Audio Watermarking for Live Performance

Ryuki Tachibana

Tokyo Research Laboratory, IBM Japan

ABSTRACT

Audio watermarking has been used mainly for digitally stored content. Using real-time watermark embedding, its coverage can be extended to live broadcasts and live performances. In general, a conventional embedding algorithm receives a host signal (HS) and outputs the summation of the HS and a watermark signal (WS). However, when applied to real-time embedding, there are two problems: (1) delay of the HS, and (2) possible interruption of the broadcast. To solve these problems, we propose a watermark generation algorithm that outputs only a WS, and a system composition method in which a mixer outside the computer mixes the WS generated by the algorithm and the HS. In addition, we propose a new composition method "sonic watermarking". In this composition method, the sound of the HS and the sound of the WS are played separately by two speakers, and the sounds are mixed in the air. Using this composition method, it would be possible to generate a watermarking sound in a concerto hall so that the watermark could be detected from content recorded by audience members who have recording devices at their seats. We report on the results of experiments and discuss the merits and flaws of various real-time watermarking composition methods.

Keywords: Audio watermarking, real-time embedding, sonic watermarking

1. INTRODUCTION

A digital audio watermark has been proposed as a means to identify the owner or distributor of digital audio data.[1–4] Proposed applications of audio watermark are copyright management, annotation, authentication, broadcast monitoring, and tamper-proofing.[2] For these purposes, transparency, data payload, reliability, and robustness of audio watermarking technologies have been improved by a number of researchers. Recently, several audio watermarking techniques that work by modifying magnitudes in the frequency domain were proposed to achieve robustness against distortions such as time-scale modification and pitch shifting.[5–7]

Of the various applications, the primary driving forces of audio watermarking research have been copy control of digital music and searching for illegally copied music.[8, 9] In these usages, it is natural to consider that an original music sample, which is the target of watermark embedding, exists as a file stored digitally on a computer. For creating a watermarked DVD-Audio disc, when most of the creation process including recording, mixing, mastering, etc. is finished, copying the finished song to a computer and embedding an audio watermark in the music files on the computer is possible. In the scenario of SDMI, an audio watermark may be embedded in a music file prepared for Internet distribution.

Though it is true that these purposes have greatly encouraged audio watermarking research so far, audio watermarking is not a technology useful only for digitally stored music. Of course, music is performed, created, stored, and listened to in many different ways, and it is much more common that music is not stored as a digital file on a computer. When applying audio watermarking technology in various musical environments, *real-time watermark embedding* is a preferred approach. For example, in music mastering studios, by embedding a watermark in real time in a sound being played, the watermarked sound could be instantly checked without saving it as a file on a computer. For broadcasting of a watermarked sound, it would become possible to embed a watermark in real time and instantly broadcast the sound of a live performance being performed in a studio, or the voice of a newscaster or a talk show.

Further author information: E-mail: ryuki@jp.ibm.com, Address: 1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken, Japan 242-8502

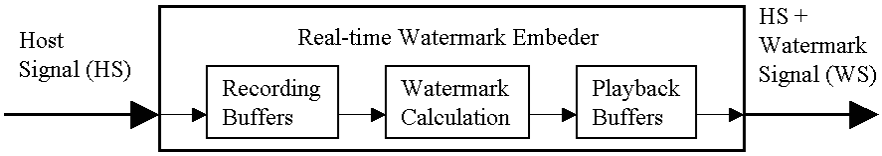

Figure 1. [C1] Naive Composition Method: A mixture of a host signal (HS) and a watermark signal (WS) is calculated for a short HS stored in a recording buffer, and then is played back from a playback buffer.

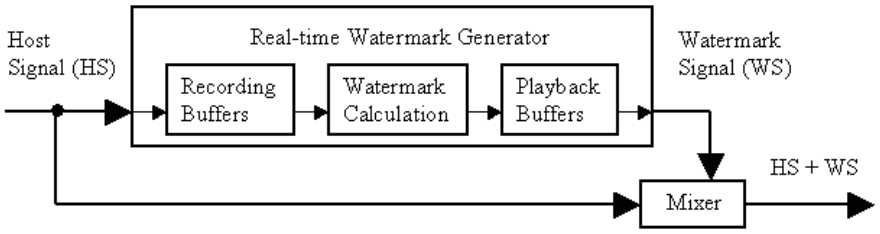

Figure 2. [C2] Analogue Watermarking: The computer outputs only the WS. Mixing is performed at a trusted conventional analogue mixer outside of the computer. Control of the volume of the WS is much easier.

It is not difficult to perform watermark embedding faster than playback using current audio watermarking technologies and computers. However, there are other problems besides the known problems for file embedding. In this paper, we classify several composition methods for real-time watermark embedding, and point out problems that occur with the composition methods, propose a real-time embedding method, and report the results of some experiments.

