Acoust. Sci. & Tech. **31**, 5 (2010)

PAPER

Audio watermarking based on subband amplitude modulation

Akira Nishimura[1,*]

[1]*Department of Media and Cultural Studies, Faculty of Informatics, Tokyo University of Information Sciences, 4–1, Onaridai, Wakaba-ku, Chiba, 265–8501 Japan*

(*Received 13 October 2009, Accepted for publication 2 April 2010*)

Abstract: A watermarking system based on amplitude modulation is proposed. Sinusoidal amplitude modulations at relatively low modulation frequencies applied to neighboring subband signals in opposite phases are used as the carrier of embedded information. The embedded information is encoded in the form of relative phase differences between the amplitude modulations applied to several groups of subband signals. Robustness against perceptual codecs, additive white noise, reverberation, pitch modifications and time scale modifications was verified by computer simulation using 100 pieces in a music genre database. Listening tests to measure the detection thresholds of watermarking and a subjective assessment of the slight reduction of audio quality due to watermarking were conducted using trained listeners. The results of both the listening tests and the computer simulation showed that the quality degradation caused by watermarking with moderate embedding intensity was perceptible but not annoying, while maintaining sufficient detectability of the watermarks.

Keywords: Data hiding, Robustness, Perceptual codecs, Reverberation, ITU-R BS.1116-1

PACS number: 43.60.Ek [doi:10.1250/ast.31.328]

1. INTRODUCTION

Audio watermarking is a method of embedding extra data into a host audio signal. The embedding data is generally utilized for copyright protection by making it difficult to separate the identification code from the audio signal after embedding, which is called a stego signal. The requirements of audio watermarking technology, which is mainly intended to be used for copyright management and copy control of commercial music that is delivered and broadcast widely by various means including the Internet, are summarized as follows. It can hide data at a rate of up to several bits per second [1] in order to embed a copy control code and a licensing code. It is expected to be robust and secure against common signal processing modifications and intentional attacks [2] on the stego signal. In addition, the quality degradation of the stego signal from the host signal should be minimized and be less than the quality degradation induced by a typical perceptual codec of 128 kbps, which is widely used to maintain the commercial value of music.

Recent studies on audio watermarking have resulted in significant progress in inaudibility and reliability. Audio watermarking techniques have achieved robustness against, for example, MPEG compression, additive noise, low-pass filtering, pitch change and time scale modification [3,4]. The inaudibility and reliability of most techniques depend on the acoustic characteristics of the host signal. However, few studies have confirmed the robustness of watermarking techniques using a number of musical samples or actual sample sounds. In other words, the choice of a small number of musical samples could demonstrate the better robustness of a watermarking system.

In terms of the inaudibility of the watermark, a number of studies have shown the watermarked signal to be transparent in sound quality compared with the host signal. However, the ability of listeners to detect sound quality degradation was not verified in most of the previous studies, although the ability depend on listener familiarity with the sound materials, experience and the amount of training on subjective quality evaluation [5].

In previous studies, robustness against emission from a loudspeaker in closed spaces, i.e., reverberation and reflection disturbance, has seldom been considered [6,7]. Robustness against reverberations and reflections is important for audio watermarking techniques used for protecting live performances and new applications of transmitting hidden data by reproduction from loudspeak-

*e-mail: akira@rsch.tuis.ac.jp

ers. Previous audio watermarking methods that adopt relatively short embedding time frames, for example, less than 1 s, such as spread-spectrum [8], echo-hiding [9], time-spread echo [10], patchwork [11], time-frequency analysis [12], and phase-manipulation-based methods [13] cannot easily be used in reverberant spaces, because the delayed signal masks and interferes with the stego signal of the following frame.

In the present paper, a new watermarking system based on subband coding using amplitude modulation is proposed. The system satisfies the conventional requirements of security and robustness. An embedding key required for watermark embedding and decoding ensures the security and concealment of the hidden data. No host signal is required for decoding the hidden data. The system is evaluated in terms of robustness against typical modifications of the stego signal, including perceptual codecs with resampling, additive noise, and pitch and time scale modifications. Robustness testing has been conducted for 100 pieces of music of various genres in order to confirm that the system is applicable to many types of musical signals. In addition, robustness testing against reverberation confirmed that the system can be applied under the aerial watermarking condition. Subjective tests to obtain the detection threshold of the watermarking intensity and to evaluate the subjective degradation of the sound quality induced by watermarking were conducted by trained listeners in order to verify the reliability of the subjective testing.

2. AUDIO WATERMARKING BASED ON AMPLITUDE MODULATION

2.1. Embedding Process

A host signal waveform $H(t)$ in a discrete time series t ($0 \le t < T_p$), whose components are below the upper-limit frequency of embedding, where T_p is the length of one data frame period, is decomposed into n pairs of subband signals $h_{2m-1}(t)$ and $h_{2m}(t)$ using an equal-bandwidth filter bank:

$$H(t) = \sum_{m=1}^{n} (h_{2m-1}(t) + h_{2m}(t)) + H_{\text{high}}(t), \qquad (1)$$

where $H_{\text{high}}(t)$ is an unwatermarked high-pass host signal whose cutoff frequency is hc Hz. A higher hc results in greater robustness of the watermarking system; however, note that the high-frequency region can be eliminated or replicated (for example, using the technique of spectral band replication in MPEG4 AAC) without significant perceptual degradation. Sinusoidal amplitude modulations (SAMs) at a relatively low modulation frequency (f Hz) are applied to the adjacent subband signals $h_{2m-1}(t)$ and $h_{2m}(t)$ in opposite phases. An embedding key produced by a known pseudorandom number generator arbitrarily classifies the n subband pairs into k subband groups. It also defines the random initial phase angles $r(m)$ of the SAMs for each subband pair by generating the seed of a pseudorandom sequence. The outputs of amplitude-modulated subband-pair signals are mixed to obtain $x_m(t)$. The amplitude modulation process is shown in Fig. 1 and Eq. (2). $x_m(t)$ is labeled by superscript i, denoting that it belongs to the ith subband group.

