Journal of Information Hiding and Multimedia Signal Processing ©2022 ISSN 2073-4212
Ubiquitous International Volume **13**, Number **2**, June **2022**

Speech Manipulation Detection Method using Audio Watermarking

Kota Muroi and Kazuhiro Kondo

Graduate School of Science and Engineering
Yamagata University
4-3-16 Jonan, Yonezawa, Yamagata 9928510, Japan
stgfst12345@gmail.com, kkondo@yz.yamagata-u.ac.jp

Received February 2022; revised May 2022

Abstract. *We propose a method for detecting manipulations in audio recordings using speech fingerprinting and time-stamping methods. The fingerprints are line spectrum pairs extracted from the audio, and the timestamp is a continuous numeral sequence created from the frame number. Encryption is applied to the timestamp data to maintain confidentiality. The fingerprints and timestamps are embedded in the audio recording using a watermark. At the decoder, the fingerprint detected from the watermark is checked against the fingerprint extracted from the audio recording itself, and the extracted timestamp data is checked for continuity. The vector distance between the fingerprint obtained from the watermark and the fingerprint obtained from the recorded audio is then calculated. If this distance is greater than a predetermined threshold, tampering is suspected; if it is smaller, the authenticity of the recording can be proven. Simultaneously, the timestamp data obtained from the watermark is checked for continuity, and if the value is continuous before and after each frame, it is determined that the speech data has not been tampered with, and the continuity of the speech data can be proven. We simulated a recording of interrogation speech in a realistic noise environment with a realistic scenario, and simulated both deletion and substitution of a key sentence, emulating incriminating editing of the interrogation speech. It was shown that it is possible to identify the tampered frame with 100% accuracy in this realistic noisy environment.*

Keywords: Manipulation Detection, Authentication, Audio Watermarking, Speech Fingerprinting, timestamp Data

1. **Introduction.** In recent years, digital multimedia (images, sounds, videos, etc.) has become an indispensable part of our lives and society with the spread of cell phones and other smart devices, as well as the advancement of various technologies such as mobile and wireless networks. Some multimedia data whose authenticity is not guaranteed or which has been tampered with is admissible as evidence, which may affect court decisions. The proof of authenticity of digital multimedia has become a new challenge as more and more sophisticated tampering takes place. While considerable progress has been made in the authentication of images and videos, the authentication of audio is still in its early stages. In general, methods for proving the authenticity of speech are based on the detailed analysis and evaluation of the speech to verify its originality and the possibility of tampering. In the past, it was easy to prove the authenticity of speech based on irregularities or sudden changes in the speech spectrum, such as environmental sounds or speakers. However, with recent advances in technology and the use of advanced tools, it is difficult to detect all forms of tampering by conventional auditory or visual analyses.

Tools that allow advanced tampering of recordings are now widely available, and they can be easily tampered with without leaving any audio traces.

Existing tamper detection methods can be classified as passive and active. Passive tamper detection is a technique that does not use any signatures or watermarking, but analyzes the signal itself for any unnatural clues. Identifying the location of tampering in tampered audio is more difficult than determining whether the audio has been tampered with or not. Also, even if it is a two-way decision, some misjudgment will occur. In a recent study, the authors of [1] used measures such as environmental noise and the magnitude of the impulse response of the acoustic channel to prove authenticity and detect substitution tampering. To evaluate their method, they used TIMIT and another database created in four different environments. Gaussian Mixture Model (GMM) was used as the classification method, but the false positive rate obtained was more than 3%.

In active tamper detection, on the other hand, tampering is detected by embedding and extracting a digital watermark in speech signals. If the embedded watermark and the extracted watermark are the same, we can prove the authenticity. In related work, [2] proposed a watermarking scheme for both the source excitation signal and the vocal tract characteristics using quantization index modulation and formant enhancement. In [3], a tampering detection scheme for speech signals based on the same formant enhancement-based watermarking was proposed by the same group. This embedding concept not only enables the proposed scheme to be inaudible but also provides the possibilities for both robustness against speech processing and sensitivity to tampering. In our previous work by our group [4, 5], we used fingerprints as the watermark data. We use Line Spectrum Pairs (LSP) [6], which is one of the features of speech, as fingerprints. The vector distance between the LSP calculated from the embedded speech signal and the LSP obtained from the watermark data is calculated, and if the distance is smaller than a specified value, no tampering is detected, otherwise tampering is suspected. When the tampering was detected on a frame-by-frame basis, the accuracy, in terms of specificity was over 99%, and the sensitivity was over 97%, which can be said to achieve levels high enough for practical applications.