2. COMPOSITION METHODS OF REAL-TIME WATERMARK EMBEDDING

In this section, we describe several system compositions for real-time embedding, and point out their merits and flaws.

2.1. [C1] Naive Composition

Most watermarking algorithms are designed to be given a host signal (HS) and to produce a signal that is mixture of the HS and a watermark signal (WS). The most naive composition method by simply applying an algorithm of this kind to real-time embedding is illustrated in Figure 1. In this composition method, a HS with a certain length is stored in a recording buffer. The stored HS is given to the watermark embedding process where a mixture of the HS and a WS is calculated. The mixture is then stored in a playback buffer and played at the proper time. Any audio watermarking algorithm can be used for real-time embedding with this composition method as long as it runs faster than the playback.

However, there are two drawbacks with this composition method. The first drawback is the inevitable delay of the HS (and the WS). The recording buffer and the playback buffer must have some length so that the mixed signal can be played steadily. If the buffers are longer, stability of playback increases at the expense of the delay of the HS. The second drawback is risk of interrupting of the playback. When a problem blocks the real-time watermark embedder, playback of not only the WS but the HS stops. This would make a big problem for a live broadcast of a watermarked signal.

2.2. [C2] Analogue Watermarking

To solve the problem of C1, another composition method illustrated in Figure 2 mixes the HS and the WS outside the computer using a trusted conventional analogue mixer after the WS is converted to an analogue signal. The computer, a real-time watermark generator, should be designed to output only the watermark

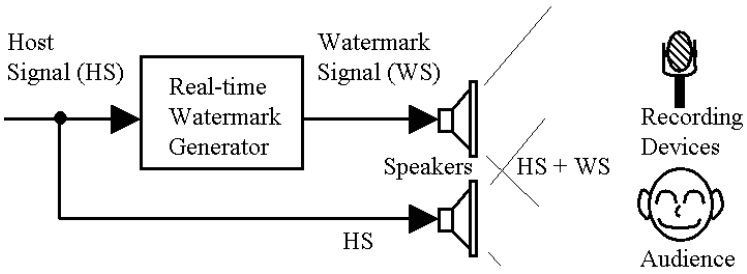

Figure 3. [C3] Sonic Watermarking: The WS sound and the HS sound enter the air from separate speakers. Both sounds mix in the air and form a watermarked sound so that the embedded message could be detected from a recorded sound.

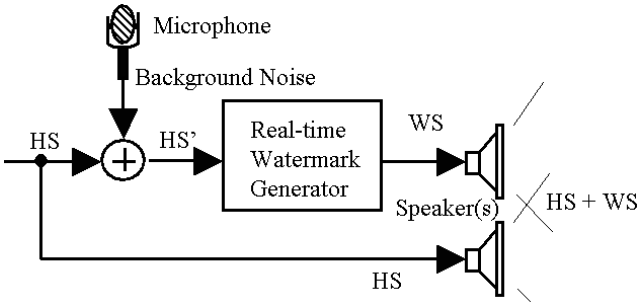

Figure 4. [C4a] Noise Recording: The background noise of the environment is recorded using a microphone and mixed with the host signal so that the frequency masking effect of the noise can be also taken into account.

signal. The reason that the computer needs to be fed the HS is that it utilizes the HS for calculating the frequency masking effect[10] of the HS. This composition does not introduce additional delay for the HS. If a problem effects the computer, while generation of the WS would stop, the broadcast of the HS is still secure. In addition, it is very easy for a sound engineer to control the volume of a WS using a familiar mixer. This would make it easy for sound engineers to use a watermark embedder.

That the WS is delayed relative to the HS because of the buffers and the watermark calculation can be a problem of this composition method.

2.3. [C3] Sonic Watermarking

Because both the HS and the WS are sounds, it is possible to mix them in the air (Figure 3). In this composition method, a real-time watermark generator the same as the one used in C2 outputs the WS, and the WS enters the air from a speaker. The WS sound and the HS sound mix in the air and make a watermarked sound.

This composition method will be useful for live concerts in which electronic instruments are used. In concerts of popular music, because the sounds of the instruments and the vocals are electronically captured and amplified, it would be easy to snatch the sound signal and feed it to a watermark generator just before amplification for the speakers. Using watermark embedding by this composition method, it would be possible to find *bootleg recordings* on the Internet, that is, illegal music files that have been recorded in concert halls by malicious audience members using recording devices.

This composition method also has the same potential problem of the delay of the WS relative to the HS. Another problem of this composition method is that background noises made by sources besides the musical instruments become disturbing factors for watermark detection. Such sounds include voices and clapping by audience members, reverberations of the hall, and rustling noises made by hands touching the recording device.

