Cho *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2015) 2015:10
DOI 10.1186/s13636-015-0053-x

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

An acoustic data transmission system based on audio data hiding: method and performance evaluation

Kiho Cho[1], Jae Choi[2] and Nam Soo Kim[2*]

Abstract

Acoustic data transmission (ADT) forms a branch of the audio data hiding techniques with its capability of communicating data in short-range aerial space between a loudspeaker and a microphone. In this paper, we propose an acoustic data transmission system extending our previous studies and give an in-depth analysis of its performance. The proposed technique utilizes the phases of modulated complex lapped transform (MCLT) coefficients of the audio signal. To achieve a good trade-off between the audio quality and the data transmission performance, the enhanced segmental SNR adjustment (SSA) algorithm is proposed. Moreover, we also propose a scheme to use multiple microphones for ADT technique. This multi-microphone ADT technique further enhances the transmission performance while ensuring compatibility with the single microphone system. From a series of experimental results, it has been found that the transmission performance improves when the length of the MCLT frame gets longer at the cost of the audio quality degradation. In addition, a good trade-off between the audio quality and data transmission performance is achieved by means of SSA algorithm. The experimental results also reveal that the proposed multi-microphone method is useful in enhancing the transmission performance.

Keywords: Acoustic data transmission; Data hiding; Information hiding; Acoustic communication; Audio watermarking; Modulated complex lapped transform

1 Introduction

Audio data hiding (or information hiding) has been widely applied in many areas such as audio watermarking for copyright protection, steganography, covert communication, and broadcast monitoring [1,2]. Apart from these traditional applications, audio data hiding techniques can be also deployed as a fundamental framework for acoustic data transmission (ADT) of which a brief implementation is illustrated in Figure 1.

ADT implies a method that sends a message signal through aerial space by playing it back using a loudspeaker at a transmitter and receives the signal by recording it using a microphone at a receiver. The ADT technique can be applied to various applications, e.g., querying an audio track on a radio, automatic check-in a store, localizing a pirate camcorder in a cinema [3], and providing additional information during a TV show or an advertisement.

One of the most important advantages of the ADT is that it can establish a simple low bit rate transmission channel without any additional communication devices. For instance, this method makes it possible to receive a data stream while listening to some music, which has been modulated to embed the intended data. What is important in this scenario is that the modulation should not modify the original music sound severely such that it can be perceived differently.

There have been a number of ADT schemes motivated by conventional audio watermarking techniques such as the spread spectrum [3-5] and echo hiding [6]. These approaches, however, cannot support sufficient data rates to transmit text messages while providing good audio quality at the same time. In [7], digital communication signals are produced and then allocated at certain frequency bands and temporal positions in order to imitate music

*Correspondence: nkim@snu.ac.kr
[2]Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, 151-742 Seoul, Korea
Full list of author information is available at the end of the article

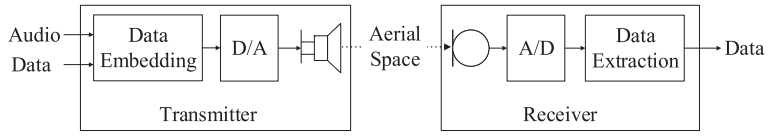

Figure 1 Schematic representation of acoustic data transmission.

. In [8], the optimal symbol length of the acoustic differential binary phase shift keying (DBPSK) signal in air was experimentally examined, and suitable ranges of data rates versus signal-to-noise ratio (SNR) were suggested. In sparse multi-carrier-based techniques [9], the cumulative distribution function of the room impulse response is utilized to design an optimal filter bank to cope with reverberant environments. As a reliable method providing a reasonable bit rate to transmit text messages, the acoustic orthogonal frequency division multiplexing (AOFDM) technique was developed [10]. Even though the transmission performance of the AOFDM has been reported to be much better than that of the previous techniques, the quality degradation of the data-embedded audio is significant. Particularly, the audio quality usually degrades seriously for speech-like signals [10].

Our previous studies [11-14] have been found to possibly overcome the limitations of the conventional approaches to ADT. In our approach to audio data hiding, the modulated complex lapped transform (MCLT) [15] is applied and the phases of MCLT coefficients are modified according to the embedded data bits. MCLT is widely used in various applications because it can reduce the blocking artifacts induced by the modification of spectrum parameters [16]. The experimental results have demonstrated that the proposed data hiding method [11] yields better performance than AOFDM in terms of both the audio quality and transmission range. Moreover, incorporating various techniques such as the masking threshold, data extraction with clustering, selecting proper frequency band and frames [12], adjusting spectral magnitude after data embedding [13], and the segmental SNR adjustment (SSA) algorithm [14] has been found to further improve the performance.

In practical ADT, however, the received audio signal usually suffers from heavy attenuation at a certain frequency band or time frame due to the destructive interference of multi-path propagation and the spectral characteristics of the audio signal. This phenomenon makes the ADT system fragile to background noises, and thus, the transmission performance can be deteriorated. To cope with the heavy attenuation of the received signal, a variety of diversity schemes have been adopted in the field of wireless communication [17].

Among various diversity techniques, we employ the framework of the spatial diversity where the transmitter and receiver communicate messages with multiple antennas located at spatially separated positions. In the context of ADT, the loudspeakers and microphones are equivalent to the antennas at the transmitter and receiver, respectively. This indicates that the spatial diversity can be implemented in ADT with multiple loudspeakers or multiple microphones.

In this paper, we present a practical ADT system based on the audio data hiding technique, which recomposes and extends our previous studies. The main contribution of the current work can be summarized as follows: First, the masking model and trade-off parameter is added to the SSA algorithm in order to achieve a good trade-off between the audio quality and the data transmission performance. Second, we propose a novel multi-channel ADT technique in which multiple microphones are used to achieve spatial diversity. In the proposed approach, the data is decoded by combining the decisions made at separate receivers with weighting factors. To obtain the weighting factors, a channel estimation technique designed based on the Wiener estimator is applied. One of the most noticeable advantages of this multi-channel technique is that it is backwards compatible regardless of the number of loudspeakers and microphones. This implies that the proposed method with multiple microphones and loudspeakers can be implemented as a simple extension of the single microphone-loudspeaker case. Finally, we evaluate the performance of the proposed ADT system in various experimental conditions.

The rest of this paper is organized as follows: In Section 2, MCLT is briefly introduced. In Section 3, the data embedding procedure is described. In Section 4, we present the modified SSA algorithm with the trade-off parameter controlling audio quality and data transmission performance, and in Section 5, we describe the data extracting procedure, which takes advantage of the multi-microphone scheme. The experimental results for audio quality and data transmission performance are shown in Section 6. Finally, Section 7 concludes this paper.

2 Modulated complex lapped transform

In this section, we present the basic MCLT formulation. MCLT generates M complex-valued coefficients from the $2M$-length frame of a real-valued input signal $x(n)$. The i-th input frame, which is shifted by M samples, is denoted by a vector $\vec{\mathbf{x}}_i = [x(iM), x(iM+1), \ldots, x(iM+2M-1)]^T$,

with T denoting the transpose of a vector or a matrix, and the MCLT coefficient vector $\vec{\mathbf{X}}_i = [X_i(0), X_i(1), \ldots, X_i(M-1)]^T$ corresponding to $\vec{\mathbf{x}}_i$ is given by [15]:

$$\vec{\mathbf{X}}_i = \vec{\mathbf{X}}_{c,i} - j\vec{\mathbf{X}}_{s,i} \quad (1)$$

$$= (\mathbf{C} - j\mathbf{S})\mathbf{W}\vec{\mathbf{x}}_i \quad (2)$$

where $\vec{\mathbf{X}}_{c,i}$ and $\vec{\mathbf{X}}_{s,i}$ represent the real (cosine) and imaginary (sine) parts of $\vec{\mathbf{X}}_i$, respectively. In (2), $\mathbf{C}$ and $\mathbf{S}$ denote the $M \times 2M$ cosine and sine modulation matrices whose (k,n)-th elements are defined by:

$$(\mathbf{C})_{kn} = \sqrt{\frac{2}{M}} \cos\left[\left(n + \frac{M+1}{2}\right)\left(k + \frac{1}{2}\right)\frac{\pi}{M}\right] \quad (3)$$

$$(\mathbf{S})_{kn} = \sqrt{\frac{2}{M}} \sin\left[\left(n + \frac{M+1}{2}\right)\left(k + \frac{1}{2}\right)\frac{\pi}{M}\right], \quad (4)$$

respectively, with $j = \sqrt{-1}$. The diagonal matrix $\mathbf{W}$ is a $2M \times 2M$ window matrix whose n-th main diagonal element is commonly designed as:

$$(\mathbf{W})_{nn} = -\sin\left[\left(n + \frac{1}{2}\right)\frac{\pi}{2M}\right]. \quad (5)$$

The inverse MCLT of $\mathbf{X}_i$ is given by:

$$\vec{\mathbf{y}}_i = \mathbf{W}\left(\beta_c \mathbf{C}^T \vec{\mathbf{X}}_{c,i} + \beta_s \mathbf{S}^T \vec{\mathbf{X}}_{s,i}\right) \quad (6)$$

where β_c and β_s are arbitrary values that satisfy $\beta_c + \beta_s = 1$. In this work, we choose $\beta_c = \beta_s = \frac{1}{2}$.

