Method of Audio Watermarking Based on Adaptive Phase Modulation

Nhut Minh NGO^{†a)}, Student Member and Masashi UNOKI^{†b)}, Senior Member

SUMMARY This paper proposes a method of watermarking for digital audio signals based on adaptive phase modulation. Audio signals are usually non-stationary, i.e., their own characteristics are time-variant. The features for watermarking are usually not selected by combining the principle of variability, which affects the performance of the whole watermarking system. The proposed method embeds a watermark into an audio signal by adaptively modulating its phase with the watermark using IIR all-pass filters. The frequency location of the pole-zero of an IIR all-pass filter that characterizes the transfer function of the filter is adapted on the basis of signal power distribution on sub-bands in a magnitude spectrum domain. The pole-zero locations are adapted so that the phase modulation produces slight distortion in watermarked signals to achieve the best sound quality. The experimental results show that the proposed method could embed inaudible watermarks into various kinds of audio signals and correctly detect watermarks without the aid of original signals. A reasonable trade-off between inaudibility and robustness could be obtained by balancing the phase modulation scheme. The proposed method can embed a watermark into audio signals up to 100 bits per second with 99% accuracy and 6 bits per second with 94.3% accuracy in the cases of no attack and attacks, respectively.

key words: audio watermarking, blind detection, adaptive phase modulation, all-pass filter, chirp-z transformation

1. Introduction

Development in multimedia and communication technologies has made life more convenient but at the same time put the security of digital audio at risk. Reproducing and redistributing multimedia data easily has led to demands for preventing copyright infringements. Conventional data-protection methods such as encryption are inconvenient for digital audio because audio data need to be decrypted before being played back. In that context, audio watermarking, a technique that hides special codes in the actual content of audio without any affecting its normal use, has been proposed to solve these problems [1]. Watermarking has become a promising technique for not only copyright protection but also issues such as copy control, tampering detection, and covert communication [2].

To be effectively applied in practice, an audio watermarking method must satisfy four basic requirements [1]: (i) *inaudibility*—keeping watermarks inaudible to listeners dur-

ing normal use, (ii) *blindness*—detecting watermarks without the need of original signals, (iii) *robustness*—keeping watermarks able to withstand signal processing operations or intentional attacks from illegal users, and (iv) *high embedding capacity*—conveying a large amount of data for applications such as covert communication. Among these requirements, inaudibility and robustness for audio watermarking are the most challenging because they conflict. It is straightforward that to provide an inaudible watermark, perceptually insensitive features of the audio should be exploited for embedding. However, robustness against attacks such as lossy compression is then difficult to achieve because the watermark can be easily removed without degrading the sound quality. We usually find a trade-off among these requirements when implementing a watermarking system for digital audio.

In general, audio watermarking methods embed watermarks directly into audio samples in the time domain or audio features in the transformed domain. Typical methods in the time domain embed watermarks by modifying least significant bits (LSB) or insert watermarks that are perceptually shaped in accordance with the human auditory system (HAS) [3], [4]. Although these methods could embed inaudible watermarks into audio signals, they are not robust against processing (e.g., LSB) and have low embedding capacity [4]. Methods in the transformed domain exploit advantages of simultaneous masking characteristics of HAS to embed inaudible watermarks such as the direct spread spectrum method (DSS) [5] and echo hiding method [6]. The DSS method is relatively robust against attacks since watermarks are spread throughout the whole frequency spectrum. However, it does not satisfy all the requirements, especially inaudibility. Echo hiding method is advantageous in inaudibility because adding the echo to the host signal essentially emphasizes the host signal, making the signal louder perceived, hence the watermark is basically inaudible to HAS. Nevertheless, it has a conflict between robustness and inaudibility with regard to echo magnitude and delay time. It has been reported that controlled phase alteration results in inaudible change in sound to HAS [7], [8]. This characteristic of HAS has been extensively exploited for inaudible audio watermarking with the technique of phase modulation and promising results have been reported [9]–[15].

Audio watermark methods based on phase modulation such as that of Dong et al. [11] and the cochlear delay method [15] adopt a static scheme of phase modulation in which all the frames of signals are treated in the same

Manuscript received March 30, 2015.

Manuscript revised August 8, 2015.

Manuscript publicized October 21, 2015.

[†]The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923–1292 Japan.

a) E-mail: nmnhut@jaist.ac.jp

b) E-mail: unoki@jaist.ac.jp

DOI: 10.1587/transinf.2015MUP0019

Fig. 1 Waveform and power spectrogram of two pieces of music: (a) waveform of a pop song, (b) waveform of a dance song, (c) spectrogram of (a), and (d) spectrogram of (b).

manner. However, audio signals in practice are usually non-stationary or, in other words, their perceptually insignificant features that are potentially suitable for embedding inaudible watermarks also vary with time. As a result, a watermark is not audible for some signals but may be for others. This property should be thoroughly considered for designing a watermark embedding algorithm in order to ensure the same watermarking performance for different kinds of signals.