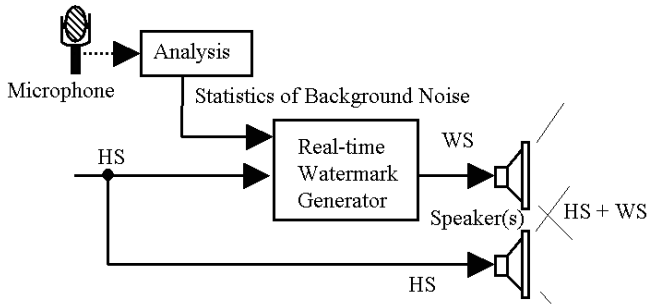

Figure 5. [C4b] Noise Prediction: If the background noise is known to be stationary, its frequency masking effect can be predicted without recording during the performance.

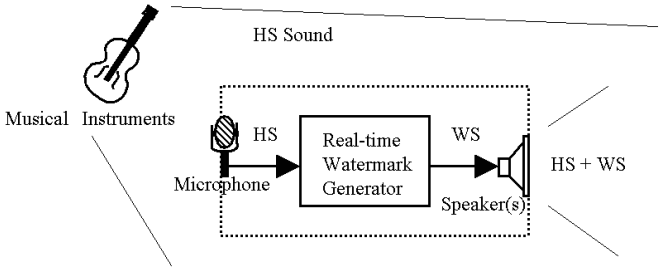

Figure 6. [C4c] Host Signal Recording: When the musical instruments do not use electronic amplification, their sound together with the background noises should be recorded by a microphone for watermark calculation.

2.4. [C4] Sonic Watermarking with Sampling

A method to make a watermark detectable even when there is a loud background noise is to also make use of the frequency masking effect of the noise. Figure 4 illustrates a composition method, C4a, where background noises are recorded using a microphone, added to the host signal played by musical instruments, and finally input to a real-time watermark generator. When there is a loud background noise, the WS also increases, and hence detection of the embedded message would be easier. However, there may be several problems in the recording of background noises. It is impossible to record noises that are made near scattered recording devices. Recoding only the background noise while excluding the sound of the musical instruments is also difficult.

If the background noise of the environment is known to be stationary, recording of the noise during the performance is not necessary. The watermark generator can predict the frequency masking effect of the noise as long as its power spectrums are analyzed prior to the performance (Figure 5). Because this composition method, C4b, does not follow the dynamic changes of the noise, the successful rate of detection of embedded message will be lower than for C4a. For both C4a and C4b, the actual masking effects of each audience member depend on their locations. Arrangements using multiple speakers will be better. When using multiple speakers, it is also necessary to consider their mutual interference.

A composition method, C4c, illustrated in Figure 6 is the only composition method that can be applied to musical instruments that do not use electronic amplification. The only difference in C4a is that sounds of the instruments together with the background noises are also recorded by a microphone for the watermark calculation. For mixing in the air, we believe that it does not matter whether the HS sound is electronically amplified or acoustic. This composition method can dynamically respond to changing background noises.

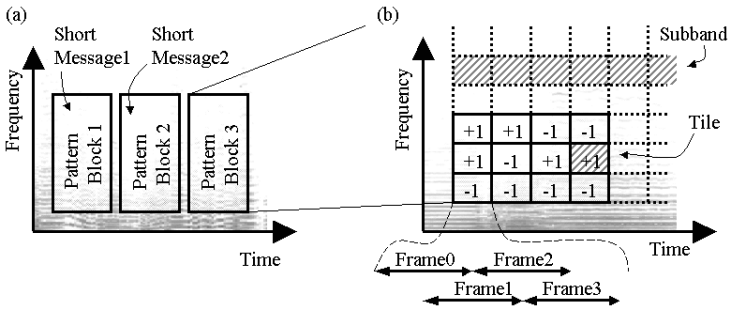

Figure 7. A pattern block consists of tiles, which are segmented areas in the time-frequency plane of the content. The embedding algorithm modifies magnitudes in the tiles according to pseudo-random numbers assigned to the tiles. The numbers in the figure are examples of the pseudo-random values.

3. PREVIOUS METHODS

In this paper, we examine the applicability of the audio watermarking method proposed in[6, 11] to real-time embedding. This is because the method has the following advantages, and it would be useful to use the same method for watermarking both of files and live performances. The advantages are:

- Even if the discrete Fourier transform (DFT) frames used for detection are displaced from the original DFT frames used for embedding, the detected watermark strength remains high enough, therefore sample-by-sample exhaustive search for the original embedding DFT frames is unnecessary.
- By using a two dimensional pseudo random array and frequency subbands containing multiple frequency bins, the detection to some extent is robust against time- and frequency-fluctuation even without a search.
- It is possible to make detection more robust against excessive distortion by using multiple stretched patterns.[12] Since most of the detection process for the multiple patterns is shared, the additional computational cost is limited.