To obtain the reconstructed frame, the inverse MCLT frames are overlapped and added by M samples (half of the length of an MCLT frame) with its adjacent frames. Let $\hat{\vec{\mathbf{y}}}_i$ be the i-th reconstructed frame. Then,

$$\hat{\vec{\mathbf{y}}}_i = \begin{bmatrix} \vec{\mathbf{y}}_{2,i-1} \\ \vec{\mathbf{0}} \end{bmatrix} + \begin{bmatrix} \vec{\mathbf{y}}_{1,i} \\ \vec{\mathbf{y}}_{2,i} \end{bmatrix} + \begin{bmatrix} \vec{\mathbf{0}} \\ \vec{\mathbf{y}}_{1,i+1} \end{bmatrix} \quad (7)$$

where $\vec{\mathbf{y}}_i = \left[\vec{\mathbf{y}}_{1,i}^T | \vec{\mathbf{y}}_{2,i}^T\right]^T$ with $\vec{\mathbf{y}}_{1,i}$ and $\vec{\mathbf{y}}_{2,i}$ being the M-length subvectors of $\vec{\mathbf{y}}_i$ and $\vec{\mathbf{0}}$ denotes an M-length zero vector.

3 Data embedding

In this section, we describe the procedure for data embedding in the proposed acoustic data transmission system. The block diagram of the data embedding procedure is shown in Figure 2. For a reliable transmission of messages, a cyclic redundancy check (CRC) algorithm and a forward error correction (FEC) technique are indispensable for detecting and correcting bit errors. An interleaver permutes the bit sequence in a pseudo-random manner. It can improve the performance of FEC by avoiding errors bursting on certain time or frequency regions and reduce the peak-to-average ratio (PAPR), which possibly degrades the quality of the data-embedded audio signal [18].

A host audio signal is first divided into consecutive MCLT frames, and the data bits are embedded by modifying the MCLT coefficients. The main strategy of data embedding is to modify the MCLT coefficients of the host audio signal in such a way that the phases of MCLT coefficients of the reconstructed frame are set as being either 0 or π. Here, without loss of generality, we presume the usage of a binary signaling scheme.

Based on the data embedding strategy, the desired value of the MCLT coefficient of the data-embedded audio $\hat{Y}_{i,\text{desired}}(k)$ for the k-th frequency element at the i-th reconstructed frame should be given by:

$$\hat{Y}_{i,\text{desired}}(k) = \max\left(|X_i(k)|, M_i(k)\right) b_i(k), \quad (8)$$

where $M_i(k)$ represents the masking threshold [19] and $b_i(k) \in \{-1, 1\}$ means the data bit. In (8), the magnitude of the coefficients is lower bounded by the level of masking threshold for improving the transmission performance while maintaining the audio quality [12].

The data-embedded MCLT coefficient $\hat{X}_i(k)$ can be derived by analyzing the relationship between $\hat{X}_i(k)$ and the MCLT coefficient of reconstructed frame $\hat{Y}_i(k)$. Referring to (7), the relationship between $\hat{X}_i(k)$ and $\hat{Y}_i(k)$ is depicted in Figure 3. When the frequency index k is not 0 or M, $\hat{Y}_i(k)$ is derived analytically as follows:

$$\hat{Y}_i(k) = \frac{1}{2}\hat{X}_i(k) + j\frac{1}{2}\left[\vec{\mathbf{A}}_{-1,k}\vec{\mathbf{X}}_{i-1} + \frac{1}{2}X_i(k-1) - \frac{1}{2}X_i(k+1) + \vec{\mathbf{A}}_{1,k}\vec{\mathbf{X}}_{i+1}\right] \quad (9)$$

where $\vec{\mathbf{A}}_{m,k}$ represents the interference weighting vector due to the overlapping with the adjacent frames. The l-th element of $\vec{\mathbf{A}}_{m,k}$ is given by:

$$A_{m,k,l} = \begin{cases} (-m)\dfrac{(-1)^{l+d}}{\pi(2d-1)(2d+1)} & \text{if } |l-k| = 2d \\ \dfrac{(-1)^l}{4} & \text{else if } |l-k| = 1 \\ 0 & \text{otherwise,} \end{cases} \quad (10)$$

where d is a nonnegative integer.

Based on the fact that $\hat{Y}_{i,\text{desired}}(k)$ in (8) and $\hat{Y}_i(k)$ in (9) should be kept the same, the data-embedded MCLT coefficient $\hat{X}_i(k)$ can be derived in the following way:

$$\hat{X}_i(k) = 2\max\left(|X_i(k)|, M_i(k)\right) b_i(k) - j\left[\vec{\mathbf{A}}_{-1,k}\vec{\mathbf{X}}_{i-1} + \frac{1}{2}X_i(k-1) - \frac{1}{2}X_i(k+1) + \vec{\mathbf{A}}_{1,k}\vec{\mathbf{X}}_{i+1}\right], k \in \mathbb{D}, \quad (11)$$

where $\mathbb{D}$ is the set of the frequency indices in which data bits are embedded.

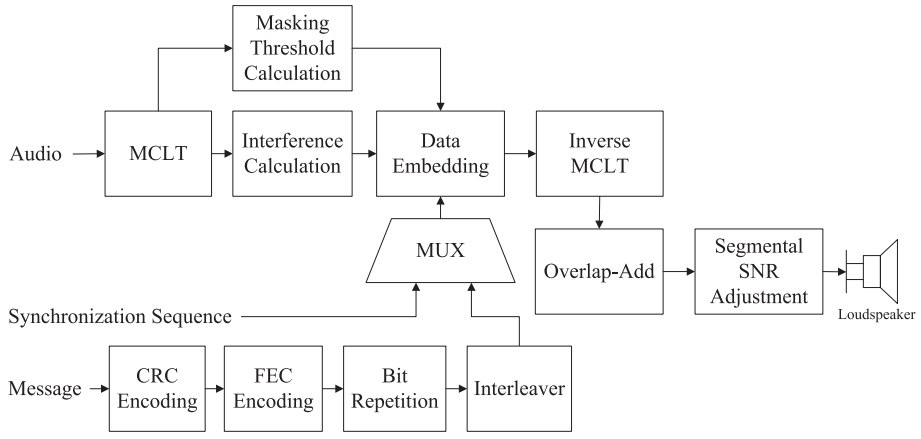

Figure 2 Data embedding procedure of the proposed acoustic data transmission system.

From (11), we can see that the two adjacent frames and two adjacent coefficients $\vec{\mathbf{X}}_{i-1}$, $\vec{\mathbf{X}}_{i+1}$, $X_i(k-1)$, and $X_i(k+1)$ are needed for deriving $\hat{X}_i(k)$. In order to avoid the interference due to these terms, it would be safe to embed the data in every other frame and frequency line as illustrated in Figure 4. The data-embedded MCLT coefficients are then converted into a time domain signal segment by applying inverse MCLT and overlapping with the previous and next MCLT frames.

For a practical ADT system, a synchronization frame is inserted between consecutive message frames. The binary data $b_i(k)$ in (11), therefore, can be a part of either the synchronization sequence or the message to be transmitted as shown in Figure 2. The synchronization sequence should be known *a priori* at the receiver not only to synchronize the message frame but also to compensate the channel effects. At message frames, the message bit is embedded in L different MCLT coefficients with L-length spreading sequence to improve robustness of the data-embedded audio signal against channel effects and additive noise [20]. As L gets larger, the robustness of the system improves accordingly at the price of reduced bit rate.