On the basis the above consideration, we preliminarily studied feasibility of an audio watermarking method based on adaptive phase modulation [16]. Depending on a local characteristic of each signal frame, a suitable phase modulator is chosen to ensure inaudibility or robustness on demand. More precisely, a portion of modification for watermark embedding in the phase spectral domain is adapted to the local signal energy distribution so that humans can perceive watermarks as the slightest distortion or watermarks can withstand most attacks. The experimental results show that the method can produce watermarked sound with better sound quality in comparison with other static watermarking methods. However, the method works under an assumption that frame positions were available in the watermark detection process, which may be unrealistic in some scenarios. Moreover, the method meets a trade-off between inaudibility and robustness.

In this paper, we propose an audio watermarking method based on adaptive phase modulation with automatic frame synchronization and good balance between inaudibility and robustness. Unlike conventional techniques for frame synchronization that embed synchronization codes along with watermarks, the proposed algorithm uses cues for detecting watermarks as a hint for detecting frame positions. Since the proposed method is implemented on a frame basis, the artifacts at frame boundaries deeply affect the sound quality, i.e., shorter frames cause larger degradation in watermarked signals. To further improve sound qual-

ity without reducing the watermark bit rate, we increase the frame size and embed multiple bits into one frame instead of embedding a single bit into a frame.

The rest of this paper is organized as follows. Section 2 describes the concept of the proposed method, the details of watermark embedding and detection processes, frame synchronization, and illustrative examples for these processes. We present the results of evaluations with regard to effectiveness of the proposed adaptive scheme, inaudibility, robustness, and comparison with conventional watermarking methods in Sect. 3. Discussions are given in Sect. 4. Section 5 gives some concluding remarks and suggests a direction for future work.

2. Proposed Method

2.1 Principle of Adaptive Phase Modulation

Real world audio signals are non-stationary, i.e., their characteristics vary with time. For this reason, acoustic signal processing applications usually divide signals into frames for non-stationary signal analysis. An obvious example of the variability characteristic of audio signals is that their power distribution over frequency components is time-variant. Figure 1 shows two pieces of music with their power distribution in the time-frequency domain. The top panels show the waveforms and the bottom ones show their power spectrogram. For the left signal, some periods have power distributed over the whole range of frequencies and the others have power distributed in only low frequencies. For the right signal, most periods have power distributed on low and high region of frequencies. The variability characteristic of audio signals should be taken into account for designing a watermarking algorithm to achieve the best performance.

Audio watermarking methods based on phase modula-

Fig. 2 Characteristics of second-order APFs with respect to θ : (a) phase response, (b) group delay, and (c) pole-zero locations in z-plane.

tion usually take advantage of controlled phase alteration resulting in relatively inaudible change to human ears [7], [8]. The phase of an audio signal can be manipulated by an all-pass filter (APF). An APF passes all the frequency components equally but changes the phase in accordance with the filter phase response. The principle of embedding a watermark into an audio signal is as follows. The audio signal is first split into frames. In accordance with the watermark in each frame, a corresponding APF is designed. The watermark is then embedded into a frame by using the designed filter to enhance the phase spectrum of the audio frame. Since enhancing the audio phase by an APF results in relatively inaudible change, the watermarked frame is perceived without any severe difference. However, an issue occurring is that sound quality is affected by the abrupt change at frame boundaries due to phase shift. Longer frames may produce fewer artifacts. These issues should be thoroughly taken into consideration when implementing a watermarking method based on phase modulation.

In practice, a second-order infinite impulse response (IIR) APF can be used to flexibly control phase change in audio signals by adjusting the modulus and angle of the filter's poles and zeros. This filter is represented as follows:

$$H(z) = \left(\frac{-\beta^* + z^{-1}}{1 - z^{-1}\beta} \right) \left(\frac{-\beta + z^{-1}}{1 - z^{-1}\beta^*} \right), \quad (1)$$

where $\beta = re^{j\theta}$ and β^* is a complex conjugate of β . It has a pair of zeros ($\frac{1}{\beta}$ and $\frac{1}{\beta^*}$) and a pair of poles (β and β^*) that are symmetric on the real axis in z-plane. A second-order IIR APF can be characterized by the modulus and angle of its poles and zeros, i.e., r and θ , respectively. Figure 2 shows the phase characteristics and pole-zero locations of second-order APFs with respect to the pole-zero frequency θ . Because of the nature of APFs, the magnitude spectrum of a filtered signal is not manipulated but its phase spectrum is changed as follows:

$$\angle Y(e^{j\omega}) = \angle X(e^{j\omega}) + \angle H(e^{j\omega}), \quad (2)$$

where $\angle X(e^{j\omega})$, $\angle Y(e^{j\omega})$, and $\angle H(e^{j\omega})$ are the phase spectrum of the input signal, the phase spectrum of the output signal, and the phase response of the APF, respectively.

