PAPER

Indoor Positioning System Using Digital Audio Watermarking

Yuta NAKASHIMA[†a)], ***Student Member***, **Ryosuke KANETO**[†*], ***Nonmember***, ***and*** **Noboru BABAGUCHI**[†], ***Fellow***

SUMMARY Recently, a number of location-based services such as navigation and mobile advertising have been proposed. Such services require real-time user positions. Since a global positioning system (GPS), which is one of the most well-known techniques for real-time positioning, is unsuitable for indoor uses due to unavailability of GPS signals, many indoor positioning systems (IPSs) using WLAN, radio frequency identification tags, and so forth have been proposed. However, most of them suffer from high installation costs. In this paper, we propose a novel IPS for real-time positioning that utilizes a digital audio watermarking technique. The proposed IPS first embeds watermarks into an audio signal to generate watermarked signals, each of which is then emitted from a corresponding speaker installed in a target environment. A user of the proposed IPS receives the watermarked signals with a mobile device equipped with a microphone, and the watermarks are detected in the received signal. For positioning, we model various effects upon watermarks due to propagation in the air, *i.e.*, delays, attenuation, and diffraction. The model enables the proposed IPS to accurately locate the user based on the watermarks detected in the received signal. The proposed IPS can be easily deployed with a low installation cost because the IPS can work with off-the-shelf speakers that have been already installed in most of the indoor environments such as department stores, amusement arcades, and airports. We experimentally evaluate the accuracy of positioning and show that the proposed IPS locates the user in a 6 m by 7.5 m room with root mean squared error of 2.25 m on average. The results also demonstrate the potential capability of real-time positioning with the proposed IPS.

key words: *indoor positioning system, real-time, digital audio watermarking, particle filter*

1. Introduction

As wireless networks become ubiquitous, a variety of location-based services (LBS) such as navigation and mobile advertising have been proposed [1], [2]. This strongly motivates development of positioning systems that accurately locate or track a service user in real-time. One of the most well-known techniques for positioning is a global positioning system (GPS); however, GPS cannot be used in indoor environments because GPS signals are unavailable.

Accordingly, many indoor positioning systems (IPSs) using WLAN [3]–[6], radio frequency identification (RFID) tags [7], [8], ultrasonic [9]–[11], audible sound [12], and so forth have been proposed [13], [14]. However, some of them require special devices dedicated only for positioning, which result in high installation costs.

Recently, Nakashima *et al*. and Lazic *et al*. have proposed IPSs that use digital audio watermarking techniques [15], [16]. Although such techniques has been extensively used for applications such as copyright protection, broadcast monitoring, and authentication, their attempts have developed a new application domain, *i.e.*, positioning. These IPSs use detection strengths (DSs) calculated to detect spread spectrum-based watermarks for positioning. They first embed a watermark into a host signal (HS) to generate a watermarked signal (WS), and the WS is emitted from a speaker. Propagating in the air delays and attenuates the WS according to its propagation path, consequently changing DSs that are defined as cross-correlations of a signal received by a microphone and pseudo-random sequences (PRSs) used to embed the watermarks. The digital audio watermarking-based IPSs utilize these changes as cues for positioning.

The main advantage of these IPSs is that they are easily deployed with low installation costs because they use only commercially available speakers that have been already installed in target environments and user's mobile device equipped with a microphone, making the digital audio watermarking-based IPSs viable. However, each of them has a drawback: The IPS [15], which uses delays of WSs measured based on DSs, requires the mobile device to constantly receive WSs from at least three speakers. This is a severe restriction for real-time positioning because the WSs attenuate rapidly as the user recedes from the speakers, and the delays cannot be measured from the attenuated WSs. The IPS [16] only provides the speaker position that is nearest to the user position based on attenuation of WSs instead of an estimate of the user position; therefore, the accuracy can be insufficient for some applications of the IPS.

In this paper, we propose an IPS based on [15] aiming at applications such as indoor navigation that requires accurate and real-time user positions (Fig. 1). To this end,

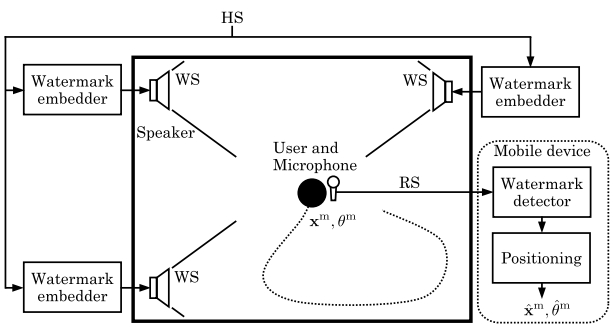

Fig. 1 An overview of the proposed IPS.

Manuscript received March 4, 2011.

[†]The authors are with the Graduate School of Engineering, Osaka University, Suita-shi, 565–0871 Japan.

[*]Presently, with Nisshin Seifun Group Inc.

a) E-mail: nakashima@nanase.comm.eng.osaka-u.ac.jp

DOI: 10.1587/transinf.E94.D.2201

we adopt a novel approach that leverages both delays and attenuation of WSs in the positioning algorithm [17]. This approach enables our proposed IPS to pinpoint the user even when the microphone does not receives WSs from three speakers constantly because the attenuation of WSs provides additional constraints on the user position. We also reduce the computational cost compared with our previous IPS [17] so that our proposed IPS can be used with low spec devices.

The following is the main contributions of this paper:

- We propose a new watermarking algorithm that is robust against excessive noises based on [18] so that watermarks can be detected even in noisy environments such as amusement arcades, considering that degradation of acoustic quality is negligible in such noisy environments.
- To leverage both delays and attenuation of WSs, we explicitly model DSs taking into account delays and attenuation of WSs. We also introduce an assumption that noises on DSs are independent and identically distributed to reduce the computational cost.
- We experimentally evaluate the accuracy of the proposed IPS in actual environments of a 45.5 m^2 rectangle room and a 435.8 m^2 octagonal cafeteria, and verify that real-time positioning is possible. The accuracy and processing time of the proposed IPS are compared with those of our previous IPS [17].

The rest of this paper is organized as follows: In the next section, we introduce related works. We provide an overview of the proposed IPS in Sect. 3. Sections 4 and 5 describe the watermarking algorithm and the positioning algorithm, respectively. Experimental results are given in Sect. 6. Section 7 concludes this paper.

2. Related Work

Many IPSs have been proposed using WLAN, RFID tags, ultrasonic, audible sound, and so forth [13], [14]. For example, Yim *et al*. developed an IPS that uses received signal strengths of WLAN [4]. Ni *et al*. proposed an RFID-based IPS that uses reference RFID tags deployed in target environments [7]. An ultrasonic-based IPS called Cricket is proposed by Priyantha *et al*. [10]. They installed ultrasonic senders in target environments and receive the signals by an inexpensive dedicated ultrasonic receiver. Compared to these techniques, our proposed IPS is characterized by its easiness of installation; it uses commercially available microphones and speakers that have been already installed in many indoor environments. In addition, IPSs that perform positioning in a remote device can suffer from a privacy issue because real-time positions of a specific user can be aggregated without the user's consent [19]–[21]. The privacy issue does not arise in our proposed IPS because positioning is done in the user's mobile device.