However, there are two practical problems with the previous study: first, there is no way to identify the type of tampering (deletion, insertion, or substitution). The second is that there is no way to accurately detect tampering when the length of the tampered audio does not match the frame length, *i.e.*, if the audio length after tampering does not exactly match the original length, it is not possible to detect tampering accurately. Since many false positives occur due to frame misalignment, the specificity may not as accurate as stated in [4].

Accordingly, we propose a method for detecting tampering in audio recordings by using both fingerprinting and temporal information together as watermarked information [8]. The fingerprinting introduced in the previous study is tolerant to noisy environments because the detection of tampering is determined by vector distance with some margin for tolerance. Although we were able to identify the discontinuous parts (substitutions and deletions) due to tampering, we were not able to detect other types of tampering and the length of the tampering. It was also not possible to detect intact parts following the tampered parts when tampering that was not in units of frames. This is because in this case, the portions after tampering were not aligned in the original frame positions. The combined use of timestamp data introduced in this research makes it possible to identify the tampering types by identifying the discontinuous frames. The timestamp data is encrypted so that the confidentiality of the tamper detection can be maintained. Next, we introduce the embedding of a synchronization signal in the audio in order to recover from frame misalignment due to tampering [9]. This makes it possible to estimate the frame position after the tampered portions, and eliminate frame misalignment during watermark detection. These will enable us to improve the analysis performance and prove the authenticity of audio recordings. In other words, this system is a highly versatile method of tamper detection.

In the performance evaluation of the proposed method, we assumed a police interrogation scenario, where the interrogation voice was tampered with in such a way that it was rewritten by the police officer to suit his intentions to incriminate. In this study, we aim to achieve a detection accuracy of more than 90% by evaluating the proposed tampering detection method on audio recordings that have been subjected to various types of tampering (deletion, insertion, and substitution).

2. **Speech Manipulation Detection Method.** Figure 1 shows the overall structure of the proposed audio tampering detection method briefly described in the previous section. First, an overview of the tampering detection is given assuming a recording of a police interrogation, to be used later as evidence in a courthouse, as an example. In a police station, where suspects are being interrogated, the audio of the interrogation is recorded using a digital recorder housed in a tightly sealed black box. The sealed box prevents tampering by the interrogators or the recorder administrator. Inside the recorder, the input audio, which is divided into frames, are recorded after being watermarked with (1) frame-wise sequential numbers, *i.e.* timestamps, and (2) a fingerprint of the audio, which is a measurement of the general characteristics of the input audio along the time axis. The timestamp data is encrypted to prevent tampering by direct manipulation of the watermark data. By not encrypting the fingerprint, we aim to increase the robustness against noise and improve the accuracy of tampering detection. The fingerprint and the encrypted timestamp data are embedded in the audio recording as a watermark. When verifying the authenticity of the interrogation speech as evidence, the fingerprints and timestamp data embedded as watermarks are extracted, and the timestamp data is decrypted using a secret key. The vector distance between (a) the fingerprints extracted from the watermark and (b) the new fingerprints extracted from the watermarked speech, is calculated and compared with a predetermined threshold. If this distance is within the threshold value, the speech is judged to have not been tampered with. However, if the vector distance exceeds the threshold, a warning is issued for possible tampering. The time information is checked for continuity of the data as an alternative means to detect tampering. If the value is continuous, it is judged not to have been tampered with, and if the value is discontinuous, it is judged to have been tampered with. All of these detection processes need to be done in a black box that is strictly sealed.

Figure 2 shows the process of embedding a watermark in a recorder. First, the speech signal is divided into frames and a synchronization signal indicating the position of the frames is embedded. The synchronization signal is created by applying inverse FFT on the phase component of the M-sequence signal combined with the amplitude component of the original sound. This synchronization signal is scaled, and added to the original signal. At the decoder, the synchronization position of the frame can be estimated by cross-correlation between the synchronization signal and the M-sequence signal.