4 Segmental SNR adjustment for audio quality enhancement

In this section, we describe the SSA algorithm incorporating the masking threshold for achieving a good trade-off between the transmission performance and audio quality. The SSA algorithm further modifies the spectral components of the data-embedded audio signal as shown in Figure 2. The block diagram of the proposed SSA algorithm is shown in Figure 5. In this algorithm, an MCLT analysis with frame length M_s, which is much shorter than the frame length for data embedding M, is needed to calculate the segmental SNR of the data-embedded audio signal.

Let $X_{i_s}(k_s)$ and $\hat{X}_{i_s}(k_s)$ denote the MCLT coefficients of the original and data-embedded audio signal, respectively, obtained from the MCLT analysis with frame length M_s. Here, i_s and k_s respectively indicate the frame and frequency indices, which are introduced in order to distinguish them from the long window-based MCLT analysis. The segmental SNR for the i_s-th MCLT frame and k_s-th frequency bin is defined as follows:

$$\mathrm{SNR}_{i_s}(k_s) = 10\log_{10}\left(\frac{|X_{i_s}(k_s)|^2}{\left(|X_{i_s}(k_s)| - |\hat{X}_{i_s}(k_s)|\right)^2}\right). \quad (12)$$

If $\mathrm{SNR}_{i_s}(k_s)$ is smaller than a target segmental SNR, $\Gamma_{i_s}(k_s)$, the magnitude of $\hat{X}_{i_s}(k_s)$ is modified such that:

$$|\tilde{X}_{i_s}(k_s)| = \begin{cases} \dfrac{\Gamma_{i_s}(k_s)+1}{\Gamma_{i_s}(k_s)}|X_{i_s}(k_s)|, & |X_{i_s}(k_s)| < |\hat{X}_{i_s}(k_s)| \\ \dfrac{\Gamma_{i_s}(k_s)-1}{\Gamma_{i_s}(k_s)}|X_{i_s}(k_s)|, & \text{otherwise,} \end{cases} \quad (13)$$

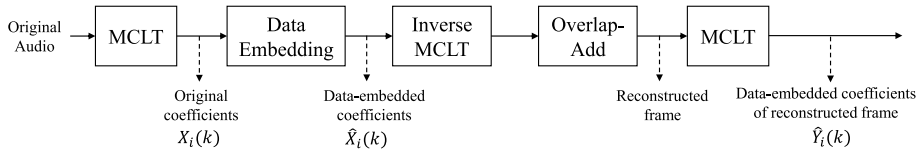

Figure 3 Relationship of the data-embedded MCLT coefficients and those of the reconstructed frame. The notations are related to those in Section 3.

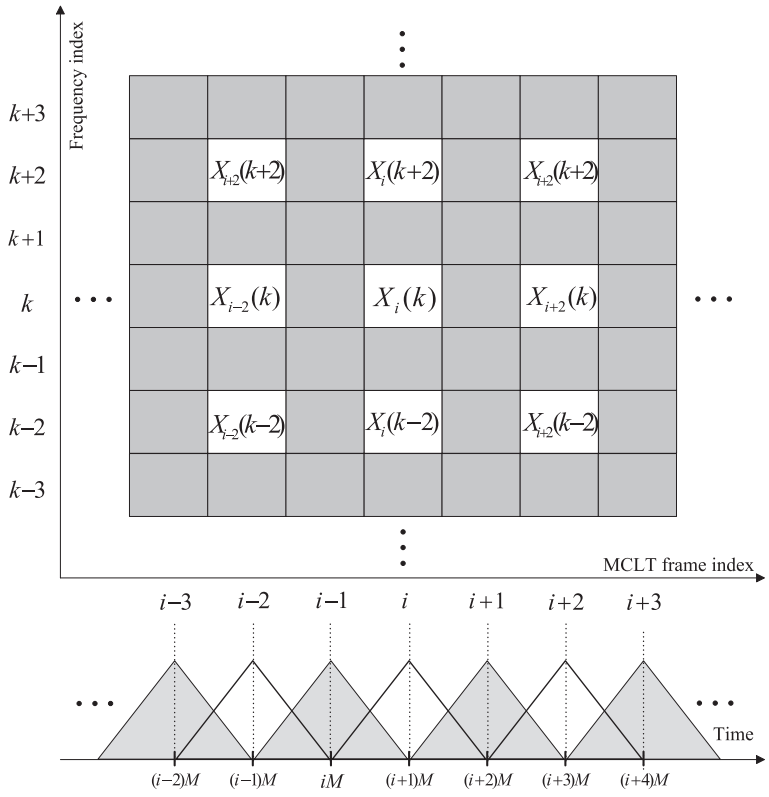

Figure 4 Time-frequency representation of MCLT indices. An equivalent window for each MCLT frame is drawn as a triangle. Data can be embedded only in white blocks.

where $\tilde{X}_{i_s}(k_s)$ denotes the adjusted MCLT coefficient of data-embedded audio signal and this adjusted MCLT vector is then transformed back to a time domain signal through overlap-add.

The target segmental SNR $\Gamma_{i_s}(k_s)$ is adaptively determined utilizing the signal-to-masking ratio (SMR). The SMR represents the ratio between the magnitude of the MCLT coefficient and the masking threshold derived from a psychoacoustic model [19]. In this paper, $\Gamma_{i_s}(k_s)$ is defined as follows:

$$\Gamma_{i_s}(k_s) = \mathrm{SMR}_{i_s}(k_s) + \Delta_{i_s}, \tag{14}$$

where the offset Δ_{i_s} is given by:

$$\Delta_{i_s} = S_t - 20\log_{10}\left[\frac{\sum_{k_s}|X_{i_s}(k_s)|}{\sum_{k_s}|X_{i_s}(k_s)|10^{(-\mathrm{SMR}_{i_s}(k_s)/20)}}\right]. \tag{15}$$

Hence, the pre-defined value S_t plays the role of making a trade-off between the audio quality and the transmission performance which can be determined experimentally; applying higher S_t indicates better audio quality with degraded transmission performance. Since the SSA algorithm adjusts the magnitude only, the phase alteration at the receiver is not severe.

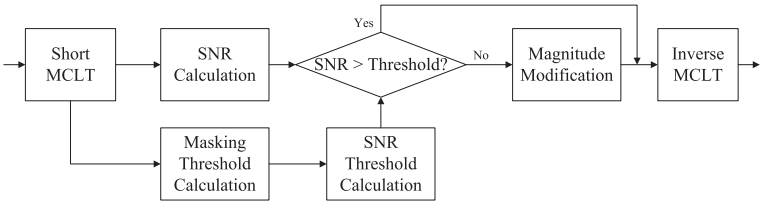

Figure 5 Block diagram of the segmental SNR adjustment (SSA) algorithm.

5 Data extraction

The data extraction procedure with the proposed multi-microphone technique is described in this section. At each microphone output, the synchronization and channel estimation are performed separately except for the final decision of the data bits. If multiple loudspeakers are used at the transmitter side, we assume that the same bits are embedded simultaneously in all the loudspeaker inputs. The block diagram of this data extraction procedure is shown in Figure 6.

5.1 Synchronization

Before extracting data from the audio signal, the received audio signal needs to be synchronized. The receiver exhaustively computes the phase correlation between the known synchronization sequence and the received MCLT coefficients. Then, it finds the time index at which the phase correlation achieves the maximum; this enables identifying the starting time index of the first message frame. The starting time index $\hat{n}$ is given by:

$$\hat{n} = \arg\max_{n} \sum_{k \in \mathbb{D}} \frac{\hat{Y}^R(k,n)p(k)}{|\hat{Y}^R(k,n)|} \tag{16}$$

where $p(k)$ is the known synchronization sequence and $\hat{Y}^R(k,n)$ is the k-th MCLT coefficient computed at the receiver when the analysis window starts at time n.

In real environments, a synchronization timing error may occur since exact timing synchronization is difficult due to a variety of acoustic interference sources and the clock mismatch between the transmitter and the receiver. This timing error of the analysis window results in a phase rotation of the received MCLT coefficients, and this may lead to decoding errors.