Various digital audio watermarking algorithms have been proposed. Yeh and Kuo proposed to modify least significant bits for embedding watermarks [22]. Bassia *et al*. proposed to adopt the spread spectrum technique in the time domain [23] for watermarking. Cvejic and Seppänen developed a spread spectrum-based algorithm in the frequency domain to improve robustness against low pass filtering and compression [24]. Tachibana *et al*. [18] as well as Kirovski and Malvar [25] proposed spread spectrum-based algorithms in the time-frequency space for further improving robustness against various attacks on WSs. Our proposed IPS adopts a digital audio watermarking algorithm based on [18] as watermarks based on this algorithm survive even after propagating in the air although other spread spectrum-based algorithms are potentially applicable.

Such digital audio watermarking techniques are adopted in a wide range of applications, *e.g.*, copyright protection, broadcast monitoring, authentication, and so forth. We use them for a very different purpose: Without watermarking, the positioning from a received signal (RS) that are mixture of multiple signals from speakers is a very tough problem if we have no knowledge on the original signals. The spread spectrum-based watermarking technique makes the problem easy because it converts the problem of positioning from the RS into the problem of positioning from DSs of which waveform is known.

Several positioning systems that use digital audio watermarking techniques have been proposed. Lazic and Aarabi proposed a digital audio watermarking algorithm and they introduced a positioning system as its application [16]. Their system provides the speaker position that is nearest to the user based on attenuation of WSs. Nakashima *et al*. proposed to use delays of WSs for accurate positioning [15], [26]–[28], aiming at a countermeasure of movie piracy in theaters. We have proposed an IPS that extends [15], which uses both attenuation of WSs and delays for real-time and accurate positioning [17]. In this paper, we construct a simplified model of DSs based on [17] for further speeding up so that the proposed IPS can work even in low spec devices.

3. An Overview of the Proposed IPS

Figure 1 shows an overview of the proposed IPS. First, a spread spectrum-based watermark derived from [18] is embedded into an HS to generate a WS. We use a music piece as an HS because our target indoor environments such as shopping malls and amusement arcades often play background music. The WS is emitted into the air from the corresponding speaker installed in a target environment, and received by a user's mobile device equipped with a microphone. Therefore, the RS is a mixture of the WSs from the speakers around the user. From the RS, DSs of each watermark are calculated. The proposed IPS locates the user using a model of the DSs given the user position and direction. To track the user position and direction, we adopt particle filter [29] because it can control the computational cost by changing the number of particles and can improve the accuracy by Bayesian updating.

The basic idea for positioning using the model of the DSs is to utilize the property of the spread spectrum-based

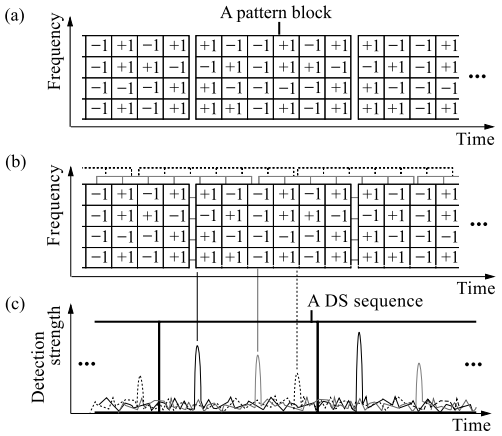

Fig. 2 An illustrative example of (a) a WS, (b) an RS, and (c) DSs.

watermarking technique that the DSs of the watermark, which are defined as a cross-correlation between PRSs and the RS in a transformed domain, form peaks as in Fig. 2 (c). The heights and positions of the peaks depend on, *e.g.*, delays and attenuation of WSs due to propagation in the air from the speakers to the microphone; therefore, the heights and positions of the peaks can be used for positioning.

4. Digital Audio Watermarking for IPS

In our watermarking algorithm, we generate a WS in the time-frequency plane of the HS constructed using the discrete Fourier transform (DFT). The energy of the watermark is spread on a region in the time-frequency plane called a pattern block (Fig. 3 (a)) that consists of $W_{\mathrm{B}} \times H_{\mathrm{B}}$ smaller regions called tiles as shown in Fig. 2 (a), where a tile consists of $2 \times H_{\mathrm{T}}$ Fourier coefficients (Fig. 3 (b)), *i.e.*, the amplitudes of Fourier coefficients in a tile at (w, h) are modified according to a PRS $\omega^c(w, h)$ for the c-th WS. Considering that degradation of acoustic quality of the WSs is not a critical problem in noisy environments, we modify the original watermarking algorithm [18] to cut off the high frequency part of the HS so that they can be robust against noises.

4.1 Watermark Embedding

The c-th WS is generated as follows:

1. The HS $x(t)$ is divided into frames, each of which consists of N samples, using the sine window defined as

$$win(t) = \begin{cases} \sin(\pi t/N) & \text{for } 0 \le t < N \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

   Adjacent frames are overlapped with each other by $N/2$ samples to avoid discontinuities. The t-th sample of the f-th frame is given by

$$\tilde{x}(f, t) = x(t + fN/2)\, win(t). \quad (2)$$

2. A frame is transformed into the frequency domain using the DFT. The k-th Fourier coefficient of the f-th frame, $X(f, k)$, is obtained as

$$X(f, k) = \mathrm{DFT}[\tilde{x}(f, t)](k). \quad (3)$$

   The amplitude and phase of the Fourier coefficient are denoted by $X^{\mathrm{A}}(f, k)$ and $X^{\mathrm{P}}(f, k)$, respectively.
3. The amplitude modification sign $sign(f, k)$, which indicates whether an amplitude in the tile at (w, h) is increased or decreased, is calculated as

$$sign^c(f, k) = \omega^c(w, h)\, m_{(f \bmod 2)}, \quad (4)$$

   where (f, k) is replaced with corresponding (w, h); $m_0 = +1$ and $m_1 = -1$ are introduced to alleviate degradation of the watermark due to the HS assuming that the amplitudes of Fourier coefficients in successive frames give similar values.
4. The amplitude of the Fourier coefficient of the WS, $Y^{\mathrm{A}}(f, k)$, is determined. For k corresponding to the frequency lower than TH, the amplitude is set to

$$Y^{\mathrm{A}}(f, k) = X^{\mathrm{A}}(f, k) + A(f, k) sign^c(f, k). \quad (5)$$

   where $A(f, k)$ is an inaudible amount of amplitude change obtained by the psychoacoustic model [30]. Otherwise, depending on $sign^c(f, k)$, the amplitude is cut off as

$$Y^{\mathrm{A}}(f, k) = \begin{cases} X^{\mathrm{A}}(f, k) & \text{if } sign^c(f, k) = +1 \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

   We empirically set TH to 6000 Hz because noises higher than this frequency seem relatively inaudible.
5. The WS in the frequency domain is transformed into time domain by the inverse DFT using the original phases of the HS as

$$\tilde{y}(f, t) = \mathrm{IDFT}\left[Y^{\mathrm{A}}(f, k) \exp\left\{jX^{\mathrm{P}}(f, k)\right\}\right](t), \quad (7)$$

   where $j = \sqrt{-1}$.
6. The final WS, $y(t)$, is generated by the overlap-and-add using the sine window as follows:

$$y(t) = \sum_f \tilde{y}(f, t - fN/2)\, win(t - fN/2). \quad (8)$$

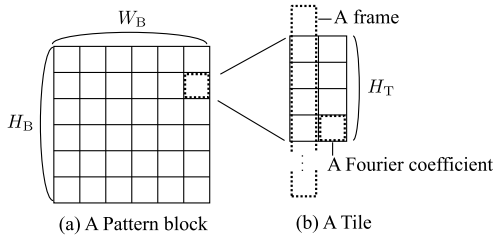

Fig. 3 (a) A pattern block consisting of $W_B \times H_B$ tiles. (b) A tile comprised of H_T amplitude spectra of two successive frames.

