# Infrastructure-less Localization from Indoor Environmental Sounds Based on Spectral Decomposition and Spatial Likelihood Model

Satoki Ogiso[1], Yoshiaki Bando[1], Takeshi Kurata[1] and Takashi Okuma[1]

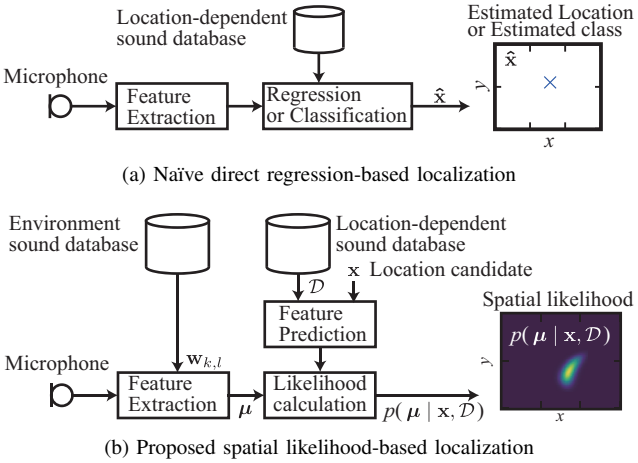
***Abstract*— Human and/or asset tracking using an attached sensor units helps understand their activities. Most common indoor localization methods for human tracking technologies require expensive infrastructures, deployment and maintenance. To overcome this problem, environmental sounds have been used for infrastructure-free localization. While they achieve room-level classification, they suffer from two problems: low signal-to-noise-ratio (SNR) condition and non-uniqueness of sound over the coverage area. A microphone localization method was proposed using supervised spectral decomposition and spatial likelihood to solve these problems. The proposed method was evaluated with actual recordings in an experimental room with a size of 12×30 m. The results showed that the proposed method with supervised NMF was robust under low-SNR condition compared to a simple feature (mel frequency cepstrum coefficient: MFCC). Additionally, the proposed method could be easily integrated with prior distribution, which is available from other Bayesian localizations. The proposed method can be used to evaluate the spatial likelihood from environmental sounds.**

## I. Introduction

Human and/or asset localization with attached sensor units is essential to understanding their activities and enhancing workspace productivity. Since indoor localization systems are installed at the expense of the beneficiaries, a number of localization methods are available according to requirements. A common method for measuring human behavior is to install devices that emit signals at known locations. This method achieves up to centimeter-grade accuracy, but the installation and maintenance of these devices are costly. Thus, the importance of the infrastructure-less localization system lies in its simple deployment.

Infrastructure-less localization has been mostly studied using existing radio waves or electromagnetic fields in the environment. Camera-based localization achieves high accuracy while consuming energy; it is sometimes prohibited in facilities for security reasons. A major approach is to use the signal strength of existing Wi-Fi base stations. This technology requires many base stations, so it is limited to large facilities with Wi-Fi, such as a train station. Using magnetic anomalies, precise localization within a few meters of area is possible. Similar anomalies may occur in every few meters, because the location-specific cue is simply a three-dimensional vector of magnetometer. Dead reckoning with IMU is completely independent of environment, whereas the absolute location needs to be provided with other methods.

This study focuses on infrastructure-less localization based on location-specific environmental sounds. Environmental sounds exist in most of the facilities and contain more information than magnetic fields. Some methods record sounds in advance to train classifiers for room or area classifiers. However, use of these sounds has two challenges: low-SNR conditions and non-uniqueness over a coverage area. Low-SNR signals are not similar to training samples, so classifiers fail. In addition, environmental sounds may not be unique at several locations, such as fans of the same type in each corridor. To solve this problem, environmental sounds are only used for coarse localization as room/area classification. It would be useful for infrastructure-free indoor localization if position with coordinate from environmental sound at low SNR could be estimated.




(a) Naïve direct regression-based localization

(b) Proposed spatial likelihood-based localization

Fig. 1: Comparison between naïve direct regression-based localization and our spatial likelihood-based localization.