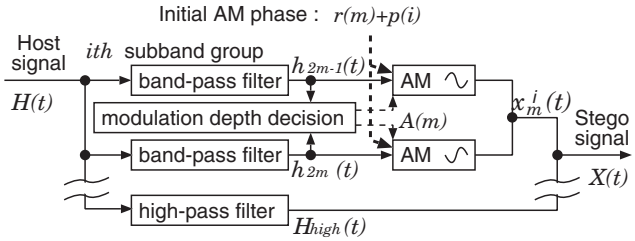

Fig. 1 Block diagram of embedding process. The amplitude modulation process for the $2m-1$ and $2m$ subband pair, which belong to the ith subband group, is represented.

$$x_m^i(t) = h_{2m-1}(t)(1 + A(m)\sin(2\pi ft + r(m) + p(i))) + h_{2m}(t)(1 - A(m)\sin(2\pi ft + r(m) + p(i))) \qquad (2)$$

Where $A(m)$ is the depth of the SAM of the mth subband pair. Embedded information is encoded by phase shift keying, defined as the difference between the phase angle of the SAM of the first subband group and that of the ith subband group, $p(1)$ and $p(i)$ ($i = 2, \ldots, k$). Four-phase shift keying encodes 2-bit information ($D_i = 0, 1, 2, 3$) to phase angles of $p(i)$ separated by $\pi/2$.

$$p(i) = \begin{cases} 0 & i = 1; \\ \dfrac{\pi D_i}{2} & i = 2, \ldots, k. \end{cases} \qquad (3)$$

As a result, $2(k-1)$ bits of information are embedded per data frame period. Multiplex watermarking can be independently applied using different modulation frequencies simultaneously. Finally, a stego signal $X(t)$ is obtained by summing all the amplitude-modulated subband pairs $x_m(t)$.

$$X(t) = \sum_{m=1}^{n} x_m(t) + H_{\text{high}}(t) \qquad (4)$$

In this equation, $x_m(t)$ is not labeled by a superscript since it can belong to any subband group ($i = 1, 2, \ldots, k$). The empirical parameter values for embedding, hc, n, k, f and T_p, which are used in the following sections, are shown in Table 1. Under this condition using these parameters, the filter bank was implemented by 4,096-tap finite impulse response filters using a fast Fourier transform.

The embedding key determines the initial phases of the SAMs applied to all pairs of adjacent subbands and also determines which adjacent subbands belong to the same

Table 1 Embedding conditions and parameters.

Parameters	Values
sampling freq.	44,100 Hz
high-pass cutoff freq. (*hc*)	11,025 Hz
subband pairs (*n*)	128
subband groups (*k*)	5
mod. frequencies (*f*)	2, 3, 5 Hz
frame period (T_p)	5 s
total bit rate of hidden data	4.8 bps

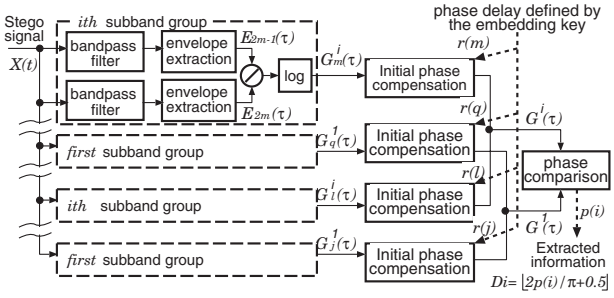

Fig. 2 Block diagram of watermark detection for the *i*th subband group.

subband group. Combined with the small SAM depth applied to each subband, the embedded watermark offers high concealment.

Synchronization of the data frames is achieved by inverting the relative phase of the SAMs between successive data frames for the first ($i = 1$) subband group.

2.2. Extraction Process

Figure 2 shows a block diagram of the detection process for the *i*th subband group. At the beginning of the extraction process, a stego signal $X(t)$ is split into $2n$ subband signals using the equal-bandwidth filter bank used in the embedding process. The filtering process is performed by calculating amplitude spectra from a three-quarters-overlapped running discrete Fourier transform (DFT) of length T_f.

$$\boldsymbol{E}(\tau) = \mathrm{abs}(\mathrm{DFT}(X(t + \tau T_f/4))),\ (0 \le t < T_f) \quad (5)$$

Where abs means absolute value and τ is the index number of the overlapped DFT process for the observation period of the stego signal. $\boldsymbol{E}(\tau)$ is the matrix of the power spectra indexed by the discrete frequency and discrete time τ. As a result, the temporal amplitude envelopes of the mth subband signal $E_m(\tau)(m = 1, 2, \ldots, 2n)$ can be derived from the DFT calculation.

In the practical extraction process applied to the modified stego signal, the amplitude envelope $E_{2m}(\tau)$ belonging to the *i*th subband group is expressed using the amplitude envelope of the $2m$-subband host signal $S_{2m}(\tau)$ and the $2m$-subband noise signal $N_{2m}(\tau)$ induced by the modification of the stego signal,

$$E_{2m}(\tau) = (1 - AM_m(\tau))S_{2m}(\tau) + N_{2m}(\tau), \quad (6)$$

where $AM_m(\tau) = A(m)\sin(2\pi f\tau + p(i) + r(m))$.