5.2 Channel estimation

After the synchronization for the i-th data frame, the received MCLT coefficient obtained at the m-th microphone $\hat{Y}_{i,m}(k)$ can be approximated as follows:

$$\hat{Y}_{i,m}(k) = H_{i,m}(k)|X_i(k)|b_i(k) + N_{i,m}(k), \tag{17}$$

where $H_{i,m}(k)$ and $N_{i,m}(k)$ respectively refer to the channel transfer coefficient and the additive noise for the k-th frequency bin at the m-th microphone, and $|X_i(k)|$ is the magnitude of the transmitted MCLT coefficient.

Due to the presence of the channel coefficients, it is needed to perform channel estimation to decode the message bits successfully. In this work, the Wiener estimation-based channel estimation algorithm [21] is applied. A key idea of this Wiener estimator is to smooth the channel measurements at the known synchronization positions and then to interpolate them for predicting the channel transfer coefficients at the message positions. In the current work, interpolation is performed along the time axis only because message frames reside between synchronization frames.

In order to obtain the channel measurements at the synchronization frames, the MCLT coefficients are multiplied by the known synchronization sequence. Then, the estimated channel coefficients for the message frames is interpolated by solving the normal equation. The channel estimation process is repeatedly performed at each frequency bin separately. Compared to the clustering-based decoding method in our previous work [12], the channel estimation and compensation has been found more suitable to deal with multiple microphone scheme and a long MCLT window.

5.3 Data decoding with multi-microphones

In this work, we propose a multi-channel method based on combining the results of the separate microphones. In the proposed multi-channel method, the signals captured at different microphones are linearly combined to increase SNR. After the channel is estimated, message bits are extracted from the corresponding message frames through the following steps.

First, the sequences of the MCLT coefficients and the channel coefficients are respectively restored by the de-interleaver performing an inverse operation with the interleaver. Next, the MCLT coefficients from each microphone are linearly combined. The combined MCLT coefficient at the i-th message frame is given by:

$$\hat{Y}_i(k) = \sum_m w_m(k)\hat{Y}_{i,m}(k), \tag{18}$$

and each weighting factor $w_m(k)$ is given by:

$$w_m(k) = \frac{\hat{H}^*_{i,m}(k)}{|\hat{H}_{i,m}(k)|}, \tag{19}$$

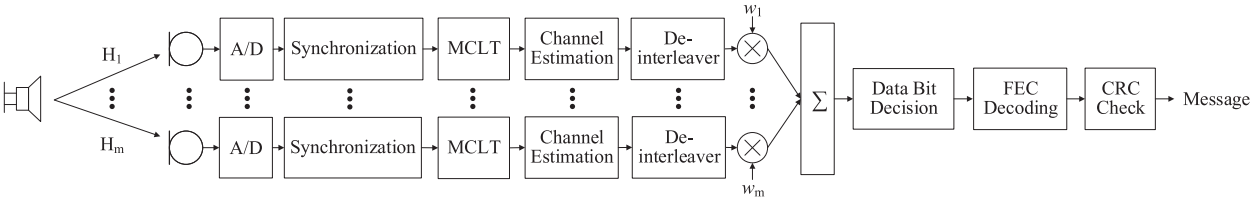

Figure 6 Data extraction procedure of the proposed acoustic data transmission system with multiple microphones.

where $\hat{H}_{i,m}(k)$ indicates complex conjugation of the estimated channel transfer coefficient and $(.)^*$ indicates complex conjugation. In addition, the message coefficient is obtained by calculating the normalized correlation between the MCLT coefficients and the L-length spreading sequence. Finally, the received message bit is obtained by examining the sign of the real part of the message coefficient, which is identical with the binary phase shift keying (BPSK) demodulation scheme. After message bits are obtained, FEC decoding and CRC checking are carried out to correct and detect errors in the message sequence.

6 Performance evaluation

In order to evaluate the performance of the proposed ADT system, we conducted several experiments concerned with audio quality and data transmission performance. First, we examined the audio quality and transmission performance while varying the MCLT length. Next, to evaluate the effects of the SSA algorithm, we made a performance comparison among the four ADT systems implemented with different configurations in a number of artificial and actual acoustic environments. Moreover, we compared the transmission performance when the FEC is applied and evaluated how much the proposed ADT systems are robust to signal processing. Finally, we compared the transmission performance in a real room with respect to the number of microphones to evaluate the effects of the proposed multi-microphone technique.

Sixteen stereo audio clips consisting of pop, rock, jazz, classical, and Latin music, each with a length of 30 s, were used in these experiments, and the sampling frequency was 44.1 kHz. The average signal power level over all the tested audio signals was adjusted to -18 dB in the digital domain in order to play back the audio signal at a similar volume level.

The audio clips were played back from a loudspeaker and then recorded by a mobile phone (Samsung Nexus S) at various distances in a room with dimension 11 m $\times$ 7 m $\times$ 3 m. The recording format was uncompressed WAV with 44.1 kHz sampling rate and 16 bits per sample. A loudspeaker and microphones were placed at positions as shown in Figure 7. As can be seen in this figure, the distances to the microphone were set to 1, 3, 5, and 7 m. The average reverberation time (RT60) of the room measured at the corresponding positions was 1.1 s [22], which indicates a severely reverberant condition. In this room environment, the average measured sound pressure level of the background noise was 40 dB and that of the audio signal was 65 dB at 1 m from the front of the loudspeaker.

6.1 Effect of MCLT length

To examine what the effect of the MCLT length is on the performance of the proposed ADT system, we evaluated the transmission performance and audio quality for various MCLT lengths. The system parameters are listed in Table 1.

6.1.1 Robustness to reverberation in a simulated room

To evaluate how MCLT length affects the transmission performance in different environments, we convolved the audio signals with the room impulse response obtained by a simulator based on the image method [23] while altering RT60 [22]. The dimensions of the simulated room were the same as the target room depicted in Figure 7. The microphone of the receiver was placed at 5-m distance from the loudspeaker. The set of RT60 of the simulated room was set to 200, 400, 600, 800, and 1,000 ms.

The bit error rates (BERs) obtained in the simulated room for various MCLT lengths are displayed in Figure 8, where each line refers to the result for a different reverberation time. In this figure, we can see the tendency that using longer MCLT window makes the ADT system more robust to the reverberant environment.

6.1.2 Objective audio quality

To measure the quality of the data-embedded audio content with varying MCLT length, we calculated the objective difference grade (ODG) using the perceptual evaluation of audio quality (PEAQ) method [24]. The ODG score ranges from -4 to 0, where each digit score indicates that the perceived audio quality is very annoying, annoying, slightly annoying, perceptible but not annoying, or imperceptible. The average ODG score obtained from the 16 test music clips is shown in Table 2, which makes it likely that the data-embedded audio is almost indistinguishable from the host audio if the MCLT length is shorter than 1,024.

The reason that the audio quality is decreasing when the MCLT window length increases can be found from the fact that the proposed ADT system modifies the phases of MCLT coefficients. The phase modification of the audio signal can incur a significant quality degradation if the length of the time-frequency transform is very long [25]. Especially, if a percussive sound made by drums, cymbals, or short unvoiced speech exists within an interval of the MCLT window, pre-echo may frequently occur over the whole interval. If the length of the pre-echo is longer than the pre-masking interval, the data-embedded audio would be annoying for the listeners [26]. The pre-echo is considered one of the most important causes of quality degradation in data-embedded audio [27].