After the addition of the synchronization signal, the fingerprints and timestamp data for each frame are calculated. We used Line-Spectral-Pairs (LSP) as fingerprints [6]. The LSP coefficients were coarsely quantized to be robust against non-malicious signal degradation while capturing the changes in the speech signal caused by malicious tampering. Through preliminary experiments, we decided that 10 LSP coefficients are optimum to measure the speech spectrum for our purpose, and each of the coefficients should be linearly quantized using 4 bits, *i. e.*, 16 levels, for a total of 40 bits per frame. The

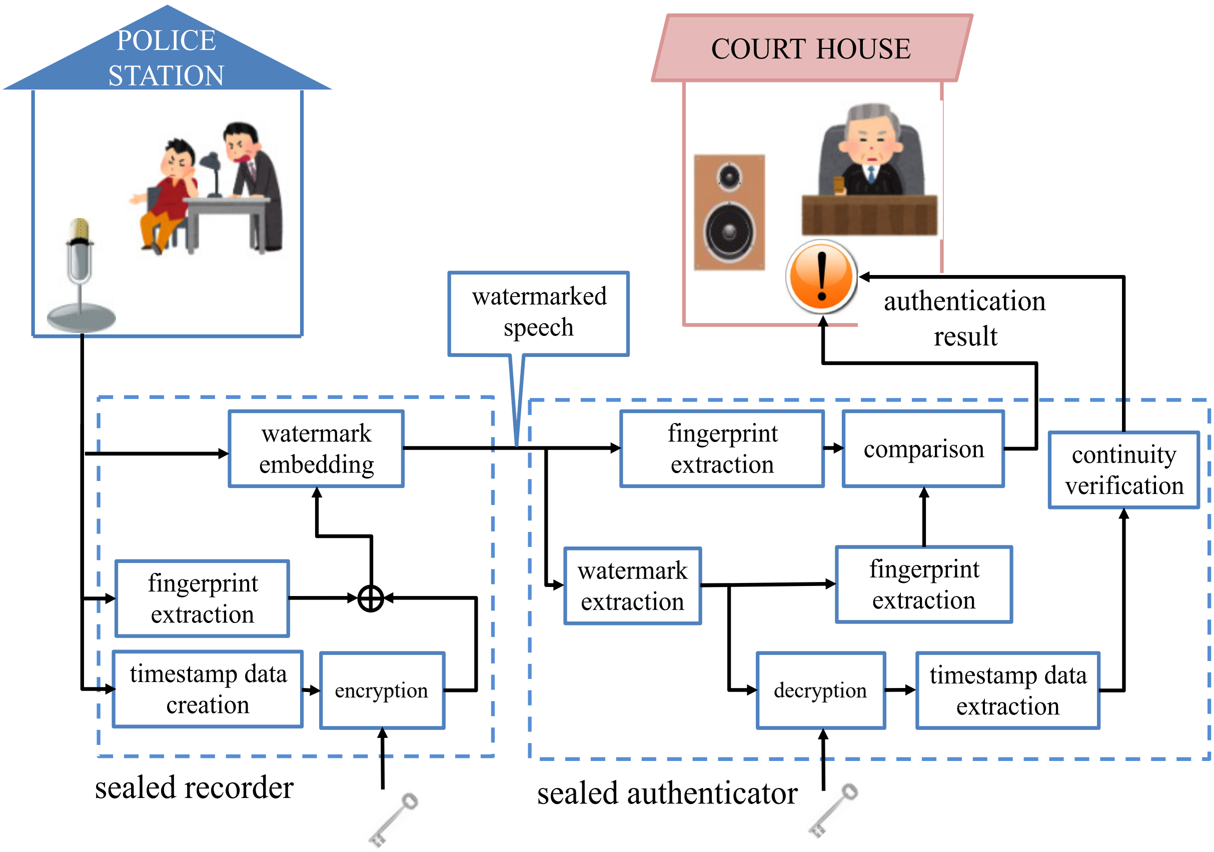

FIGURE 1. Block diagram of speech tampering detection method

timestamp data is essentially the frame number. The frames were numbered in the order in which they were processed, and used as the timestamp data. For the timestamp data, 32 bits per frame are used so that a sufficiently long recording time can be uniquely identified using the timestamp assuming actual operation. Encryption was applied to the timestamp data on a frame-by-frame basis to maintain confidentiality. We used the AES encryption [7], which is known to be a modern and secure frame-by-frame encryption method. The fingerprints extracted from the audio and the encrypted timestamp data were then combined into a single data sequence. Figure 3 shows the fingerprint and timestamp data merged as watermark data for a single frame. We employed the conventional DSS (Direct-Spread-Spectrum) [10] as the watermarking method. In this method, the watermark signal is spread over a wide frequency range by multiplying this signal with pseudo-random spreading codes to minimize the influence of the watermark on speech quality. If noise is introduced into the watermark signal during detection, the power of the noise can be spread over a wide frequency range during inverse spreading, there-by thinly spreading the effect of this noise over a wide frequency range, making this watermarking method resistant to additive noise.