The method can be summarized as follows. The method embeds a multiple-bit message in the content by dividing it to short messages and embedding each of them in *a pattern block*. The pattern block is defined as a two-dimensional segmented area in the time-frequency plane of the content (Figure 7), which is constructed from the sequence of power spectrums calculated using short-term DFTs. A pattern block is further divided into *tiles*. We call the tiles in row a *subband*. A tile consists of four consecutive overlapping DFT frames. A pseudo-random number is selected corresponding to each tile. We denote the value of the pseudo-random number assigned to the tile at the b-th subband in the t-th frame by $\omega_{t,b}$, which is +1 or -1. We also denote the value of the bit assigned to the tile by $B_{t,b}$, which is 1 or 0. The values of the pseudo-random numbers and the tile assignments of the bits are determined by a symmetric key which is shared by the embedder and the detector.

3.1. Embedding

The embedding algorithm calculates the complex spectrum, $c_{t,f}$, of the f-th frequency bin in the t-th frame of a pattern block of the content by using the DFT analysis of a frame of the content. We denote the magnitude and the phase of the bin by $a_{t,f}$ and $\theta_{t,f}$, respectively. Then the algorithm calculates the inaudible level of the magnitude modification by using a psychoacoustic model based on the complex spectrum. We indicate this amount of the f-th frequency of the t-th frame in a pattern block by $p_{t,f}$. We use this amount for the magnitude in the f-th frequency bin of the watermark signal (WS).

A sign, $s_{t,b}$, which determines whether to increase or decrease the magnitudes of the host signal (HS) in a tile is calculated from the pseudo-random value, $\omega_{t,b}$, the bit value, $B_{t,b}$, and the location, t, of the frame in the block. If the frame is in the first two frames of a row of tiles, that is, if the reminder of dividing t by 4 is less than 2, $s_{t,b} = \omega_{t,b}(2B_{t,b} - 1)$. Otherwise $s_{t,b} = -\omega_{t,b}(2B_{t,b} - 1)$. This is because, by embedding opposite signs in the first and last two frames of a tile and by detecting the watermark using the difference of the magnitudes, cancellation of the host signal can make the detection robust. In the tiles where the calculated sign, $s_{t,b}$, is positive, the phase of the HS, $\theta_{t,f}$, is used for the phase, $\phi_{t,f}$, in the f-th frequency bin of the WS, while we assume the f-th frequency is in the b-th subband. In the tiles with a negative sign, the opposite phase $-\theta_{t,f}$ is used for $\phi_{t,f}$.

So far, the magnitude $p_{t,f}$ and the phase $\phi_{t,f}$ of the WS have been calculated. The WS is converted to the time domain using inverse DFTs, and added to the HS. Consequently, this procedure increases the magnitudes of the HS by $p_{t,f}$ in the tiles with a positive sign, and decreases the magnitudes in the tiles with a negative sign, while preserving the phases of the HS.

3.2. Detection

The detection algorithm calculates the magnitudes for all tiles of the content and correlates them with the pseudo-random array.

The magnitude $a_{t,f}$ of the f-th frequency in the t-th frame of a pattern block of the content is calculated by the DFT analysis of a frame of the content. A frame overlaps the adjacent frames by a half window. The magnitudes are then normalized by the average of the magnitudes in the frame. We denote a normalized magnitude by $\widehat{a}_{t,f}$. The difference between the logarithmic magnitudes of a frame and the next non-overlapping frame is taken as $P_{t,f} = \log \widehat{a}_{t,f} - \log \widehat{a}_{t,f+2}$. The magnitude $Q_{t,b}$ of a tile located at the b-th subband of the t-th frame in the block is calculated by averaging the $P_{t,f}$s in the tile. The detected watermark strength for the j-th bit in the tile is calculated as the cross-correlation of the pseudo-random numbers and the normalized magnitudes of the tiles by

$$X = \frac{\sum_{assigned\ (t,b)} \omega_{t,b}(Q_{t,b} - \overline{Q})}{\sqrt{\sum_{assigned\ (t,b)} \left\{\omega_{t,b}(Q_{t,b} - \overline{Q})\right\}^2}}\,, \tag{1}$$

where $\overline{Q}$ is the average of $Q_{t,b}$, and the summations are calculated for the tiles assigned for the bit. Similarly, the synchronization strength is calculated for the synchronization signal. Watermark strength for a bit is calculated after a synchronization to search for the first frame of a block is done. The synchronization process consists of a global synchronization and a local adjustment. In the global synchronization, it is assumed that several consecutive blocks having synchronization positions are separated by the same number of frames. In the local adjustment, the frame with a local maximum synchronization strength is chosen from a few neighboring frames.

