Journal of Information Hiding and Multimedia Signal Processing ©2017 ISSN 2073-4212
Ubiquitous International Volume **8**, Number **5**, September **2017**

A Frame Synchronization Method for Audio Watermarks Robust Against Analog Aerial Transmission

Kazuhiro Kondo

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

Joji Yamada

Alpine Giken Inc.
1-58 Yoshima-Kogyodanchi, Iwaki, Fukushima 9701144, Japan

Received December, 2016; revised June, 2017

Abstract. *We evaluated the effect of frame synchronization on the detection accuracy of embedded data using audio data hiding. We investigated a simple frame synchronization signal which is added to the host audio signal. The synchronization signal was generated by using an M-sequence for the phase, and shaping its magnitude relative to the host signal spectrum. Frame synchronization at the detector was achieved using the cross-correlation function with the same M-sequence used to generate the synchronization signal, and detecting the peak of this function. This synchronization signal was added to a spread-spectrum watermarked audio signal. This signal was then transmitted over an analog channel, by connecting the analog signal from the digital to analog converter to the analog to digital converter by an electrical cable, and also by playing the analog signal through a loudspeaker and recording this signal using a microphone. The addition of the synchronization signal allowed the frame position to be mostly recovered at the decoder, and enabled the detection of the watermark through analog channels at acceptable accuracy. The frame synchronization rate was also shown to strongly influence the watermark bit error rate (BER).*

Keywords: Audio watermark, frame synchronization, M-sequence, Cross-correlation function, Analog aerial transmission

1. **Introduction.** Watermarking methods for audio signals have been studied for some time now. Initially, these audio watermarks were intended to be used to control the illegal distribution of copyrighted material. However, some of the recent proposed methods are intended to be used to hide supplementary data into audio which will add value to the digital audio content. For example, Ito et al. [1] attempted to hide supplementary data that will potentially improve the quality of the high frequency band that is estimated from the low frequency band signal of a narrowband audio signal, i.e., bandwidth extension of a narrowband signal. They attempted to achieve this by hiding Line Spectrum Pair (LSP) coefficients that accurately describes the spectrum envelope of the high frequency band into the narrowband signal. The extracted LSP coefficients were then used to shape

the spectrum envelope of the high frequency band signal created by bandwidth extension from the received narrowband signal to improve the audio quality.

Aoki [2, 3] also attempted to hide data that will facilitate concealment of lost audio segments in packet audio. His proposal hides the pitch and gain variation pattern of audio in the previous packet. If the previous packet is lost, these patterns are recovered from the current packet, and are used to compensate the pitch and gain variation of the recreated signal by repeating a pitch interval in the packet preceding the lost packet, potentially improving the audio quality of the recreated signal.

These two examples obviously do not assume that the data-hidden audio signal will be transmitted over an analog channel. Thus, detection of hidden data from analog transmitted signal is apparently not considered in their methods.

However, there are examples where the data-hidden signal is assumed to be transmitted over some analog channels, and the detection of hidden data is attempted on this signal. For example, Matsuoka et al. [4, 5] attempted to hide the URL of the web site of the artist of the host musical signal. One of their goals is to implement their method on a smartphone, and have the user record a portion of the music that they may be interested in, being played out from a loudspeaker. This will trigger a web browser which will show promotional information about this piece of music, such as artist or album information.

There are also existing products which attempt to extract data from audio being played out from digital signages. Dai Nippon Printing Co. (DNP) is marketing QUick and Easy Media Access (QUEMA) [6] which is implemented on a smartphone, and can extract data from audio being played out from digital signages. Digimarc Corp. also is marketing the Digimarc Discover Platform [7], which is a multimedia platform that can extract hidden information from video, audio and images. This again is implemented on multiple platforms, including smartphones, and can be used to extract data from cameras and microphones on the smartphones.

Nishimura also attempted to embed the lyrics of a Karaoke song using data hiding [8]. The lyrics can then be decoded and displayed on a mobile terminal located near the singer by continuously recording the Karaoke being played out from the loudspeaker, at the exact timing of the song when the lyrics should be sung.

The detection of hidden data from a watermarked audio signal that go through analog channels may be a challenge for a number of reasons. Various types of noise may be added during the analog transmission. Non-linear distortion can also be added to the signal due to the analog circuitry associated with the A/D and the D/A conversion. Timing skews also may be added due to unstable clock signals. Since most watermarks embed data in frames, maintaining frame position at the detector is essential, and may lead to a large number of errors in the detected watermark if not recovered accurately. However the framing within the host signal is lost due to analog conversion, and the frame position must be recovered at the detector.