We set a time delay offset for embedding the watermark. This is outlined in Fig. 4. As shown in this figure, the frame position from which the fingerprint is extracted does not match the frame position used for the watermark. By using this offset, a part of the samples to extract fingerprints from the watermarked audio will be lost due to deletion or substitution tampering. This makes it impossible to extract the matching fingerprint from the audio and the watermark, and tampering can be detected. On the other hand, without this offset, the substitution cannot be detected because the substituted samples

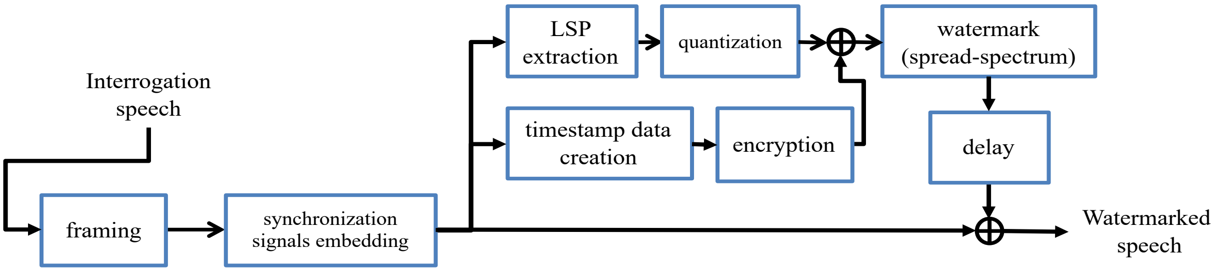

FIGURE 2. Block diagram of watermarking at the recorder

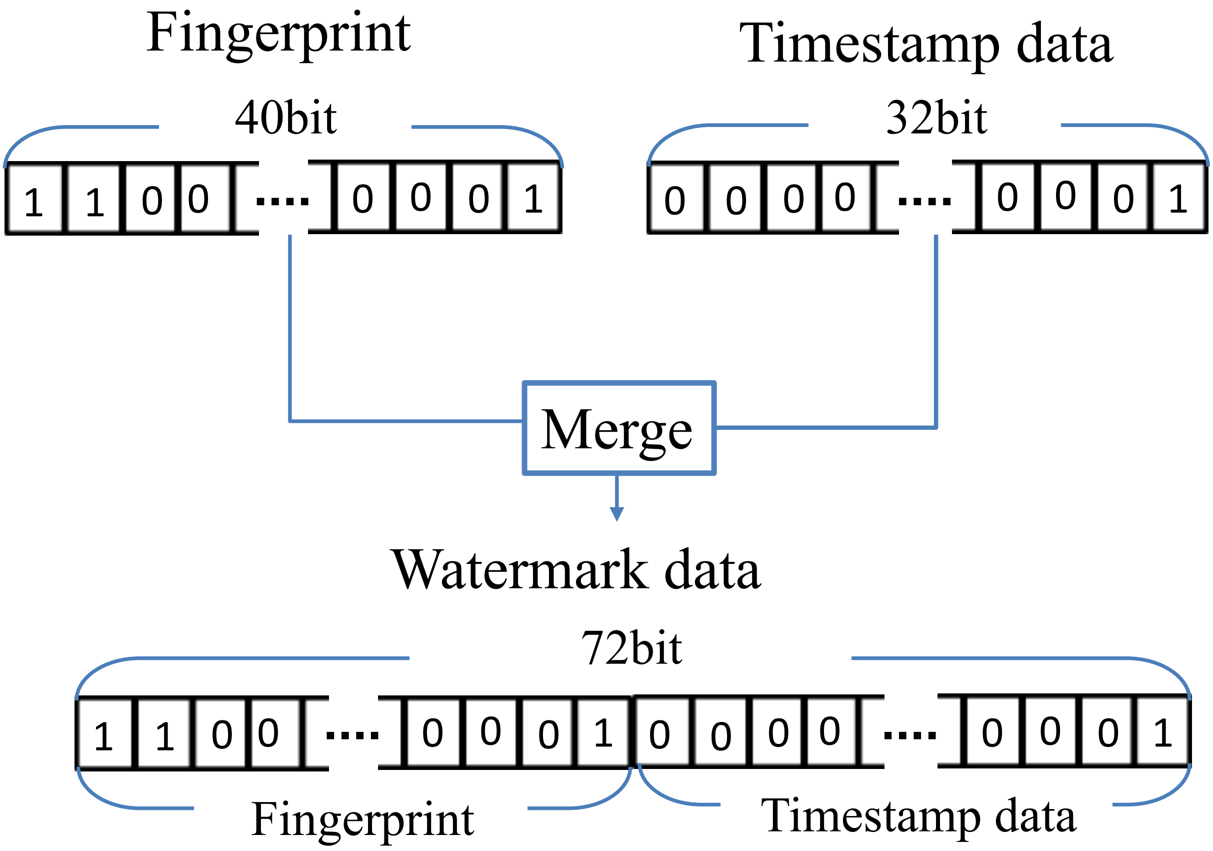

FIGURE 3. Merging fingerprint with timestamp data

produce fingerprints matching those extracted from the watermark. In this paper, we used 1/2 frame as an arbitrary offset.

Figure 5 shows the structure of the tampering detection process. First, the frame position is estimated using the cross-correlation of the M-sequence signal with the watermarked audio in which the synchronization signal is embedded. The audio signal with embedded watermark is divided into frames, and the clean speech signal is estimated from the embedded watermarked speech signal. In this paper, we used a powerful noise removal filter called the iterative Wiener filter [11, 12] to remove the watermark component from the watermarked signal and obtain the estimated clean speech. Blind watermark detection is then performed on the difference signal between the watermarked speech and the estimated clean speech. The subtraction of the estimated clean speech was necessary for blind detection since the speech component itself may be regarded as noise during the despreading processing to extract the watermark. The fingerprint extracted from the