The absolute phase response and the group delay have a peak at the frequency that corresponds to the angle of the filter pole-zero θ (referred to hereafter as pole-zero frequency). According to Eq. (2), this frequency can be interpreted as the location where the signal is modified the most in the phase spectrum. We have observed that the region around the pole-zero frequency is more affected than the outside region. In other words, if a signal has its energy concentrating at the pole-zero frequency then the filtered signal has faster disruption at frame boundaries than the signal having its energy outside of the pole-zero frequency. From these observations, we found that we can design an IIR APF with its pole-zero frequency lying on a low energy region in the spectrum domain to achieve the best sound quality. On the other hand, artifacts at frame boundaries could be reduced by increasing frame size. To further improve sound quality while keeping the same bit rate, we consider embedding more bits in a frame. These are the key principle for our proposed watermarking method based on frame-adaptive phase modulation.

2.2 Watermark Embedding

Second-order IIR APFs are used to realize the adaptive phase modulation scheme. N_{bpf} watermark bits are embedded into a frame. The modulus r of the APF is m -ary modulated with the embedded bits, i.e., $r_0, r_1, \dots, r_{(m-1)}$ are used ($m = 2^{N_{\text{bpf}}}$) for the corresponding bits. The angle θ is adapted to power distribution in sub-bands of audio frames. In each frame, the pole-zero angle θ is determined by the center frequency of the lowest power sub-band.

Figure 3 depicts a block diagram of the proposed embedding process. Watermark $s[i]$ is embedded into an original signal, $x[n]$, as follows.

Step 1. Original signal $x[n]$ is first split into frames

Fig. 3 Watermarking based on phase modulation: embedding process.

$x_i[n]$, in which the frame length $N = f_s/N_{\text{bps}}$; f_s is the sampling frequency and N_{bps} is the number of watermark bits per second (bps). Each frame is filtered by the APF, which is designed in the next two steps.

Step 2. Frame $x_i[n]$ is then decomposed into K sub-bands with equivalent bandwidth. The power of each sub-band is calculated and the lowest power sub-band index, k , is identified by Eq. (3). The output of this step is the parameter θ , i.e., the center frequency of the k -th sub-band that is calculated by Eq. (4).

$$k = \arg \min_{\ell} \left\{ 10 \log_{10} \left(\sum_{\omega} |X_{\ell}(e^{j\omega})|^2 \right) \right\}, \quad (3)$$

where $X_{\ell}(e^{j\omega})$ is the Fourier transform of the ℓ -th sub-band of the frame $x_i[n]$; $k \in \{1, 2, \dots, K\}$.

$$\theta_k = \frac{(2k-1)\pi}{2K} \quad (\text{rad}) \quad (4)$$

Step 3. The watermark bits represented by the filter parameter r are embedded into the frame in this step. The values of parameter r , r_0 , r_1 , $\dots$, and $r_{(m-1)}$, are determined in advance by experimental analysis. In accordance with watermark bits $s[i]$, r is set to r_b , where $b = \text{base}_{10}(s[i])$ ($\text{base}_{10}(\cdot)$ is the function that converts a binary number into a decimal number). For example, if $m = 4$ and $s[i] = '10'$, then $r = r_2$. **Step 4.** Filter $H(z)$ is designed with two parameters, r and θ , by Eq. (1). Then, frame $x_i[n]$ is filtered by $H(z)$.

Steps 2–4 are repeated until we reach the final frame.

Step 5. Finally, the filtered frames are joined together by the non-overlap and add (non-OLA) technique to yield watermarked signal $y[n]$.

2.3 Watermark Detection

To detect embedded watermark bits in a frame, each watermarked frame is analyzed to determine the filter used to process that frame in the embedding process. Before that, frame synchronization is performed to identify frame positions. Note that the filtering process can be performed in the frequency domain as follows:

$$Y(z) = X(z)H(z), \quad (5)$$

Fig. 4 Watermarking based on phase modulation: detection process.

where $X(z)$ and $Y(z)$ are z -transform of the original frame and the watermarked frame, respectively. Let z_0 be a zero of the filter $H(z)$. As the definition of a zero of an IIR filter, $H(z_0) = 0$. Hence, $Y(z_0) = 0$. This is a hint for blind watermark detection. We can use chirp- z transformation (CZT), a tool for evaluating z -transform at any M -arbitrary points in the z -plane, [17] to calculate $Y(z_0)$ as:

$$Y(z) = \sum_{n=0}^{N-1} y[n] A^n W^{nk}, \quad (6)$$

where $k = 0, 1, \dots, M-1$, $A = A_0 \exp(j2\pi\theta_0)$, $W = W_0 \exp(j2\pi\phi_0)$, and θ_0 and ϕ_0 are the initial phases. The case $A = 1$, $M = N$, and $W = \exp(-j2\pi/N)$ corresponds to the discrete Fourier transform (DFT). The CZT can be used in various types of signal processing such as a zoom DFT, finer spectrum analysis, and arbitrary poles in transfer analysis [18].

CZT over the zero location of the APF exposes a minimum at the pole-zero frequency that corresponds to the pole-zero angle θ in the log-magnitude spectrum. On the basis of the zero location over which the minimum occurs, we can determine which APF was used and then detect the embedded bits.

Figure 4 depicts a block diagram of the proposed detection process. A watermark is detected in the five following steps.