4.2 Watermark Detection

In watermark detection, we calculate DSs for each WS in the RS $z(t)$ for every Δ samples as follows:

1. The RS is divided into frames $\tilde{z}_i(f,t)$ by the sine window so that $\tilde{z}_i(0,t)$ starts at the $i\Delta$-th sample of $z(t)$, *i.e.*,
$$\tilde{z}_i(f,t) = z(t + i\Delta + fN/2)\, win(t). \quad (9)$$
2. The frame is transformed into the frequency domain by the DFT as
$$Z_i(f,k) = DFT[\tilde{z}_i(f,t)](k). \quad (10)$$
3. The amplitude of $Z_i(f,k)$ is normalized as
$$\bar{Z}_i^{\mathrm{A}}(f,k) = \frac{Z_i^{\mathrm{A}}(f,k)}{\frac{1}{N/2}\sum_{k=0}^{N/2-1} Z_i^{\mathrm{A}}(f,k)}. \quad (11)$$
4. The difference between log amplitudes of two frames is calculated as
$$D_i(w,k) = \log \bar{Z}_i(2w,k) - \log \bar{Z}_i(2w+1,k). \quad (12)$$
As mentioned in the previous section, this alleviates degradation of the watermark due to the HS because the amplitudes of the HS in the successive frames are canceled while the watermark is enhanced.
5. The sum of $D_i(w,k)$ is computed by
$$\rho_i(w,h) = \sum_k D_i(w,k), \quad (13)$$
where summation is computed for k in the tile at (w,h) of the pattern block that is assumed to start at $i\Delta$-th sample of the RS.
6. The i-th DS for the c-th WS is calculated as
$$s^c(i) = \frac{\sum_{w=1}^{W_{\mathrm{B}}}\sum_{h=1}^{H_{\mathrm{B}}} \omega^c(w,h)\left[\rho_i(w,h) - \bar{\rho}_i\right]}{\sqrt{\sum_{w=1}^{W_{\mathrm{B}}}\sum_{h=1}^{H_{\mathrm{B}}} \left\{\omega^c(w,h)\left[\rho_i(w,h) - \bar{\rho}_i\right]\right\}^2}}, \quad (14)$$
where
$$\bar{\rho}_i = \frac{1}{W_{\mathrm{B}}H_{\mathrm{B}}}\sum_{w=1}^{W_{\mathrm{B}}}\sum_{h=1}^{H_{\mathrm{B}}} \rho_i(w,h). \quad (15)$$

From the central limit theorem, $s^c(i)$ follows the Gaussian distribution. If the RS does not contain the watermark embedded with $\omega^c(w,h)$, $s^c(i)$ asymptotically follows the standard Gaussian distribution because the standard deviation of the numerator of (14) is given by the denominator.

5. Positioning

The DSs form a peak at the time position where a pattern block starts as shown in Fig. 2 (c). This peak position is determined by the delay of WS that is proportional to the length of the propagation path as in Fig. 4. The peak height, corresponding to the value of $s^c(i)$ 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, as well as background noises, WSs themselves behave like noises that decrease the peak height. Based on these observations, we construct a DS model. The user is located using the DS model and particle filter [29], which allows us to control the computational cost and to improve accuracy.

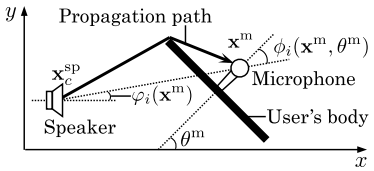

Fig. 4 A propagation path of a WS.

5.1 Detection strength model

To construct a DS model, we first divide $s^c(i)$ into DS sequences so that a DS sequence has a single peak as in Fig. 2 (c), *i.e.*,
$$\boldsymbol{s}_l^c = [s^c(l\Lambda), s^c(l\Lambda+1), \cdots, s^c((l+1)\Lambda-1)]^\top \quad (16)$$
where Λ is the length of a pattern block in sample divided by Δ and $\top$ represents transpose.

The peak position in $\boldsymbol{s}_l^c$ that depends on the length of propagation path from the speaker for the c-th WS to the microphone is modeled as follows. Let $\boldsymbol{x}_c^{\mathrm{sp}} = (x_c^{\mathrm{sp}}, y_c^{\mathrm{sp}})$, $\boldsymbol{x}^{\mathrm{m}} = (x^{\mathrm{m}}, y^{\mathrm{m}})$, and θ^{m} denote the position of the speaker for the c-th WS, the position of the user, and the user direction with respect to the x-axis, respectively, as shown in Fig. 4. A WS emitted from the speaker can be diffracted by the user body when the user body is on the direct path from the speaker to the microphone. Assuming that the user body is a plane with the width of W_{U}, we model the length of the propagation path in the case of diffraction by
$$r'_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) = \sqrt{\left\{r_c(\boldsymbol{x}^{\mathrm{m}}) - \frac{W_{\mathrm{U}}\xi}{2}\right\}^2 + \left\{\frac{W_{\mathrm{U}}\zeta}{2}\right\}^2} + \frac{W_{\mathrm{U}}}{2}, \quad (17)$$
where $r_c(\boldsymbol{x}^{\mathrm{m}}) = \|\boldsymbol{x}^{\mathrm{m}} - \boldsymbol{x}_c^{\mathrm{sp}}\|$. ξ and ζ are obtained by
$$\xi = \sin\phi_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) \quad (18)$$
$$\text{and} \quad \zeta = \cos\phi_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}), \quad (19)$$
where $\phi_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) = |\theta^{\mathrm{m}} - \varphi_c(\boldsymbol{x}^{\mathrm{m}})|$ and $\varphi_c(\boldsymbol{x}^{\mathrm{m}})$ is the angle between $(\boldsymbol{x}^{\mathrm{m}} - \boldsymbol{x}_c^{\mathrm{sp}})$ and the x-axis. Considering that the WS is not diffracted if the user faces to the speaker, the length of the propagation path $R_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})$ from the speaker for the c-th WS to the microphone is given by
$$R_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) = \begin{cases} r'_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) & \text{if } \phi_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) < \pi/2 \\ r_c(\boldsymbol{x}^{\mathrm{m}}) & \text{otherwise} \end{cases}. \quad (20)$$
Using this length, we model the peak position in $\boldsymbol{s}_l^c$ as
$$\tau_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0) = \frac{F_{\mathrm{S}} R_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})}{V_{\mathrm{S}}\Delta} + \tau_0 \quad (21)$$
where τ_0 is a parameter that 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_l^c. The

peak height decreases due to propagation in the air and screening caused by the user body. The WSs and background noises also decrease the peak height. Taking these into account, we model the peak height as

$$h_l^c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) = \frac{[\alpha G(\phi_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})) + \beta]/R_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})}{\sum_{c' \neq c} 1/R_{c'}(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}) + \epsilon}, \quad (22)$$

where $G(\phi)$ represents the screening due to the user body defined using a predetermined parameter ν as

$$G(\phi) = e^{-\nu(\phi-\pi)^2}. \quad (23)$$

In (22), we assume that the peak height for the c-th WS is inversely proportional to $R_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})$. The effect of the screening by the user body is controlled by the parameters α and β. The denominator of (22) represents the reduction of the peak height: The first and second terms of the denominator correspond to reduction of the peak height due to the WSs and background noises, respectively.