4. PROBLEMS

As long as the composition method C1 is chosen for real-time embedding, there is no problem in applying the previous method. In that case, nothing is different from handling files stored on a computer. However, applying to compositions methods C2 to C4 results in an inevitable delay of the watermark signal (WS) relative to the host signal (HS) which causes a serious problem.

The delay is caused by the recording buffers, playback buffers, and WS calculation (Figure 8). Though the length of the playback buffers and the recording buffers can be reduced using ASIO* software and hardware, it is impossible to reduce them to zero. A total of 128 samples for each buffer is required for stable real-time

*ASIO is the Steinberg Audio Stream Input/Output architecture for low latency high performance audio handling.

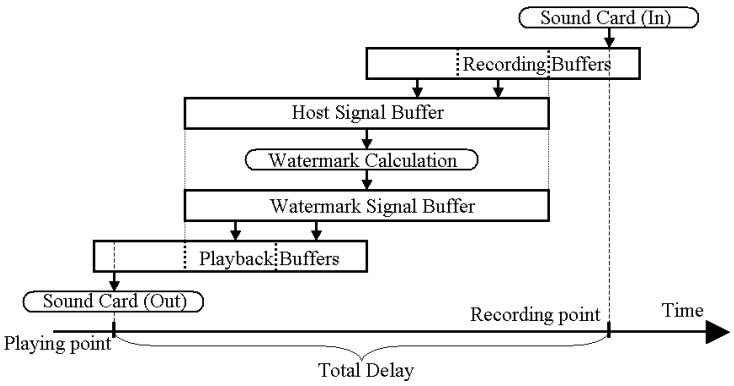

Figure 8. A watermark signal (WS) is delayed relative to a host signal (HS) because of recording buffers, watermark calculation, and playback buffers.

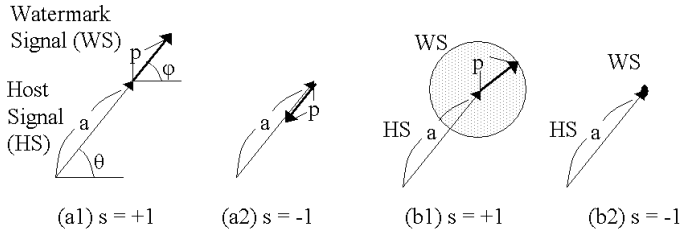

Figure 9. The host signal and the watermark signal for the conventional method (a) and for the proposed method (b)

embedding for our experiment[†]. The WS calculation causes two kinds of delay. The first is that it is necessary to store one DFT frame of the HS to calculate its power spectrums. The second is the elapsed time for the WS calculation. The watermark calculation takes approximately 10% of the playback time in our experiment. For example, if the length of a DFT frame is 512 samples, the elapsed time for the WS calculation corresponds approximately to playback time for 51.2 samples. Hence, using 512-sample DFT frames, the total delay is $128 + 128 + 512 + 51.2 = 819.2$ samples, which is about 18.5 ms for 44.1 kHz sampling.

This delay makes perfect synchronization of the WS and the HS impossible. The embedding algorithm of the previous method embeds the watermark by modifying amplitudes in the frequency domain while preserving the phases (Figure 9a). When the watermark signal is delayed by at least 18.5 ms, since the phase of the HS drastically changes during the delay, the phases of the HS and the WS become almost independent. Accordingly, decreasing the amplitudes in the tiles where negative pseudo-random numbers are assigned becomes impossible.

Figure 10(a) is the result of an experiment where we calculated a WS for a HS using the previous method, delayed the WS by N_d samples, overlapped it with the HS, and observed the watermark strengths detected from the mixture. The horizontal axis of the figure is the given delay, N_d. The vertical axis is the average of detected watermark strengths. The distribution of the strengths for unwatermarked content can be regarded as a normal distribution whose variance is 1.14. The reason the variance for unwatermarked content is not unity is that, since the detected strengths for bits are also used for the synchronization, the positions with large strengths are selected in the searching process. It can be seen that as small a delay as eight samples (approximately 0.02 ms) made detection impossible. Consequently, the previous algorithm cannot be used for compositions methods C2 to C4.

[†]We used a Mobile Pentium III 1.13 GHz notebook computer with a docking station equipped with a LynxOne sound card by Lynx Studio Technology.