The embedded AM waveform is extracted from $G_m(\tau)$, which is the logarithmic ratio of the amplitude envelopes extracted from adjacent subband signals, given by Eq. (7). $G_m(\tau)$ is approximated using up to the second term of the Maclaurin series, that is, $\log(1 + x) \approx x - x^2/2$, where $x = AM_m(\tau) + N_{2m}(\tau)/S_{2m}(\tau)$ and usually $|x| < 1$.

$$\begin{aligned} G_m(\tau) &= \log\frac{E_{2m-1}(\tau)}{E_{2m}(\tau)} \quad (7)\\ &= \log\frac{S_{2m-1}(\tau)}{S_{2m}(\tau)} + \log\frac{1 + AM_m(\tau) + \dfrac{N_{2m-1}(\tau)}{S_{2m-1}(\tau)}}{1 - AM_m(\tau) + \dfrac{N_{2m}(\tau)}{S_{2m}(\tau)}}\\ &\approx \log\frac{S_{2m-1}(\tau)}{S_{2m}(\tau)} + \left(2 - \frac{N_{2m}(\tau)}{S_{2m}(\tau)} - \frac{N_{2m-1}(\tau)}{S_{2m-1}(\tau)}\right)AM_m(\tau)\\ &\quad + \left(\frac{N_{2m-1}(\tau)}{S_{2m-1}(\tau)}\right)^2 - \left(\frac{N_{2m}(\tau)}{S_{2m}(\tau)}\right)^2 \end{aligned}$$

The benefit of the logarithmic ratio calculation is the reduction of the signal-level difference between the adjacent subbands for the subsequent process of synchronous addition of the AM envelope of the same subband group. After temporal compensation of the initial phase difference $r(m)$ for each $G_m(\tau)$, $G^i(\tau)$ is generated by the synchronous addition of $G_m(\tau)$ signals that belong to the *i*th subband group. This emphasizes $AM_m(\tau)$, since the terms including $AM_m(\tau)$ are added in phase and other terms including $S(\tau)$ and $N(\tau)$ are considered to be added with random phases. Even if a noise signal modulated at the embedding AM frequency is added to the stego signal, the compensation of the random initial phase $r(m)$ for all subband pairs gives the additive AM noise a random phase.

The initial phase difference between the first and *i*th subband groups, that is, the embedded information, is obtained by comparing the phase angles of the DFT spectra calculated from $G^1(\tau)$ and $G^i(\tau)$.

2.3. Decision Process of Watermarking Intensity

Figure 3 shows a schematic diagram of the process used to decide the watermarking intensity. The intensity of a watermark, that is, the depth of the SAM $A(m)$ for the mth subband pair, is determined relative to the inherent logarithmic ratio between the subband amplitude envelopes of the host signal in every decision period T_g.

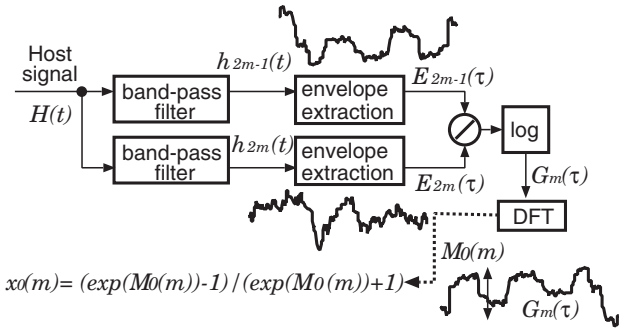

Fig. 3 Schematic diagram of watermarking intensity decision.

First, watermark extraction processing is conducted for the host signal with no watermarking to obtain the amplitude $M_0(m)$ of the inherent logarithmic ratio waveform $G_m(\tau)$ given by Eq. (7). $M_0(m)$ is the amplitude of the spectrum at the modulation frequency f obtained from the DFT spectra of $G_m(\tau)$ for period T_g. The watermarking intensity is determined by regarding the AM depth that realizes $M_0(m)$ as 0 dB.

Let the amplitude modulation depth given to the subband pair be $x_0(m)$ when it realizes a logarithmic modulation amplitude ratio of $M_0(m)$. Let the amplitude of the direct current of the amplitude envelope extracted from the $(2m-1)$th subband be a times larger than that of the $2m$th subband. If the amplitude of the direct current component of the AM observed from both subbands is assumed to be constant during the modulation period, the valley of the AM envelope of the $(2m-1)$th subband is expressed using $x_0(m)$ and Eq. (7) as follows:

$$D - M_0(m) = \log \frac{a(1 - x_0(m))}{1 + x_0(m)}, \tag{8}$$

where $D = \log a$. Also, the peak of the AM envelope is expressed as

$$D + M_0(m) = \log \frac{a(1 + x_0(m))}{1 - x_0(m)}. \tag{9}$$

Then, $x_0(m)$ is expressed in terms of $M_0(m)$ as follows:

$$x_0(m) = (\exp(M_0(m)) - 1)/(\exp(M_0(m)) + 1). \tag{10}$$

The watermarking intensity for the mth subband pair is determined by the AM depth $A(m)$ and is expressed relative to $x_0(m)$ in dB in the form of $20\log_{10}(A(m)/x_0(m))$.

