NCSP'24 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2024
February 27 - March 1, 2024, Honolulu, Hawaii

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

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

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: {2310401, candylim, isoyama-t, unoki}@jaist.ac.jp

Abstract

A reliable speech watermarking technique must balance satisfying the four requirements: inaudibility, robustness, blind-detectability, and confidentiality. The previous study proposed an LP-DSS scheme that could simultaneously satisfy these four requirements. However, the inaudibility issue happened due to the blind detection scheme with frame synchronization. In this paper, we investigate the feasibility of utilizing a psychoacoustical model to control the suitable embedding level of the watermark signal to resolve the inaudibility issue that arises in the LP-DSS scheme. A psychoacoustical model simulates the auditory masking phenomenon that "mask" signals below the masking curve to be imperceptible to human ears. Results of the evaluation confirmed that the controlled embedding level from the psychoacoustical model balanced the trade-off between inaudibility and detection ability with payload up to 64 bps.

1. Introduction

The development of information and communication technology (ICT) and the increase of multimedia information usage via the Internet has positively affected societies and communities in many ways. However, multimedia big data, which might contain personal data, faces a high risk of illegal distribution and misuse through non-authentic media. Digital audio watermarking, also known as speech watermarking, has established itself as a dependable technology for secure communication [1, 2]. There are four requirements for speech watermarking techniques: inaudibility, robustness, blind-detectability, and confidentiality.

Our previous study proposed an approach for a reliable speech watermarking technique, namely the LP-DSS scheme [3, 4], which enhanced the non-blind speech watermarking method based on direct spread spectrum (DSS) using a linear prediction (LP) residue [3]. Compared to the DSS method as one of the most popular techniques for digital watermarking [5], LP-DSS uses the LP residue as its spreading signal of the embedded watermark instead of the pseudorandom noise (PN) signal. This method successfully embeds and detects speech code into the host signal while satisfying the four requirements of speech watermarking. However, the inaudibility has reappeared since the blind scheme using frame synchronization has been considered in this scheme [4]. Moreover, as this scheme has used spread spectrum as its core idea, it is prone to sound quality if the spreading level of the message is neglected. Therefore, it is essential to control the spreading level of the message in the watermarked signal [2, 6].

Auditory masking is one of the most important phenomena in our hearing system. It describes the phenomena when a louder inaudible sound (the masker) causes a fainter but still audible sound (the maskee) to become inaudible [5]. We have been motivated by this idea to improve the speech watermarking technique's inaudibility by implementing the masking approach. The knowledge of auditory masking has been modeled through a psychoacoustical model, which is also used in another study of inaudible speech watermarking [7].

This paper aims to investigate the feasibility of the psychoacoustical model to improve inaudibility in the LP-DSS scheme. We investigate the embedding-strength level adjustment using the masking threshold (i.e., masking curve) of the host signal derived from the psychoacoustical model. We hypothesize that the adjustment of the embedding-strength level is suitable for improving the inaudibility of the watermarked signal since embedding-strength can control the message's energy spread throughout the host signal's spectrum. Incorporating the psychoacoustical model is considered a novel idea delivered through this study.

2. Watermarking Based on LP-DSS Scheme

The most widely used digital watermarking method is spread spectrum (SS) watermarking [5], with one of its types is the direct spread spectrum (DSS) watermarking. It spreads the message across the host signal's spectrum, making it difficult to identify the energy contained in each frequency bin, resulting in high robustness and security [5]. LP-DSS is the advancement method of DSS, which adopts the most basic speech coding method, linear predictive coding (LPC). The speech signal's sound source is represented by the LP residue, and the spectral envelope is represented by the LP coefficient, which the LPC provides. To create the watermark signal

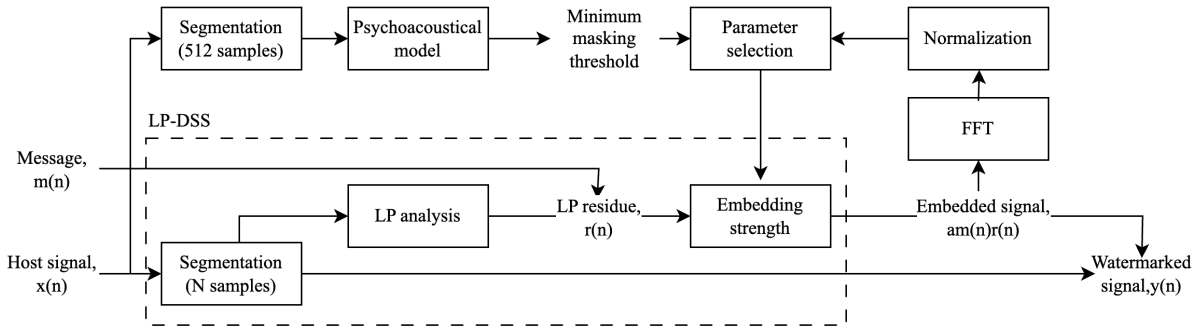

Figure 1: Embedding process using the psychoacoustical model and the LP-DSS embedding scheme

$m(n)r(n)$, the message $m(n)$ is modulated by the LP residue $r(n)$, then is subsequently added to the host signal $x(n)$ per frame to create the watermarked signal $y(n)$ as follows.

