PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Robust long-distance aerial audio data hiding: Comparison between amplitude modulation-based hiding and bilateral time-spread echo hiding

Akira Nishimura1(1)

(1)Tokyo University of Information Sciences, Japan, akira@rsch.tuis.ac.jp

Abstract

Voice evacuation and mass notification systems that use outdoor loudspeaker systems play an important role in propagating emergency information in Japan. However, informing the hearing impaired and the elderly who have difficulty in hearing as well as tourists and foreigners who do not understand Japanese remains a serious problem. Therefore, aerial data transmission combined with voice evacuation messages with the aid of information hiding technology should be considered. The bilateral time-spread echo hiding method has been proposed for such a system, in which speech signals are broadcast by the outdoor loudspeakers of a voice evacuation and mass notification system. In addition, amplitude modulation-based watermarking has been shown to be robust against reverberation and background noise. Evaluations of both methods for speech signals were conducted using computer simulations, including several disturbances caused by the long-distance (70—800 m) aerial transmission of sounds. The frequency response of a distant horn-array loudspeaker system, the absorption of sound by the atmosphere, reverberation, a single long-path echo, and additive background noise are simulated as disturbances. The results demonstrated that the bilateral time-spread echo hiding method was more robust against the disturbances than the amplitude modulation-based method.

Keywords: Watermarking, Bit error rate, Reverberation, Background noise, Air absorption

1 INTRODUCTION

This study aims to compare audio data hiding methods—the improved time-spread echo hiding method [8] and the method based on subband amplitude modulation [7]—for speech signals broadcast by the outdoor loudspeakers of a voice evacuation and mass notification system. Evaluations of the data hiding system for speech signals are conducted by computer simulations, including several disturbances caused by the long-distance (several hundreds of meters) aerial transmission of sounds.

Large-scale disasters, such as earthquakes, tsunamis, terrorism, and the firing of ballistic missiles, are occasionally a public threat in Japan. Voice evacuation and mass notification systems that use outdoor loudspeaker systems play an important role in propagating emergency information. However, one severe problem is informing the hearing impaired and the elderly who have difficulty in hearing as well as informing tourists and foreigners who do not understand Japanese. Audio data hiding technologies combined with mass notification systems are promising systems that can transmit both hidden information and voice messages via speech sounds emitted from a loudspeaker. The received and decoded information is displayed on a digital terminal, such as a smartphone.

Some aerial audio watermarking technologies have been developed for use in an evacuation and mass notification system [12, 11, 4]; however, they are not suitable for outdoor long-distance transmission. They do not take into account strong background noise, frequency response of outdoor horn loudspeakers, or the absorption of sound by the atmosphere.

The improved time-spread echo hiding method has been demonstrated to be effective for outdoor long-distance aerial transmission [8]. In addition, an audio watermarking technique based on amplitude modulation [7] can be applied to speech signals and is robust against aerial transmission [5, 6].

The following section briefly describes the two audio data hiding methods. The subsequent section describes a realistic simulated environment that considers typical disturbances in the long-distance aerial transmission

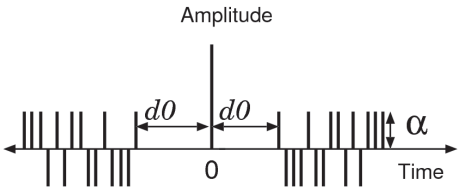

Figure 1. An example of bilateral echo kernel: $L = 15$.

of sounds emitted from outdoor loudspeakers of voice evacuation and mass notification systems [9, 8]. The effectiveness of the methods in long-distance aerial transmission for the voice evacuation system are evaluated by computer simulation.

2 AUDIO DATA HIDING

This section introduces the improved bilateral time-spread echo hiding method [8] and amplitude modulation method [7].

