a Psychoacoustical Auditory Model and Spread Spectrum Theory, AES E-Library, 1999.

[8] P. Bassia, I. Pitas and N. Nikolaidis: Robust audio watermarking in the time domain, IEEE Trans. Multimed., Vol. 3, No. 2, pp. 232–241, 2001.

[9] ISO/IEC 11172-3:1993, Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s – Part 3: Audio.

[10] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe and H. Kuwabara: Speech Database User's Manual, ATR Technical Report, ATR-Promotions, 2010.

Aulia Adila received her B.S. in Computer Science from Institut Teknologi Bandung (ITB) in Bandung, Indonesia. She is currently pursuing her Master of Science degree in speech processing at the School of Information Science at the Japan Advanced Institute of Science and Technology (JAIST). Her primary research interests include auditory signal processing, speech privacy, and machine learning.

Candy Olivia Mawalim received her B.S. in Computer Science from Institut Teknologi Bandung (ITB), Bandung, Indonesia. She received her M.S. and Ph.D. in the School of Information Science from the Japan Advanced Institute of Science and Technology (JAIST) in 2019 and 2022, respectively. She was selected as a research fellow for young scientists DC1 (JSPS) in FY2020-2022. Since April 2022, she works as an assistant professor at the School of Information Science and Research Center for Biological Function and Sensory Information, JAIST. She is also on the education team of the ISCA special interest group of security and privacy in speech communication (SIG-SPSC) committee. Her main research interests are speech signal processing, hearing perception, voice privacy preservation, and machine learning.

Takuto Isoyama received his Ph.D. from the Japan Advanced Institute of Science and Technology, Ishikawa, Japan, in 2024. He is currently a postdoctoral researcher at the same university. His research interests include auditory signal processing.

Masashi Unoki received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His main research interests are auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher from 1999 to 2000 and a visiting research associate at the Centre for the Neural Basis of Hearing (CNBH) in the Department of Physiology at the University of Cambridge from 2000 to 2001. He has been on the faculty of the School of Information Science at JAIST since 2001 and is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize from the ASJ in 1999, 2010, and 2013 for an Outstanding Paper and the Yamashita Taro "Young Researcher" Prize from the Yamashita Taro Research Foundation in 2005. He is a Fellow of IEICE, a vice president of ASJ, and a Member of IEEE and ASA.

(Received May 15, 2024; revised July 12, 2024)