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

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

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

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

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

where $E\{\cdot\}$ is the expected value of "$\cdot$". In the detection process, the fast Fourier Transform (FFT) is used to detect the sign of $E\{y(n)r(n)\}$ in each frame, and then $m(n)$ is calculated by Eq. (4) using the sign obtained.

3. Proposed Method

The psychoacoustical model is a quantitative model which mimics the human hearing mechanism. From the many phenomena in the hearing process, one crucial task for this model is the simultaneous frequency masking [5, 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 in without being perceived. The masking condition is achieved when the first tone, known as "maskee," is barely audible in the presence of "masker" as the second tone. The difference in sound pressure level between the "masker" and "maskee" is defined as "masking level" [7]. The psychoacoustical model processes the audio information to derive the final masking threshold, that is, the minimum masking threshold (MMT).

In this paper, our approach is to adopt the psychoacoustical model to the LP-DSS scheme by using the calculated MMT of the host signal $x(n)$ to control the shape of the watermark signal $m(n)r(n)$. The scaling factor a, which corresponds to the embedding-strength, is the selected parameter that is adjusted accordingly to meet the condition below the masking threshold so that it could be imperceptible. The watermarking scheme used in this work is the LP-DSS method [3, 4].

As shown in Fig.1, the watermarking embedding process consists of two parts. The first part which marked with the dotted line box is the watermarking embedding process using LP-DSS method. The second part is the parameter selection of embedding-strength based on a psychoacoustical model. In this work, we adopt the psychoacoustical model 1 (ISO/IEC MPEG-1 Standard) [8] to derived the minimum masking threshold (MMT) of the host signal which then used as a criterion for selecting the parameter scaling factor a.

To obtain the masking curve, the host signal is divided into K frames with using a fixed length of 512 samples. FFT is performed to the segmented signal for the accurate analysis of frequency components. Then, power spectral density (PSD) is calculated and normalized to a sound pressure level (SPL) of 96 dB. The normalized PSD is used to discern frequency components as tonal (more sinusoid-like) and nontonal (more noise-like) maskers. The invalid tonal and nontonal maskers are removed, i.e., the maskers below the threshold in quiet (human hearing threshold) and the maskers with lower SPL comparing to other maskers within the distance of 0.5 Bark. The individual masking threshold is computed for each remaining tonal dan nontonal masker. The global masking threshold is calculated as the combination of individual masking threshold and the threshold in quiet. Finally, the MMT of 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 MMT of the host signal. At first, we determined a constant a by calculating the average a value obtained from the original LP-DSS method in Eq. (2), using the predefined embedding-strength level L_{SSL}. The constant a is used as the scaling factor of the watermark signal $m(n)r(n)$, which results in an embedded signal $am(n)r(n)$. The embedded signal is transformed into a frequency domain signal using FFT, and then its PSD is normalized using the same normalization as applied in the psychoacoustical model. The normalized embedded signal is then compared to the host signal's MMT

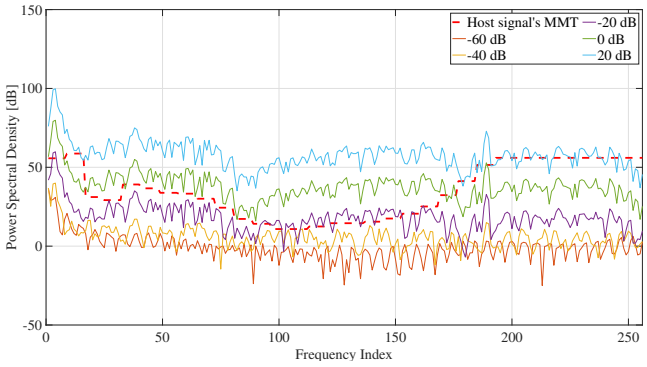

Figure 2: Watermark signal comparison with the host signal

to investigate the compatibility of constant a as the scaling factor corresponding to the embedding level. Moreover, we calculated an adaptive a as the comparison and applied the following steps for the constant a. In the adaptive a settings, we control the scaling factor a according to the power of the host signal and the watermark signal in each signal frame. Figure 2 shows how we investigate the suitable embedding level using the psychoacoustical model.