2.1 Bilateral time-spread echo hiding

Time-spread echo hiding [3] is an extended method of echo hiding [2], providing greater security, robustness, and perceptual transparency. The embedding process divides the cover signal $s(n)$ into $F + T_r$ samples, with an overlap of T_r samples. The stego signal $r(n)$ is obtained by the convolution of the framed cover signal $s(n)$ with the impulse response $k(n)$, consisting of the Dirac delta function $\delta(n)$ and an echo kernel $P(n)$ of length L and delay time $d0$. Echo kernel $P(n)$ is defined by a pseudo-random (PN) sequence or M-sequence signal.

$$k(n) = \delta(n) + \alpha P(n - d0), \tag{1}$$

$$r(n) = s(n) * k(n), \tag{2}$$

where $*$ denotes convolution and α is the echo gain. Subsequently, the framed stego signals are concatenated with initial and final overlaps of $T_r/2$ samples each. The payload data are embedded in every segmented frame signal.

The current hiding scheme circularly shifts $P(n)$ to obtain $P'(n)$, depending on the integer value m encoded by payload bits, as follows:

$$P'(n) = \begin{cases} P(n+L-m) & (\ 1 \le n \le m\ ), \\ P(n-m) & (\ m+1 \le n \le L\ ). \end{cases} \tag{3}$$

where m is quantized by m' steps, i.e., $m \in \{0, m', 2m', ..., \lfloor L/m' \rfloor m'\}$, to be robust against frequency shifts of the stego signal. Consequently, the amount of payload is $\log_2(\lfloor L/m' \rfloor + 1)$ per segmented frame.

Chou and Hsieh [1] applied the bilateral symmetric echo kernel technique to the time-spread echo kernel. Figure 1 presents an example of the impulse response of the bilateral symmetric time-spread echo kernel, which is, henceforth, called the bilateral method. The detection gain, which represents the gain in the peak of cross-correlation exhibits a maximum value when the echo gain $\alpha = 1/(2\sqrt{L})$. It achieves approximately 1.7 times greater detection gain than that of the conventional unilateral condition, resulting in better performance [9, 1]. A drawback of the bilateral echo kernel is the degradation of imperceptibility; however, that is less important in the broadcasting of voice messages in this study.

In the detection process, the real cepstrum transform of the stego signal is calculated using the absolute spectrum obtained from DFT (discrete Fourier transform). The cross-correlation of the real cepstrum and echo kernel sequence $P(n)$ exhibits a peak in the amount of circular shifting m.

Frame synchronization can be realized by observing the maximum amplitude of the cross-correlation function. The recorded sound is segmented to the frame signal of length F samples. Then, cross-correlation functions are calculated by shifting the frame location by $F/8$. The decoded data from the frame that exhibits the maximum peak in the cross-correlation function obtained over eight frames is the most reliable data to be decoded because if the frame signal to be decoded spreads over the different sub-frame pairs, then the peak of the cross-correlation function is attenuated by the cancellation in the cepstrum domain.

For speech signals, the middle frequency region (200 Hz – 3 kHz), where the first and second formants exist, tends to have relatively strong power. Under low SNR conditions, suppressing low- and high-frequency regions in the logarithmic spectrum domain is effective for decoding. The suppression is realized by multiplying the logarithmic magnitude spectrum below the low-cutoff frequency (200 Hz) and above the high-cutoff frequency (3 kHz) by 0.1 [8].

2.2 Audio data hiding based on amplitude modulation [7]

At the beginning of the embedding process, a cover signal is filtered by a filterbank of equal bandwidth. Sinusoidal amplitude modulations (SAMs) at relatively low modulation frequencies applied to the neighboring subband signals in the opposite phase are used as carriers of embedded information. A key defined by a known PN generator classifies subband pairs into several groups. Embedded information is encoded by phase-shift keying (PSK), defined as the phase differences between SAMs applied to the base group and the other groups. For example, 4-PSK encodes 2-bit information for every $\pi/2$ phase difference. Multiple watermark embedding is achieved by applying different modulation frequencies simultaneously.

Extraction of the SAM waveform from the watermarked signal is performed by calculating the logarithmic ratio of the amplitude envelopes extracted from the neighboring subband signals. The amplitude envelopes of each subband can be derived from the power spectrum of the half-overlapped running FFT. The synchronized addition of the SAM waveforms extracted between 200 Hz and 3 kHz and the same subband group emphasizes the SAM waveform. The embedding intensity is defined as the degree of SAM. In the present study, the embedding intensity for all sub-bands is identical and is simply determined as a constant parameter value.

Synchronization of the embedded data frames is achieved by inverting the relative phase of the amplitude modulations between neighboring data frames.