In this study, a sound-based infrastructure-less localization method is proposed that can deal with low-SNR conditions and ill-conditioned measurements that cannot determine a unique location. The comparison between conventional and proposed methods is shown in Fig. 1. This study addresses the issues listed above. The proposed method first extracts source root-mean-square (RMS) values from an observed mixture signal using supervised non-negative matrix factorization (NMF) [1] and a Wiener filter. The supervised NMF and Wiener filter reduce out-of-domain noise, by decomposing the sound into components which corresponds to the premeasured environmental sounds. Then spatial likelihood is calculated by comparing them with predicted distribution from a Gaussian process (GP) regression model. GP provides a sound model in the environment as probabilistic distribution by combining with the supervised NMF. The proposed method enables

[1]National Institute of Advanced Industrial Science and Technology (AIST), Japan {s.ogiso, y.bando, t.kurata, takashi-okuma}@aist.go.jp

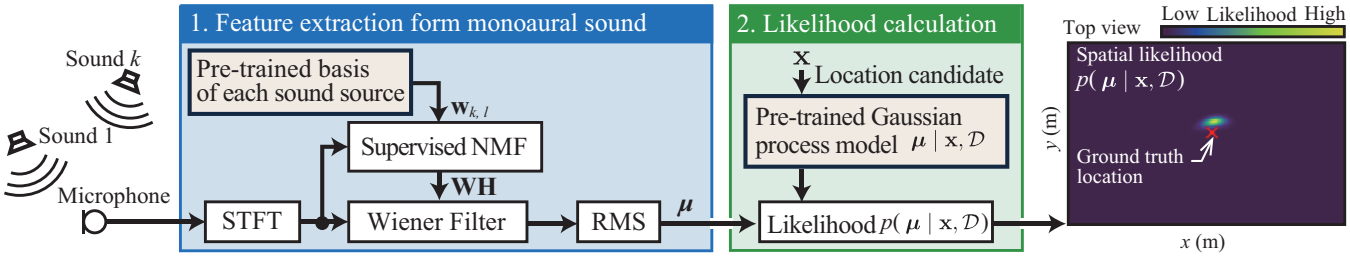



Fig. 2: Architecture of the proposed localization method.

maximum a posteriori (MAP) estimation with any Bayes-based indoor localization.

## II. Related Work

### A. Infrastructure-assisted methods

The majority of the indoor localization methods require special infrastructure in the environment. The most common method uses the received signal strength indicator (RSSI) of Bluetooth low energy (BLE) or radio-frequency identifier (RFID) [2], which achieves 2–3 m accuracy [3], [4]. The major drawback of these methods is that they require numerous BLE beacons to be deployed. For example, a case study in a five-story building deployed 218 beacons [5]. Other infrastructure-based methods such as ultra-wideband (UWB) or ultrasound [6], [7] require wiring between base stations to achieve accurate localization. Installing and removing of these beacons or base stations are costly. Moreover, maintaining the health of these beacons, such as battery life or fault, is exhausting work.

### B. Infrastructure-less methods

To address this challenge, infrastructure-less indoor localization methods have been proposed. These methods use existing environmental cues or self-contained sensors. A camera [8] or light detection and ranging (LiDAR) sensor [9] achieves the most accurate localization. However, these sensors have a high power requirement, and are sometimes not allowed for security reasons. WiFi-based methods [2] use Wi-Fi base stations that are already installed as beacons. However, several Wi-Fi stations must be simultaneously detected, which is not possible in Wi-Fi-restricted locations such as factories. Another useful environmental cue is the local magnetic anomalies caused by the ferromagnetic materials in a building [10]. These magnetic anomalies are unique within a few meters, but not throughout the entire building. This is due to the fact that a magnetic anomaly is merely a three-dimensional vector that can be almost identical in every few meters. Another example is IMU-based indoor localization, which uses the estimated gravity vector, linear velocities, and angular velocities [11]. They use specific acceleration pattern [11] or zero-velocity data [12]. This method achieves high short-term accuracy, but absolute localization is required for long term use. Since these methods have different advantages and disadvantages, the Bayesian integration of these sensors has also been actively researched [13], [14].