Ono has attempted to embed the frame position by adding an M-sequence with spectral shaping to match the spectrum of the host signal [9]. This is similar to the method that we will propose and evaluate in this paper. However, Ono seems to have not evaluated its performance when the embedded signal is transmitted over analog channels, which is the focus of this paper.

So far, there have been very few efforts reported on the investigation of the effect of analog transmission on embedded watermarks. For example, Hiratsuka et al. investigated the addition of synchronization codes to the host signal for frame alignment of a patchwork watermark [10]. They evaluated the synchronization accuracy with analog loop back (ALB), but not aerial transmission (AT). Karnjana et al. have recently evaluated the effectiveness of synchronization codes embedded in the LSBs of the host signal, and also

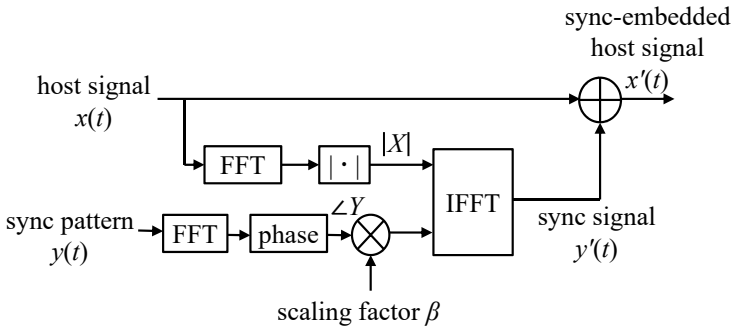

FIGURE 1. Synchronization Pattern Embedding Method

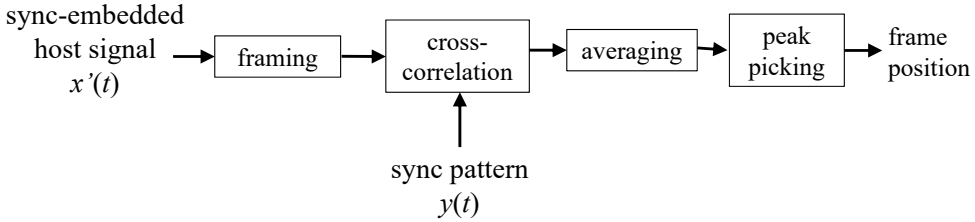

FIGURE 2. Frame Synchronization Recovery Method

addition of random synchronization patterns [11, 12]. However, they have not evaluated their methods in analog channels. Xiang proposed a synchronization signal robust to D/A and A/D conversions, in which pseudo-random code is embedded into the low frequency subbands of Wavelet coefficients [13]. However, this method has not been evaluated for robustness to aerial transmission.

In this paper, we investigate a simple frame position recovery method which adds random synchronization pattern in the phase of the host signal. We study the effectiveness of this method by trying to recover the frame position from received signals after analog transmission, including AT. We also study the effect of frame position recovery on the BER of the recovered watermark data.

This paper is organized as follows. The next section will describe the frame synchronization method investigated in this paper. This will be followed by the evaluation of robustness to analog channels, including aerial transmission, and its results. Finally, the conclusion and issues will be given.

2. **Frame Synchronization for Audio Data Hiding.** In this paper, we investigate a frame synchronization method for audio watermarks based on additive M-sequence synchronization pattern. As shown in Fig. 1, a synchronization signal is added to the host signal. The phase of the additive signal is the M-sequence pattern, while the envelope of the signal is shaped according to the host signal, thereby improving the imperceptibility of the added signal.

The recovery process of the frame position is depicted in Fig. 2. The received signal with unknown frame position is split into interim frame positions with a fixed frame length, which we assume is known here. Then, the cross-correlation between these frame signals and the same M-sequence added in Fig. 1 is calculated. This cross-correlation is averaged over multiple frames, smoothed to remove variations due to noise signals, and the peak position is picked from this signal to detect the frame starting position.

2.1. **Cross-correlation functions for frame synchronization.** The cross-correlation function, $g_{xy}(\tau)$, between the received signal, $x'(t)$, and the M-sequence for synchronization, $y(t)$, can be defined as follows:

$$g_{xy}(\tau) = \int x'(t)y(t+\tau)\,dt \tag{1}$$

where τ is the lag time.