In the extraction process, two sequential rectangular windows (W_a and W_b) of the same length as a data frame are applied iteratively to the SAM waveform extracted from the base subband group. All SAM waveforms whose center frequency is between 200 Hz and 3 kHz in the W_a window are summed (ΣW_a), and ΣW_b is derived in the same manner. Then, Fourier amplitudes Aw of $\Sigma W_a - \Sigma W_b$ at the embedded modulation frequency are calculated while the position of the windows is shifted along the width of a data frame. When the positions of the windows completely overlap the positions of the data frame, Aw reaches a maximum. The correct positions of the data frames can then be detected.

3 SIMULATION OF LONG-DISTANCE AERIAL SOUND TRANSMISSION

Figure 2 shows a block diagram of the simulated environment of long-distance aerial transmissions. The frequency response of a distant horn-array loudspeaker system, absorption of sound by the atmosphere, reverberation, and a single long-path echo, a constant frequency shift caused by the Doppler shift induced by the movement of the receiver, and additive background noise are simulated. Modeling and implementations of the signal processing for each disturbance are described in detail in the literature [9]. Table 1 shows the disturbances and their simulated parameters for implementation. All signal processing units are connected in series.

'STRAFFIC/ch01.wav', which was recorded at a busy traffic intersection; 'SPSQUARE/ch01.wav', which was recorded at a public town square with many tourists; and 'SCAFE/ch01.wav', which was recorded from the terrace of a cafe in a public square, are selected from DEMAND (Diverse Environments Multichannel Acoustic

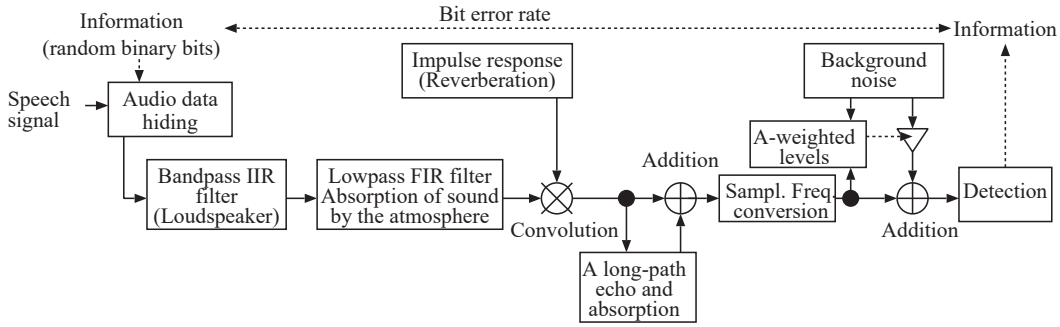

Figure 2. A block diagram of the simulated environment.

Table 1: Disturbances and their simulated parameters for long-distance aerial sound transmission.

Disturbance	Implementation	Parameters
Loudspeaker characteristics including 70-m absorption	Bandpass IIR filter	Highpass: 3rd-ord. Butterworth, 300 Hz Lowpass: 2nd-ord. Butterworth, 2 kHz
Absorption of sound by the atmosphere	Highpass FIR filter	ISO-9613-1, Temp: 15 °C Humidity: 60 %, Pressure: 1013 hPa
Reverberation	Synthesized impulse response using Gaussian noise with Exp. decay	Reverberation time: 1 s, Direct to reverberant ratio: 4 dB, Speech transmission index [10]: 0.6
A long-path echo	A single delay with reverberation and absorption	–6 dB, delay time is randomly chosen from 0.1—1.0 s
Frequency shift caused by Doppler shift	Re-sampling	+0.1 %
Background noise	DEMAND database	A-weighted SNR: 10, 5, 0 dB

Noise Database) [13] as background noises. These noises are sampled at 16-kHz, 16-bit quantization, and a single channel. A randomly selected segment is mixed with the stego speech signal for several signal to noise ratios (SNRs). In this study, the SNR is calculated based on the A-weighted sound pressure levels, defined by the international standard IEC 61672:2003, which is commonly used for measuring environmental noise.