Finally, we integrate the models of the peak position and the peak height to construct a DS model. The characteristic waveform of a DS sequence as in Fig. 5 (a) comes from the watermarking algorithm, *i.e.*, in watermark embedding, a pseudo-random number in a PRS is embedded by modifying the amplitudes of successive two Fourier coefficients toward opposite signs, and this forms two valleys at the both sides of the peak. We compute the averaged waveform of a DS, $\boldsymbol{a} = [a_0, a_1, \cdots, a_{\Lambda-1}]$, as shown in Fig. 5 (b). Let $\boldsymbol{a}^{[\tau]}$ denote the vector whose elements are the circular shift of those of $\boldsymbol{a}$ by the floor of τ, *i.e.*,

$$\boldsymbol{a}^{[\tau]} = [a_{\Lambda-\lfloor\tau\rfloor+1}, \cdots, a_\Lambda, a_1, a_2, \cdots, a_{\Lambda-\lfloor\tau\rfloor}]. \quad (24)$$

Assuming that noises on $\boldsymbol{s}_l^c$ are independent and identically distributed and follow a Gaussian distribution, $\boldsymbol{s}_l^c$ can be modeled as the multivariate Gaussian distribution of which mean is given by $\boldsymbol{\mu}_l^c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0) = h_l^c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})\boldsymbol{a}^{[\tau_c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0)]}$, *i.e.*,

$$p(\boldsymbol{s}_l^c \mid \boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0) = \mathcal{N}(\boldsymbol{s}_l^c \mid \boldsymbol{\mu}_l^c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0), \Sigma) \quad (25)$$

where $\mathcal{N}(\cdot)$ represents the multivariate Gaussian distribution and $\Sigma = \mathrm{diag}(\sigma^2, \cdots, \sigma^2)$ is a $\Lambda \times \Lambda$ diagonal matrix.

5.2 Parameter Estimation

The DS model given by (25) depends on the parameters α, β, ϵ, and σ^2. We estimate the values of these parameters using a set of DS sequences for known $\boldsymbol{x}^{\mathrm{m}}$ and θ^{m} with various speaker arrangements. Let $\boldsymbol{s}_n^c$ denote the n-th DS sequence for the c-th WS in the set received at $\boldsymbol{x}_n^{\mathrm{m}}$ and θ_n^{m}. The peak position $\tau_{0,n}$ is unknown even in the set because it depends on the timing to start receiving that is hard to precisely control. Therefore, we estimate $\tau_{0,n}$ by

$$\tilde{\tau}_{0,n} = \arg\max_{\tau_{0,n}} \sum_c \log p(\boldsymbol{s}_n^c \mid \boldsymbol{x}_n^{\mathrm{m}}, \theta_n^{\mathrm{m}}, \tau_{0,n}). \quad (26)$$

This estimation can be more accurate by concatenating DS sequences that share the same value of $\tau_{0,n}$, *i.e.*, DS sequences obtained from a single RS.

Using the estimates of the peak positions, we define a log likelihood function $L'_{\alpha,\beta,\epsilon,\sigma^2}$ of the set as

$$L'_{\alpha,\beta,\epsilon,\sigma^2} = \sum_{n,c} \log p(\boldsymbol{s}_n^c \mid \boldsymbol{x}_n^{\mathrm{m}}, \theta_n^{\mathrm{m}}, \tilde{\tau}_{0,n}). \quad (27)$$

We can estimate the parameters by maximizing $L'_{\alpha,\beta,\epsilon,\sigma^2}$. However, our preliminary study indicated that the maximum of $L'_{\alpha,\beta,\epsilon,\sigma^2}$ cannot be uniquely determined; therefore, assuming that ϵ is small, we introduce a penalty term as

$$L'_{\alpha,\beta,\epsilon,\sigma^2} - \lambda\epsilon^2. \quad (28)$$

For given ϵ, the maximization is reduced to a linear least square problem with respect to α and β, which can be easily solved because σ^2 is irrelevant to α, β, and ϵ. Therefore, we exhaustively search for ϵ that minimizes (28) with solving the linear least square problem and then estimate σ^2 that maximizes $L'_{\alpha,\beta,\epsilon,\sigma^2}$.

5.3 Positioning Using Particle Filter

By applying the sampling importance resampling (SIR) particle filter [29], we estimate the distribution of $\boldsymbol{x}^{\mathrm{m}}$, θ^{m}, and τ_0 from $\boldsymbol{s}_l^c$. Each particle has a state vector $\boldsymbol{v} = (\mathbf{x}^{\mathrm{m}}, \theta^{\mathrm{m}}, \tau_0)$. In each iteration, a particle is weighted by the likelihood

$$L(\boldsymbol{v}) = \prod_c p(s_l^c|\boldsymbol{v}) \quad (29)$$

calculated using (25), and the state vector $\boldsymbol{v}$ is updated by

$$\boldsymbol{v}' = \boldsymbol{v} + \boldsymbol{\eta} \quad (30)$$

where $\boldsymbol{\eta}$ is a Gaussian distributed noise whose mean is $\mathbf{0}$ and variance is $\Sigma_{\mathrm{U}} = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_\theta^2, \sigma_\tau^2)$. We compute the average of $\mathbf{x}^{\mathrm{m}}$ and θ^{m} over all particles as an estimate. Compared to our previous IPS [17], the proposed IPS is computationally efficient: In [17], the weight for each particle requires to calculate the logarithm, which is computationally expensive, for each element in a DS sequence. In contrast, the proposed IPS uses only the four arithmetic operations to calculate (29) in our implementation using logarithms because the variance of $p(\boldsymbol{s}_l^c|\boldsymbol{v})$, which is a constant, can be ignored.

To improve the accuracy of positioning, we can use the ensemble mean of $\boldsymbol{s}_l^c$ instead of $\boldsymbol{s}_l^c$ given by

$$\langle \boldsymbol{s}_l^c \rangle = \frac{1}{Q} \sum_{q=0}^{Q-1} \boldsymbol{s}_{l-q}^c, \quad (31)$$

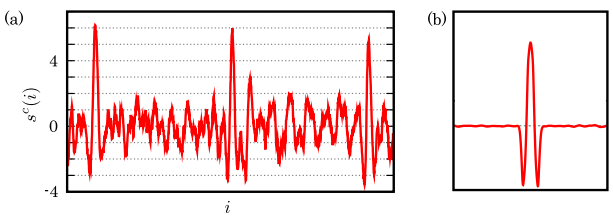

Fig. 5 (a) An example of $s^c(i)$ and (b) the averaged waveform of a DS sequence.

with modification of σ^2. This reduces the noises of s_l^c if the user is static. However, if the user rapidly moves, the ensemble mean introduces extra noises; therefore, we should set Q as small as possible for real-time positioning.

6. Experiments

We experimentally evaluated the accuracy of the proposed IPS for the static and moving user cases.

6.1 Experimental Setup

We evaluated our IPS in the following two environments with various speaker arrangements.

- ROOM: An office room of 6.5 m × 7.0 m (Fig. 6 (a)).
- CAFE: An octagon-shaped cafeteria each edge of which is 9.5 m (Fig. 6 (b)).