The Fourier Transform of the above function gives the cross-power spectrum between x' and y, $G_{xy}(\omega)$. The Generalized Cross-Correlation Function (GCCF) can then be defined as follows:

$$R_{xy} = \int \Psi(\omega)G_{xy}(\omega)e^{i\omega t}\,d\omega \tag{2}$$

where $\Psi(\omega)$ is the generalized weighting function.

The weighting function needs to be selected appropriately so that the peaks indicating the correct frame position within the interim frames will become apparent. The following weighting functions were compared.

(1) Cross-power Spectrum Phase (CSP)

The CSP method has been commonly used to detect the sound source direction from the cross-correlation function. The weighting function for this method can be defined as follows:

$$\Psi(\omega) = \frac{1}{|G_{xy}(\omega)|} \tag{3}$$

Since this method normalizes the cross-correlation function with its magnitude, the resultant GCCF mainly reflects the mutual phase between the two signals.

(2) Smoothed Coherence Transform (SCOT)

The SCOT method uses the coherence function, and is also a commonly used method for sound source direction estimation. The weights used for the SCOT method is as follows:

$$\Psi(\omega) = \frac{1}{\sqrt{G_{xx}(\omega)G_{yy}(\omega)}} \tag{4}$$

This method can also be regarded as whitening of the cross-correlation function. Since the correlation is normalized using the power spectrum of the two signals, we can expect the resultant function to be able to detect the out-of-phase components between the signals even under low SNR conditions.

(3) No weighting

For comparison with CSP and SCOT, we also included using the cross-correlation function with no weighting, i.e.,

$$\Psi(\omega) = 1 \tag{5}$$

We used 70 clips from the SQAM database published by the European Broadcasting Union (EBU) [14] to compare the above weighting functions for the GCCF. All clips were re-sampled at 16 kHz, 16 bits per sample, with a single channel (i.e., monaural). We selected 10 s non-silence portions from each clip. The synchronization pattern used here

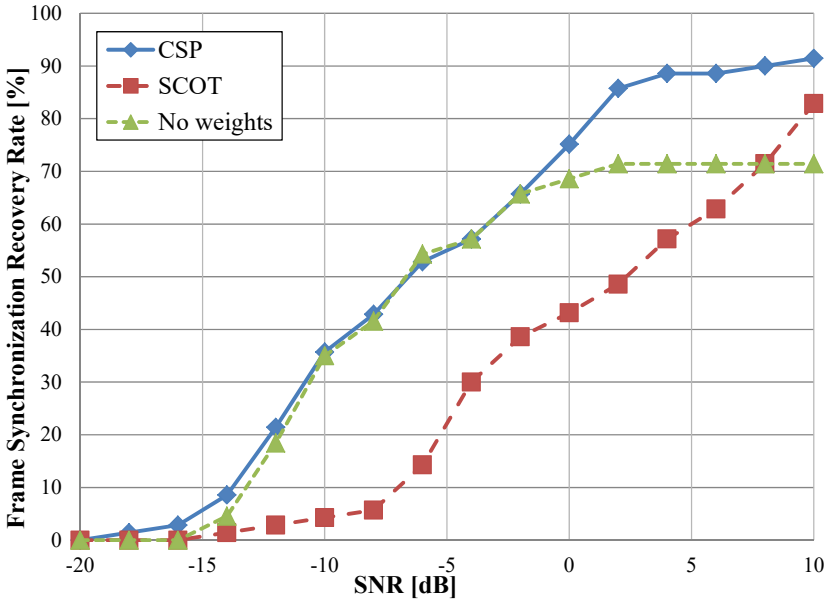

FIGURE 3. Comparison of GCCF Weighting for Frame Synchronization Recovery

was an M-sequence with a length of 2047 bits, and the frame length was set to 1 s (16,000 samples).

The synchronization-embedded signal was converted to analog, and fed back to the D/A circuitry as an electrical signal to be re-digitized. We also added white noise to this re-digitized signal at various SNR levels to emulate the additive noise during analog transmission. The frame synchronization recovery was attempted on this signal from the GCCF with the three weighting functions described above. Fig. 3 shows the SNR vs. frame synchronization recovery rate for the three weighting functions.

As can be seen in this figure, the CSP shows the highest recovery rate over all SNR levels tested. The SCOT method, which was expected to perform well with poor SNR signals, showed a low recovery rate over all SNR range. On the other hand, using no weights showed a comparable recovery rate with the CSP method, except when the SNR was relatively high. CSP seems to show better recovery rate than no weighting at higher SNR because when the noise level is low, the weighting function suppresses the frequency component of the host signal, enabling the extraction of the synchronization signal in the phase more stable, while at lower SNR, where additive noise becomes dominant, the weighting does not make a difference. From these results, we will be using CSP in all further evaluation.