Step 1. Frame synchronization is performed. The details of the algorithm for frame synchronization are described in the next subsection.

Step 2. Watermarked signal $y[n]$ is first split into the frames $y_i[n]$ in which the frame length is the same as that in the embedding process.

Step 3. Parameter θ_k is determined as Step 2 in the embedding process. Since the APF does not change the magnitude spectrum of the output signal, the power distribution of the watermarked frames in sub-bands could be obtained that was the same as that of the original frames.

Step 4. Frame $y_i(n)$ is then analyzed by m types of CZT over m contours with radii of $1/r_{ir}$, $ir = \{0, 1, \dots, m-1\}$. The outputs are $Y_{ir}(z)$, $ir = \{0, 1, \dots, m-1\}$.

Step 5. The log-magnitudes at θ_k , $|Y_{ir}(\theta_k)|$, $ir = \{0, 1, \dots, m-1\}$, are compared together to detect watermark $\hat{s}[i]$ as follows.

$$\hat{s}[i] = \text{base}_2(\arg \min_{ir} |Y_{ir}(\theta_k)|), \quad (7)$$

Fig. 5 Overlap of group delay between two successive bands with respect to K : (a) $K = 5$, (b) $K = 10$, and (c) $K = 20$.

where $Y_{ir}(\theta_k) = Y(z)|_{z=\frac{1}{r}e^{j\theta_k}}$ and $\text{base}_2(\cdot)$ is the function that converts a decimal number into a binary number.

Steps 3–5 are repeated for the total number of frames.

2.4 Parameter Settings

To realize the proposed watermarking scheme, we have to determine the specific parameter values for the number of sub-bands, K and the pole-zero modulus, r .

2.4.1 Determination of Number of Sub-Band K

We proposed an audio watermarking method based on adaptive phase modulation, in which the parameter θ is adapted to the power distribution of the audio signal in sub-bands. The suitable number of sub-bands, K , needs to be specified so that the adaptation of phase modulation is effective. Essentially, the group delay when θ is set to the center frequency of a band, is overlapped with those of other successive bands. Figure 5 illustrates the overlap of group delays between two successive bands. The overlap is shown in the gray area. In the case $K = 20$, the overlap of the group delays is wide, resulting in no discrimination between two bands. In other words, the phase modification of these two bands are similar. When K decreases, the overlap becomes narrower. The value of K should be chosen so that the overlap of the group delays between two successive bands is not too large. By contrast, the too small K causes less variation of the band, unable to efficiently take advantage of adaption. Accordingly, the recommended values for K are from 4 to 8. Parameter K was set to five in this paper.

Fig. 6 Discontinuity caused by phase modulation with respect to parameter r_0 : (a) $r_0 = 0.2$, (b) $r_0 = 0.5$, and (c) $r_0 = 0.8$.

2.4.2 Determination of Optimal Parameter r

By definition, $r > 0$ and $r \neq 1$. The filter phase responses with $r > 1$ is symmetric with those with $0 < r < 1$, hence we investigate $0 < r < 1$. The parameters r should be chosen so that the phase modulation does not degrade the sound quality of the audio signal severely. As the nature of second-order IIR AFP, when the value of r decreases, the region of phase modification become larger, affecting more to the sound quality. Figure 6 shows the discontinuity caused by the phase modulation at a frame boundary in the watermarked signal. When r increases, the discontinuity decreases and hence the sound quality becomes better. Therefore, the values of r should be kept large. We fixed the highest value of r at 0.99 in this paper.

On the other hand, if two r s, e.g., r_0 and r_1 , are two close to each other, the tracking of r in the detection process become difficult, resulting in incorrect detection of the watermark. Therefore, to detect the watermark precisely, we have to determine the optimal difference between r s. We carried out bit detection tests to investigate the relationship between the accuracy of the watermark detection and the difference between r s. All the 102 tracks of RWC music database [19] were used as the original signals. Each music track has a sampling frequency of 44.1 kHz and 16-bit quantization and is cut to be 60 seconds long. The bit rate was set to 6 bps. Parameter N_{bpf} was set to one; r_1 was set to 0.99; r_0 was set to 0.989, 0.988, ..., and 0.945. The watermarked signals were processed by MP3 128 kbps before the watermark was detected. Figure 7 shows the bit detection rate of the watermark with respect to r_0 . When r_0 decreases, the bit detection rate increases. When r_0 is less than or equal to 0.955, the bit detection rates are almost constant. It suggested that the difference between r_0 and r_1 should be kept

Fig. 7 Bit detection rate with respect to parameter r_0 , where $r_1 = 0.99$.

at 0.035 to ensure that the watermark can be most accurately detected.

Accordingly, we set the parameters r to 0.885, 0.92, 0.955, and 0.99 for the 4-ary scheme, and 0.745, 0.78, 0.815, 0.85, 0.885, 0.92, 0.955, and 0.99 for the 8-ary scheme. For the 2-ary scheme, although r s can be set to 0.955 and 0.99, we set them to 0.92 and 0.99 for higher accuracy while the sound quality is not affected too much.