In order to reduce the audibility of AM in the components of relatively high intensity, the AM depth for subband pairs that exhibit a large level difference is controlled. If the level difference ΔL between the subband pair exceeds 20 dB, the AM depth for the subband of the higher level is multiplied by $1 - 0.2\log(10^{\Delta L/20})$ and that of the lower level is multiplied by $1 + 0.2\log(10^{\Delta L/20})$. These multiplications of AM depth maintain the same extracted AM depth in Eq. (7) while suppressing the audibility of watermarking for the higher-level components. The threshold of the level difference of 20 dB was chosen so that the higher-level components sufficiently mask the lower-level components to which a large AM is applied. The threshold of the level difference and the additional AM depth were determined by performing preliminary listening tests on the audibility of watermarking.

Although an auditory model that predicts the audible threshold of AM [14] has been proposed, there is no appropriate model to predict the audibility of AM, when it is applied to the signal of a complex spectrum. Therefore, the appropriate decision period for the depth of AM to be applied to a music signal whose spectra vary continuously is not clear. The length of the decision period T_g was chosen by performing preliminary listening tests on the audibility of watermarking. In the present study, T_g was two modulation periods when the modulation frequency was no more than 3 Hz and four modulation periods when the modulation frequency was above 3 Hz. Also, the maximum modulation depth was limited to 0.316 in order to prevent over modulation.

2.4. Finding the Starting Point of the Data Frame

Before decoding the phase shift keying data, the starting point of the embedded data frame must be detected. A rectangular temporal window of the data frame of length $T = \lfloor 4T_p/T_f \rfloor$ is iteratively applied to the modulation waveform $G^1(\tau)$ extracted from the first subband group. Equation (11) defines a vector $\boldsymbol{R}_u$ with starting point at the window denoted by u. Then, $F(u)$ is derived by subtracting the synchronized sum of the modulation waveforms in the odd-order windows from the synchronized sum of the modulation waveforms in the even-order windows.

$$\boldsymbol{R}_u = \{G^1(u), G^1(u+1), \ldots, G^1(u+T-1)\} \tag{11}$$

$$F(u) = \sum_{v=0} \boldsymbol{R}_{u+2vT} - \sum_{v=0} \boldsymbol{R}_{u+(2v+1)T} \tag{12}$$

The Fourier amplitude of $F(u)$ corresponding to the modulation frequency f, denoted by $\mathrm{AMP}_f(F(u))$, exhibits a maximum when the position of the window overlaps completely with the position of the frame. Consequently, the starting point of data frame y is given by

$$y = \underset{u}{\operatorname{argmax}}\, \mathrm{AMP}_f(F(u)). \tag{13}$$

The present implementation utilizes a stego signal length of $8T$ to calculate $F(u)$.

3. COMPUTER SIMULATION OF WATERMARKING AND EXTRACTION

Evaluation of the present watermarking system began by determining a watermarking intensity that ensures

reliable detection, that is, the embedded watermarks are successfully detected in various types of stego music signals modified by common methods of signal processing, since reliability is the most important feature of audio watermarking used for copyright management. Afterward, the perceptual detection threshold and the degree of quality degradation due to watermarking are precisely measured by trained listeners as reported in the next section.

A computer simulation was conducted to confirm that the watermarking was successful for 100 pieces of music in a music database containing various types of music (RWC-MDB-G-2001) [15]. Watermarks of random bits were embedded in the initial 60 s left-channel signal of each piece of music. Watermarks of 2, 3 and 5 Hz SAMs were simultaneously applied to the host signal in order to obtain a threefold higher capacity of the hidden data. Details of the simulation conditions and the parameters are shown in Table 1. 128 subband pairs are divided into five subband groups based on the remainder after division by 5.

The modifications to the watermarked music were MP3 transcoding, RealAudio8 transcoding, reverberation, additive Gaussian noise, time scale modification and pitch conversion. Reverberation was applied by convolving a random Gaussian sequence with exponential decay. The level of additive Gaussian noise was determined relative to the average overall level of each piece of music. Time stretching was achieved by increasing the modulation frequencies and shortening the embedding period in detection processing. Time shrinking was achieved by performing the opposite operation to time stretching. Pitch conversion was achieved by varying the DFT size from 4,072 to 4,120 in 8 steps during the detection processing, while the exact size of the DFT was 4,096 points.

The bit detection rates, defined as the ratio of the number of correct bits to the total number of embedded bits, are shown in Figs. 4 to 9. Circles connected with broken lines represent median detection rates for 100 tunes at 0 dB watermarking intensity. Squares connected with solid lines represent median detection rates at −5 dB watermarking intensity. Error bars show the maximum and minimum detection rates. Upward and downward triangles show the 10th and 90th percentiles of the detection rates, respectively.

A practical watermarking system should adopt an error correction scheme such as block or convolutional coding of the watermarks in order to reduce detection errors and false positive detection, although this is beyond the scope of this paper. Under the present embedding conditions, for example, BCH codes [16] consisting of 63 bits of code, 36 bits of information and 5 bits of error correction achieve 2.4 bps of embedded information. If the BCH coding is combined with the extended soft decoding scheme [17], it extends the limit of error correction twofold. Therefore, a bit detection rate above 85% is here considered to be successful detection.

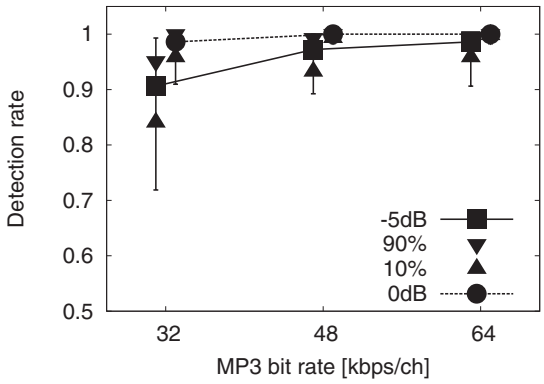

Fig. 4 Median bit detection rates after MP3 transcoding. Error bars denote the minimum and maximum detection rates in 100 music pieces.