Understanding speech sounds emitted from the outdoor loudspeakers of a voice evacuation and mass notification system is often hampered by long-path echoes reflected from large-scale structures and geographical features. To improve an articulation index, an announcer reads a text with approximately 1-s pauses between phrases. In addition to the original speech signals from the database, speech signals inserting 0.8-s pauses between phrases are also simulated. Figure 3 shows the original speech signal from the database on top, a simulated speech signal with 0.8-s pauses between segmented phrases in the middle, and the actual speech signal with pauses used for an actual mass notification system at the bottom.

4 EVALUATION

4.1 Simulation conditions

The bilateral method and amplitude modulation method are evaluated using the computer simulation. Watermarks of a random bit, which simulates coded information for refugees for disasters in small numbers of infor-

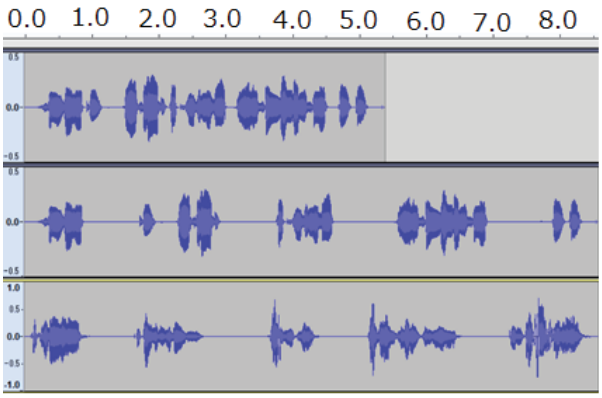

Figure 3. Top: original speech signal from the database, middle: simulated speech signal with 0.8-s pauses between segmented phrases, bottom: actual speech signal with pauses used for mass notification system.

Table 2: Parameter values for the bilateral time-spread echo hidings and amplitude modulation method. The amount of payload bits are 4.4 bps and 4.0 bps, respectively.

Bilateral time-spread echo hiding				
delay time	frame overlap	frame length	length of PN series	payload bits per frame
$d0$	$T_r/2$	F	L	
80 samples (5 ms)	50 samples	32,768 (2.05 s)	2,047	9
Audio data hiding based on amplitude modulation				
subband pairs	subband groups	frame period	modulation frequencies	max. modulation degree
255	3	3 s	1.0, 1.67, 2.67 Hz	0.7

mation bits, are embedded into the speech signals. The amount of payload is 4.4 bps for the bilateral method, and 4.0 bps for the amplitude modulation method without error-correction code. Table 2 shows the parameter values of both methods tested in this study.

A total of 753 speech files spoken by 10 male speakers, in which the two speech files are concatenated, and 903 speech files spoken by 12 female speakers are served as cover data. These files are recorded in the Continuous Speech Database for Research (Vol. 1), published by the Acoustical Society of Japan. All speech files are sampled at 16 kHz and 16-bit quantization.

4.2 Results

The bit error rates (BERs) of the extracted bits from the simulated long-distance transmission are evaluated as an index of the performance of the audio data hiding method. Figure 4 compares BERs obtained from both the male and female voice, and the different types of background noise for both methods under 0 dB SNR and loudspeaker distance of 400 m. The results show that the bilateral method is superior to the amplitude modulation method under all conditions. In addition, the median BERs obtained from the male speakers are generally better than those obtained from the female speakers by 0.5% to 2.5% for the bilateral method and by 4.2 % to 6.3 % for the amplitude modulation method. The lower fundamental frequencies of the male speakers are better cover signals because of their higher density of spectral components. The lower BER performance obtained under each condition did not depend on the speech signals but rather on the segment of background noise. Although SNR was kept constant among the simulated conditions, segmental SNRs change dynamically, such that occasional and large degradation on BER was observed.