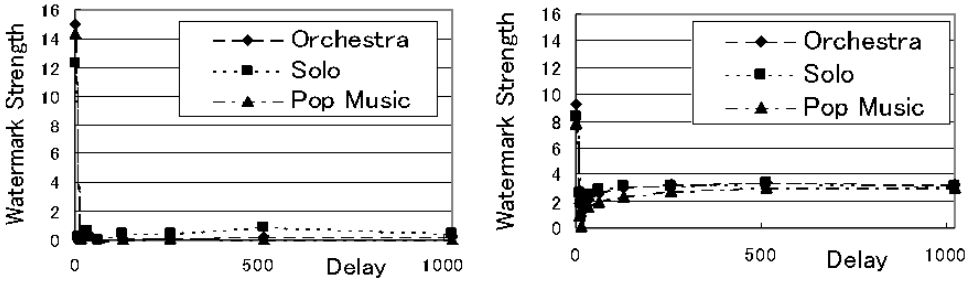

Figure 10. Average of detected watermark strengths from mixtures of a watermark signal (WS) and a host signal (HS) for the previous method (a) and the new method (b). The WS is delayed by N_d samples relative to the HS. A delay as small as eight samples ruins detection.

5. NEW METHOD

We altered the embedding algorithm of the previous method as described below so that modification of amplitudes is possible under the situation where the watermark signal is delayed relative to the host signal.

The only change is to make the values of the magnitudes $p_{t,f}$ of the WS zero for the tiles with a negative sign (Figure 9b2). This is because it is impossible to decrease the magnitudes of the HS. As for the tiles with a positive sign, the magnitudes and the phases of the WS are given as in the previous method. However, because of the delay, to give the WS the same phases as the HS has almost the same effect as giving the WS a random phase (Figure 9b1).

This change makes the power distribution of the content non-uniform, and hence makes detection possible. However, because the efficiency of amplitude modification is much worse than in the previous algorithm, a decrease of the detected watermark strength is inevitable. It is necessary to use a stronger watermark signal than the previous method uses.

Note that the alteration is only for embedding, and that the same previous detection algorithm can detect the watermark from the content whether the previous algorithm or the altered algorithm is used for watermarking.

6. EXPERIMENT

In this section, we report results of experiments applying the new method to the real-time embedding composition methods C2 to C4.

6.1. Implementation

We implemented the method in a software system that can embed and detect 64-bit messages in 30-second pieces of music. The message is encoded in 448 bits by adding 8 Cyclic Redundancy Check (CRC) parity bits, using Turbo Coding, and repeating it twice. Each pattern block has 3 bits and a synchronization signal embedded, and the block has 24 columns and 8 rows of tiles. Each of the 24 frequency subbands is given an equal bandwidth of 6 frequency bins. The frequency of the highest bin used is 12.7 kHz. While 48 tiles out of the 192 tiles are dedicated for the local adjustment of the pattern block synchronization, the tiles assigned for bits are also used for the global synchronization. For the global synchronization, It is assumed that 16 consecutive blocks have consistent synchronization positions. The length of a DFT frame is 512 samples to shorten the delay. The false alarm error ratio is theoretically controlled to be under 10^{-5} by the threshold of the square means of the detected bit strengths. Another threshold on the estimated watermark SNR is set to keep the code word error ratio under 10^{-5}. The reason to use double thresholds is described in Ref..[13] Nine pieces of monaural music from pop music, orchestral music, and solo instruments were used for the analysis. All of the signals were sampled at a frequency of 44.1 kHz and with a bit resolution of 16 bits, and each piece is 100 seconds long.

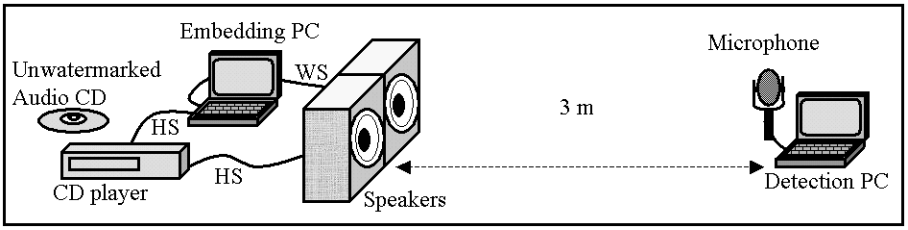

Figure 11. An experimental system for testing C3 and C4b was set up in a soundproof room where a 30 dB(A) background noise was heared when nothing was played by the speakers.

6.2. Psychoacoustic Model

The ISO-MPEG 1 Audio Psychoacoustic Model 2 for Layer III[14] is used as the basis of the psychoacoustic calculation for the experiments with some alterations.

- An absolute threshold was not used for these experiments. We believe this is not suitable for practical watermarking because it originally depends on the listening volume and is too small in the frequencies used for watermarking.
- A local minimum of masking values within each frequency subband was used for all frequency bins in the subband. Not only excessively changing the WS magnitudes does not contribute to the watermark strength but it causes lower acoustic quality by increasing the WS without effect.
- A 512-sample frame, 256-sample IBLEN, and a sine window were used for the DFT for the psychoacoustic analysis to reduce the computational cost.