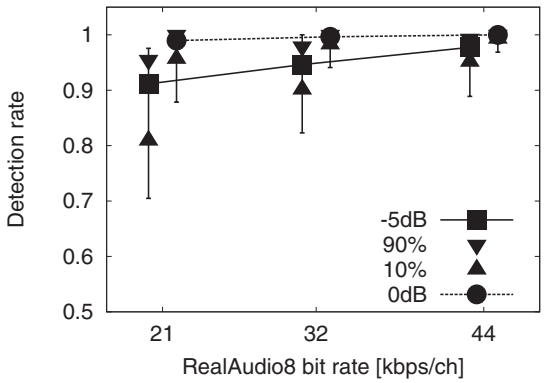

Fig. 5 Detection rate after RealAudio8 transcoding.

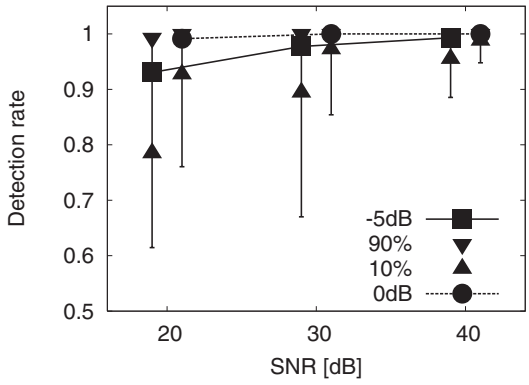

Fig. 6 Detection rate after addition of Gaussian noise.

The results revealed that the watermarking was robust for the perceptual audio codings, such as MP3 and RealAudio. More than half of the pieces of music retained 100% correct watermarks after encoding and decoding at greater than 44 kbps for an embedding intensity of 0 dB. 0 dB watermarking is valid for extreme modifications in which perceptual codings, reverberation and additive Gaussian noise of 30 dB SNR are applied. −5 dB water-

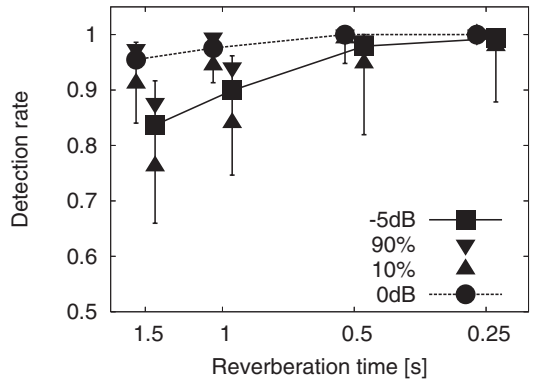

Fig. 7 Detection rate after applying reverberation.

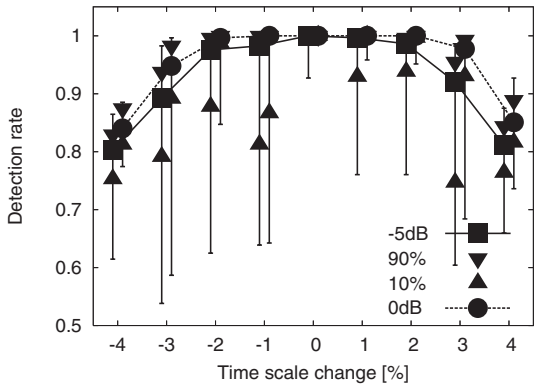

Fig. 8 Detection rate after time stretching.

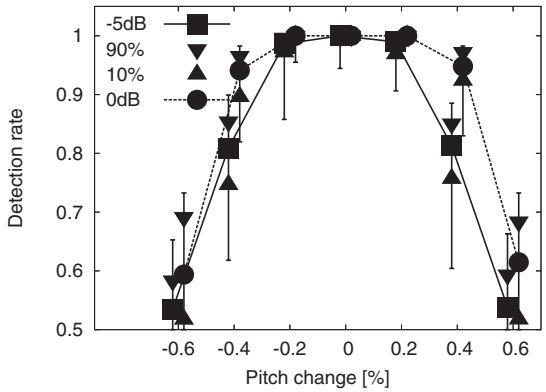

Fig. 9 Detection rate after pitch shift.

marking is robust against moderate levels of the above modifications. On the other hand, the watermarking system is weak with respect to pitch change. However, since the combination of a general digital-to-analog (DA) converter and an analog-to-digital (AD) converter shows a mismatch of the sampling frequency of below 0.1%, the watermarking system is robust against ordinary DA and AD conversion.

The specific feature of the AM-based watermark is robustness against reverberations since it applies relatively slow amplitude modulation in a long embedding frame of several seconds. Generally speaking, most audio watermarking techniques that adopt relatively short embedding time frames, for example, below 1 s, cannot be used in reverberant spaces, since the delayed signal frame masks and interferes with the following frame. Such watermarking techniques include spread-spectrum, echo-hiding and phase-manipulation techniques.

−10 dB watermarking was additionally tested under the same conditions as those described above. In this case, over 85% correct detection was achieved for more than 80% of pieces under the conditions of the smallest degree of each modification, except for RealAudio8, which is so fragile that it is not suitable for the purpose of copyright management.

