DETERMINING RECORDING LOCATION BASED ON SYNCHRONIZATION POSITIONS OF AUDIO WATERMARKING

Yuta Nakashima[1], *Ryuki Tachibana*[2], *Masafumi Nishimura*[2], *Noboru Babaguchi*[1]

[1]Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka, Japan
[2]Tokyo Research Laboratory, IBM Japan, 1623-14 Shimotsuruma, Yamato, Kanagawa, Japan

ABSTRACT

In this paper, we propose a novel application of digital watermarking, determination of recording locations. This application enables us to determine the seat location in an auditorium where a recording was made. Precisely measured synchronization positions of the spread-spectrum watermarks are used for the determination. To avoid use of mismeasured synchronization positions, the algorithm discards synchronization positions with the corresponding normalized correlation values below a threshold. The experiments with our implementation resulted in accurate determinations; almost all of the locations can be determined within the error of 0.5 m. These experimental results successfully show the potential applicability of our application.

Index Terms— Acoustic signal processing, position measurement, copyright protection

1. INTRODUCTION

Rapid development of digital video cameras has enabled us to make high quality recordings. However, this also resulted in increases of illegal camcording. These illegal copies violate the copyright of the content owner.

In recent years, digital watermarking techniques to deter such criminal acts have been proposed. Haitsma et al. [1] and Nguyen et al. [2] have introduced methods to embed IDs, which reveal the auditorium and date when the movie was presented, as a watermark (WM), and to detect it. However, these ID-based techniques are not sufficient to prevent the illegal camcording since the information obtained from the ID cannot specify the precise recording location. A more precise location of the recording in the theater should be obtained for prevention of the illegal camcording.

In this paper, we address a novel application of digital audio watermarking, determination of recording location [3], performing further experimental considerations. The precise location where an illegal recording was made can be determined from the WMs embedded into the audio signals, adding additional deterrence against the illegal camcording. To determine the recording location, we use the synchronization positions of a spread-spectrum (SS) watermarking [4, 5, 6]. When the watermark signal is too weak compared to the host signal and noise, the error of the synchronization can result in unignorable error of the determined recording location. To avoid this, we introduce a threshold to evaluate the reliability of the determined recording location. In addition, we can estimate the variance of the error from the value of normalized correlation calculated in the watermark detector.

The rest of this paper is organized as follows: We introduce the determination of recording location in the next section. In Section 3, our watermarking method is briefly described. We give a description on the method to determine the recording locations in Section 4. The results of experiments with our method are then shown in Section 5. The summary of the paper is Section 6.

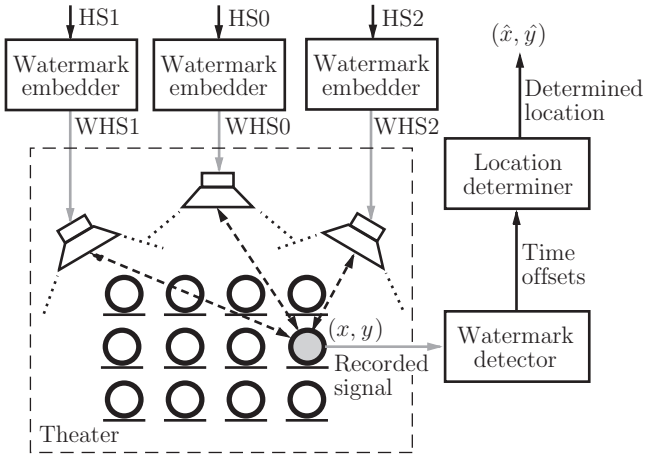

Fig. 1. An overview of the proposed system.