In the detection process, the watermarked signal $y(n)$ is divided into K frames using the same frame processing as in the embedding process. We employ the same detection properties as indicated in Eq. (4). Therefore, the sign of $E\{y(n)r(n)\}$ in each frame is determined using FFT. Regarding the LP-DSS approach, the message $m(n)$ can be derived by using the following Eq. (4) after the FFT is used to determine the sign of $E\{y(n)r(n)\}$ in each frame.

4. Evaluation

We evaluated our proposed method using two significant steps. First, we investigated which range of L_{SSL} is the most suitable according to its comparison with the host signal's MMT. Then, we measured the robustness and inaudibility by carrying out objective tests, which are bit-error-rate (BER), log-spectrum distortion (LSD), and perceptual validation of speech quality (PESQ) ITU-T P.862 on 12 utterances in the ATR speech dataset (B set) [9] which is sampled at 44,1 kHz. Each signal in this dataset has an 8.1-sec duration length. The BER calculates the number of incorrectly detected watermark bits over all embedded watermark bits, which we evaluated under normal conditions. 10% is the standard BER level for speech watermarking. LSD was conducted to determine how well the watermarked signal was perceived compared to the host signal. The typical threshold for LSD in speech watermarking is 1 dB. As for PESQ, which is expressed as the mean opinion score (MOS), it has a scale of 1 (bad) to 5 (excellent), with a standard threshold of 3 (fair or slightly annoying) for speech watermarking. We use LP-DSS for the watermarking scheme, with the payload of 4, 8, 16, 32, and 64, respectively. The scaling factor a was determined using the embedding-strength level L_{SSL} ranges from -60 dB to 20 dB with 5 dB increments, with the settings of a constant and an adaptive scaling factor a. The messages embedded in the host signal were random bit streams, with no error correction schemes used in the proposed method.

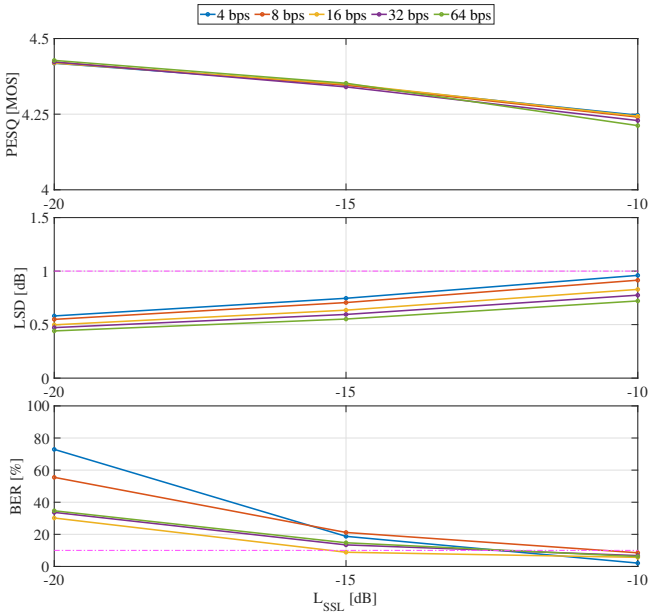

Figure 3: Relationship between embedding-strength level in the constant scaling factor a

Before conducting objective tests to evaluate the speech watermarking inaudibility and robustness, we compared each resulting watermark signal $m(n)r(n)$ with the host signal by examining the proportion of data samples in the watermark signal conditioned under the host signal's minimum masking threshold. As we tried to experiment with both SS watermarking methods (i.e., DSS and LP-DSS), our experiments indicate that the watermark signal with the LP-DSS scheme has a more significant signal portion under the masking level, which is estimated to offer better inaudibility than the DSS scheme at any embedding levels. Embedding-strength levels ranging from -10 dB to -20 dB resulted in the scaling factors a that control the watermark signal's conditioning under the masking level, with a portion ranging from 70% to 90%. Our evaluation continues by measuring the selected watermark signals with the L_{SSL} ranging from -10 dB to -20 dB with 5 dB increments using three objective tests on the constant and adaptive a settings to find the suitable embedding-strength level L_{SSL} and the payload accordingly.