6.2 Effects of the SSA algorithm

Based on the previous experimental results, we conducted additional experiments in order to investigate the effects of the SSA algorithm. In these experiments, we compared the comprehensive ADT systems with four different configurations denoted by S_S, S_L, S_{LA1}, and S_{LA3}. S_S refers to

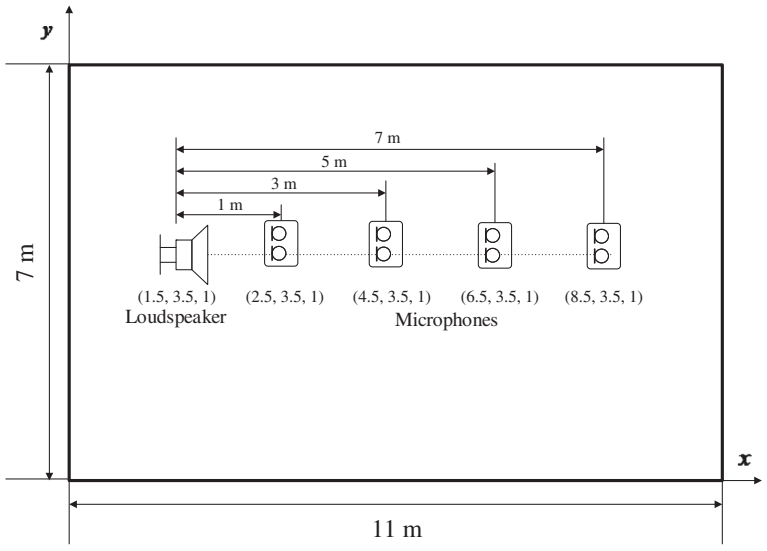

Figure 7 Room environment with location of a loudspeaker and a receiver. The height of the room is 3 m. Cartesian coordinates are written below the positions.

the system with MCLT length 512. S_L represents the system with MCLT length 8,192 without the SSA algorithm, and S_{LA1} and S_{LA3} denote the systems using the SSA algorithm setting S_t to 1 and 3 dB, respectively. The length of the short MCLT window M_s for the SSA algorithm was set to 512. The system parameters of each configuration are specified in Table 3, and the other parameters are the same as the ones in Table 1.

6.2.1 Audio quality tests

For subjective audio quality evaluation, a MUSHRA test [28] was conducted. In the MUSHRA test, each listener compared the sixteen reference signals (host audio signal) with eight differently processed test audio clips for each reference signal: hidden reference (no modification), data-embedded audio signals processed through five different configurations, and two anchor signals obtained by low-pass filtering (LPF) with 3.5 and 7.0 kHz cutoffs [28]. Thirteen listeners who have experiences of various listening tests participated in this test.

Table 1 Parameters of the system configurations tested

Parameters	Values
Sampling frequency	44.1 kHz
MCLT length (M)	512, 1,024,…, 8,192
Message frames between synchronization frames	2 frames
Data frequency	6.5 to 9.2 kHz
Bit repetition (L)	4
Bit rate	231 bps

The results are shown in Figure 9, where the average scores of the test audio clips are displayed in conjunction with 95% confidence intervals. The scores for anchor signals with 3.5-kHz LPF are omitted in the figure for a clear display. As can be seen in this figure, the average MUSHRA scores of S_S were slightly higher than those of S_{LA1} and S_{LA3}, which, however, indicate good audio quality except for S_L. Comparing between S_{LA1} and S_{LA3}, we can see that using higher S_t usually gives rise to less quality degradation. The average score obtained from S_L was quite worse than those of other configurations.

Because the average MUSHRA score of the proposed configurations are very high and the difference among them is very little except for S_L as can be seen from Figure 9, we additionally performed the PEAQ test for each music clip. The obtained ODG scores are shown in Table 4. From the results, we can see that the ODG scores demonstrate tendencies similar to the MUSHRA scores.

In summary, the results of quality evaluation tests have shown that the SSA method is useful for maintaining the quality of data-embedded audio. The main reason can be found from the fact that the SSA algorithm attenuates the pre-echo. An example of the pre-echo attenuated by the SSA algorithm is displayed in Figure 10.

6.2.2 Transmission performance of recorded signals in indoor environments

The transmission performance of the proposed approaches was evaluated in actual room environments with various distances between the loudspeaker and microphone. The experiment is performed in the same manner as in Section 6.5. In this experiment, the BERs of

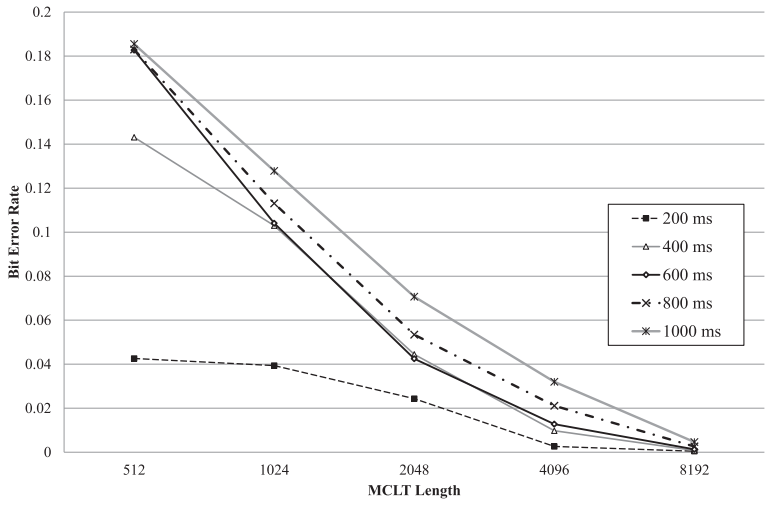

Figure 8 Bit error rate (BER) at 5-m distance in the simulated room. The horizontal axis represents the MCLT length. Each line refers to a different reverberation time of the simulated room.

four different system configurations obtained from the actual room environments are shown in Figure 11.

As can be seen from these figures, all configurations showed good data transmission performance at 1-m distance from the loudspeaker. At 5 and 7 m, however, S_L, S_{LA1}, and S_{LA3} showed better data transmission performance than S_S. This also demonstrates the tendency that using a longer MCLT window makes the algorithm more robust to reverberant environments. Especially, we can see that S_{LA1} showed better transmission performance than S_{LA3}.

Summarizing the results, we can see that the usage of the long MCLT frame outperforms the short MCLT frame especially when the distance between the loudspeaker and microphone is relatively farther. Moreover, it can be said that using a long MCLT frame length in conjunction with the SSA algorithm can enhance the transmission performance without significant audio quality degradation. In the SSA algorithm, applying a different value for S_t can achieve the trade-off between the audio quality and the data transmission performance; higher S_t makes the audio quality better at the cost of increasing BER.

Table 2 Objective difference scores for various MCLT lengths, calculated using the PEAQ algorithm

	512	1,024	2,048	4,096	8,192
M1	−0.292	−0.412	−0.685	−1.354	−1.321
M2	−0.267	−0.346	−0.515	−0.729	−0.805
M3	−0.422	−0.915	−1.665	−1.806	−1.992
M4	−0.531	−0.863	−1.343	−1.583	−1.653
M5	−0.239	−0.359	−0.549	−0.647	−0.654
M6	−0.566	−1.107	−2.092	−2.377	−2.350
M7	−0.243	−0.292	−0.486	−0.567	−0.623
M8	−0.326	−0.418	−0.844	−0.966	−1.314
M9	−0.134	−0.271	−0.404	−0.528	−0.609
M10	−0.250	−0.314	−0.620	−0.792	−0.854
M11	−0.352	−0.615	−1.088	−1.376	−1.475
M12	−0.140	−0.240	−0.359	−0.482	−0.477
M13	−0.328	−0.658	−1.061	−1.540	−1.567
M14	−0.389	−0.778	−1.505	−1.587	−1.529
M15	−0.246	−0.346	−0.535	−0.758	−0.803
M16	−0.199	−0.248	−0.375	−0.463	−0.490
Average	−0.308	−0.511	−0.883	−1.097	−1.157

6.3 Error correction using convolutional coding

In order to investigate the effect of FEC coding, the BERs corrected by convolutional coding with code rate 1/3 was obtained from the same recorded audio clips as in the previous experiment. The convolutional coding is one of the most famous FEC coding algorithms. The results are displayed in Figure 12.