2.5 Frame Synchronization

The detection process works with an assumption that the frame positions are synchronized. In practice, this assumption may be invalid, so we have to identify the correct starting point before detecting watermarks. On the basis of the mechanism of watermark detection, we can see that CZT of a correct watermarked frame over the zero location of the IIR APF theoretically exposes a negative infinite at the pole-zero frequency in the log-magnitude spectrum. The negative infinite can be used as a hint for realizing frame synchronization.

We can search for a correct frame position over a frame length of signal. Obviously, if we select a correct frame $y_i[n]$, the log-magnitude at θ_k of CZT of this frame over one of the contours is approximate to a negative infinite. This log-magnitude can be obtained by:

$$|Y_p(\theta_k)| = \min |Y_{ir}(\theta_k)|, \quad ir \in \{0, 1, \dots, m-1\}, \quad (8)$$

Then the average differences between $|Y_p(\theta_k)|$ and $|Y_{ir}(\theta_k)|$, $ir \neq p$, which is calculated by:

$$d = \frac{1}{m-1} \sum_{b=0}^{m-1} \left| |Y_p(\theta_k)| - |Y_b(\theta_k)| \right|, \quad (9)$$

is very large. By contrast, an incorrectly selected frame has a small d . We can shift the frame sample by sample and calculate d for each frame. The frame with d which is the maximum is a correct frame. Once the correct starting point is identified, all the frame positions can be synchronized.

Figure 8 illustrates frame position detection. We computed d for a ten-frame period of signal with the initial point randomly selected in two cases: (b) without and (c) with

Fig. 8 Frame synchronization in the case frame length = 167 ms: (a) watermarked signal, (b) synchronization on original signal, and (c) synchronization on watermarked signal.

Fig. 9 Example on modulation of r with embedded bits and adaptively adjusting θ : (a) four audio frames and the value of r in each frame; (b), (c), (d), and (e) power spectra of frame #1, frame #2, frame #3, and frame #4 and determination of θ in each frame.

a watermark embedded. There is no obvious maximum in Case (b) while high peaks occur at the correct frame positions in Case (c).

2.6 Example

Figure 9 shows an example of watermark embedding at the bit rate of 4 bps and $N_{\text{bpf}} = 1$. Watermark bits $s = '1001'$ were embedded into four frames. Parameter r was set to either r_0 or r_1 based on $s[i]$. The total number of sub-bands, K , is five. Parameter θ was set differently for each frame on the basis of its power distribution in sub-bands. For example, frame #1 has $\theta_{\#1} = \frac{3\pi}{10}$ and frame #2 has $\theta_{\#2} = \frac{\pi}{2}$ as shown in Figs. 9 (b) and (c).

Figure 10 shows an example of detecting the above embedded bits in the first two frames. The pole-zero frequency, θ , was determined in the same way as in the embedding process. The CZT of $y_1[n]$ over $\frac{1}{r_1}$ exposes a minimum at $\theta_{\#1}$

Fig. 10 Minima in CZT spectra of watermarked signals: (a) and (b) show CZT of original frames #1 and #2 over $\frac{1}{r_0}$, (c) and (e) show CZT of watermarked frame #1 over $\frac{1}{r_0}$ and $\frac{1}{r_1}$, respectively, and (d) and (f) show CZT of watermarked frame #2 over $\frac{1}{r_0}$ and $\frac{1}{r_1}$, respectively. Bits of '1' and '0' were embedded into $y_1[n]$ and $y_2[n]$, respectively. Minima at $\theta_{\#1}$ in (d) and at $\theta_{\#2}$ in (e) indicate that the watermark bits were correctly detected.

that decodes bit '1' while the CZT of $y_2[n]$ over $\frac{1}{r_0}$ exposes a minimum at $\theta_{\#2}$ that decodes bit '0'. The two bits were correctly detected.

3. Evaluations

3.1 Database and Conditions

We carried out several experiments to evaluate the proposed method with the Real World Computing (RWC) music database [19], which consists of 102 music tracks of various music genres. Each music track has a sampling frequency of 44.1 kHz and 16-bit quantization and is cut to be 60 seconds long. The watermark was set as random binary data. The bit rate, N_{bps} , was chosen from 6 to 400 bps. We investigated the m -ary modulation scheme with $m = 2, 4$, and 8 (1, 2, and 3-bit encoding in a frame). The total number of sub-bands, K , was set to five. The parameters r were set to 0.92 and 0.99 for 2-ary scheme, 0.855 to 0.99 with the step of 0.035 for 4-ary scheme, and 0.745 to 0.99 with the step of 0.035 for 8-ary scheme.

Inaudibility of the proposed method was objectively tested by a perceptual evaluation of audio quality (PEAQ) [20]. PEAQ was used to measure degradation in audio in accordance with the objective difference grade (ODG), which consists of five grades: -4 (very annoying), -3 (annoying), -2 (slightly annoying), -1 (perceptible but not annoying), and 0 (imperceptible). The evaluation criterion for PEAQ was set to an average ODG of -1. The watermark detection accuracy is measured by bit detection rate (BDR), which is defined as the ratio between the number of correctly detected bits and the total number of embedded bits.