We use EDIROL UA-101 as an audio interface from a PC to speakers YAMAHA HS-50M. The microphone that receives WSs is SHURE SM63L and is connected to another EDIROL UA-101. The RS is stored to a PC for evaluating the accuracy for various parameter setting. Two music pieces are used for evaluation: popular music with a low dynamic range (POP) and instrumental music with a high dynamic range (INST). In both ROOM and CAFE environments, we played the WSs at the several sound volumes, *e.g.*, 0 dB, −3 dB, −6 dB, −9 dB, and −12 dB, and received them at the predetermined positions and directions for parameter estimation. The value of λ is set to the number of DS sequences used for parameter estimation, and the number of the particles are set to 1500. The other parameter values are empirically determined as in Table 1.

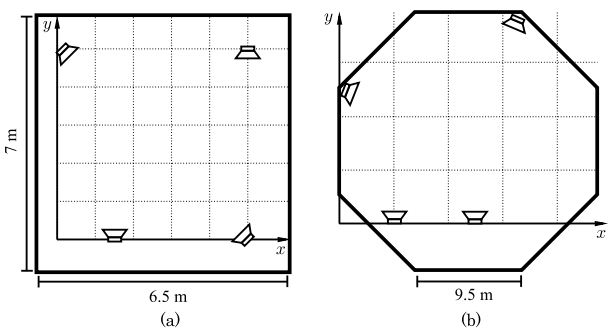

Fig. 6 The experimental environments (a) ROOM and (b) CAFE with examples of speaker arrangements. The intervals between dashed lines for (a) and (b) is 1 m and 5 m, respectively.

Table 1 Parameter values used in our experiments.

W_B	H_B	H_T	N	TH [Hz]	Δ	F_S [Hz]
15	42	6	512	6000	8	44100
V_S [m/sec]	W_U [m]	ν	σ_x^2	σ_y^2	σ_θ^2	σ_τ^2
340	0.6	0.3	0.08	0.08	0.2	0.002Λ

6.2 Static User Case

We evaluated the estimation accuracy at the randomly chosen positions and directions (R1–R9) listed in Table 2 in ROOM to show influences of sound volumes, the number of DS sequences Q for ensemble mean, and speaker arrangements. The speaker arrangements used in this evaluation are SPR1, SPR2, and SPR3 summarized in Table 4. These speaker arrangements are different from those for parameter estimation. We used excerpts of POP and INST whose durations are 10 seconds. Since the estimation results depend on initial values of state vectors of particles that are randomly chosen, we stored each RS and repeated applying positioning 100 times. The accuracy is evaluated by the root mean squared error (RMSE) of the last estimate of each repetition.

Figures 7 (a)–(d) show the results. For POP, the RMSE values for positions are approximately 1 m in SPR1, but increase according to the number of the speakers decrease as SPR2 and SPR3. For directions, differences among the RMSE values for SPR1 and SPR2 are small although SPR3 gives larger RMSE values. The effects of the sound volume and the value of Q are insignificant for both positions and directions. For INST, most of the RMSE values exceed 2 m, and they slightly increase along with decrement of the sound volume. Especially for SPR1, the RMSE values are unstable in sound volume. These results indicate that our IPS is

Table 2 Positions and directions for evaluation in ROOM.

Label	R1	R2	R3	R4	R5	R6	R7	R8	R9
x^m [m]	2	1.1	2.5	3.2	2.4	3.3	4.3	3.6	4.5
y^m [m]	1.9	2.2	2.5	1.3	4.2	3.5	3.3	4.1	4
θ^m [deg.]	205	115	270	175	105	210	0	340	90

Table 3 Positions and directions for evaluation in CAFE.

Label	C1	C2	C3	C4	C5	C6	C7	C8
x^m [m]	5.8	6.8	12.4	14.4	15.1	13.9	17.5	16.7
y^m [m]	11.6	11.2	10.9	10.3	15.4	18.1	17.2	18.2
θ^m [deg.]	45	260	165	80	225	190	80	340

Table 4 Speaker arrangements. The values of θ^{sp} are the angles with respect to the x-axis.

	ROOM			CAFE	
	SPR1	SPR2	SPR3	SPC1	SPC2
x_1^{sp} [m]	1.5	1.5	1.5	16.0	16.0
y_1^{sp} [m]	0.0	0.0	0.0	19.0	19.0
θ^{sp} [deg.]	90	90	90	247	247
x_2^{sp} [m]	5.0	5.5	5.5	7.0	0.5
y_2^{sp} [m]	0.0	2.5	3.0	19.0	12.5
θ^{sp} [deg.]	135	180	180	292	337
x_3^{sp} [m]	4.5	0.0	—	7.0	5.3
y_3^{sp} [m]	5.0	5.0	—	8.5	0.0
θ^{sp} [deg.]	270	315	—	90	90
x_4^{sp} [m]	0.0	—	—	17.0	12.5
y_4^{sp} [m]	5.0	—	—	8.5	0.0
θ^{sp} [deg.]	315	—	—	135	90

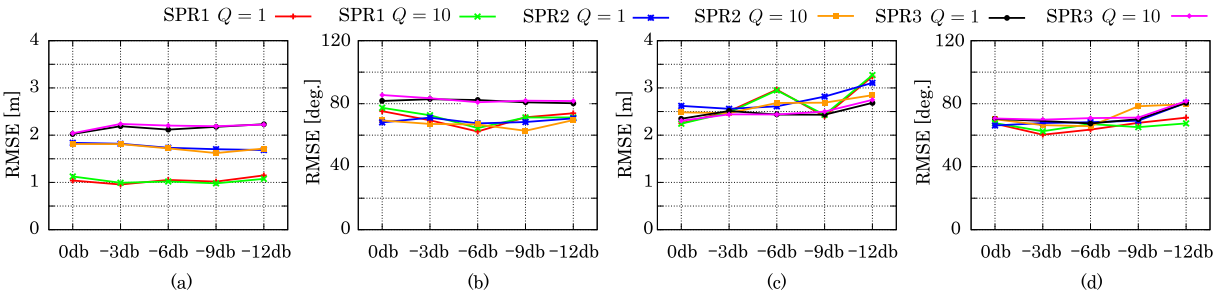

Fig. 7 RMSEs for (a) positions and (b) directions of POP; and (c) positions and (d) directions of INST with respect to various sound volumes.

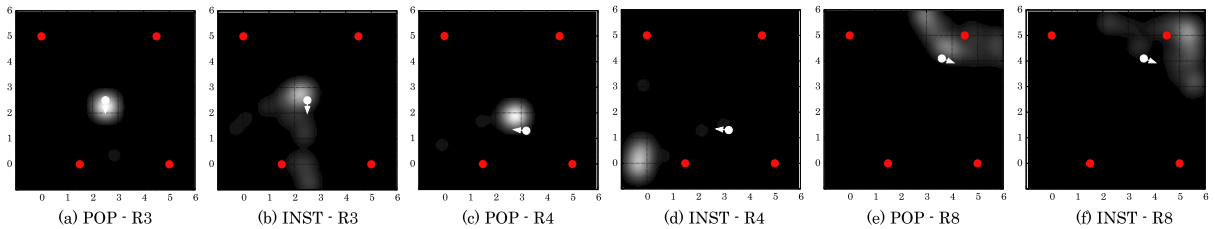

Fig. 8 Example distributions of estimates for SPR1.