Figure 3 shows evaluation results for PESQ, LSD, and BER, respectively, for the constant scaling factor a. It is observed that the BER decreases as L_{SSL} increases, and distortions increase as L_{SSL} increases. As from the figure, the L_{SSL} was determined to be -10 dB for all the payloads, which resulted in the BER less than 10%, LSD less than 1 dB, and MOS score greater than 3. Furthermore, we investigated how the different settings strategy of the scaling factor a would

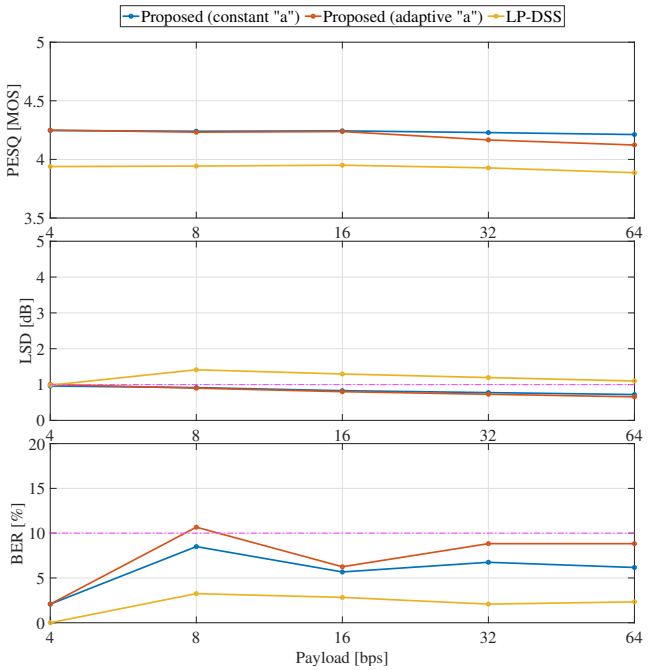

Figure 4: Objective evaluation results with different scaling factor a settings

affect the speech watermarking result.

Figure 4 shows the result of PESQ, LSD, and BER of our proposed method compared with the LP-DSS scheme [3, 4], where the horizontal axis represents the payload in bps and the vertical axis represents PESQ, LSD, and BER, respectively. The dotted lines indicated the typical threshold for each evaluation test regarding speech watermarking. Compared to the LP-DSS scheme, our proposed method resulted in a better inaudibility, as denoted by the higher PESQ and lower LSD score, due to the suitable L_{SSL} selection based on the psychoacoustical model. But it has a higher BER, considering the cost of improved inaudibility. However, because it is approximately less than or equal to the speech watermarking BER threshold, it still has an acceptable BER. Moreover, we also investigated that the different scaling factor a setting has no significant difference in the watermarked signal's inaudibility. Despite this, the payload of 8 bps and 32 bps had a higher BER due to the adaptive scaling factor a settings.

5. Conclusions

This paper proposed a new approach to determine a suitable speech watermark embedding level using the psychoacoustical model. Our evaluation confirmed that the selected embedding level of -10 dB could result in an inaudible and robust watermarked signal under normal conditions. The resulting speech has a low sound distortion and an acceptable bit detection rate of equally under 10% for payloads of 4, 8, 16, 32, and 64 bps. Compared to the LP-DSS scheme, it has improved its inaudibility but slightly decreased the robustness. We also investigated two settings of scaling factor a, which corresponds to the embedding level settings by setting the constant a for all signal frames and adaptive a for each signal frame. The results showed that constant a has a lower BER than adaptive a in 8 and 32 bps. However, there is no significant difference in terms of inaudibility.

As our future direction, we will consider another approach to adjust the watermarked signal embedding level by utilizing the psychoacoustical model to improve the inaudibility while still maintaining the robustness. We will also enhance the performance of the proposed method in dealing with attacks by conducting various robustness tests.

Acknowledgment

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

References

[1] N. Cvejic and T. Seppänen, 'Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks," 01 2008.

[2] 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, 04 2016.

[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, pp. 63–66, 01 2020.

[4] T. Isoyama, S. Kidani, and M. Unoki, "Blind Speech Watermarking Method with Frame Self-Synchronization Based on Spread-Spectrum Using Linear Prediction Residue," Entropy, vol. 24, pp. 677, 05 2022.

[5] Y. Lin and W.H. Lin, Audio watermark: A comprehensive foundation using MATLAB, pp. 1-199, 01 2005.

[6] L. Boney, Ahmed Tewfik, and K.N. Hamdy, "Digital watermarks for audio signals," IEEE Proceedings Multimedia, pp. 473-480, 07 1996.

[7] R.A. Gracia, "Digital watermarking of audio signals using a psychoacoustical auditory model and spread spectrum theory," AES E-Library, 1999.

[8] 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

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