Fig. 11 Objective evaluation on PEAQ of the proposed method in different music genres.

3.2 Effectiveness of Proposed Method

To confirm the adaptation effectiveness of the proposed method, we comparatively evaluated the proposed scheme based on adaptive phase modulation with schemes based on non-adaptive phase modulation. Non-adaptive schemes were implemented by setting a fixed value of APF parameter θ instead of frame-variant parameter θ in the adaptive scheme. Parameters θ of $\pi/32$, $\pi/16$, $\pi/8$, and $\pi/4$ were chosen in the non-adaptive schemes.

Figure 11 shows the average ODG in different music genres for the proposed method in comparison with the non-adaptive schemes. The bit rate was set to 6 bps in this experiment. The proposed method has good sound quality ($\text{ODG} \geq -1$) for all the genres except for classical, while some non-adaptive schemes have bad quality for a few genres. For all the genres, the proposed method has better sound quality than the non-adaptive schemes. For the proposed method, total average ODGs are greater than -0.47 and more than 89% of all the music tracks have ODG greater than -1. The non-adaptive schemes seem to have a trend that when θ increases, the ODGs also increase. However, for the jazz, march, and vocal genres, the sound quality is best when θ is equal to $\pi/8$.

Figure 12 depicts the average ODGs with respect to the embedding bit rate. For the bit rates less than 100 bps, the adaptive scheme satisfies the evaluation criterion and has better sound quality performance than the non-adaptive schemes. The proposed method has the best sound quality with the 4-ary scheme. When the bit rate increases, the sound quality decreases for both the adaptive and non-adaptive schemes. This is because the more bits embedded into audio signals, the more degraded the quality of the watermarked signals.

3.3 Subjective Evaluation

The subjective listening test was carried out in a sound-

Fig. 12 Objective evaluation on PEAQ of the proposed method with respect to embedding bit rate.

proof room. The audio was played via an amplifier (Audio-Technica Stereo Headphone Amplifier AT-HA5000) and headphones (Sennheiser HDA 200). Twenty tracks (nos. 5, 9, 10, 12, 14, 21, 22, 23, 26, 27, 29, 58, 63, 86, 88, 90, 95, 97, 98, and 99) in the RWC music database were used. The watermark is the same random binary data. The bit rate was set to 4 bps. These experimental configurations were those referred to by Unoki and Hamada [21].

Eight voluntary subjects participated in the experiment. The subjects are asked to listen to a pair of signals (a reference signal and a test signal) and then score their similarity in accordance with a subjective scale consisting of four scores: 0 (completely the same), 1 (probably the same), 2 (probably different), and 3 (completely different). For each track, there were four pairs in a random order, in which the first track was the original (Org) and the other is the original track or an watermarked track created by the proposed method with 4-ary (APM), the cochlear delay watermarking method with blind detection (Cdb) [15], or the direct spread spectrum (DSS) method [5]. Each subject evaluated the similarity of 80 test pairs in total, 20 tracks multiplied by four pairs (Org-Org, APM-Org, Cdb-Org, and DSS-Org).

Figure 13 shows the mean scores of four test pairs for each track. The mean scores of the pairs APM-Org are approximate to those of the pairs Org-Org, suggesting that the distortion caused by the watermark embedding of the proposed method is barely perceptible. By contrast, the distortion caused by the Cdb method and the DSS method is slightly perceptible and perceptible for most of the tracks.

3.4 Robustness Evaluation

We carried out experiments to evaluate robustness of the proposed method against the following processing: addition of white Gaussian noise (AWGN) with an SNR of 36 dB, MP3 compression at 128 kbps, MPEG4 HE-ACC (MP4) compression at 96 kbps, re-quantization to 8 bits, re-sampling to 22 kHz and 16 kHz, and bandpass filtering (BPF) with the passband between 100 Hz and 6 kHz and the stopband of -12 dB per octave. The embedding bit rate was

Fig. 13 Results of subjective listening test.

Table 1 Inaudibility and robustness evaluation with two configurations.

Measure	Config. 1			Config. 2		
	2-ary	4-ary	8-ary	2-ary	4-ary	8-ary
PEAQ (ODG)	-0.42	-0.35	-0.47	-1.23	-0.69	-0.77
BDR (%)						
No processing	99.08	99.56	99.51	99.75	99.96	99.76
Resample 22k	44.35	49.71	49.89	99.22	99.11	93.19
Resample 16k	43.99	49.52	49.78	98.27	98.21	89.08
AWGN 36 dB	83.07	81.23	79.91	97.43	97.98	96.60
Requantize	73.97	74.59	72.30	92.06	95.33	91.61
MP3 128 kbps	58.17	55.65	54.99	94.98	95.14	89.20
MP4 96 kbps	44.08	50.13	49.91	94.51	94.30	81.88
BPF	46.29	49.91	50.18	61.26	64.06	55.49

set to 6 bps in this experiment. We set up the robustness test with two different configurations of adaptive phase modulation schemes to investigate the trade-off between the inaudibility and robustness of the proposed method. With default configuration (Config. 1), parameter θ is adjusted adaptively to the lowest power sub-band. Since the least sensitive portion of audio signal is modified, the sound quality is the best but the robustness may be quite low. We investigated another scheme of adaptation in which parameter θ is adjusted adaptively to the highest power sub-band (Config. 2).