Compared with the result in Figure 11, the BERs in Figure 12 are usually decreased and it can be concluded that the convolutional coding is effective to reduce the bit error. However, it is also observed that the BER

Table 3 Parameters of the system configurations used for testing the SSA algorithm

Parameters	S_S	S_L	S_{LA1}	S_{LA3}
MCLT length (M)	512	8,192	8,192	8,192
SSA algorithm	X	X	O (S_t = 1 dB)	O (S_t = 3 dB)

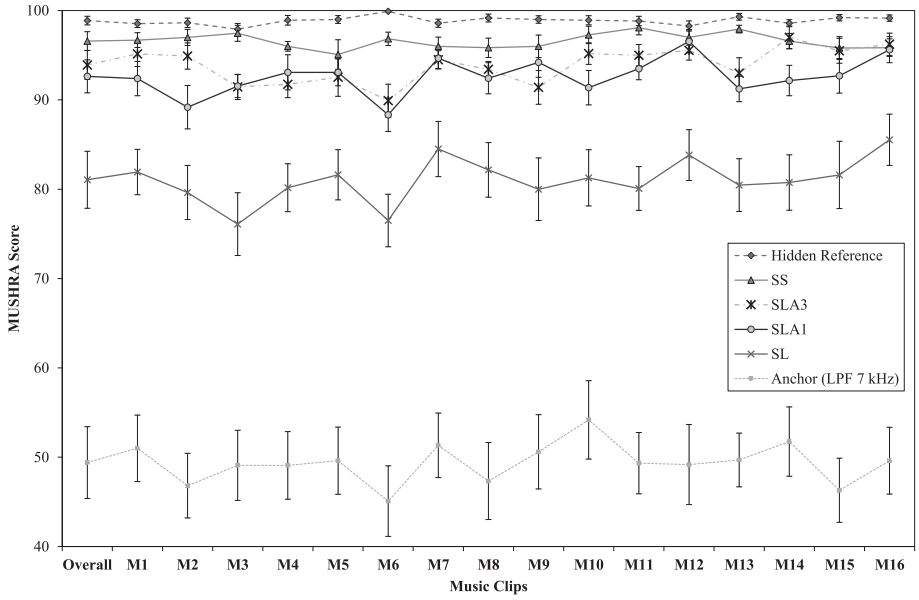

Figure 9 MUSHRA test scores for test music clips evaluating the effects of the SSA algorithm. M1 to M16 on the horizontal axis represents the name of each music clip. The numbers on the vertical axis represent the MUSHRA scores for the test music clips in different experimental conditions. Vertical lines on the top of bars denote the 95% confidence intervals.

is increasing dramatically with the use of the convolutional coding when the BER is greater than a certain threshold.

The reason can be analyzed by investigating the performance of the convolutional coding. The performance of the FEC coding techniques, however, is usually given in the form of the statistical relationship of BER versus SNR [29], which is inappropriate for this work. Therefore, we investigated the relationship between the BER with and without convolutional coding and the results are depicted in Figure 13.

In this figure, it is observed that the BER is increasing dramatically with the use of the convolutional coding when it is greater than a certain threshold (0.12 in this work), which is a common phenomenon in digital communication [29]. For designing a practical application, the BERs at target places should be lower than this threshold, which will be decided with respect to the different FEC schemes.

Table 4 Objective difference scores for test audio clips with different system configurations, for testing the SSA algorithm

	S_S	S_L	S_{LA1}	S_{LA3}
M1	−0.292	−1.321	−0.477	−0.461
M2	−0.267	−0.805	−0.448	−0.423
M3	−0.422	−1.992	−0.565	−0.519
M4	−0.531	−1.653	−0.636	−0.580
M5	−0.239	−0.654	−0.423	−0.381
M6	−0.566	−2.350	−0.664	−0.613
M7	−0.243	−0.623	−0.383	−0.366
M8	−0.326	−1.314	−0.532	−0.491
M9	−0.134	−0.609	−0.430	−0.414
M10	−0.250	−0.854	−0.492	−0.463
M11	−0.352	−1.475	−0.535	−0.495
M12	−0.140	−0.477	−0.339	−0.334
M13	−0.328	−1.567	−0.600	−0.550
M14	−0.389	−1.529	−0.585	−0.532
M15	−0.246	−0.803	−0.492	−0.469
M16	−0.199	−0.490	−0.361	−0.354
Average	−0.308	−1.157	−0.496	−0.465

6.4 Robustness to signal processing

For a practical ADT system, the data embedded in the audio signal should be robust to signal processing because the data-embedded audio signal can be distributed after being passed through several signal processing modules such as quantization, compression, and additive noise. The BER of distorted audio clips was measured, and the results are shown in Table 5. In this table, we can see that S_{LA1} and S_{LA3} showed slightly higher BER's than others because the SSA algorithm modifies a part of the amplitude spectrum. Nevertheless, we can see that all of the configurations of the proposed method are robust to requantization, inversion, additive noise, and MP3 codecs, which is comparable with the robustness of other audio data hiding techniques [30,31].

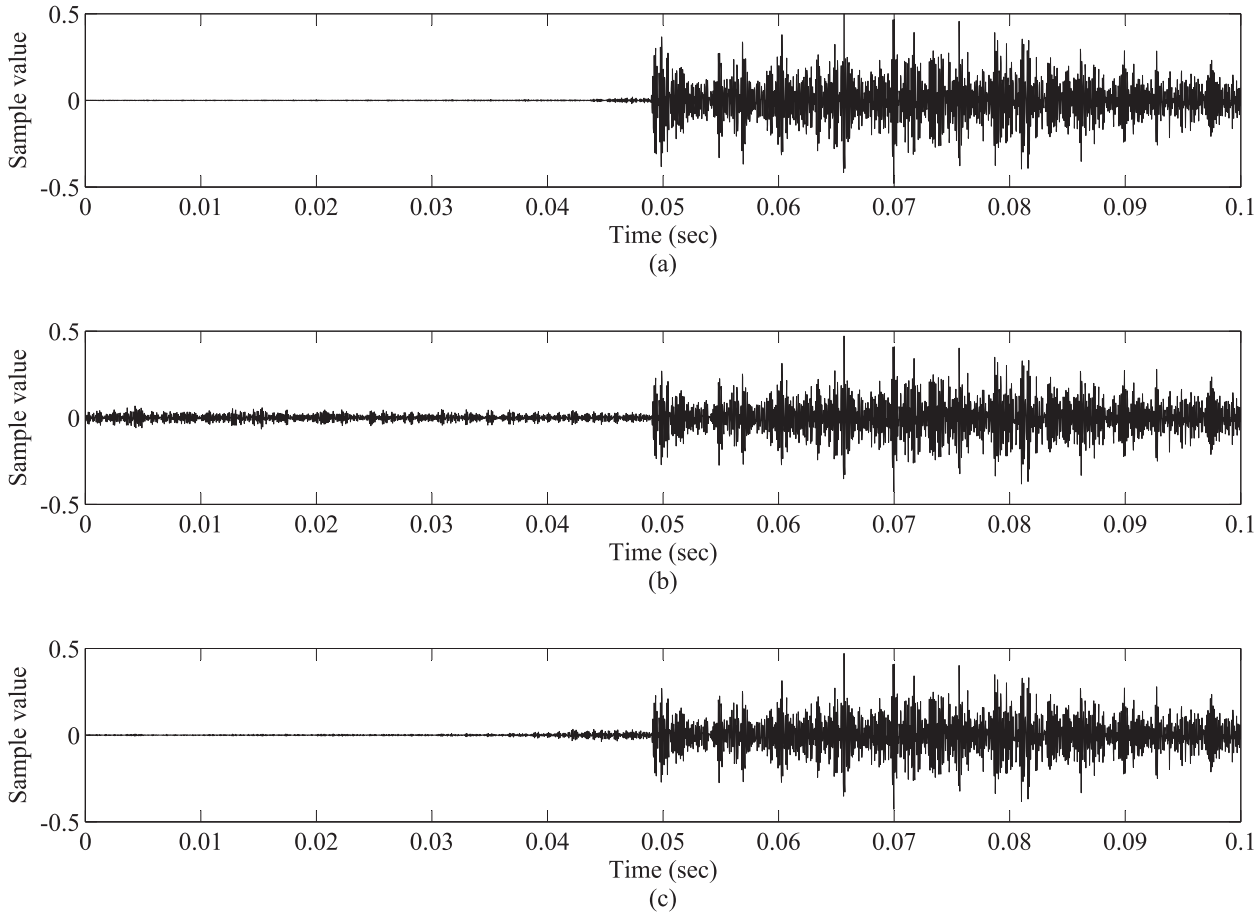

Figure 10 Example of pre-echo where the length of the MCLT frame is 0.37 s. **(a)** Host signal, **(b)** data-embedded signal, **(c)** data-embedded signal with segmental SNR adjustment algorithm.