2.2. **Smoothing of cross-correlation functions.** The GCCF estimated as stated above still contains additive noise and host signal components that result in false peaks in the function. Thus, we attempt to mitigate the effect of these false peaks by smoothing the GCCF. Since the additive noise seems to result in non-biased white noise-like characteristics in the resultant GCCF, we can expect simple averaging to reduce the effect of this noise. From preliminary experiments, we found that the Weighted Moving Average (WMA), which applies a linearly decreasing weight on forward and backward neighboring samples gives best results. We empirically found that using two samples in both the forward and backward direction as shown in the following formula gives a reasonable amount of smoothing.

$$\hat{s}_i = \frac{s_{i-2} + 3s_{i-1} + 5s_i + 3s_{i+1} + s_{i+2}}{13} \tag{6}$$

where s_i is the raw GCCF at lag time instance i, and $\hat{s}_i$ is the smoothed GCCF.

3. **Evaluation of Frame Synchronization for Analog Transmission.** We evaluated the effect of frame synchronization on the detection of data embedded using audio data hiding technology in analog transmission. We evaluated the effect in both the analog loopback (ALB), which does not include aerial transmission, and aerial transmission (AT) of analog signals. The frame synchronization rate with the synchronization method, and also the BER of the detected watermark which is embedded simultaneously with the synchronization signal was evaluated.

3.1. **Experimental conditions.** The audio sources used here were 70 clips (tracks 1 through 70) from the SQAM database [14] published by the EBU, from which initial 10 s non-silent portions were selected. All samples were re-sampled at 16 kHz, 16 bits per sample, monaural.

The synchronization pattern used was the M-sequence with a length of 2047 bits. The frame length was 1 s, and the phase components of the first 2047 samples were set to a scaled value of this M-sequence, while the remaining samples were set to 0. The spectrum envelope of the host signal was used as the spectrum envelope of this synchronization signal as well. Then, this spectrum is converted into the time domain, and added to the host signal. The scale factor was set empirically to 0.08 to maintain the inaudibility of the synchronization signal. The average Objective Difference Grade (ODG) estimated using the Perceptual Evaluation of Audio Quality (PEAQ), ITU standard ITU-R BS.1387 [15], over 20 randomly selected samples out of the 70 clips, was -1.92. This level of ODG is in the subjective category of "slightly annoying".

At the receiving end, we computed the GCCF with CSP weighting, smoothed this with WMA, and detected the peak in the resulting function to identify the frame position within the interim frame.

We also embedded watermarks to this synchronization signal embedded audio using the classical direct spread spectrum (DSS) method [16], attempted to recover the frame position, and evaluated the BER of the detected watermark. DSS watermarking was selected since this method is one of the methods known to be prone to frame mis-synchronization [17, 18]. An M-sequence with a length of 4095 bits was used as the spread code for DSS, with a chip rate of 10, and the generated DSS signal was added to the host signal after scaling. The scaling factor again was set to 0.08 empirically, optimized for its effect on audio quality.

The experimental set up for the ALB experimentation is shown in Fig. 4. The synchronization signal and watermark embedded audio signal is converted to analog using the D/A converter on the PC sound card, played out from the line-out plug, fed back to the line-in plug using an audio cable, and re-converted to digital signal using the A/D converter on the same card. Thus, no electrical-acoustic conversion is applied. White noise was added at various SNR to the re-digitized signal to emulate additive environmental noise. The noise level was adjusted so that the SNR would range from 0 to 50 dB in 5 dB intervals.

Fig. 5 shows the experimental setup for the aerial transmission experimentation. In this experiment, the analog signal is actually played out from the loudspeaker, recorded with a microphone placed 10 cm from the loudspeaker, and re-digitized. Thus, the analog electrical signal is converted to an audio signal and vice-versa. The rest of the setup is the same as the ALB setup. The audio source used in this experiment was the female sentence speech from the SQAM database [14]. A 10 s non-silent interval was selected