Table 1 shows the average BDR of the watermark detector in two configurations. With Config. 1 (2-ary, 4-ary, and 8-ary), the BDRs are all lower than 90% while the PEAQs are close to imperceptible grade. By contrast, with Config. 2 (2-ary and 4-ary), the BDRs are greater than 94.3% for all the attacks except for bandpass filtering while the PEAQs are less than those with Config. 1. The PEAQ with 4-ary is greater than that with 2-ary because 4-ary uses longer frames and hence reduces artifacts at frame boundaries. On the other hand, the PEAQs and the BDRs of 8-ary seem to be inferior to those of 2-ary and 4-ary. This is because 8-ary using various values of parameter r for phase modulation causes more abrupt change at frame boundaries and incorrect tracking of the parameter r in detection process. Encoding 4-ary is the most suitable for the proposed method to ensure high sound quality and high robustness. These results suggest that the inaudibility could be traded for the robustness by using Config. 2 (4-ary) with the proposed method. Depending on the requirements of applica-

Fig. 14 Comparative evaluation on PEAQ and BDR with respect to embedding bit rate.

tions, a suitable configuration could be chosen for high inaudibility or high robustness.

3.5 Comparative Evaluation

We compared the proposed method with three other conventional methods: the CDb method [15], the DSS method [5], and the echo hiding method (ECHO) [6]. All these methods detect watermarks using a blind detection scheme. In this experiment, the default adaptive scheme (Config. 1) was used and no processing is applied to watermarked signals. Figure 14 plots the average ODG and BDR of the proposed method and the compared methods with respect to the bit rate. For all the conditions of bit rates, ODGs of the proposed method are higher than those of the compared methods. The CD-based method has the best BDR with all the BDRs greater than 99%. The BDRs of the proposed method with 4-ary and 8-ary are greater than 99% for the bit rates less than or equal to 100 bps. The DSS and ECHO methods have the BDRs greater than 90% for the bit rates less than 50 bps and 18 bps, respectively.

The results of PEAQ and BDR suggested that the proposed method can embed watermarks into various kinds of audio signals with good inaudibility and blindly detect embedded watermarks. The proposed adaptive watermarking method provides better sound quality for watermarked signals than the conventional methods.

4. Discussion

Phase modulation itself does not actually degrade sound quality of each frame but introduces discontinuity at frame boundaries in the whole signal. Discontinuity creates the spread of spectrum like pulsed sound's spectrum in watermarked signals, which could be perceived as annoying distortion. The OLA technique is usually employed to eliminate discontinuity in conventional signal processing applications. However, OLA does not work with the proposed

method because it destroys the audio samples around the frame boundaries which are a key factor for blind detection by CZT. For that reason, the proposed method modifies the least significant portion to embed a watermark instead of using the OLA technique for high sound quality.

The proposed method meets a trade-off between inaudibility and robustness. To achieve the best sound quality, the proposed method chooses the least perceptually insensitive portion of audio signals for embedding watermarks. Robustness could be achieved by choosing the resistant portion of audio signals for embedding but at the expense of reducing sound quality. To achieve a reasonable trade-off between robustness and inaudibility, we consider reducing artifacts at frame boundaries by enlarging frames or, in other words, embedding more bits into a frame. However, larger-ary phase modulation produces more abrupt change at frame boundaries and makes parameter r in detection process harder to determine. Therefore, 4-ary scheme is the most suitable for the proposed method.

5. Conclusion

In this paper, we proposed an audio watermarking method with a reasonable trade-off between inaudibility and robustness on the basis of the concept of adaptive phase modulation. The parameter of phase modulator was adaptively adjusted to the center frequency of the lowest or highest power sub-band of an audio frame to achieve high sound quality or high robustness. The experimental results verified that the proposed method could achieve high inaudibility for many kinds of audio signals and precisely detect watermark from watermarked audio signals by a blind scheme. Embedding bit rates of 100 bps with 99% accuracy and 6 bps with 94.3% accuracy were obtained in the case of no processing and the case of attacks, respectively. It was revealed that the proposed method could satisfy three requirements: *inaudibility*, *blindness*, and *robustness*.

Future work will leverage inverse filtering [15] for a reversible watermarking scheme that satisfies inaudibility, blindness, robustness, and high embedding capacity.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (B) (No.23300070) and an A3 foresight program made available by the Japan Society for the Promotion of Science.