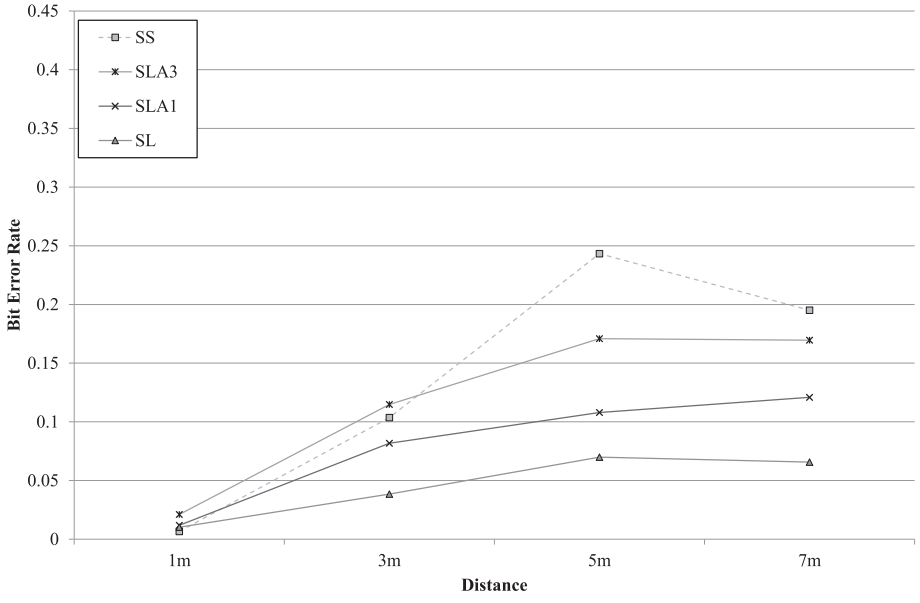

Figure 11 BER of the data transmission systems evaluating the effects of the SSA algorithm. The horizontal axis represents the distance between loudspeaker and microphone in a real room. Each line represents a different system configuration with parameters listed in Table 3.

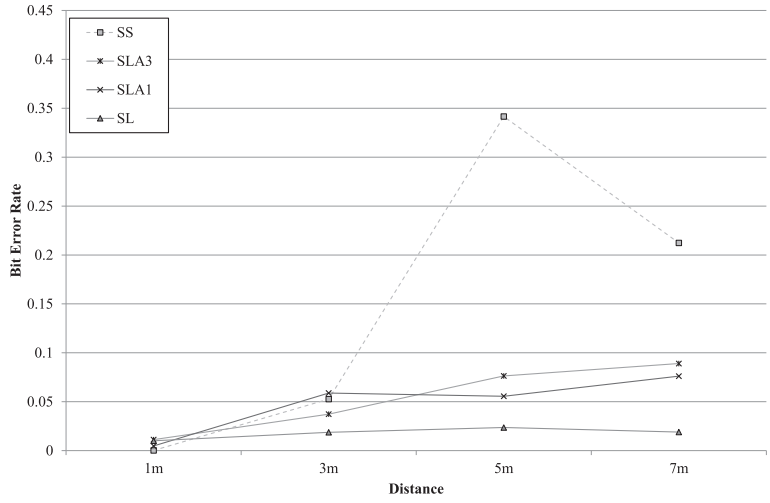

Figure 12 BER of the data transmission systems with 1/3 convolutional coding, comparing the effects of channel coding. The horizontal axis represents the distance between loudspeaker and microphone in a real room. Each line represents a different system configuration with parameters listed in Table 3.

6.5 Effect of multiple microphones

In this experiment, the transmission performance with respect to the number of microphones was investigated. As a performance measure, the BERs with one-, two-, and three-microphone configurations were compared for S_L. The second and third microphones were placed such that the distance between adjacent microphones along the y-axis became 20 cm and the results are shown in Figure 14.

From the results, we can see that the proposed multi-channel method outperformed the single microphone case. In addition, we can also see that exploiting more microphones further reduces the bit error. Consequently, we can see that the combining method is effective in transmitting data more reliable.

If statistical independence between the channels can be achieved, the transmission performance can be improved further and it is usually attempted by placing the microphones at the receiver with sufficient distance [17]. However, practically there exists an upper limit for the distance between the microphones because most of the ADT-related applications should be implemented on small devices such as mobile phones. For this reason, we depicted the BERs of 2 microphones with distances 20 and 40 cm in Figure 14 and we can see that there are no significant differences. From the result, it can be concluded

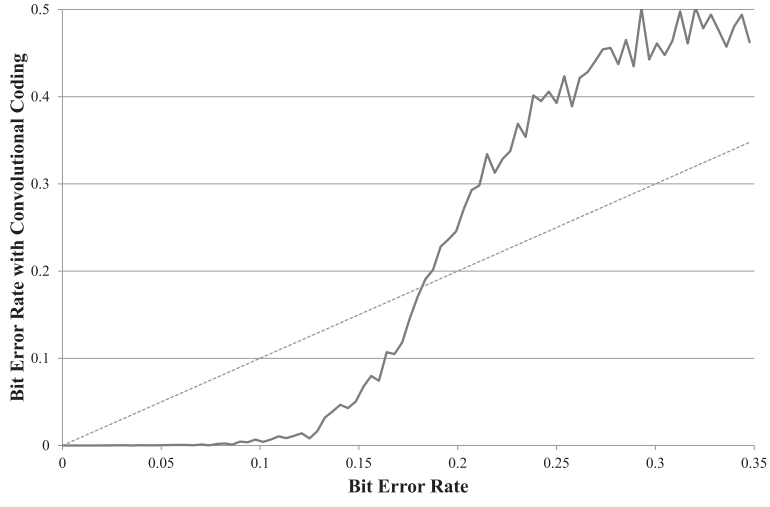

Figure 13 BER with 1/3 convolutional coding as a function of BER (solid line). The horizontal and vertical axes represent BER before and after error correction, respectively. The dotted line denotes the identity line.

Table 5 Bit error rate for distorted test audio clips with various types of signal processing applied

	S_S	S_L	S_{LA1}	S_{LA3}
No Attack	0.0000	0.0000	0.0008	0.0013
Requantize (8 bit)	0.0000	0.0000	0.0018	0.0023
Invert	0.0000	0.0000	0.0008	0.0013
MP3 (192 kbps)	0.0000	0.0000	0.0015	0.0022
MP3 (128 kbps)	0.0005	0.0006	0.0055	0.0069
MP3 (64 kbps)	0.0051	0.0049	0.0168	0.0210
AWGN (0 dB)	0.0668	0.0492	0.0880	0.0951
AWGN (5 dB)	0.0217	0.0088	0.0231	0.0274
AWGN (10 dB)	0.0000	0.0000	0.0061	0.0071

that it is hard to achieve a dramatic improvement in statistical independence between channels when the size of the receiver is restricted to the dimension of small mobile devices.

7 Conclusions

In this paper, we have proposed an ADT method based on audio data hiding which modifies and extends our previous works. Moreover, we evaluated the performance of the proposed ADT method with various MCLT lengths, number of microphones, and with and without SSA algorithm.

From the series of experiments, the MCLT frame length has been found to be one of the most important parameters in ADT system. The transmission performance improves as the length of the MCLT frame gets longer at the cost of the audio quality. Because the length of MCLT window is a fixed parameter known to both the embedder and receiver, it should be determined carefully depending on the target applications.

To determine a proper length of an MCLT frame, the reverberation time of the channel should be considered. A long MCLT window, for example, can achieve a good transmission performance in a reverberant environment such as a living room or cafeteria. On the other hand, for the applications dedicated to very short distance such as device-to-device data transmission, a short MCLT window might be preferable because it yields better audio quality. In the highly reverberant conditions, however, it is considered doubtful to find an optimal MCLT length guaranteeing both good audio quality and transmission performance without the SSA algorithm.

In this respect, the SSA algorithm would be indispensable when implementing a practical ADT system suitable for highly reverberant conditions. Using a long MCLT window with the SSA algorithm makes the audio data hiding system more suitable for the ADT applications. In addition, a good trade-off between the audio quality and data transmission performance is achieved by adjusting only a single parameter in the SSA algorithm.