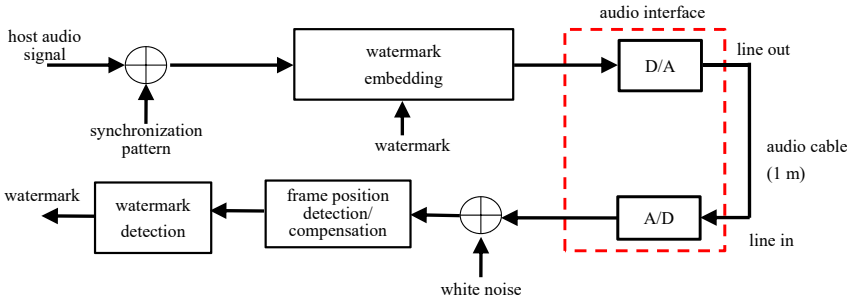

FIGURE 4. Experimental Setup for Analog Loop Back (ALB)

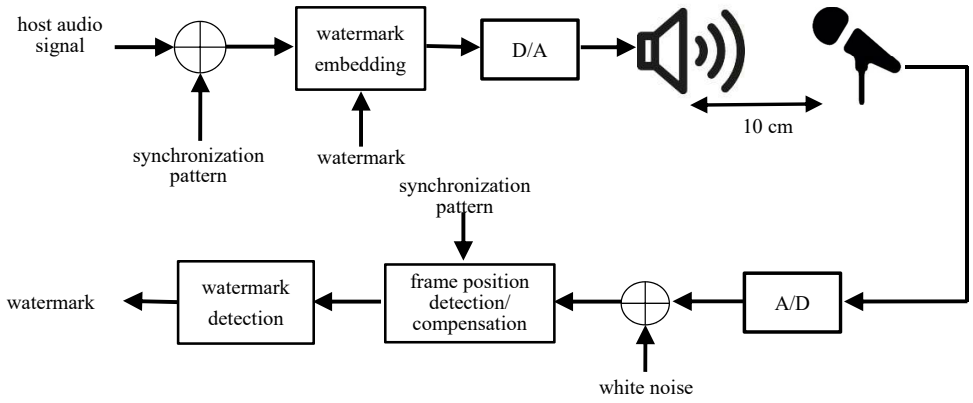

FIGURE 5. Experimental Setup for Aerial Transmission (AT)

from this clip. The D/A and the A/D sampling rate was 44.1 kHz, 16 bits per sample, monaural.

3.2. **Results and discussions.**

3.2.1. *Robustness to analog loopback.* Fig. 6 shows the frame recovery rate for ALB vs. SNR of the received signal with additive white noise. We define the frame synchronization recovery rate as the ratio of the number of sources for which the frame position was detected within ±10 samples from the "true" frame position vs. the total number of tested sources. A tolerance of ±10 samples were allowed here since this level of synchronization error was shown not to decrease the watermark BER significantly through preliminary experiments, and also because many of the detected frame positions were just a few samples from the "true" positions. As can be seen, with ALB, the frame position can be recovered for close to 100% of the time if the SNR is above 20 dB, while this rate gradually decreases as SNR becomes lower, eventually to about 65% at SNR 0 dB.

Fig. 7 compares the detected watermark BER, with and without synchronization. For detection without synchronization, we simply took the first recorded sample as the first sample in the frame, up until the length of the frame, processed this as the first frame, and so on. The BER is about 50% without synchronization, which basically means that virtually no watermark can be recovered. With synchronization, however, the BER becomes lower than 15% at SNR of 20 dB, and below 10% if the SNR is above 30 dB. At this level of BER, we can expect the bit errors to be virtually eliminated using error correction codes. Thus, we can state that the use of synchronization signals is very effective in lowering the BER of embedded watermarks. We can also see that the frame recovery rate is correlated with the watermark BER, i.e., the higher the recovery rate, the lower the BER. This result is with the classic DSS watermarks, which is known to be affected by frame synchronization. However, other data hiding methods are also affected