References

- [1] N. Cvejic and T. Seppänen, *Digital Audio Watermarking Techniques and Technologies*, IGI Global, 2007.
- [2] C.I. Podilchuk and E.J. Delp, "Digital watermarking: algorithms and applications," *IEEE Signal Process. Mag.*, vol.18, no.4, pp.33–46, 2001.
- [3] B. Chen and G.W. Wornell, "Quantization index modulation: A class of provably good methods for digital watermarking and information embedding," *IEEE Trans. Inf. Theory*, vol.47, no.4,

- pp.1423–1443, 2001.
- [4] P. Bassia, I. Pitas, and N. Nikolaidis, “Robust audio watermarking in the time domain,” *IEEE Trans. Multimedia*, vol.3, no.2, pp.232–241, 2001.
 - [5] I.J. Cox, J. Kilian, F.T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for multimedia,” *IEEE Trans. Image Process.*, vol.6, no.12, pp.1673–1687, 1997.
 - [6] D. Gruhl, A. Lu, and W. Bender, “Echo hiding,” *Proc. Information Hiding 1st Workshop*, pp.295–315, 1996.
 - [7] H.L.F. Helmholtz, *On the Sensations of Tone*, 2nd ed., Dover Publications, 1954.
 - [8] D.A. Nelson and R.C. Bilger, “Pure-tone octave masking in normal-hearing listeners,” *J. Speech and Hearing Research*, vol.17, no.2, p.223, 1974.
 - [9] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for data hiding,” *IBM Syst. J.*, vol.35, no.3-4, pp.313–336, 1996.
 - [10] A. Takahashi, R. Nishimura, and Y. Suzuki, “Multiple watermarks for stereo audio signals using phase-modulation techniques,” *IEEE Trans. Signal Process.*, vol.53, no.2, pp.806–815, 2005.
 - [11] X. Dong, M.F. Bocko, and Z. Ignjatovic, “Data hiding via phase manipulation of audio signals,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol.5, pp.377–380, 2004.
 - [12] M. Arnold, P.G. Baum, and W. Voelbing, “A phase modulation audio watermarking technique,” *Proc. Inf. Hiding Workshop, LNCS 5806*, pp.102–116, 2009.
 - [13] X.-M. Chen, G. Doerr, M. Arnold, and P.G. Baum, “Efficient coherent phase quantization for audio watermarking,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp.1844–1847, 2011.
 - [14] M. Unoki and D. Hamada, “Method of digital-audio watermarking based on cochlear delay characteristics,” *International Journal of Innovative Computing, Information and Control*, vol.6, no.3(B), pp.1325–1346, 2010.
 - [15] M. Unoki and R. Miyauchi, “Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay characteristics,” *IEICE Trans. Inf. & Syst.*, vol.E98-D, no.1, pp.38–48, 2015.
 - [16] N.M. Ngo and M. Unoki, “Watermarking for digital audio based on adaptive phase modulation,” *Proc. International Workshop on Digital-forensics and Watermarking, LNCS 9023*, pp.105–119, 2015.
 - [17] L.R. Rabiner, R.W. Schafer, and C.M. Rader, “The chirp z-transform algorithm,” *IEEE Trans. Audio Electroacoust.*, vol.17, no.2, pp.86–92, 1969.
 - [18] T.T. Wang, “The segmented chirp z-transform and its application in spectrum analysis,” *IEEE Trans. Instrum. Meas.*, vol.39, no.2, pp.318–323, 1990.
 - [19] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “Music genre database and musical instrument sound database,” *Proc. ISMIR 2003*, pp.229–230, 2003.
 - [20] P. Kabal, “An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality,” *Tech. Rep.*, Dept. Elect. Comp. Eng., 2002.
 - [21] M. Unoki and D. Hamada, “Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics,” *International Journal of Innovative Computing, Information and Control*, vol.2, no.1, pp.1–23, 2011.

Nhut Minh Ngo is currently a Ph.D. student in School of Information Science, Japan Advanced Institute of Science and Technology (JAIST). He received the B.S. degree in Information Technology from Ho Chi Minh City University of Science, Vietnam, in 2010 and M.S. degree in Information Science from JAIST, in 2012. His research interests are in audio watermarking and audio signal processing.

Masashi Unoki received his M.S. and Ph.D. (Information Science) from the Japan Advanced Institute of Science and Technology (JAIST) in 1996 and 1999. His main research interests are in auditory motivated signal processing and the modeling of auditory systems. He was a Japan Society for the Promotion of Science (JSPS) research fellow from 1998 to 2001. He was associated with the ATR Human Information Processing Laboratories as a visiting researcher from 1999–2000, and he was a visiting research

associate at the Centre for the Neural Basis of Hearing (CNBH) in the Department of Physiology at the University of Cambridge from 2000 to 2001. He has been on the faculty of the School of Information Science at JAIST since 2001 and is now an associate professor. He is a member of the Research Institute of Signal Processing (RISP), the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Acoustical Society of America (ASA). He is also a member of the Acoustical Society of Japan (ASJ), and the International Speech Communication Association (ISCA). Dr. Unoki received the Sato Prize from the ASJ in 1999, 2010, and 2013 for an Outstanding Paper and the Yamashita Taro “Young Researcher” Prize from the Yamashita Taro Research Foundation in 2005.