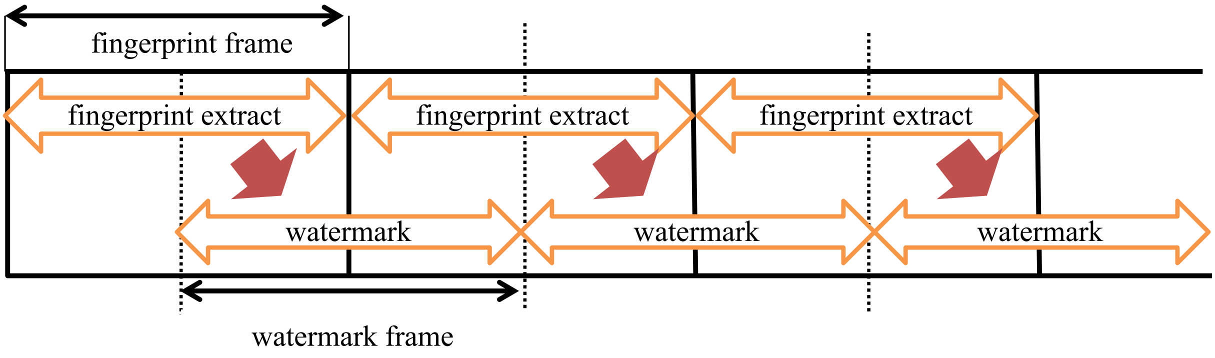

FIGURE 4. Delayed processing

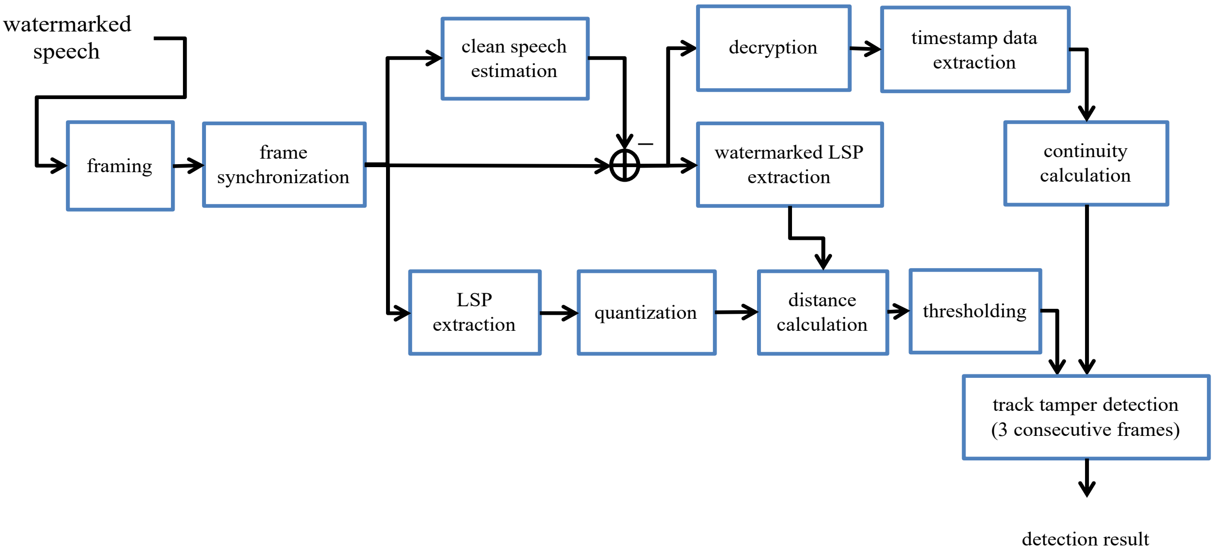

FIGURE 5. Block diagram of the tampering detection process

watermark is compared to the fingerprint newly extracted from the watermarked speech signal. The distances between the fingerprints are calculated for each frame. We used the Euclidean distance, $\dot{L}$, shown in the following equation:

$$\dot{L} = \sqrt{\sum_{i=1}^{N}(l_i - l'_i)^2} \tag{1}$$

Here, l_i is the fingerprint (LSP) extracted from the i-th frame speech, l'_i is the fingerprint extracted from the watermark, and N is the order of the LSP, for which we used 10 in this paper. The computed distance is compared with a predetermined threshold value, and the presence of tampering is detected if the distance exceeds the threshold value. The extracted frame numbers were judged to be tamper-free if they were consecutive compared to one previous frame, and tampered with if they were not. If tampering is detected for three or more consecutive frames, the frames are synchronized since frame misalignment may occur due to tampering, and tampering is detected again.

3. **Detection Accuracy Evaluation.** In this section, we will evaluate the detection accuracy for audio recorded in a real environment. We used audio recordings mimicking a

police interrogation scenario, in which a police officer interrogated a suspect according to a prepared script. In addition, considering the environmental noise in a police interrogation room, duct noise was mixed into the audio. Other naturally introduced environmental noises include the sound of a PC, the dragging sound of a chair, and reverberation were also included intentionally. The aim