It is known that postmasking stays as high as simultaneous masking for a 5 ms delay and that it starts an almost exponential decay with a time constant of 10 ms after the first 5 ms.[10] Hence, a shorter DFT frame is expected to result in better acoustic quality because of the shorter delay. However, the poor frequency resolution caused by a too short DFT frame reduces the detected watermark strength. This is the reason a 512-sample DFT frame was selected for the implementation.

Based on the psychoacoustic model, the root mean square power of the WS is 20.4 dB lower than that of the HS on average. Though we have not conducted systematic testing of the acoustic quality of the method, we noticed that the watermarked songs tend to sound a little hoarse. We think that one reason for the hoarseness is the fast switching of the WS in short DFT frames.

6.3. Experiment of Analog Watermarking (C2)

Figure 10(b) is the result of the same experiment as in Figure 10(a) using the new method. It can be seen that the detected watermark strength of this method with a delay of the WS is greater than in the previous method.

A test of analogue watermarking was conducted. The WS was mixed with the HS with a delay of around 830 samples by using an analogue mixer. The mixed analogue signal was converted to a digital signal using a sound card in a notebook PC to detect the watermark. The "Analogue Watermarking" row of Table 1 shows the average of the detected watermark strengths and the correct detection rate for the message.

6.4. Experiment of Sonic Watermarking (C3)

The system illustrated in Figure 11 was set up in a soundproof room for experiments with sonic watermarking. Though the room was soundproofed, a 30 dB(A)[‡] background noise was observed when nothing was played

[‡]dB(A) is a unit for the A-weighted sound level.[15]

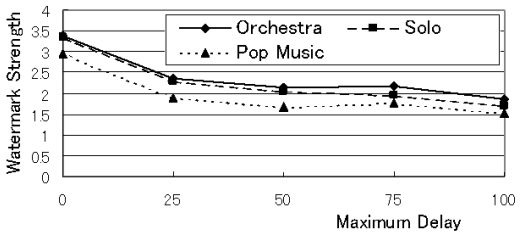

Figure 12. Average of detected watermark strengths after echo addition. The horizontal axis is the value of a maximum delay. Neither sonic watermarking nor analogue conversion was not used.

by the speakers. Two speakers with a built-in amplifier were placed next to each other to play the HS and the WS, respectively. The playback volumes of the speakers were set at the levels playing the same signal with the same sound level. The mixed sound of the HS and the WS was captured by a microphone and sent to the sound card of the notebook PC where the signal was converted to a digital signal to detect the watermark. The distance between the speakers and the microphone was 3 m. When a loud popular music song was played using the speakers, the sound level was from 70 dB(A) to 75 dB(A) at the position of the microphone.

The results of the experiment are shown in the "Sonic Watermarking" column of Table 1. The reason the detected strengths for solos and orchestras are too small to stably detect the messages is because the solo instrumental music and orchestral music have wider dynamic ranges than popular music does, and they contain more low volume portions. In particular, a solo performance of a piano has most of its power in its low frequencies, but almost no power in its middle and high frequencies. Because of the existence of the 30 dB(A) virtually white background noise, the host signal to noise ratio in those portions became too low to detect the watermark. This is the same problem described in Section 2.3 about background noises including voices and applause from the audience.

It is likely that the echo of the room will cause a detection problem in practical situations. The averages of watermark strengths detected after echo addition are shown in Figure 12. The horizontal axis is the value of a maximum delay used for echo addition. The value of 0.5 was used for the feedback coefficient for echo addition. Though the strengths were greatly decreased, correct messages were detected from every one of the tested degradations. Neither sonic watermarking nor analogue conversion was not used for this test.

6.5. Experiment of Microphone Recording (C4)

Composition methods C4a to C4c are expected to be more robust in an environment with background noises. To implement these composition methods, a calibration step is necessary between an amplitude digitally represented in the software and the corresponding actual sound played from the speakers. Without the calibration, it is impossible to generate the WS to be masked by the frequency masking effect of the sounds in the air. Calibration can be done by playing a known reference sound containing several pure tones at various frequencies, recording on the computer with a microphone the sound coming from the speakers, and calculating the ratio of the amplitudes of the played sounds to those of the recorded sounds.

We selected the composition method C4b to investigate the effect of utilizing the masking effect of the background noise, because the background noise of our experimental room can be considered to be stationary.

The background noise when no music was being played was recorded by the microphone. The power spectrums were calculated using a DFT and stored on the computer after multiplying by the calibration ratios. When real-time watermark generation is in progress, the stored power spectrums of the background noise are added to the power spectrums of the host signal for calculation of the frequency masking. The result of the experiment is shown in the "Sonic Watermarking with Sampling" row of Table 1. It can be seen that the correct detection rates for the solos and the orchestras were improved. When nothing or quiet music was being played, we could hear the sound of the WS coming out of the speaker. However, the reason we could

Table 1. The averages of detected watermark strengths (Mean) and the rates (CDR) by which the correct 64-bit message was detected.