2. SYSTEM OVERVIEW

Figure 1 shows an overview of the proposed system. A multi-channel sound, such as the sound track of a movie, is first decomposed into the multiple single-channel host signals (HSs). These signals are stored in a computer as uncompressed sound files. Then the WM embedder for each channel embeds a WM into the HS. We call the host signal into which the WM has been embedded the watermarked host signal (WHS). In the auditorium, the WHSs are emitted into the air from loudspeakers for each channel. If a monaural recording of the content is made and becomes available via the Internet or some other channels, the content owner can feed the recorded signal into the WM detector. The recorded signal is a mixture of the WHSs with time offsets proportional to the distances from the loudspeakers to the microphone (dashed arrows in Fig. 1). Although these offsets can vary because of cropping or the starting time of a recording, their differences, or time differences of arrival (TDOAs), remain unaffected. The WM detector calculates the TDOAs, forwarding them to the location determiner. Then, the location determiner iteratively solves nonlinear simultaneous equations for the recording location.

This system requires the watermark to fulfill the following two requirements: (i) Each WHS in the recorded signal can be detected independently. (ii) The time offsets of each WHS are measurable. Considering these requirements, We view SS watermarking methods to be suitable for this application. This is because SS WMs can

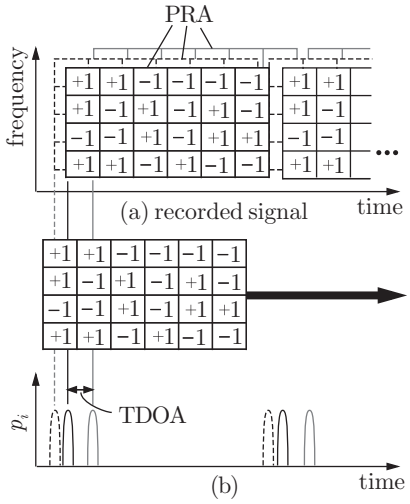

Fig. 2. The recorded signal is the mixture of the WHSs with the delays (a). Moving the PRA for the ith channel, the inner products, p_i, are calculated. Each p_i gives the maximum value when the PRA and the recorded signal are exactly synchronized (c).

be detected independently as long as different pseudo-random arrays (PRAs) are used for each WM. Furthermore, we can exploit the synchronization between the recorded signal and the PRA. In other words, the precisely measured synchronization positions can be used as the time offsets of the WHSs.

Although the framework to determine the recording location described above requires the locations of loudspeakers, in this paper, we assume that those locations are already known. An ID contained by another WM or PRAs associated with the locations of the loudspeakers, may give that information.

3. MEASURING TIME DIFFERENCE OF ARRIVAL

In this section, we briefly describe our watermarking method and measurement of the TDOAs. We based our modified watermarking method on [4]. The significant modifications are the elimination of the message bits and the smaller shifts of the PRAs in the WM detector. The elimination of the message bits increases the chances of synchronization, because all of the WM bandwidth is used for synchronization. The smaller shifts of the PRAs allow for more precise synchronization of the original PRAs and the PRAs in the recorded signal.

Each WM embedder first transforms the HS into the frequency domain by using the discrete Fourier transform (DFT), and constructs the time-frequency plane of the HS. Then the PRA is arranged on the time-frequency plane. The magnitudes of the HS are altered according to the values of the PRA with imperceptible magnitudes changes obtained from the psychoacoustic models [7, 8, 9].

The recorded signal is the mixture of all of the WHSs with the time offsets (Fig. 2(a)). Moving the PRA by Δ samples, the WM detector calculates a normalized correlation, $p_i = \mathbf{h}\mathbf{w}_i^t / \|\mathbf{h}\|\|\mathbf{w}_i\|$, between the recorded signal, $\mathbf{h}$, and the original PRA for the ith WM, $\mathbf{w}_i$, where $\mathbf{w}_i^t$ is the transposition of $\mathbf{w}_i$. The normalized correlation yields a much larger value if the original PRA and the PRA in the recorded signal are exactly synchronized. The detector finds the synchronization position where p_i is maximal. Since the peak of p_i was shifted due to the delay of the WHS caused by the distance from the loudspeaker to the microphone, the differences between the synchronization positions of one channel and another are considered to be the TDOAs between these channels.