of this experiment was to see if the proposed tampering detection method can be used under the actual environmental noise. Three types of tampering were tested to evaluate the sensitivity and specificity of the proposed method, which are defined as the following:

$$\text{Sensitivity} = \frac{\text{True Positive Frames}}{\text{Number of Manipulated Frames}} \times 100 \tag{2}$$

$$\text{Specificity} = \frac{\text{True Negative Frames}}{\text{Number of Non-manipulated Frames}} \times 100 \tag{3}$$

Sensitivity indicates how accurately the initial portions of the tampered intervals were identified, where the initial interval portion is defined as being up to three frames from the frame where the tampering started. The specificity is how accurately the frames which were not tampered with are identified. The number of true positive frames represents the total number of frames for which the system was able to correctly identify tampered frames, and the number of true negative frames represents the total number of frames for which the system was able to correctly identify the tamper-free frames.

4. **Experimental Conditions.** We manipulated the audio of the police interrogation in three different modes. (1) In tampering pattern 1, the portion with the suspect denying the accused crime was deleted. (2) In tampering pattern 2, the suspect's denial was deleted and was substituted with the voice of the same suspect admitting to the accused crime. (3) In tampering pattern 3, the suspect's denial was deleted, and was changed to another speaker admitting to the crime. The length of the tampering was 2 to 3 s. We investigated the sensitivity and specificity of the proposed method to detect tampered frames.