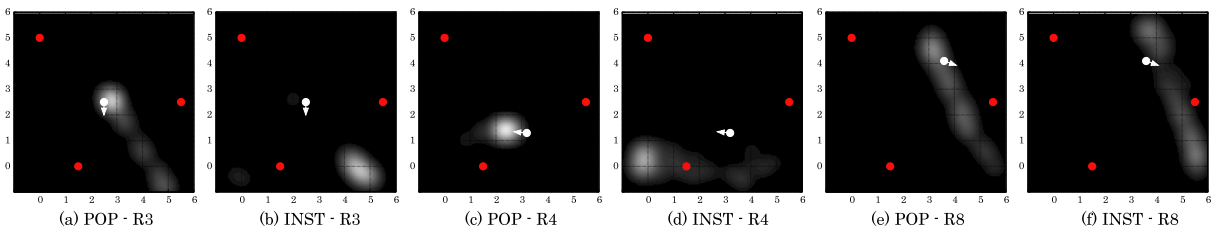

Fig. 9 Example distributions of estimates for SPR2.

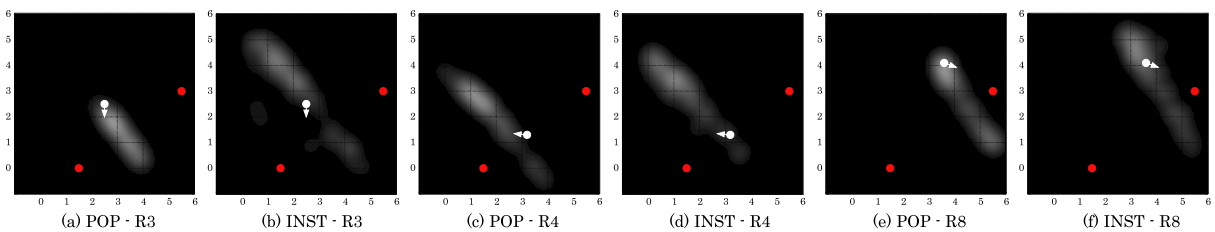

Fig. 10 Example distributions of estimates for SPR3.

suitable for low dynamic range music pieces and ensemble means are useless to improve the accuracy.

Figures 8–10 are examples of distributions of estimates for R3, R4, and R8 when $Q = 1$ and the sound volume is −3 dB. The unit of each axis is meter. Red circles represent speakers and white circles with arrows represent user positions and directions. To generate the distributions, we used kernel density estimation with the kernel $ker(\boldsymbol{x}) = e^{-\boldsymbol{x}\cdot\boldsymbol{x}/\kappa^2}$ where $\kappa = 0.25$. As can be seen in these figures, the distributions of estimates when the number of speakers is small are elongated. One of the reasons can be as follows: When the speakers are sparsely deployed, peak heights in DS sequences for WSs from speakers far from the microphone are low, and thus, the peak positions are useless for positioning. However, constraints subjected by peak heights, which are expected to be useful in such a case, are insufficient to accurately locate the user because there can be many $\boldsymbol{x}^{\mathrm{m}}$ and θ^{m} that give the same value of $h_l^c(\boldsymbol{x}^{\mathrm{m}}, \theta^{\mathrm{m}})$. Therefore, to improve accuracy, we need other constrains by, *e.g.*, incorporating directionality of speakers in the peak height model.

To compare the proposed IPS with [17], we re-implemented [17] so as to incorporate ensemble mean in [17] and applied it to our stored RSs with the parameter values listed in Table 1. The averaged RMSE values of positions and directions when $Q = 1$ and 10 are shown in Fig. 11. The averaged RMSE value of positions for [17] is

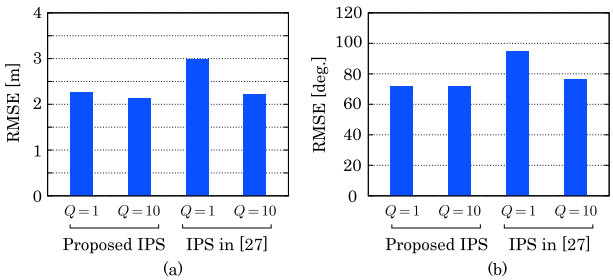

Fig. 11 Averaged RMSE values of (a) positions and (b) directions for our proposed IPS and IPS in [17]. The averaged RMSE values are calculated for Q = 1 and 10.

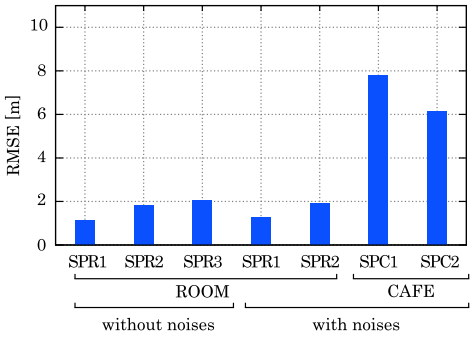

Fig. 12 RMSEs for POP in various environments (sound volume and Q are set to 0 dB and 10, respectively).

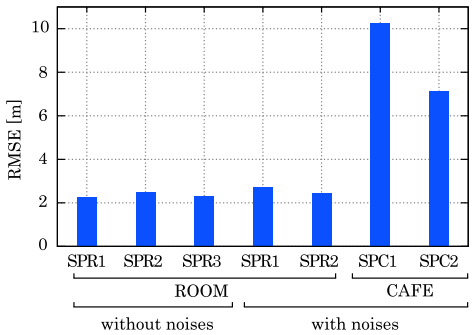

Fig. 13 RMSE for INST in various environments (sound volume and Q are set to 0 dB and 10, respectively).

larger than that in [17]. This is because that the RMSE value in [17] is calculated using all estimates obtained from a single RS without repetition. For both positions and directions, the IPS in [17] for $Q = 1$ gives the largest RMSE. Therefore, the simplified DS model proposed in this paper achieves better accuracy. The main reason is that the IPS [17] uses a log likelihood as a weight of each particle, which suppresses differences among weights of particles, resulting in estimates around the center of the target environment. Since the averaged RMSE for $Q = 10$ is smaller than that for $Q = 1$, use of larger Q alleviates the problem of log likelihood in [17].

To demonstrate the estimation accuracy in various environments, we asked two and four persons to walk around and make noises in ROOM and CAFE, respectively. For ROOM, we tested with SPR1 and SPR2. For CAFE, we employ two speaker arrangements, *i.e.*, SPC1 and SPC2, as summarized in Table 4. In SPC1, four speakers are arranged in relatively small area of CAFE while they are arranged in large area of CAFE in SPC2. Positions and directions evaluated in CAFE are summarized in Table 3. We set Q and sound volume to 10 and 0 dB, respectively.

Figures 12 and 13 show the results. The reduction of accuracy due to noises is small for both POP and INST. However, our proposed IPS gives large errors in CAFE. We consider that the causes of these errors are categorized into the following two cases: (1) For positions near a single speaker and the user faces towards the speaker such as C1, C2, and C8 in SPC1, the WS from the nearest speaker dominates the RS and WSs from the other speakers are masked, resulting in estimates around the nearest speaker as in Figs. 14 (a) and (b). (2) For positions far from any speaker such as C3 and C5 in SPC1 and most positions in SPC2, DSs of any WSs have no significant peaks. In this case, the distributions of estimates become elongated as Fig. 14 (c) or broad as Figs. 14 (d), (e), and (f). This is caused by the same reason as, *e.g.*, Fig. 9 (b). In addition to this, auto- and cross-correlations of PRSs used in watermark embedding can also cause these errors. That is, the auto- and cross-correlations form false peaks in DS sequences, resulting in local maxima of the likelihood (29). Therefore, we need to reduce auto- and cross-correlations of PRSs. Since IPSs are not for security purpose in contrast to existing applications of digital watermarking techniques that often uses PRSs as a key to detect watermarks, it is feasible to optimize PRSs with respect to auto-and cross-correlations.