4. PERCEPTUAL DETECTABILITY

4.1. Detectable Thresholds Using AXB Discrimination Task

A pilot listening test on the audibility of the watermarked music pieces using a double-blind triple-stimulus test with a hidden reference paradigm (see Sect. 4.2) revealed that five naive listeners could hardly discriminate seven music pieces watermarked with an intensity 0 dB (No. 5, 6, 17, 31, 66, 84 and 86 in RWC-MDB-G-2001) from the original pieces. A careful preliminary test conducted by the author to discriminate the 100 musical pieces in the music genre database after 0 dB watermarking from the original pieces revealed that three 5 s musical signals (No. 45, 69 and 87) were relatively discriminable after watermarking compared with other pieces of music. These pieces were used as test signals.

In order to examine the audibility of the stego music signal in the worst case, that is, using the discriminable music pieces and trained listeners, perceptual detection thresholds of the watermarking intensity were obtained by the transformed up-down method with the AXB discrimination task. A 70.7% correct detection threshold was estimated using a two-down one-up adaptive procedure with feedback provided at the end of each three-interval forced-choice trial. The step size of the watermarking intensity was 4 dB at the first three turning points and 2 dB after the third turning points. Each run terminated after 12 turning points, and the estimate for each run was obtained by averaging the levels of the last eight turning points. The equivalent sound pressure levels for the three 5 s stereo music pieces were 78 dB, 76 dB and 71 dB, respectively. All stimuli were presented diotically through headphones (STAX Lambda Nova Classic) in a soundproof room. The embedding conditions and parameters were the same as those for the computer simulation and are shown in Table 1. Four trained listeners, who had experienced the preliminary experiment of the same procedure for more

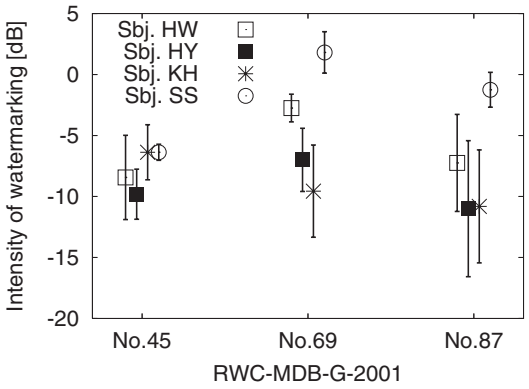

Fig. 10 Detection thresholds of watermarking intensity. Error bars show ±1 standard deviation.

than 4 h before the formal experiment, participated in the experiment. The average detection thresholds obtained from at least four runs are shown in Fig. 10.

The results revealed that the detection threshold of watermarking was approximately −10 dB for the most detectable musical sound for the trained listeners. The following section deals with the amount of perceptual quality degradation induced by suprathreshold watermarking.

4.2. Subjective Assessment of Quality Degradation

The subjective assessment of quality degradation of the audio signal must be carried out carefully in order to avoid artifacts related to the experimental procedure. ITU-R BS.1116-1 [5] recommends numerous factors that should be considered to conduct reliable listening tests on the subjective assessment of small impairments in audio systems. It includes, for example, guidelines on experimental design, the selection and training of listeners, the test method, the program material, reproduction devices, listening conditions and the statistical analysis of the results [18]. Listening tests based on ITU-R BS.1116-1 were conducted since the subjective quality degradation of the musical signals after suprathreshold watermarking was considered to be quite small.

The major differences between BS.1116-1 and the current experimental design were that a small number of listeners (five against the recommended 20) participated owing to the difficulty of obtaining sensitive listeners and that the program materials were selected from RWC-MDB-G-2001 for consistency with the robustness testing.

The test involves a double-blind triple-stimulus procedure with a hidden reference. Sixty second periods of the degraded music and the reference music (hidden reference) were randomly assigned to buttons labeled 'A' and 'B' on a computer screen. A button labeled 'X' was also displayed on the screen and was assigned to the reference. The subject can freely listen to each piece of music by pressing one of the three buttons during a trial. The reproduced sound was not interrupted when a button was pressed until the sound ended because the three sounds were synchronized in the same time scale. The listener was able to listen to each piece repeatedly and was asked to assess the sound quality impairment of 'A' or 'B' compared with 'X' using a five-grade impairment scale with a resolution of one decimal place. The correspondence between the grade and the impairment is as follows: 5.0 is imperceptible, 4.0 is perceptible but not annoying, 3.0 is slightly annoying, 2.0 is annoying and 1.0 is very annoying. All sounds were reproduced via headphones (STAX Lambda Nova Classic) connected to an audio card (M-AUDIO Delta 1010) through an amplifier (DENON AVC-1890).

The degradation of the musical signals was produced under four independent conditions, that is, watermarking with two different intensities and MP3 transcoding with two different bit rates. The MP3 sounds without watermarking were introduced in order to compare the degradation of sound quality due to MP3 transcoding and that due to watermarking. The bit rates of MP3 encoding were 128 kbps and 96 kbps. The embedding intensities of the watermarks were −5 dB and 0 dB for the three pieces of music (No. 45, 69, 87) used in the experiment to obtain the thresholds of watermarking intensity. In addition, the embedding intensities of the watermarks were 0 dB and +10 dB for No. 99 from RWC-MDB-G-2001. The reason why No. 99 was selected is that it had the lowest bit detection rate among the 100 tunes in the previous computer simulation; the bit detection rate of No. 99 at 0 dB embedding corresponded to the detection rate of the other tunes at −5 dB embedding. No. 99 was not used for the experiment described in Sect. 4.1 since the 0-dB-watermarked piece was not easy to discriminate from the original piece in the preliminary test conducted by the author. Five trained listeners, four of which participated in the discrimination test, participated in the experiment.