When a receiver can utilize multiple microphones, the proposed multi-channel methods have been found effective in enhancing the transmission reliability while preserving backward compatibility with conventional ADT systems. From the results, we can see that exploiting more microphones enhances the transmission reliability further. However, achieving greater statistical independence between the channels is difficult when the distance

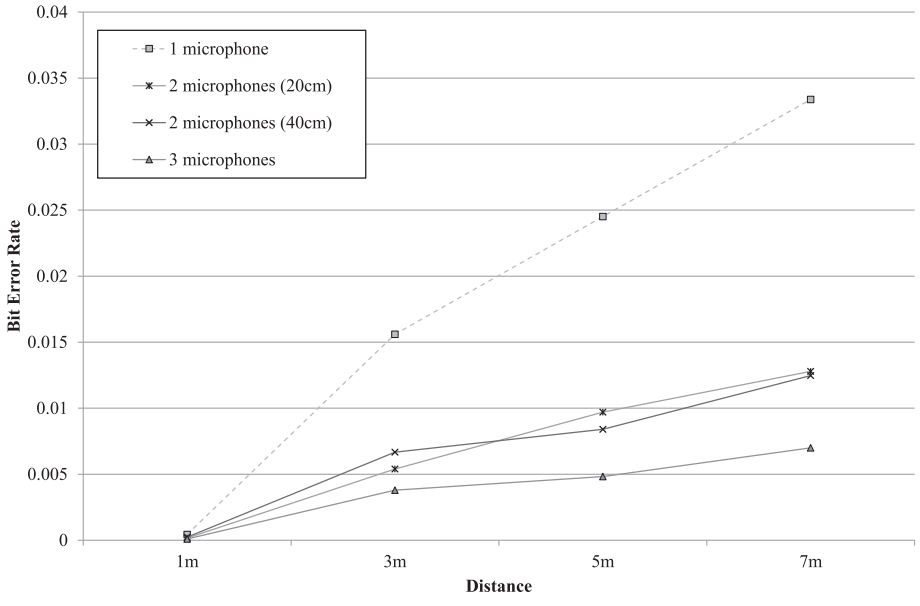

Figure 14 BER of the proposed combining method with an MCLT length of 8,192. The horizontal axis represents the distance between loudspeaker and microphones in a real room. Each line represents a different number of microphones.

between microphones at the receiver is restricted to the range of typical mobile devices. Despite of the dependency between each channel, the proposed multi-channel method has shown to be useful when applied in practical applications.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & communications Technology Promotion).

Author details
[1]Samsung Electronics co. Ltd., 129 Samsung-ro, Yeongtong-gu, 443-742 Suwon, Korea. [2]Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, 151-742 Seoul, Korea.

Received: 29 November 2014 Accepted: 20 March 2015
Published online: 18 April 2015

References

1. N Cvejic, T Seppanen, *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. Applications and Benchmarks. (IGI Global, Hershey, PA, USA, 2007)
2. I Cox, M Miller, J Bloom, J Fridrich, T Kalker, *Digital Watermarking and Steganography*, 2nd edn. (Morgan Kaufmann, Burlington, MA, USA, 2007)
3. Y Nakashima, R Tachibana, N Babaguchi, Watermarked movie soundtrack finds the position of the camcorder in a theater. Multimedia IEEE Trans. **11**(3), 443–454 (2009)
4. N Lazic, P Aarabi, Communication over an acoustic channel using data hiding techniques. Multimedia IEEE Trans. **8**(5), 918–924 (2006)
5. PW Chen, CH Huang, YC Shen, JL Wu, in *Acoustics, Speech and Signal Processing (ICASSP), 2009. IEEE International Conference On*. Pushing information over acoustic channels (Taipei, Taiwan, 2009), pp. 1421–1424
6. Y Suzuki, R Nishimura, H Tao, in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2006. IEEE International Conference On*. Audio watermark enhanced by LDPC coding for air transmission (Pasadena, CA, USA, 2006), pp. 23–26
7. CV Lopes, PMQ Aguiar, in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop On*. Aerial acoustic communications (New Paltz, NY, USA, 2001), pp. 219–222
8. K Mizutani, N Wakatsuki, K Mizutani, Acoustic communication in air using differential biphase shift keying with influence of impulse response and background noise. Jpn. J. Appl. Phys. **46**(7B), 4541–4544 (2007)
9. GD Galdo, J Borsum, T Bliem, A Carciun, S Krägeloh, in *Acoustics, Speech and Signal Processing (ICASSP), 2011. IEEE International Conference On*. Audio watermarking for acoustic propagation in reverberant environments (Prague, Czech Republic, 2011), pp. 2364–2367
10. H Matsuoka, Y Nakashima, T Yoshimura, Acoustic OFDM system and its extension. Vis. Comput. **25**(1), 3–12 (2008)
11. HS Yun, K Cho, NS Kim, Acoustic data transmission based on modulated complex lapped transform. Signal Process. Lett. IEEE. **17**(1), 67–70 (2010)
12. K Cho, HS Yun, NS Kim, Robust data hiding for MCLT based acoustic data transmission. Signal Process. Lett. IEEE. **17**(7), 679–682 (2010)
13. HS Yun, K Cho, NS Kim, Spectral magnitude adjustment for MCLT-based acoustic data transmission. Trans. Inform. Syst. IEICE. **E95-D**(5), 1523–1526 (2012)
14. K Cho, J Choi, YG Jin, NS Kim, in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2012. IEEE International Conference On*. Quality enhancement of audio watermarking for data transmission in aerial space based on segmental SNR adjustment (Piraeus, Greece, 2012), pp. 122–125
15. HS Malvar, Fast algorithm for the modulated complex lapped transform. Signal Process. Lett. IEEE. **10**(1), 8–10 (2003)
16. JJ Garcia-Hernandez, C Feregrino-Uribe, R Cumplido, C Reta, On the implementation of a hardware architecture for an audio data hiding system. J. Signal Process. Syst. **64**(3), 457–468 (2011)
17. YS Cho, J Kim, WY Yang, CG Kang, *MIMO-OFDM Wireless Communications with MATLAB*. (John Wiley & Sons, Chichester, UK, 2010)
18. H Schulze, C Lueders, *Theory and Applications of OFDM and CDMA*. (John Wiley & Sons, Chichester, UK, 2005)
19. International Standard ISO/IEC 11172-3 (MPEG), Information technology - coding of moving pictures and associated audio for digital storage media up to about 1.5 mbit/s. Part 3: Audio (1993)
20. K Fazel, S Kaiser, *Multi-Carrier and Spread Spectrum Systems*. (John Wiley & Sons, Chichester, UK, 2008)
21. P Hoeher, S Kaiser, P Robertson, in *Acoustics, Speech, and Signal Processing (ICASSP), 1997. IEEE International Conference On*. Two-dimensional pilot-symbol-aided channel estimation by Wiener filtering (Munich, Germany, 1997), pp. 1845–1848
22. MR Schroeder, New method of measuring reverberation time. J. Acoustics Soc. Am. **37**(3), 409–412 (1965)
23. SG McGovern, Fast image method for impulse response calculations of box-shaped rooms. Appl. Acoustics. **70**(1), 182–189 (2009)
24. P Kabal, An examination and interpretation of ITU-R BS. 1387: perceptual evaluation of audio quality. TSP Lab Technical Report, Dept. Electrical & Computer Engineering, McGill University (2002)
25. KK Paliwal, LD Alsteris, On the usefulness of STFT phase spectrum in human listening tests. Speech Commun. **45**(1), 153–170 (2005)
26. BCJ Moore, *An Introduction to the Psychology of Hearing* 6th edn. (BRILL, Leiden, Netherlands, 2012)
27. D Kirovski, HS Malvar, Spread-spectrum watermarking of audio signals. Signal Process. IEEE Trans. **51**(4), 1020–1033 (2003)
28. ITU-R Recommendation BS. 1534, Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA) (2003)
29. CH Chung, SH Cho, S Kang, YW Lee, in *Personal, Indoor and Mobile Radio Communications IEEE International Symposium On*. Performance of convolutional coded and uncoded DS/CDMA system in Nakagami fading channels (Toronto, Canada, 1995), pp. 502–506
30. BS Ko, R Nishimura, Y Suzuki, Robust Watermarking based on time-spread echo method with subband decomposition. Fundamentals, IEICE Trans. **E87-A**(6), 1647–1650 (2004)
31. Y Xiang, D Peng, I Natgunanathan, W Zhou, Effective Pseudonoise Sequence and decoding function for imperceptibility and robustness enhancement in time-spread echo-based audio watermarking. Multimedia IEEE Trans. **13**(1), 2–13 (2011)