Real-Time User Position Estimation in Indoor Environments Using Digital Watermarking for Audio Signals

Ryosuke Kaneto  Yuta Nakashima  Noboru Babaguchi
Department of Information and Communications Technology
Graduate School of Engineering Osaka University
2-1 Yamadaoka, Suita, Osaka, Japan
{kaneto, nakashima, babaguchi}@nanase.comm.eng.osaka-u.ac.jp

Abstract—In this paper, we propose a method for estimating the user position where a user is holding a microphone in an indoor environment using digital watermarking for audio signals. The proposed method utilizes detection strengths, which are calculated while detecting spread-spectrum-based watermarks. Taking into account delays and attenuation of the watermarked signals emitted from multiple loudspeakers and other factors, we construct a model of detection strengths. The user position is estimated in real-time using the model. The experimental results indicate that the user positions are estimated with 1.3 m of root mean squared error on average for the case where the user is static. We demonstrate that the proposed method successfully estimates the user position even when the user moves.

Keywords—Real-time position estimation; digital watermarking for audio signals

I. INTRODUCTION

A position estimation method is demanded to provide location-based services such as navigation in recent years. Such services require the real-time position of the user. Global positioning system (GPS) is one of the most well-known position estimation methods. However, GPS cannot accurately determine the position in indoor environments. For indoor environments, many position estimation methods using WLAN [1], RFID [2], ultrasonic [3], and so forth have been proposed. However, these methods require installing specialized devices to the environment.

Nakashima et al. have proposed a method for estimating a recording position using a spread-spectrum-based digital watermarking for audio signals [4]. Their method only requires multiple loudspeakers, which are installed in many places such as malls, airports, and so on, to estimate the position. In their method, watermarks are embedded into host signals (HSs) using pseudo-random sequences (PRSs) and each watermarked signal (WS) is emitted into the air from the loudspeaker. Then, the WSs are recorded by a microphone and detection strengths, which are correlation values between the watermarks in the recorded signal and the PRSs used to embed the watermarks, are calculated. The recording position is estimated using a model of the detection strengths. However, this model takes into account only the delays of the WSs, and thus, their method requires that the RS contains WSs from at least three loudspeakers and is recorded at a static position for a long duration. Therefore, their method was not applicable to location-based services requiring a real-time position.

In this paper, we propose a method which extends [4] aiming at real-time estimation of the user position. To this end, we construct a model of the detection strengths which takes into account not only the delay but also attenuation of the WS and other factors. We experimentally evaluate the estimation accuracy for the case where the user holding the microphone is static, and demonstrate the position estimation for a moving user. In addition, we confirm that the real-time user position estimation is feasible in our implementation.

The rest of the paper is organized as follows: Section 2 describes the proposed method and experimental results are given in Section 3. We conclude the paper in Section 4.

II. USER POSITION ESTIMATION

Figure 1 shows an overview of the proposed method. The proposed method consists of three steps; watermark embedding, watermark detection, and position estimation. First, watermarks are embedded into an HS to generate WSs. Each of the WSs is then emitted into the air from a loudspeaker installed in the environment, and is received with a microphone held by a user. From the received signal (RS), detection strengths are calculated while detecting the watermarks. A user position is estimated using a model of these detection strengths.

Figure 1. An overview of position estimation method.
A. Watermark Embedding

We employ a method proposed by Tachibana et al. [5] to embed a watermark. For the \( i \)-th loudspeaker \((i = 1, 2, \ldots, N_{CH})\), the time-frequency plane of the HS is constructed by using the discrete Fourier transform (DFT), and the amplitudes of the HS are modified according to a PRS arranged in the time-frequency plane of the HS as shown in Figure 2(a) to generate the \( i \)-th WS. A different PRS is used for each loudspeaker so that each watermark can be detected independently. The PRS is repeated every \( T \) samples as shown in Figure 2(a) where \( T \) is the length of the PRS.

B. Watermark Detection

The WSs emitted from the loudspeakers are received using a microphone held by a user. The RS is thus a mixture of the WSs as shown in Figure 2(b). The watermark is detected by calculating correlation between the RS and the PRS with moving the PRS by \( \Delta \) samples. The correlation value is referred to as a detection strength. Then, a detection strength sequence, containing \( K = T/\Delta \) successive detection strengths, are fed into position estimation. The \( t \)-th detection strength sequence is defined as \( c^t_i = \{c^t_{i,1}, \ldots, c^t_{i,K}\} \) as shown in Figure 2(c) where \( c^t_{i,k} \) denotes the \( k \)-th detection strength of the \( i \)-th loudspeaker \((k = 1, 2, \ldots, K)\).

C. Position Estimation

The detection strength sequence forms a single peak at the time position where the PRS starts as shown in Figure 2(c). This peak position is determined by the delay of WS which is proportional to the length of the propagation path. The peak height, corresponding to the value of \( c^t_{i,k} \) at the peak position, depends on attenuation of the WS due to the propagation in the air and screening caused by the user body. In addition, the peak height also depends on noises including the WSs other than the \( i \)-th WS. Based on these observations, we model the detection strength and estimate the position using this model.