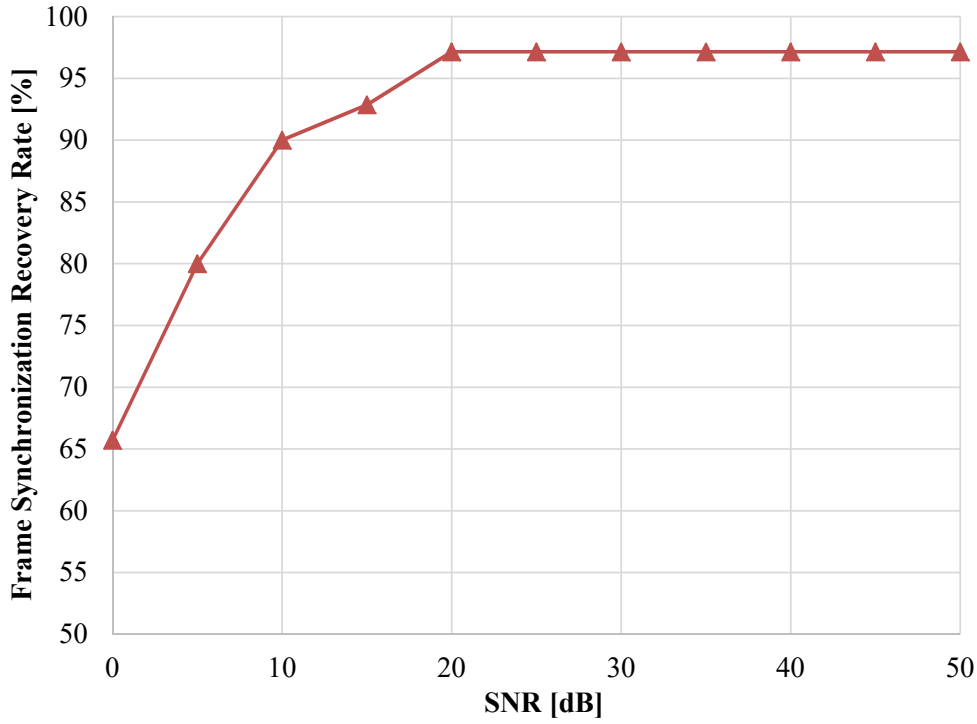

FIGURE 6. Frame Synchronization Recovery Rate for ALB with Additive White Gaussian Noise

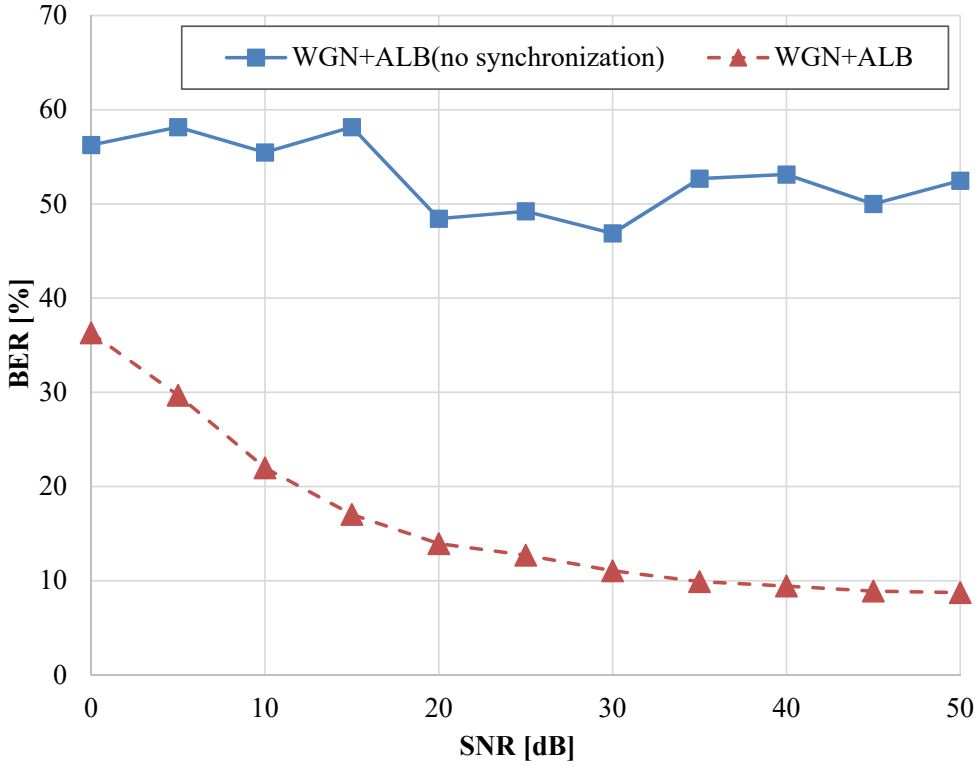

FIGURE 7. BER for ALB with Additive White Gaussian Noise

to some extent by frame synchronization, and so successful frame synchronization recovery should also be advantageous in the successful data recovery with these methods as well.