Trials under 16 experimental conditions, consisting of the combinations of four pieces of music and four types of degradation, were repeated five or six times. The subjective difference grades (SDGs), obtained by subtracting the grade given to the hidden reference from that given to the degraded object, were corrected and analyzed statistically. One-tailed t-tests conducted for each listener revealed that all listeners detected the degraded versions of music with statistically significantly results ($p < 0.06$) for at least 13 experimental conditions. In other words, all listeners had the ability to distinguish sound quality degradation.

The mean SDGs obtained from the four listeners for each degraded condition are shown in Fig. 11. The error bars show ± one standard deviation. A two-way analysis of variance (ANOVA) on the SDGs was performed for three

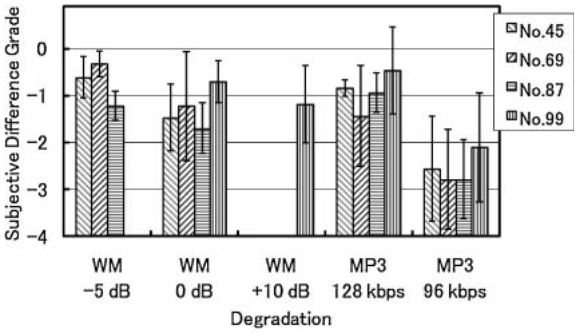

Fig. 11 SDGs obtained by ITU-R BS.1116-1-based procedure. WM denotes degradation due to watermarking. Error bars show ±1 standard deviation.

tunes (No. 45, 69, 87) with the four types of degradation as factors. The effect of degradation was highly significant ($F(3, 48) = 11.28$) ($p < 0.001$). No significant effect was found for the tunes ($F(2, 48) = 0.47$) or the interaction ($F(6, 48) = 0.53$). The Bonferroni test for the post hoc comparison of means revealed that the differences in the SDG between all pairs of degradation were significant. The SDG increased in the order of −5 dB watermarking, 128 kbps MP3, 0 dB watermarking and 96 kbps MP3. Since the three tunes were selected for their high detectability of the degradation in the preliminary tests, no significant difference in SDGs was found among these music signals. However, as can be seen in the SDGs of No. 99, which exhibits smaller degradation than the other tunes in Fig. 11, it is clear that the SDG depends on the type of music signal.

5. DISCUSSION

Since the watermarking method is still relatively weak with respect to malicious attacks such as pitch change and time scale modification, there is room for improvement. Theoretically, a pitch change is more serious than time scale modification, because a pitch change shifts the frequency boundaries between neighboring subbands and causes the negation of amplitude modulation in the subbands. The amount of pitch change can be estimated by the decrease in the intensity of the extracted modulation in the high-frequency region. The estimated pitch change can be compensated by shifting the frequencies of the filter bank. Efficient step-by-step searching by expansion and compression along both the frequency and time axes to extract salient amplitude modulations may improve the detectability of the watermark, although this requires a lot of computing power. Logarithmic frequency spacing of the filter bank, instead of linear frequency spacing, is another way of improving robustness against pitch changes.

Although the robustness against spectral modification, the addition of colored noise and low-pass filtering has not been practically confirmed in the present study, these modifications will not seriously affect the present watermarking method. Spectral modification may change the relative levels between neighboring subbands; however, such a change only affects the direct current component in the extracted SAM waveform after the calculation of the logarithmic ratio between the amplitude envelopes of neighboring subbands. Since embedded amplitude modulations are applied to a wide frequency range of the host signal, watermarks survive after low-pass filtering or the addition of colored noise.

In unofficial listening tests on several pieces of watermarked music, the perceptual quality degradation of musical signals that contain sounds of several musical instruments at the same time was difficult to detect. Perceptual quality degradation was not detectable even when the intensity of watermarking was greater than 0 dB. The present scheme for deciding the modulation depth is not based on the perception of the human auditory system. A conventional scheme for predicting the masked threshold for each frequency region, such as MPEG psychoacoustic modeling, is reasonably effective for deciding an inaudible watermarking intensity. A more effective method of deciding the perceptually optimized modulation depth is to utilize modulation detection interference (MDI) [19] of the human auditory system. MDI is a phenomenon in which the ability to detect amplitude modulation of one tonal carrier is disrupted when a tone with a very different carrier frequency is similarly amplitude modulated and added to the probe tone. Therefore, MDI makes the perceptual detection of embedded amplitude modulation difficult, and the wide variation in the detection thresholds observed among the various types of music may have been partly due to MDI. Considering the results of the perceptual detectability test, detailed modeling of the MDI in the human auditory system [14] will enable more information to be embedded without the perceptual deterioration of the sound quality.

A method for objectively measuring the perceived audio quality (ITU-R BS.1387), referred to as PEAQ, uses a number of psychoacoustical measures that are combined to provide a measure of the difference in quality between two instances of a signal (a reference signal and a test signal). The PEAQ measurement outputs an objective difference grade (ODG), which corresponds to the SDG in ITU-R BS.1116-1 [20]. The average ODG for the 100 pieces of music in Sect. 3 was −0.34 for −5 dB watermarking and −0.71 for 0 dB watermarking. These values were smaller than those corresponding to 'perceptible but not annoying' and the average ODG of −0.89 obtained by MP3 128 kbps transcoding. However, since the PEAQ was designed to evaluate perceptual audio codecs, it is not clear whether or not the PEAQ is effective for evaluating AM-based watermarking. An objective evaluation of the sound

quality degradation induced by the watermarking will be a future research topic [18].