Details of the parameters used in the experiment are shown in Table 1. A combination of one police officer and three suspects were prepared, for a total of three separate recordings. The sound pressure of added noise was set to 50 dB, 60 dB, and 70 dB, respectively.

5. **Results and Discussions.** Tables 2, 3, and 4 show the performance of tampering detection for tampering patterns 1 through 3.

In tampering pattern 1, which deletes part of the audio content of the interrogation, the sensitivity was 100% for the detection using timestamp data. With fingerprints, the sensitivity ranged from 78% to about 90%. False negatives occurred only in the boundary frames between the tamper-free frames and the tampered frames. In some boundary frames, false negatives occurred because the error distance did not exceed the threshold value due to a mixture of tampered and tamper-free samples. However, we were able to identify the correct tampered position within an error of about one frame. Next, the specificity was over 87% for SNRs 0 dB and 10 dB, but it dropped significantly at -10 dB. It was found that it was difficult to identify the unaltered frames with many false positives at this noise level. The reason for this large number of false positives is that the noise is larger than the speech, so the clean speech cannot be estimated with enough accuracy, and the watermark cannot be detected correctly. However, we note that the speech at -10 dB SNR was completely inaudible, making this sample completely unusable as evidence. Thus, the proposed method is effective for samples that may be used as evidence in court.

Table 1. Experimental conditions

Item		**Condition**
Speech source	Recorded voice	Police officer: 1 Suspect: 3 Total: 3 speaker combinations
	Speech volume	60 dB
	Sampling rate	16 kHz
	Channel	Monaural
	Frame length	512 samples
Added noise	Noise type	Duct noise
	Level	50 dB to 70 dB in increments of 10
Watermark	Method	Direct Spread Spectrum
	Spread code	M-sequence
	Frame length	512
	Bitstream	72
	Chip rate	7

Table 2. Performance with Tampering Pattern 1

SNR [dB]	**SEN [%]**		**SPC [%]**	
	LSP	**TIME**	**LSP**	**TIME**
-10	**77.78**	**100.00**	**31.74**	**0.85**
0	**77.78**	**100.00**	**98.41**	**87.68**
10	**88.89**	**100.00**	**99.24**	**94.90**

Table 3. Performance with Tampering Pattern 2

SNR [dB]	**SEN [%]**		**SPC [%]**	
	LSP	**TIME**	**LSP**	**TIME**
-10	**77.78**	**100.00**	**31.71**	**0.71**
0	**77.78**	**100.00**	**98.41**	**87.63**
10	**88.89**	**100.00**	**99.23**	**94.84**

In the case of tampering pattern 2, where a part of the interrogation speech was rewritten by the same speaker, the results were similar to those of tampering 1 in both sensitivity and specificity. In terms of sensitivity, we were able to identify the tampered position using either timestamp data or fingerprints. In terms of specificity, the accuracy decreased significantly when the noise became larger (SNR -10 dB).

In the case of tampering pattern 3, where a part of the interrogation speech was substituted for another speaker, the sensitivity was 100% for both timestamp data and fingerprints at all SNRs. Thus, fingerprints were also able to accurately identify tampered frames. Next, the specificity resembles the trend with tampering pattern 2. However, some decrease is seen at 0 dB compared to tampering 2. This may be caused by the sudden change of speaker and added noise characteristics, resulting in discontinuities between altered and unaltered frames.

Overall, it can be concluded that timestamps are effective for identifying tampered frames since they result in 100% sensitivity with all levels of noise tested here. However, fingerprints may be able to identify tamper-free frames more accurately than timestamps

Table 4. Performance with Tampering Pattern 3

SNR [dB]	SEN [%]		SPC [%]	
	LSP	TIME	LSP	TIME
-10	**100.00**	**100.00**	**28.38**	**0.77**
0	**100.00**	**100.00**	**90.80**	**57.77**
10	**100.00**	**100.00**	**99.43**	**96.23**

when the noise level is higher. Thus, these two types of watermark data may be able to complement each other in different aspects.