It should be noted that if the maximum value of p_i is not sufficiently large, the synchronization position can be incorrect. Therefore, it is necessary to avoid using mismeasured synchronization positions. Assuming that the accuracy of the synchronization positions depends only on p_i, then the detector uses the synchronization positions if $p_i \geq T$ where T is a predetermined threshold.

Generally, the duration of the PRA is much shorter than the durations of the HSs. Thus, even though we employ the linear assumption method introduced in [4], the detector gives multiple synchronization positions for a recorded signal. Let z_{i0} denote the TDOA between the ith and the $0th$ channel. The detector calculates z_{i0} only if $p_i \geq T$ and $p_0 \geq T$, and computes the averages, $\overline{z_{i0}}$, of z_{i0} for every N_A measurements of the synchronization positions. Therefore, the TDOAs, whose number is up to N_A, are averaged. This operation reduces the variance of the TDOAs by up to $1/N_\mathrm{A}$. The detector forwards these values to the location determiner.

Table 1. The music samples. The original music was cropped to be 300 seconds long.

JPOP1	recording of a live performance of japanese popular music
ROCK	recording of a live performance of rock music
JPOP2	recording of a live performance of japanese popular music
POP	popular music
JAZZ	recording of a live performance of jazz

Table 2. The testlocations. "TL" stands for testlocation.

TL No.	0	1	2	3	4
Location	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(2, 2)
TL No.	5	6	7	8	9
Location	(3, 2)	(4, 2)	(3, 3)	(4, 3)	(4, 4)

4. DETERMINING RECORDING LOCATION

We formulate the nonlinear simultaneous equations for the recording location based on [10]. Let the ith loudspeaker and the microphone be arranged at $\mathbf{s}_i = (x_i, y_i)$ where $i = 0, 1, \cdots, N_\mathrm{S} - 1$ (N_S is the number of the loudspeakers), and at $\mathbf{m} = (x_\mathrm{m}, y_\mathrm{m})$, respectively. With no loss of generality, we can map the 0th loudspeaker to the origin of the planer in the auditorium. The law of cosines applied to the 0th loudspeaker, the ith loudspeaker ($i = 1, 2, \cdots N_\mathrm{S} - 1$), and the microphone gives

$$\|\mathbf{s}_i\|^2 - \hat{d}_{i0}^2 - 2\|\mathbf{m}\|\hat{d}_{i0} - 2\mathbf{s}_i^t\mathbf{m} = 0 \qquad (1)$$

where $\mathbf{s}_i^t$ is the transposition of $\mathbf{s}_i$. In this equation, $\hat{d}_{i0}$ is an estimate of the difference of $\|\mathbf{s}_0 - \mathbf{m}\|$ and $\|\mathbf{s}_i - \mathbf{m}\|$, and given by $\hat{d}_{i0} = \overline{z_{i0}}V$ where V is the speed of sound. Thus, (1) defines the nonlinear simultaneous equations for the recording location. The determiner solves the equations with the Gauss-Newton algorithm and outputs the determined location $\hat{\mathbf{m}}$.

5. EXPERIMENTS

5.1. Setting the Threshold

First, we must decide the threshold, T, under the assumption that the accuracy of the synchronization positions is dependent only on p_i. Since the distribution of the synchronization positions is not specified, we empirically determined the threshold. We performed the

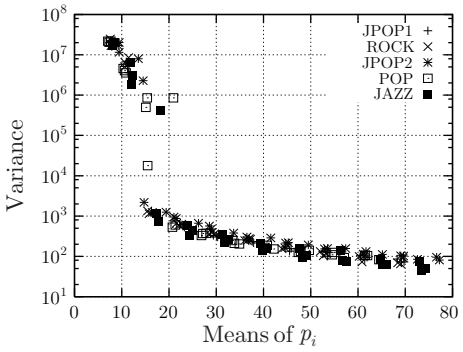

Fig. 3. The plot of the means of p_i versus variances of the synchronization positions.