First, we construct a model of the length of the propagation path. The WS emitted from the loudspeaker can diffract the user body when the user body is on the direct path from the loudspeaker to the microphone. Let \( x^m, x^{sp}, \) and \( \theta \) denote the user position, the position of the \( i \)-th loudspeaker, and the user direction with respect to the \( x \)-axis, respectively, as shown in Figure 3. Assuming that the user body is a plane with the width of \( w \), we model the length of the propagation path in case of diffraction by

\[
r_i(x^m, \theta) = \sqrt{\left( \frac{r_i(x^m)}{2} - \frac{m\xi}{2} \right)^2 + \left( \frac{m\xi}{2} \right)^2 + \frac{w}{2}},
\]

where \( r_i(x^m) = ||x^m - x^{sp}|| \). \( \xi \) and \( \zeta \) are obtained by

\[
\xi = \sin \phi_i(x^m, \theta) \quad (2)
\]

\[
\zeta = \cos \phi_i(x^m, \theta),
\]

where \( \phi_i(x^m, \theta) = |\theta - \varphi_i(x^m)| \) and \( \varphi_i(x^m) \) is the angle between \((x^m - x^{sp})\) and \( x \)-axis. Considering that the WS does not diffract if the user faces to the loudspeaker, we model the peak position in \( c^t_i \) as

\[
\tau_i(x^m, \theta, \tau_0) = \begin{cases} 
\frac{F_{s} r_i(x^m, \theta)}{V_S} + \tau_0 & \text{if } \phi_i(x^m, \theta) < \frac{\pi}{2} \\
\frac{F_{s} r_i(x^m, \theta)}{V_S} + \tau_0 & \text{otherwise}
\end{cases}
\]

where \( \tau_0 \) is a parameter which depends on when the reception of the WSs is started; \( F_S \) is the sampling frequency; and \( V_S \) is the speed of sound.

Next, we construct a model of the peak height \( h_i^t \). The peak height decreases due to the propagation in the air and screening caused by the user body. In addition, the WSs other than the \( i \)-th WS, which can be deemed as noises for the \( i \)-th WS, also decrease the peak height. Taking these into account, we model the peak height as the Gaussian distribution whose mean is given by

\[
\mu_i(x^m, \theta) = \frac{\alpha(-\cos \phi_i(x^m, \theta) + \beta)/r_i(x^m)}{\sum_{j \neq i} 1/r_j(x^m) + \epsilon}
\]
where $\alpha$ is the parameter to control the mean of the peak height. The denominator of (5) represents the reduction of the peak height where the first term of the denominator corresponds to the noises due to the WSs, and $\epsilon$ corresponds to the background noises. $(-\cos \phi_i(x^m, \theta) + \beta)$ determines the reduction of the peak height caused by the screening where $\beta$ determines the effect of the screening. The probability density function of $h^i_t$ given $x^m, \theta$ and the variance $\sigma^2_H$ is represented by

$$p(h^i_t|x^m, \theta) = \mathcal{N}(h^i_t|\mu_i(x^m, \theta), \sigma^2_H)$$  \hspace{1cm} (6)

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the Gaussian distribution whose mean and variance are $\mu$ and $\sigma^2$, respectively.

Finally, the model of the detection strength is constructed by integrating the above two models. Assuming that $c^t_{i,k}$ follows the Gaussian distribution with variance $\sigma_D^2$, we can write the conditional probability density function of $c^t_{i,k}$ as

$$p(c^t_{i,k}|x^m, \theta, \tau_0, h^i_t) = \mathcal{N}(c^t_{i,k}|h^i_t g(k - \tau_i(x^m, \theta, \tau_0)), \sigma_D^2)$$  \hspace{1cm} (7)

where $g(k)$ is the averaged curve of $c^t_{i,k}$ calculated by using various PRSs. From (6) and (7), the marginal distribution of $c^t_{i,k}$ is given by

$$p(c^t_{i,k}|x^m, \theta, \tau_0) = \int_{-\infty}^{\infty} p(c^t_{i,k}|x^m, \theta, \tau_0, h^i_t) p(h^i_t|x^m, \theta) dh^i_t.$$  \hspace{1cm} (8)

Using this marginal distribution, we define a log likelihood function as

$$L(x^m, \theta, \tau_0) = \sum_{i=1}^{N_{CH}} \sum_{k=0}^{K-1} \log p(c^t_{i,k}|x^m, \theta, \tau_0).$$  \hspace{1cm} (9)

By applying the sampling importance resampling particle filter [6], the distribution of $x^m, \theta, \tau_0$ and $v^m$ is estimated from $c^t_{i,k}$ where $v^m$ is the velocity of the user. Each particle has $(x^m, \theta, \tau_0, v^m)$ as the state vector and is weighted by $L(x^m, \theta, \tau_0)$. An estimated user position is the average of $x^m$ over all particles.