6. Conclusion.

We proposed a method for detecting tampering of audio recordings using both fingerprints and timestamp data as watermarks. In the recorder, fingerprints and timestamp data extracted from the recorded audio are embedded in the recorded audio as a tamper-evident watermark. At the decoder, the fingerprints were compared with those extracted from the watermarked speech, and the extracted timestamp data was tested for continuity in order to detect tampering. For the detection of the watermarks, the estimated clean speech was used in place of clean speech to achieve blind detection. The vector distance between the fingerprints obtained from the watermark and the fingerprints obtained from the recorded speech was calculated. If this distance was larger than a predetermined threshold, it was determined to be tampered with, and if it was smaller, it was determined to be the authenticity of the recording. The timestamp data obtained from the watermark was checked for continuity, and if the value was continuous before and after each frame, it was determined that the data had not been tampered with, proving the continuity of the audio data.

The effectiveness of the proposed method for a realistic interrogation scene between a police officer and a suspect was tested. To simulate various noise generated in the actual environment, duct noise was played during the recording. Next, the audio recording of the interrogation was manipulated to change the content to be more convenient for the police officer. We performed tampering detection on the tampered speech and found that the detection accuracy was at least 90% in a realistic noise environment where the speech evidence is audible, which is sufficient for practical use. It was found that timestamps can identify tampered frames at all levels of noise, while it may be able to detect tamper-free frames when the noise level was low. However, the detection using fingerprinting was proven to be more resistant to noise. Thus, these two types of data can complement each other under all levels of noise in the environment.

Three major issues need to be further addressed. The first is to reduce the effect of the synchronization signal on the sound quality. The fact that the synchronization signal is added to all frames is considered to be a major factor in the sound quality degradation. As a countermeasure, there is a method of adding the synchronization signal to several frames at 1 s intervals. The second issue is the evaluation of the system in various recording situations and noises. Other types of noise, especially non-stationary, as well as reverberations may need to be evaluated. Finally, we would like to develop a system that can automatically and accurately identify the type of tampering applied to the sample.

REFERENCES

[1] H. Zhao, Y. Chen, R. Wang, and H. Malik, *Audio splicing detection and localization using environmental signature*, Multimedia Tools Appl, pp. 1–31 (2016).

[2] S. Wang, W. Yuan, J. Wang and M. Unoki, *Speech Watermarking Based on Source-filter Model of Speech Production*, JIH-MSP, Vol. 10, No. 4, pp. 517- 534, (Dec. 2019).

[3] S. Wang, R. Miyauchi, M. Unoki and N. S. Kim, *Tampering Detection Scheme for Speech Signals using Formant Enhancement based Watermarking*, JIH-MSP, Vol. 6, No. 6, pp. 1264-1283 (Nov. 2015).

[4] Takahashi. S, Kondo. K, *An Interrogation Speech Manipulation Detection Method using Speech Fingerprinting and Watermarking*, IIH-MSP 2018, SIST 110, pp. 55-62, Springer (2019).

[5] Muroi, K., Kondo, K., *Speech Manipulation Detection Method using Audio Watermarking*, IEEE Global Conf. on Consumer Electr. (Oct. 2021)

[6] Sugamura. N, Itakura. F, *Speech data compression by LSP analysis-synthesis technique*, Trans. of the Inst. of Electronics, Inf., and Commun. Eng. J64-A(8) (Aug.1981), in Japanese

[7] NIST, *Advanced Encryption Standard (AES)*, FIPS 197 (Nov. 2001)

[8] Muroi, K., Kondo, K., *Speech Manipulation Detection Method Using Speech Fingerprints and Timestamp Data*, IIH-MSP 2021 (Oct. 2021)

[9] Kondo, K, Yamada, J., *A Frame Synchronization Method for Audio Watermarks Robust Against Analog Aerial Transmission*, J. Inf. Hiding and Multimedia Sig. Process., 8(5), pp. 1043-1053 (Sept. 2017)

[10] Boney. L, Tewkfik. A.H, Hamdy. K.N, *Digital watermarks for speech signals. In: Proc. IEEE Int. Conf. on Multimedia Comp. and Sys. IEEE*, Kaohsiung, Taiwan, Hiroshima (1996).

[11] Miyazaki, R., Saruwatari, H., Shikano, K., Kondo, K., *Musical-noise-free speech enhancement based on iterative Wiener filtering* 12th IEEE Intern. Symp. on Signal Process. and Inf. Tech., (Dec. 2012)

[12] S.V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, Second Edition, Wiley, 2000.