Table 3. The mean, $\mu_{\rm w}$ and variance $\sigma^2_{\rm w}$ of determination errors in the worst case ($T = 17.5$, $N_{\rm A} = 10$).

TL No.	0	1	2	3	4
μ_w	0.086	0.102	0.157	0.443	0.125
σ^2_w	0.002	0.003	0.008	0.223	0.005
TL No.	5	6	7	8	9
μ_w	0.188	0.334	0.253	0.387	0.502
σ^2_w	0.013	0.053	0.025	0.073	0.133

following procedure to specify the threshold for correlation values that discriminate mismeasured synchronization positions from correct synchronization positions: (i) Decompose the three-channel music file into three single-channel music files. (ii) Embed the WM into the files. The magnitude changes generated by the psychoacoustic models are multiplied by various factors. (iii) Calculate the means of p_i for each factor in the three single-channel music files, and the means and variances of the corresponding synchronization positions. (iv) Repeat (i)—(iii) for the music samples listed in Table 1. The threshold is set so that almost all of the synchronization positions are correct.

The plot of the means of p_i versus the variances of the synchronization positions of each music sample are shown in Fig. 3. The similarity between the results for each music sample supports the assumption that the synchronization position depends only on p_i. The experimental results imply that the synchronization positions whose values of p_i exceed approximately 15 give correct values, though there are some exceptions. In other words, the system of the application should be designed so that the p_i can go beyond 15 if this implementation is used. Thus, introducing a margin of 2.5, we set T to 17.5. This margin offers a trade-off between the number of determined locations and their accuracy.

If the variance of the synchronization positions is the only cause of the determination error, we can see the determination errors in the worst case as follows. The worst case is the case where the normalized correlation, p_i, constantly equals the threshold, T. Let the variance of a synchronization position be $\sigma^2_{{\rm s},T}$. Since a TDOA, z_{i0}, is a difference of two synchronization positions, its variance is $2\sigma^2_{{\rm s},T}$. From the central limit theorem, the average of TDOAs, $\overline{z_{i0}}$, roughly approximates to the normal distribution $\mathcal{N}(\mu_{\rm s}, 2\sigma^2_{{\rm s},t}/N_{\rm A})$ where $\mu_{\rm s}$ is the mean of the TDOAs. The determination error cannot be easily computed in analytical way because the recording location is determined by an iterative method. Thus, we conducted a simulation for each testlocation listed in Table 2, assuming that $T = 17.5$ and $N_{\rm A} = 10$. According to Fig. 3, $\sigma^2_{{\rm s},T}$ is approximately 1000 in this case. We randomly generated the averaged TDOAs, $\overline{z_{i0}}$ from its distribution with these parameters, and determined the recording locations based on the averaged TDOAs by solving the nonlinear equations. Then the means and the variance of the determination errors were calculated. The results are summarized in Table 3. The further the testlocation is from the center of loudspeakers, the larger the determination error becomes. This indicates that the accuracy of the determined locations depends on the actual recording location.

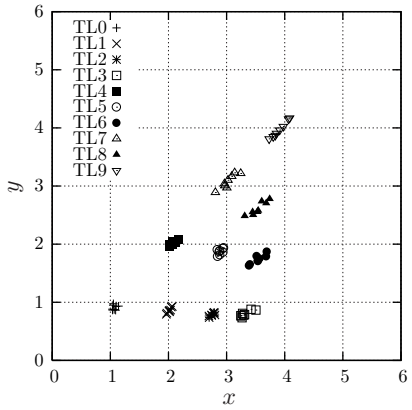

Fig. 4. The results of the simulation for ROCK music are plotted on a floormap of the room.