### III. EXPERIMENTAL RESULTS

We evaluated the accuracy of the proposed method for static users, i.e., a user stayed at the same position during reception of the WSs. In addition, we estimated the position for moving users.

The experiments were conducted in a room of $6.5 \text{ m} \times 7.0 \text{ m}$. The number of loudspeakers was set to $N_{CH} = 2, 3, \text{ or } 4$. The volumes of the loudspeakers were set to five values, namely, 0 db, $-3 \text{ db}$, $-6 \text{ db}$, $-9 \text{ db}$, and $-12 \text{ db}$. The popular music (Sample1) and the instrumental (Sample2) were used as the HSs. The parameters used in the experiments were listed in Table I, and the other parameters, $\alpha$, $\beta$, $\epsilon$, and $\sigma_D^2$, were determined by the maximum likelihood method.

To evaluate the estimation accuracy for a static user, a microphone held by a user received the WSs at the randomly chosen positions and directions listed in Table II. Figures 4 and 5 shows the results. Static user positions were successfully estimated with 1.3 m of root mean squared error (RMSE) on average over all test cases. From Figures 4 and 5, the accuracy depends on the number of the loudspeakers. The reason is that more constraints on the user position can be obtained according to the increment of the number of the loudspeakers, and this results in accurate estimation. The estimation accuracy for Sample2 was worse than that for Sample1 due to the large dynamics of Sample2. The volume of the loudspeaker did not influence the accuracy so much because the background noises were sufficiently small in these experiments. Therefore, we need more experiments to clarify the influence of the background noises.

To demonstrate the performance of the position estimation for moving users, a user holding a microphone moved as the actual trajectory in Figure 6. We supposed that a user will not move so fast when he/she uses a location-based service such as a navigation system, and the user walked at the speed of

![Figure 4. The accuracies of static position estimation for Sample1.](image)

![Figure 5. The accuracies of static position estimation for Sample2.](image)

---

<table>
<thead>
<tr>
<th>$F_0$ [Hz]</th>
<th>$V_0$ [m/s]</th>
<th>$T$ [sample]</th>
<th>$\Delta$ [sample]</th>
<th>$\sigma_D$</th>
<th>$\psi$ [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>44100</td>
<td>3.4</td>
<td>7680</td>
<td>8</td>
<td>1</td>
<td>0.6</td>
</tr>
</tbody>
</table>

---

Table I

THE PARAMETERS.
Figure 6. The actual trajectories and the estimated trajectories. The arrows indicate the direction of the moving user.

Table II
THE TEST CASES.

<table>
<thead>
<tr>
<th></th>
<th>x [m]</th>
<th>y [m]</th>
<th>θ [deg.]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.0</td>
<td>1.9</td>
<td>205</td>
</tr>
<tr>
<td>2</td>
<td>1.1</td>
<td>2.2</td>
<td>115</td>
</tr>
<tr>
<td>3</td>
<td>2.5</td>
<td>2.5</td>
<td>270</td>
</tr>
<tr>
<td>4</td>
<td>3.2</td>
<td>1.3</td>
<td>175</td>
</tr>
<tr>
<td>5</td>
<td>3.3</td>
<td>3.5</td>
<td>105</td>
</tr>
<tr>
<td>6</td>
<td>4.3</td>
<td>3.3</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3.6</td>
<td>4.1</td>
<td>340</td>
</tr>
<tr>
<td>8</td>
<td>4.5</td>
<td>4.0</td>
<td>90</td>
</tr>
</tbody>
</table>

0.4 m/s in our experiments. The volumes of the loudspeakers were set to 0 db. The trajectories of estimated user positions are also shown in Figure 6. Although the number of the loudspeakers affects the estimation accuracy as same for the static user case, almost all of the trajectories of estimated user positions traced the actual trajectory. However, Sample2 in the two loudspeakers case gave large errors because of the same reasons for the static user case. In addition, once the user position is estimated with large error, the property of the particle filter that an estimate depends on the previous estimate prevents the method from estimating accurate user positions.

Our implementation of the proposed method on a PC with Windows XP Service Pack 3, equipping Intel Xeon 3.73 GHz CPU and 3 GBytes RAM, took 2.812 seconds for the watermark detection and 5.187 seconds for the position estimation from the received signals whose duration was of 10 second. Therefore, estimation of the user position in real-time is feasible.

IV. CONCLUSION

In this paper, we proposed a method for estimating a user position in real-time using digital watermarking for audio signals. The proposed method estimates the user position using a model of detection strengths. The experimental results showed that the estimation accuracy for a static user in a room of 6.5 m × 7.0 m was 1.3 m RMSE on average. The position estimation of a moving user was also possible. In addition, we confirmed that the proposed method can estimate a user position in real-time. From these results, we consider that the proposed method is potentially applicable to location-based services. In the future, the estimation accuracy in various environments and in a condition where many people exist in an environment should be evaluated. We should also clarify the relationship between the area of the environment and the number of the loudspeakers to achieve a desired accuracy. This work is partly supported by a Grant-in-Aid for scientific research from the Japan Society for the Promotion of Science.

REFERENCES