	Pop Music		Orchestra		Solo	
	Mean	CDR	Mean	CDR	Mean	CDR
File Watermarking	2.93	100.0%	3.37	100.0%	3.34	100.0%
Analogue Watermarking	3.82	100.0%	3.09	100.0%	2.51	100.0%
Sonic Watermarking	3.49	100.0%	1.46	66.7%	1.21	66.7%
Sonic Watermarking with Sampling	3.56	100.0%	1.97	100.0%	1.80	100.0%

hear the noise was mainly because we could localize the source of the noise at the speaker, and the noise itself was not unpleasant.

7. SUMMARY

In this paper we proposed several composition methods for real-time audio watermark embedding, one of which was "sonic watermarking" to mix the host signal and the watermark signal in the air. To solve a problem found in applying a magnitude modification method to the composition, we proposed an alternate method for the embedding algorithm. Because the change is only for embedding, the same detection algorithm can be used for content watermarked by the previous method and content watermarked by the new method. The successful results of the experiments showed the possibility of the compositions methods. However, there are still large problems to solve with these methods. The method is worse in robustness and transparency compared to the previous embedding method. We have not conducted a systematic test of transparency. There would be new problems applying the method in practical applications. For example, when it is necessary to set up multiple watermark generators in a concert hall, mutual interference of the multiple watermark sounds would be a problem. If a person plays music and another person generates a watermark sound, an ownership conflict could occur. Nevertheless, we believe watermarking analogue live performances is a new and challenging domain.

ACKNOWLEDGMENTS

The author wishes to thank Seiji Kobayashi for information on psychoacoustic models.

REFERENCES

1. W. Bender, D. Gruhl, and N. Morimoto, "Techniques for data hiding," tech. rep., MIT Media Lab, 1994.
2. D. Gruhl, A. Lu, and W. Bender, "Echo hiding," in *Information Hiding*, pp. 293–315, 1996.
3. L. Boney, A. H. Tewfik, and K. N. Hamdy, "Digital watermarks for audio signals," in *IEEE Int. Conf. on Multimedia Computing and Systems*, pp. 473–480, Hiroshima, Japan, June 1996.
4. M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, "Robust audio watermarking using perceptual masking," *Signal Processing* **66**, pp. 337–355, 1998.
5. J. Haitsma, M. van der Veen, F. Bruekers, and T. Kalker, "Audio Watermarking for Monitoring and Copy Protection," in *Proc. of ACM Multimedia*, pp. 119–122, Marina del Rey, USA, November 2000.
6. R. Tachibana, S. Shimizu, T. Nakamura, and S. Kobayashi, "An audio watermarking method robust against time- and frequency-fluctuation," in *Proc. of SPIE Int. Conf. on Security and Watermarking of Multimedia Contents III*, **4314**, pp. 104–115, San Jose, USA, January 2001.
7. D. Kirovski and H. Malvar, "Spread-spectrum audio watermarking:requirements, applications, and limitations," in *IEEE Forth Workshop on Multimedia Signal Processing*, pp. 219–224, Cannes, France, October 2001.
8. Secure Digital Music Initiative, "http://www.sdmi.org/."

9. Japan Society for the Rights of Authors, Composers and Publishers, "Final selection of technology toward the global spread of digital audio watermarks." http://www.jasrac.or.jp/step2001/presse.htm, October 2001.
10. E. Zwicker and H. Fastl, *Psychoacoustics*, Springer, 1999.
11. R. Tachibana, S. Shimizu, S. Kobayashi, and T. Nakamura, "An audio watermarking method using a two-dimensional pseudo-random array," in *Signal Processing*, **82**, pp. 1455–1469, October 2002.
12. R. Tachibana, "Improving audio watermark robustness using stretched patterns against geometric distortion," in *Advances in Multimedia Information Processing*, Y.-C. Chen, L.-W. Chang, and C.-T. Hsu, eds., *Lecture Notes in Computer Science (LNCS) 2532*, pp. 647–654, Springer, Hsinchu, Taiwan, December 2002. Proc. of the 3rd IEEE Pacific-Rim Conference on Multimedia (PCM2002).
13. S. Shimizu, "Performance analysis of information hiding," in *Proc. of SPIE Int. Conf. on Security and Watermarking of Multimedia Contents IV*, **4675**, pp. 421–432, San Jose, USA, January 2002.
14. ISO/IEC, "Information technology – coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s – part 3: Audio," Tech. Rep. 11172-3, ISO/IEC, 1993.
15. M. J. Crocker, "Rating measures, descriptors, criteria, and procedures for determining human response to noise," in *Encyclopedia of Acoustics*, M. Crocker, ed., **2**, ch. 80, pp. 943–965, John Wiley & Sons, Inc., 1997.