The median BERs of the bilateral method obtained from the background noise SCAFE are the worst among the

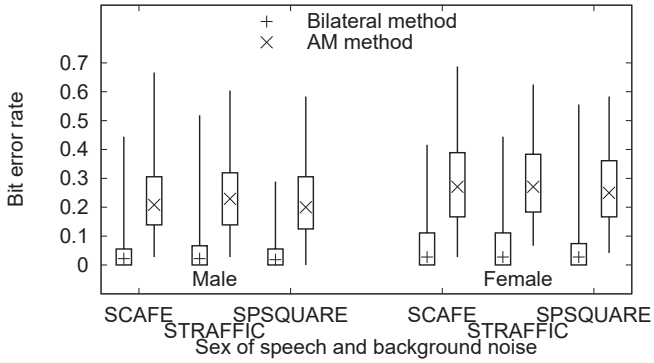

Figure 4. BERs obtained from the bilateral method and the amplitude modulation method as a function of background noise and the sex of the speaker at an SNR of 0 dB and loudspeaker distance of 400 m. Box and error bars show minimum, 10 percentile, median, 90 percentile, and maximum values.

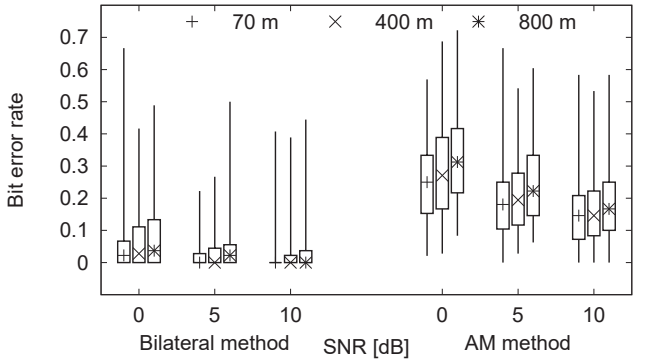

Figure 5. The effects of the loudspeaker distance and SNR on BERs.

three noise files, because the sound of SCAFE mainly consists of human speech noise, whereby the spectrum overlaps with that of the stego message voice. To investigate the effectiveness of speech watermarking under the worst outdoor conditions using the preferred audio data hiding method, the following results show data obtained from the female speakers and the background noise SCAFE.

Figure 5 compares the BERs obtained from the conditions of the simulated loudspeaker distances of 70 m, 400 m, and 800 m as a function of SNR. The results show that shorter loudspeaker distance and larger SNR induce better BER performance. In addition, the remarkable result is that 87 % of the speech signals can transmit more than 90 % of the hidden 4.4 bps message by the bilateral method under the 800 m distance and 0 dB SNR condition.

Figure 6 shows the effects of the disturbance factors on BERs. Conditions in which one of the three factors, reverberation, long-path echo, and frequency shift, was removed, were tested for the bilateral method and the amplitude modulation method. In addition, the speech signals with 1-s pauses between phrases were tested.

The results show that the performance of the bilateral method is mainly degraded by frequency shift; the optimal BER is obtained from the condition without frequency shift. Reverberation and long-path echoes do not affect the performance of the bilateral method; meanwhile, they increase BERs obtained from the amplitude modulation method by 4–5 %. The speech signals with 1-s pauses between phrases slightly degrade BER performance for both methods by 3–5 %.

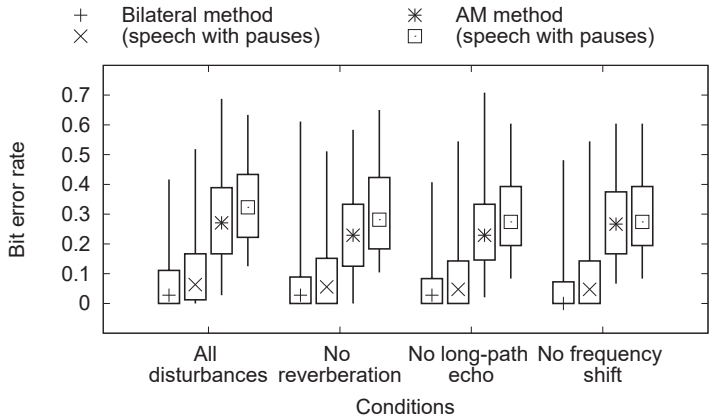

Figure 6. The effects of disturbance factors on BERs obtained from female speech signals with and without 1-s pauses between segmented phrases.