5.2. Simulation Experiments

Simulation experiments were performed to show the accuracy of the determined locations in an ideal environment. The simulation procedure was as follows: (i) The WM embedder embedded the WMs into the three-channel music listed in Table 1. The average of the watermark-to-host-signal ratio was -11.0 dB. (ii) The WHSs were mixed with theoretically calculated delays for each testlocation listed in Table 2. The loudspeakers were supposed to be located at (0,0), (0,3) and (3,0). (iii) The detection and determination process were performed for each of the mixed signals. $N_{\rm A}$ was set to ten.

The results are listed in Table 4, and the sample results for the ROCK music are plotted on a floormap of the room in Fig. 4 for an example. In the table, μ and σ^2 are the mean and variance of each $\|\hat{\mathbf{m}} - \mathbf{m}\|$, respectively. CDR stands for the correct determination ratio, defined as $K_{\rm C}/K_{\rm D}$ where $K_{\rm C}$ is the number of determined locations whose determination errors are less than 0.5 m, and $K_{\rm D}$ is the number of determined locations. That is, a determined location $\hat{\mathbf{m}}$ is considered to be correct if it is within a circle with radius 0.5 m centered at $\mathbf{m}$. The radius is set so that the suspects for the illegal camcording activity can be reduced to a few persons.

High CDRs in Table 4 indicate that almost all of the testlocations were successfully determined within the error of 0.5 m. However, the location determiner could not output any determined location for JAZZ. This is because the normalized correlation, p_i, in some channel did not exceed the threshold T. The means show that the testlocations on $x = 4$ give a larger errors than the others. The results plotted in Fig. 4 also have the same predisposition.

5.3. Experiments in a Real Environment

Experiments in a real environment were conducted to examine the actual performance of our implementation. The room where the experiments were conducted was approximately 6 m $\times$ 6 m rectangle

Table 4. The results of the simulation experiments.

	JPOP1			ROCK			JPOP2			POP			JAZZ		
TL	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2
0	8/8	0.122	0.001	8/8	0.097	0.001	8/8	0.116	0.001	8/8	0.086	0.000	0/0	—	—
1	8/8	0.149	0.001	8/8	0.158	0.001	8/8	0.152	0.007	8/8	0.117	0.000	0/0	—	—
2	8/8	0.327	0.002	8/8	0.280	0.004	8/8	0.327	0.003	8/8	0.241	0.004	0/0	—	—
3	0/8	0.708	0.009	3/8	0.560	0.019	0/7	0.797	0.005	3/8	0.514	0.010	0/0	—	—
4	8/8	0.089	0.003	8/8	0.101	0.003	8/8	0.182	0.006	8/8	0.123	0.004	1/1	0.191	0.000
5	8/8	0.168	0.003	8/8	0.163	0.005	8/8	0.156	0.007	8/8	0.093	0.002	0/0	—	—
6	2/8	0.530	0.015	5/8	0.457	0.031	4/7	0.455	0.028	7/8	0.424	0.011	0/0	—	—
7	8/8	0.153	0.011	8/8	0.123	0.008	7/8	0.311	0.018	7/8	0.238	0.032	0/0	—	—
8	3/8	0.604	0.026	6/8	0.205	0.011	6/7	0.360	0.009	6/8	0.382	0.020	0/0	—	—
9	8/8	0.181	0.007	8/8	0.354	0.079	4/8	0.588	0.185	6/8	0.367	0.077	0/0	—	—

Table 5. The results of the experiments in a real environment.