6.3 Moving User Case

To demonstrate our proposed IPS for moving user case under various parameter settings in ROOM with SPR1, SPR2, and SPR3, we asked a subject who holds a microphone to stay at (4, 4) for 10 seconds and then to walk from (4, 4) to (1, 1) for subsequent 10 seconds. We stored the RSs and applied our proposed IPS for 100 times to each RS as in the previous section. From the latter 10 seconds, we extracted a sequence of estimates. For each estimate in the sequences, we calculated a sample mean and sample variance-covariance matrix over the 100 sequences. We also found the sequences that give maximal and minimal RMSE values assuming that the subjects walked in a constant speed.

Figures 15 and 16 show the results. The blue circles in each graph represent speaker positions. The sample variance-covariance matrices are represented by the ellipses in gray. For POP, we consider that the sequences of estimates well trace the actual trajectory although there are some erroneous sequences such as the sequence with maximal RMSE value in Fig. 15 (d). For INST, most of the sequences of estimates give large error. In addition, when estimates at the initial positions give large errors, the proposed IPS cannot recover them. This can be caused by the problem of sample impoverishment [29]; all particles are con-

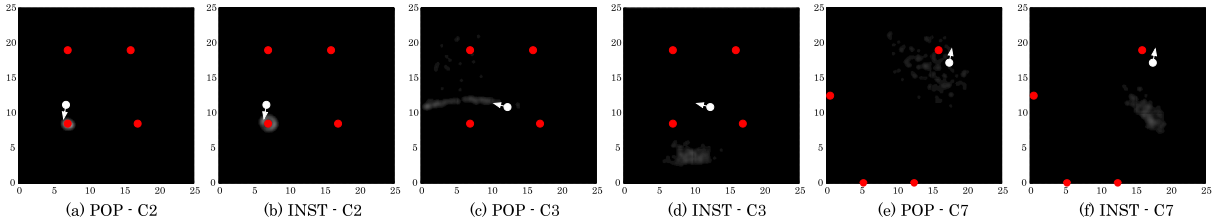

Fig. 14 Distributions of estimates in CAFE. (a)–(d) are in SPC1. (e) and (f) are in SPC2.

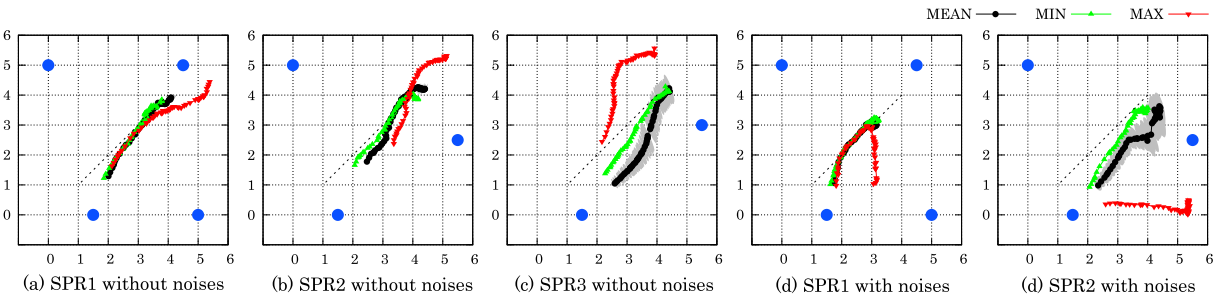

Fig. 15 Trajectories of estimates for POP where $Q = 1$. MEAN, MIN, and MAX represent sample means, the sequence with maximum RMSE, and the sequence with minimum RMSE, respectively.

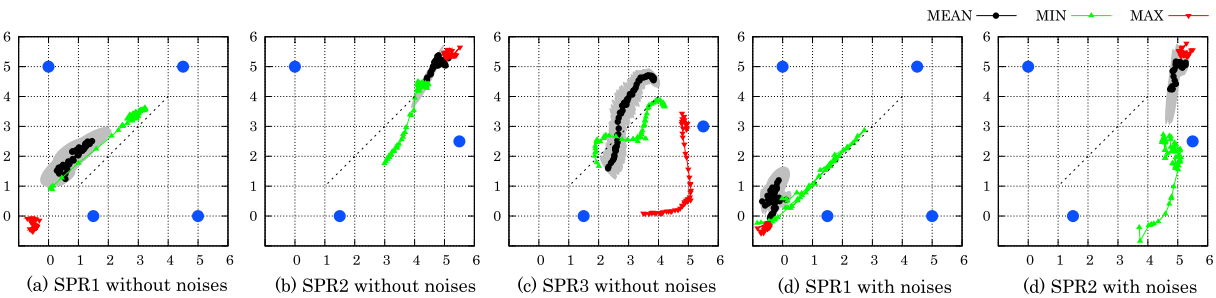

Fig. 16 Trajectories of estimates for INST where $Q = 1$. MEAN, MIN, and MAX represent sample means, the sequence with maximum RMSE, and the sequence with minimum RMSE, respectively.

centrated at erroneous positions and lose their diversity to find the actual position even after the user begin to walk. The problem of sample impoverishment also occurs in the static user case. We can overcome this problem by using a larger variance Σ_U of the noise in (30) and increasing the number of particles. However, there is a trade-off between the number of particles and the computational cost.

6.4 Verification of Real-Time Positioning

To verify that the proposed IPS works in real-time, we applied it for 10 times to a stored RS of which duration is 10 seconds and measured elapsed time on a laptop PC with Intel Core2Duo L9300 1.6 GHz CPU, 2.0 GBytes memory, and Microsoft Windows 7 (32 bit) OS. The average and standard deviation are 6.0 seconds and 0.2 seconds, respectively. Since the computational burden due to reception of audio signal is low compared to watermark detection and positioning, this result indicates that the proposed IPS works in real-time. Actually, our real-time implementation using Steinberg's ASIO API[†] works properly.

For comparison, we also measured elapsed time of the IPS [17] on the same PC. The average and standard deviation are 11.2 seconds and 0.1 seconds, respectively. Therefore, the modification on the likelihood (29) shorten the elapsed time to 53%, enabling us real-time positioning even on a relatively low spec PC.

7. Conclusion

In this paper, we proposed an indoor positioning system (IPS) using a spread spectrum-based digital audio watermarking technique, which is easily deployed by installing several off-the-shelf speakers to target environments. The experimental results indicated that our IPS is suitable for uses in small environments; it locates user positions with 2.25 m of RMSE on average. In addition, our IPS can find the speaker nearest to the user even in larger environments. Although the estimates from our IPS is unstable in initial states of particles and in high dynamics of host signals, the IPS is promising because it works in real-time. We believe

[†]Currently found at http://www.steinberg.net/en/company/developer.html

that our IPS is potentially applicable to location-based services that require accurate user positions. The future work includes to construct a model of detection strengths stable in dynamics of host signals and initial states of particles, and to introduce other constraints such as directionality of speakers.