6. SUMMARY

A new watermarking system based on amplitude modulation was proposed. The system is robust against perceptual audio coding, reverberation and additive Gaussian noise. Detecting the existence of the watermark is difficult without knowing the key used in the encoding process. No host signal is required for the decoding process. The watermarking intensity is adaptively determined relative to the inherent amount of AM in the host signal. The transformed up-down method was used to determine perceptual thresholds of the watermarking intensity for several perceptible musical sounds. The thresholds were approximately -10 dB to -5 dB for trained listeners. A subjective evaluation test of audio quality degradation, in accordance with ITU-R BS.1116-1, was conducted for several musical pieces with a perceptually detectable watermark or that had been degraded by MP3 transcoding. The results revealed that the sound quality of -5-dB-watermarked music was comparable to or better than that of 128 kbps MP3-encoded music. The SDGs of 0-dB-watermarked music indicate an audio quality corresponding to MP3 encoding of between 96 kbps and 128 kbps.

ACKNOWLEDGMENT

The present study was supported in part by a Grant-in-Aid for Scientific Research (C) No. 20560365.

REFERENCES

[1] 4C Entity, "4C 12 BIT watermark specification," http://www.4centity.com/docs/4cspec.pdf (1999).

[2] N. Cvejic and T. Seppänen, "Introduction to digital audio watermarking," in *Digital Audio Watermarking Techniques and Technologies, Applications and Benchmarks*, N. Cvejic and T. Seppänen, Eds. (Information Science Reference, New York, 2008), pp. 1–10.

[3] R. Tachibana, S. Shimizu, S. Kobayashi and T. Nakamura, "An audio watermarking method robust against time- and frequency-fluctuation," in *Proc. Security and Watermarking of Multimedia Contents III, SPIE*, Vol. 4314, pp. 104–115 (2001).

[4] W. Li and X. Xue, "Audio watermarking based on music content analysis: Robust against time scale modification," in *Digital Watermarking: Second Int. Workshop, IWDW 2003, LNCS 2939*, T. Kalker, I. J. Cox and Y. M. Ro, Eds. (Springer-Verlag, Berlin, 2004), pp. 289–300.

[5] ITU-R Recommendation, "Method for the subjective assessment of small impairments in audio systems including multichannel sound systems," **BS.1116-1** (1997).

[6] H. Matsuoka, Y. Nakashima and T. Yoshimura, "Acoustic communication with OFDM signal embedded in audio," in *29th AES Int. Conf.: Audio for Mobile and Handheld Devices* (2006).

[7] T. Modegi, "Audio watermark embedding technique applying auditory stream segregation: G-encoder mark, able to be extracted by mobile phone," in *Proc. 3rd Int. Conf. on Mobile Computing and Ubiquitous Networking, ICMU 2008*, pp. 33–40 (2008).

[8] H. S. Malvar and D. A. Florencio, "Improved spread spectrum: a new modulation technique for robust watermarking," *IEEE Trans. Signal Process.*, **51**, 898–905 (2003).

[9] D. Gruhl, A. Lu and W. Bender, "Echo hiding," in *Proc. 1st Int. Workshop on Information Hiding, LNCS 1174*, pp. 295–315 (1996).

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

[11] R. Tachibana, "Improving audio watermarking robustness using stretched patterns against geometric distortion," in *Advances in Multimedia Information Processing, PCM 2002,* (*LNCS*) *2532*, (Springer, Hsinchu, 2002), pp. 647–654.

[12] S. Krishnan, B. Ghoraani and S. Erkucuk, "Time-frequency analysis of digital audio watermarking," in *Digital Audio Watermarking Techniques and Technologies, Applications and Benchmarks*, N. Cvejic and T. Seppänen, Eds. (Information Science Reference, New York, 2008), pp. 187–204.

[13] A. Takahashi, R. Nishimura and Y. Suzuki, "Multiple watermarks for stereo audio signals using phase-modulation techniques," *IEEE Trans. Signal Process.*, **53**, 806–815 (2005).

[14] T. Dau, B. Kollmeier and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. II. spectral and temporal integration," *J. Acoust. Soc. Am.*, **102**, 2906–2919 (1997).

[15] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. 4th Int. Conf. on Music Information Retrieval* (*ISMIR 2003*), pp. 229–230 (2003).

[16] T. Yamada, *Essentials of Error-Control Coding Techniques* (Academic Press, San Diego, 1990), pp. 46–49.

[17] Y. Fujii, I. Echizen, T. Yamada, S. Tezuka and H. Yoshiura, "An improvement of error correction coding for digital watermarking," *Trans. Inf. Process. Soc. Jpn.*, **45**, 1980–1997 (2004).

[18] M. Arnold, P. Baum and W. Voeßing, "Subjective and objective quality evaluation of watermarked audio," in *Digital Audio Watermarking Techniques and Technologies, Applications and Benchmarks*, N. Cvejic and T. Seppänen, Eds. (Information Science Reference, New York, 2008), pp. 260–277.

[19] W. A. Yost and S. Sheft, "Modulation detection interference: Across-frequency processing and auditory grouping," *Hear. Res.*, **79**, 48–58 (1994).

[20] P. Kabal, "An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality," *TSP Lab. Tech. Rep., Dep. Electr. Comput. Eng., McGill Univ.*, pp. 1–89 (2002).