5 DISCUSSION

The benefit of time-spread echo hiding is that the timbre formed by the echo kernel is similar to that formed by random reflections observed in indoor and outdoor environments. In addition, the amplitude modulation applied to the speech signal does not greatly degrade the speech quality. Nishimura [5] conducted identification tests for watermarked VCV (vowel-consonant-vowel) syllables to investigate the effect of amplitude modulation-based watermark on articulation scores. The mean articulation scores of the 125 VCV syllables, for which an amplitude modulation of 0.6 modulation degrees was applied, was 0.86 under a 10-dB SNR condition. Because speech intelligibility is generally higher than that suggested by the articulation scores of syllables, no serious deterioration in the intelligibility of the watermarked speech is expected. Thus, the current study did not confirm the subjective quality and clarity of the stego speech signal.

Embedding data combining both audio hiding methods simultaneously may improve overall BER performance because both detection methods represent a theoretically orthogonal relationship between the frequency-domain (bilateral method) and the time-domain (amplitude modulation method). In this case, the subjective quality and clarity of the stego speech signal should be confirmed. These topics will be addressed in further work.

6 CONCLUSIONS

Two audio data hiding methods, which are robust against aerial transmission—the bilateral method and amplitude modulation method— were tested by the simulation in which speech signals are broadcast by the outdoor loudspeakers of a voice evacuation and mass notification system. The computer simulations include several disturbances caused by the long-distance (70–800 meters) aerial transmission of sounds. The frequency response of a distant horn-array loudspeaker system, the absorption of sound by the atmosphere, reverberation, a single long-path echo, a constant frequency shift induced by Doppler effect, and additive background noise were simulated as disturbances. The results show that the bilateral method was more robust against disturbances than the amplitude-modulation-based method.

ACKNOWLEDGMENTS

This study was partially supported by TOA Corporation and Evixer Inc.

REFERENCES

[1] S. A. Chou and S. F. Hsieh. An echo-hiding watermarking technique based on bilateral symmetric time spread kernel. In *Proceedings of ICASSP 2006 III*, pages 1100–1103, 2006.

[2] D. Gruhl, A. Lu, and W. Bender. Echo hiding. In *Proceedings of the First International Workshop on Information Hiding, LNCS 1174*, pages 295—315, 1996.

[3] B.-S. Ko, R. Nishimura, and Y. Suzuki. Time-spread echo method for digital audio watermarking. *IEEE Trans. on Multimedia*, 7:212–221, 2005.

[4] T. Munekata, T. Yamatuchi, H. Handa, R. Nishimura, and Y. Suzuki. A portable acoustic caption decoder using IH technique for enhancing lives of the people who are deaf or hard-of-hearing — system configuration and robustness for airborne sound —. In *Proc. of IIHMSP2007*, pages 406—409, 2007.

[5] A. Nishimura. Data hiding for speech sounds using subband amplitude modulation robust against reverberations and background noise. In *Proceedings of IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 7–10. IEEE, 2006.

[6] A. Nishimura. Audio data hiding that is robust with respect to aerial transmission and speech codecs. *International Journal of Innovative Computing Information and Control*, 6(3(B)):1389–1400, 2010.

[7] A. Nishimura. Audio watermarking based on subband amplitude modulation. *Acoustical Science and Technology*, 31(5):328–336, 2010.

[8] A. Nishimura. Improvement and evaluation of time-spread echo hiding technology for long-distance voice evacuation system. In *Digital Forensics and Watermarking*, number 10431, pages 391–405. Springer LNCS, 2017.

[9] A. Nishimura. Simulation of long-distance aerial transmissions for robust audio data hiding. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing Part I, Springer Smart Innovation Systems and Technologies*, number 81, pages 361–369. Springer, 2017.

[10] M. R. Schroeder. Modulation transfer functions: Definition and measurement. *Acustica*, 49:179–182, 1981.

[11] K. Tetsuya, O. Akihiro, and P. Udaya. Properties of an emergency broadcasting system based on audio data hiding. In *Proc. IIHMSP2015*, pages 142–145, 2015.

[12] K. Tetsuya, K. Kan, and P. Udaya. A disaster prevention broadcasting based on audio data hiding technology. In *Proceedings of Joint 8th International Conference on Soft Computing and Intelligent Systems and 17th International Symposium on Advanced Intelligent Systems*, pages 373–376, 2016.

[13] J. Thiemann, N. Ito, and E. Vincent. DEMAND : Diverse environments multichannel acoustic noise database, 2013.