References

[1] A. Butz, J. Baus, A. Krüger, and M. Lohse, "A hybrid indoor navigation system," Proc. 6th Intl. Conf. Intelligent User Interfaces, pp.25–32, Jan. 2001.
[2] L. Aalto, N. Göthlin, J. Korhonen, and T. Ojala, "Bluetooth and WAP push based location-aware mobile advertising system," Proc. 2nd Intl. Conf. Mobile systems, applications, and services, pp.49–58, June 2004.
[3] E. Marti, O. Vinyals, G. Friedland, and R. Bajcsy, "Precise indoor localization using smart phones," Proc. 18th ACM Intl. Conf. Multimedia, pp.1–4, Oct. 2010.
[4] J. Yim, C. Park, J. Joo, and S. Jeong, "Extended Kalman filter for wireless LAN based indoor positioning," Decision Support Systems, vol.45, no.4, pp.960–971, Nov. 2008.
[5] L. Mengual, O. Marbán, and S. Eibe, "Clustering-based location in wireless networks," Expert Systems with Applications, vol.37, no.9, pp.6165–6175, Sept. 2010.
[6] P. Bahl and V.N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," Proc. IEEE Nineteenth Annual Joint Conf. IEEE Computer and Communications Societies, pp.775–784, March 2000.
[7] L.M. Ni, Y. Liu, Y.C. Lau, and A.P. Patil, "LANDMARC: Indoor location sensing using active RFID," Wireless Networks, vol.10, no.6, pp.701–710, Nov. 2004.
[8] J. Hightower, G. Borriello, and R. Want, "SpotON: An indoor 3D location sensing technology based on RF signal strength," Tech. Rep. 2000-02-02, University of Washington CSE Technical Report, Feb. 2000.
[9] M. Minami, Y. Fukuju, K. Hirasawa, S. Yokoyama, M. Mizumachi, H. Morikawa, and T. Aoyama, "DOLPHIN: A practical approach for implementing a fully distributed indoor ultrasonic positioning system," UbiComp 2004: Ubiquitous Computing, Lecture Notes in Computer Science, vol.3205, pp.347–365, Sept. 2004.
[10] N.B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," Proc. 6th Annual Intl. Conf. Mobile Computing and Networking (MobiCom'00), pp.32–43, Aug. 2000.
[11] A. Ward, A. Jones, and A. Hopper, "A new location technique for the active office," IEEE Pers. Commun., vol.4, no.5, pp.42–47, Oct. 1997.
[12] A. Mandal, C. Lopes, T. Givargis, A. Haghighat, R. Jurdak, and P. Baldi, "Beep: 3D indoor positioning using audible sound," 2005 Second IEEE Consumer Communications and Networking Conf., pp.348–353, Jan. 2005.
[13] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," IEEE Trans. Syst. Man Cybern. C, Appl. Rev., vol.37, no.6, pp.1067–1080, Nov. 2007.
[14] V. Zeimpekis, G.M. Giaglis, and G. Lekakos, "A taxonomy of indoor and outdoor positioning techniques for mobile location services," ACM SIGecom Exchanges, vol.3, no.4, pp.19–27, Dec. 2002.
[15] Y. Nakashima, R. Tachibana, and N. Babaguchi, "Watermarked movie soundtrack finds the position of the camcorder in a theater," IEEE Trans. Multimed., vol.11, no.3, pp.443–454, April 2009.
[16] N. Lazic and P. Aarabi, "Communication over an acoustic channel using data hiding techniques," IEEE Trans. Multimed., vol.8, no.5, pp.918–924, Oct. 2006.
[17] R. Kaneto, Y. Nakashima, and N. Babaguchi, "Real-time user position estimation in indoor environments using digital watermarking for audio signals," Proc. 2010 Intl. Conf. Pattern Recognition, pp.97–100, Aug. 2010.
[18] R. Tachibana, S. Shimizu, S. Kobayashi, and T. Nakamura, "An audio watermarking method using a two-dimensional pseudo-random array," Signal Processing, vol.82, no.10, pp.1455–1469, Oct. 2002.
[19] O.K.U. Rerrer, "Suitability of positioning techniques for location-based services in wireless LANs," Proc. 2nd Workshop on Positioning, Navigation and Communication 2005, pp.51–56, 2005.
[20] G. Myles, A. Friday, and N. Davies, "Preserving privacy in environments with location-based applications," IEEE Pervasive Comput., vol.2, no.1, pp.56–64, Jan. 2003.
[21] M. Gruteser and D. Grunwald, "Enhancing location privacy in wireless LAN through disposable interface identifiers: A quantitative analysis," Mobile Networks and Applications, vol.10, no.3, pp.315–325, June 2005.
[22] C. Yeh and C. Kuo, "Digital watermarking through quasi m-arrays," Proc. 25th Annual Conf. IEEE Industrial Electronics Society, pp.459–461, Oct. 1999.
[23] P. Bassia, I. Pitas, and N. Nikolaidis, "Robust audio watermarking in the time domain," IEEE Trans. Multimed., vol.3, no.2, pp.232–241, June 2001.
[24] N. Cvejic and T. Seppänen, "Spread spectrum audio watermarking using frequency hopping and attack characterization," Signal Processing, vol.84, no.1, pp.207–213, Jan. 2004.
[25] D. Kirovski and H. Malvar, "Spread-spectrum watermarking of audio signals," IEEE Trans. Signal Process., vol.51, no.4, pp.1020–1033, April 2003.
[26] Y. Nakashima, R. Tachibana, M. Nishimura, and N. Babaguchi, "Estimation of recording location using audio watermarking," Proc. 8th Workshop on Multimedia and Security, pp.108–113, Sept. 2006.
[27] Y. Nakashima, R. Tachibana, M. Nishimura, and N. Babaguchi, "Determining recording location based on synchronization positions of audio watermarking," Proc. Intl. Conf. Acoustic, Speech, Signal Processing, pp.II253–II256, April 2007.
[28] Y. Nakashima, R. Tachibana, and N. Babaguchi, "Maximum-likelihood estimation of recording position based on audio watermarking," Proc. Intl. Conf. Intelligent Information Hiding and Multimedia Signal Processing, pp.255–258, Nov. 2007.
[29] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. Signal Process., vol.50, no.2, pp.174–188, Feb. 2002.
[30] ISO, "Information technology–coding of moving pictures and associated audio for digital storage media up to about 1.5 mbits/s—part 3: Audio," 1993.

Yuta Nakashima received the B.E. and M.E. degrees in communication engineering from Osaka University, Osaka, Japan, in 2006 and 2008, respectively. He is currently a research fellow of the Japan Society for the Promotion of Science, and pursuing the doctoral degree at Osaka University. His research interests include digital watermarking, video content analysis and processing. He is a student member of the IEEE.

Ryosuke Kaneto received the B.E. and M.E. degrees in communication engineering from Osaka University, Osaka, Japan, in 2008 and 2010, respectively. He is currently with Nisshin Seifun Group Inc.

Noboru Babaguchi received the B.E., M.E. and Ph.D. degrees in communication engineering from Osaka University, in 1979, 1981 and 1984, respectively. He is currently a Professor of the Department of Communication Engineering, Osaka University. From 1996 to 1997, he was a Visiting Scholar at the University of California, San Diego. His research interests include image analysis, multimedia computing and intelligent systems, currently content based video indexing and summarization. He has published over 100 journal and conference papers and several textbooks. Dr. Babaguchi received Best Paper Award of 2006 Pacific-Rim Conference on Multimedia (PCM2006). He is a senior member of the IEEE, and a member of the ACM, the IPSJ, the ITE and the JSA.