	JPOP1			ROCK			JPOP2			POP			JAZZ		
TL	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2	CDR	μ	σ^2
0	1/1	0.465	0.000	0/0	—	—	2/2	0.338	0.009	0/0	—	—	1/1	0.369	0.000
1	0/0	—	—	0/1	1.957	0.000	0/0	—	—	0/0	—	—	0/1	2.168	8.880
2	0/7	3.008	0.002	0/0	—	—	0/6	3.246	0.021	0/7	3.106	0.014	0/8	2.923	0.002
3	0/0	—	—	0/0	—	—	0/0	—	—	0/0	—	—	0/0	—	—
4	8/8	0.269	0.003	7/8	0.357	0.023	7/7	0.332	0.002	6/6	0.216	0.019	1/1	0.319	0.000
5	0/8	1.858	0.001	0/8	1.935	0.002	0/8	1.950	0.062	0/7	1.902	0.001	0/3	2.171	0.005
6	0/0	—	—	0/0	—	—	0/0	—	—	0/0	—	—	0/0	—	—
7	8/8	0.267	0.022	7/8	0.254	0.028	2/6	0.612	0.069	5/7	0.378	0.048	0/0	—	—
8	0/8	2.981	1.080	0/8	3.341	2.045	0/6	3.406	1.992	0/7	2.744	0.837	0/0	—	—
9	4/8	1.386	1.863	0/6	3.128	6.072	1/2	1.951	3.320	0/2	8.027	41.431	0/0	—	—

and was not anechoic. The loudspeakers were arranged at the same locations as in the simulations. In the WM embedder, the watermark-to-host-signal ratio was increased by about $+3.5$ dB from the simulation setting to improve the robustness against noises. The microphone was placed at each testlocation. We recorded the sound emitted from the loudspeakers with the microphone. Then the detection and the location determination were done on the recorded signals.

The results are summarized in Table 5. The same results as for the simulations are shown in the table. Even though echoes and attenuation degraded the WMs, some testlocations such as TLs 4, and 7 were accurately determined. However, the means of μ show larger errors than in the simulation, especially for the TLs 6, 8, and 9. The locations of TL 3 could not even be determined.

6. SUMMARY

In this paper, we proposed a novel application of the digital audio watermarking, determination of a recording location, and described an implementation of the application. The determinations are based on the synchronization positions of the SS watermarks. The simulation experiments showed that many testlocations can be determined within the error of 0.5 m. Though the results of experiments in a real environment are not as good as those of the simulations, we think that the method still has the potential applicability to illegal camcording. In future, we need to reveal the accuracy of estimation that the system is used in a larger room like a theater.

7. REFERENCES

[1] Jaap Haitsma and Ton Kalker, "A watermarking scheme for digital cinema," in *Proc. of International Conference on Image Processing*, October 2001, vol. 2, pp. 487–489.

[2] Philippe Nguyen, Raphaele Balter, Nicolas Montfort, and Severine Baudry, "Registration methods for non blind watermark detection in digital cinema applications," in *Proc. of Security and Watermarking of Multimedia Contents V*, June 2003, vol. SPIE vol. 5020, pp. 553–562.

[3] Yuta Nakashima, Ryuki Tachibana, Masafumi Nishimura, and Noboru Babaguchi, "Estimation of recording location using audio watermarking," in *Proc. of ACM Multimedia and Security Workshop 2006, Geneva*, 2006.

[4] Ryuki Tachibana, Shuichi Shimizu, Seiji Kobayashi, and Taiga Nakamura, "An audio watermarking method using a two-dimensional psuedo-random array," *Signal Processing*, vol. 82, pp. 1455–1469, 2002.

[5] Ryuki Tachibana, "Sonic watermarking," *EURASIP Journal on Applied Signal Processing*, vol. 13, pp. 1955–1964, 2004.

[6] Darko Kirovski and Henrique S. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Transaction on Signal Processing*, vol. 51, no. 4, pp. 1020–1033, 2003.

[7] Mitchell D. Swanson, Bin Zhu, Ahmed H. Tewfik, and Laurence Boney, "Robust audio watermarking using perceptual masking," *Signal Processing*, vol. 66, pp. 337–335, 1998.

[8] Eberhand Zwicker and Hogo Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag, 1999.

[9] ISO/IEC, "Information technology–coding of moving pictures and associated audio for digital storage media up to about 1.5mbits/s—part 3: Audio," Tech. Rep., ISO/IEC, 1993.

[10] Julius O. Smith and Jonathan S. Abel, "Closed-form least-square source location estimation from range-difference measurements," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 1987.