3.2.2. *Robustness to analog aerial transmission.* Fig. 8 compares the detected watermark BER vs. SNR with ALB and AT. With both ALB and AT, the BER generally decreases

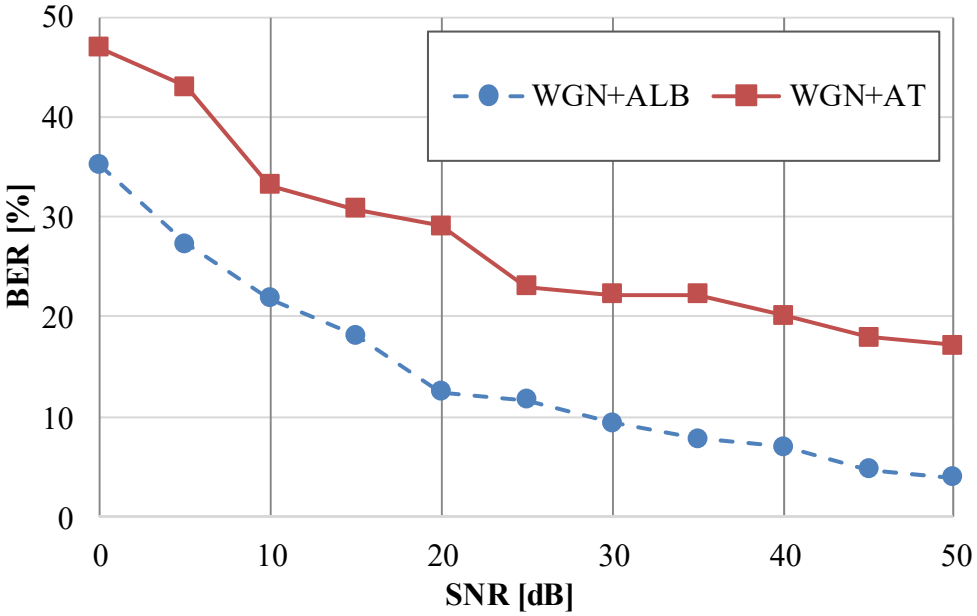

FIGURE 8. BER for AT and ALB with Additive White Gaussian Noise

TABLE 1. BER Comparison Between ALB and AT With and Without Frame Synchronization

frame synchronization	transmission mode	BER [%]
yes	ALB	2.34
	AT	16.41
no	ALB	52.34
	AT	50.78

as the SNR increases. The BER for AT is constantly lower by about 10% compared to ALB for all SNR. However, it is clear that frame synchronization recovery is very effective for AT as well, even though AT includes various non-linear distortions associated with electrical-acoustic conversions.

Table 1 tabulates the detected watermark BER with and without synchronization, both for ALB and AT transmission modes. In all cases, white noise was not added. As can be seen, for both ALB and AT, the BER is above 50% if synchronization signal is not added. However, with synchronization, the BER dramatically decreases, for both ALB and AT. Again, the BER is about 14% lower for ALB compared to AT. This again shows that the effect of non-linear distortion caused by the electrical-acoustic conversion affects the BER significantly.

4. **Conclusion.** We investigated the effect of frame synchronization on the BER of detected audio watermarks. Frame synchronization signal was added to the host audio signal. The phase of this signal was an M-sequence, and its spectrum envelope was shaped according to the envelope of the host signal. This signal was scaled, converted to time-domain, and added to the host signal. In order to recover the frame position at the decoder, a generalized cross-correlation function (GCCF) between the received signal and the same M-sequence used to generate the synchronization signal at the transmitter was calculated, smoothed, and its peak is detected to recover the frame position. This frame synchronization signal was added to a watermarked host signal to test the effect of synchronization on the BER of this watermark. We evaluated the BER using the classic

direct spread-spectrum (DSS) watermark, which is known to be affected by frame misalignment. Two modes of analog transmission were tested; (1) analog loopback (ALB) which feeds back the analog signal as electrical signal and does not include an electrical-acoustic conversion, and (2) analog aerial transmission (AT), which converts the analog signal to an acoustic signal using loudspeakers, and reconverts this signal to an electrical signal using a microphone placed 10 cm from the loudspeaker.

The frame recovery rate with and without synchronization in ALB was compared. It was found that with synchronization, if the SNR is above 20 dB, the frame position can be recovered almost perfectly.

We then tested the BER of the detected DSS watermark at the decoder with analog conversion. Without synchronization, it was not possible to detect any watermarks, with both ALB and AT. However, if we do use synchronization, the BER can be decreased to about 2% for ALB, and 16% for AT. This difference is apparently caused by the non-linear distortion associated with the electrical-acoustic conversion with AT. It was also found that the frame synchronization rate is correlated with BER, showing that the effect of synchronization is significant on the watermark BER.

We would further like to try to add error correction codes along with the synchronization signal to further decrease the BER. We also would like to further test other cross-correlation functions that enable us to reliably detect the frame position as peaks. Evaluation with other up-to-date data hiding methods is also necessary. We would also like to test with actual additive acoustical noise, including reverberation, to test its effect on the BER.

Acknowledgment. Part of this work was carried out under the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University.

REFERENCES

[1] A. Ito and Y. Suzuki, Advanced Information Hiding for G.711 Telephone Speech, *Multimedia Information Hiding Technologies and Methodologies for Controlling Data*, K. Kondo (ed.), Hershey, PA, USA, IGI Global, pp. 129–163, 2012.

[2] N. Aoki, A Packet Loss Concealment Technique for VoIP using Steganography, *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, Awaji Island, Japan, pp. 470–473, 2003.

[3] N. Aoki, Enhancement of Speech Quality in Telephone Communication by Steganography, *Multimedia Information Hiding Technologies and Methodologies for Controlling Data*, K. Kondo (ed.), Hershey, PA, USA, IGI Global, pp. 164–181, 2012.

[4] H. Matsuoka, Y. Kakashima, T. Yoshimura and T. Kawahara, Acoustic OFDM: Embedding High Bit-Rate Data in Audio, *Advances in Multimedia Modeling: 14th International Multimedia Modeling Conference*, S. Satoh, F. Nack and M. Etoh (eds.), Berlin/Heidelberg, Springer, pp. 498–507, 2008.

[5] H. Matsuoka, Acoustic OFDM Technology and System, *Multimedia Information Hiding Technologies and Methodologies for Controlling Data*, K. Kondo (ed.), Hershey, PA, USA, IGI Global, pp. 90–103, 2012.

[6] DNP, QUEMA for smartphone support site, http://www.quema.info/site/support_en.htmlD Browsed in Dec., 2016.

[7] Digimarc Corp., Digimarc Discover Platform audio technical overview, https://www.digimarc.com/docs/default-source/solution-briefs/audiotechnicalbrief.pdf?sfvrsn=4. Browsed in Dec., 2016.

[8] A. Nishimura, Presentation of Information Synchronized with the Audio Signal Reproduced by Loudspeakers Using an AM-based Watermark, *Proc. of the International Conference on Intelligent Information Hiding and Multimedia Signal Processsing*, pp. 275–278, Kaohsiung, Taiwan, 2007.

[9] N. Ono, Robust audio information hiding based on stereo phase difference in time-frequency domain, *Proc. of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 263–260, Kitakyushu, Japan, 2014.

[10] K. Hiratsuka, K. Kondo, and K. Nakagawa, On the accuracy of estimated synchronization positions for audio digital watermarks using the modified patchwork algorithm on analog channels, *Proc. of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 628–631, Harbin, China, 2008.

[11] J. Karnjana, P. Nhien, S. Wang, N.M. Ngo, and M. Unoki, Comparative study on robustness of synchronization information embedded into an audio watermarked frame, *IEICE Technical Report*, no. EMM2015-83, IEICE, pp. 41–44, Yakushima, Kagoshima, Japan, March 2013.

[12] J. Karnjana, M. Unoki, P. Aimmanee, and C. Wutiwiwatchai, Audio watermarking scheme based on singular spectrum analysis and psychoacoustic model with self-synchronization, *Journal of Electrical and Computer Engineering*, vol. 2016, 2016.

[13] S. Xiang, Audio watermarking robust against D/A and A/D conversions, *EURASIP Journal on Advances in Signal Processing*, vol.2011, no.1, Dec. 2011.

[14] European Broadcasting Union, Sound Quality Assessment Material (SQAM), CD.

[15] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, PEAQ - the ITU standard for objective measurement of perceived audio quality, *Journal of the Audio Engineering Society*, pp. 3–29, Jan.–Feb. 2000.

[16] L. Boney, A.H. Tewfik, and K.N. Hamdy, Digital watermarks for audio signals, *Proc. Int. Conf. on Computing and Systems*, pp. 473–480, Hiroshima, Japan, 1996.

[17] K. Kondo and K. Nakagawa, A digital watermark for stereo audio signals using variable inter-channel delay in high-frequency bands and its evaluation, *International Journal of Innovative Computing, Information and Control*, vol. 6, no. 3 (B), pp. 1209–1220, 2010.

[18] K. Kondo, A Data Hiding Method for Stereo Audio Signals Using the Polarity of the Inter-Channel Decorrelator, *Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 86–89, Kyoto, Japan, 2009.