### C. Environmental sound-based methods

Environmental sounds are another potential location-specific cue promising for sensor localization. Environmental sound localization can also be divided into two methods: direction of arrival (DoA)-based methods and fingerprint-based methods. The DoA-based methods use a microphone array to estimate the DoA of sound sources with sub-meter accuracy [15], [16]. However, these methods require a microphone array with sufficiently large microphone spacing [17]. The fingerprint-based methods first samples the sounds in the locating area and then use classification or regression to estimate the location of the microphone [18]–[20]. The advantages of the sound fingerprint-based methods are similar to those of the infrastructure-less magnetic methods. In addition, they are more promising than the magnetic methods because the audio spectrum has far more information (usually hundreds of dimensions) than the three-dimensional magnetic information. However, the fingerprint-based methods have two main problems. The first is that they ignore individual sound sources and simply use the features of mixed sound samples [21]. Since mixed signals vary drastically according to the recording locations, they may be suceptible to the out-of-domain problem. The second problem is that they only solve classification or regression problems to obtain a single estimation value. The environmental sounds may not be unique at several locations, such as fans of the same type in each corridor. The conventional methods cannot afford to represent this multi-modal likelihood as a regression result. The proposed method estimates spatial likelihood based on environmental sound, and it is compatible with the Bayesian estimations mentioned above.

## III. Proposed Method

As shown in Fig. 2, the proposed method consists of two major steps: NMF-based decomposition step and GP-based likelihood calculation step. This two-stage framework leads to the high interpretability of the inference and enables us to easily integrate other modalities in a Bayesian manner.

### A. Feature extraction with noise-aware supervised NMF

To extract spectral features from an observed mixture signal, the mixture signal with supervised NMF was decomposed. Let $y_{ft} \in \mathbb{C}$ be the observed mixture in the time-frequency domain obtained by the short-time Fourier transform (STFT), where $f = 1, \ldots, F$ and $t = 1, \ldots, T$ are the frequency and time indices, respectively. A mixture signal $y_{ft}$ was assumed

to consist of $K$ landmark source signals $s_{k,ft} \in \mathbb{C}$ and a noise signal $n_{ft} \in \mathbb{C}$ as follows:

$$y_{ft} = \sum_{k=1}^{K} s_{k,ft} + n_{ft} \tag{1}$$

As described later, NMF decomposes the mixture $y_{ft}$ into source signals $s_{k,ft}$ whose spectral patterns are known in advance and residual noise (e.g., non-environmental sounds). The feature vector for localization $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K] \in \mathbb{R}^K$ is then calculated as log-RMS values of source estimates $\hat{s}_{k,ft}$ as follows:

$$\mu_k = \frac{1}{2} \log \left( \sum_{f=1}^{F} \sum_{t=1}^{T} |\hat{s}_{k,ft}|^2 \right). \tag{2}$$

To extract the source signals from a mixture, we utilize supervised NMF [1], [22]. NMF assumes that each source signal $s_{k,ft}$ follows a local Gaussian model (LGM) [1] whose power spectral density is represented by $L_k$ spectral basis (template) vectors $\mathbf{w}_{k,l} = [w_{k,1f}, \ldots, w_{k,L_k f}] \in \mathbb{R}_+^F$ and their temporal activations $\mathbf{h}_{k,l} = [h_{k,1t}, \ldots, h_{k,L_k t}] \in \mathbb{R}_+^T$ as follows:

$$s_{k,ft} \sim \mathcal{N}_{\mathbb{C}} \left( 0, \sum_{l=1}^{L_k} w_{k,lf} h_{k,lt} \right), \tag{3}$$

where $\mathcal{N}_{\mathbb{C}} \left( \mu, \sigma^2 \right)$ is a complex Gaussian distribution with a mean parameter $\mu$ and a variance parameter $\sigma^2$. The basis vectors of each source are obtained in advance from isolated recordings of the landmark sound sources by a maximum likelihood estimation of Eq.(3) [1].

In the localization phase, an NMF-based LGM on the noise signal $n_{ft}$ was assumed with unknown basis vectors $\mathbf{w}_{0,l} = [w_{0,1f}, \ldots, w_{0,L_0 f}] \in \mathbb{R}_+^F$ and their activations $\mathbf{h}_{0,l} = [h_{0,1t}, \ldots, h_{0,L_0 t}] \in \mathbb{R}_+^T$ as:

$$n_{ft} \sim \mathcal{N}_{\mathbb{C}} \left( 0, \sum_{l=1}^{L_0} w_{0,lf} h_{0,lt} \right). \tag{4}$$

The unknown parameters $\mathbf{w}_{0,l}$ and $\mathbf{h}_{k,l}$ $(k = 0, \ldots, K)$ were estimated to maximize a likelihood function of the mixture signal. From Eqs. (1), (3), and (4), the likelihood function is derived as follows:

$$y_{ft} \sim \mathcal{N}_{\mathbb{C}} \left( 0, \sum_{k=0}^{K} \sum_{l=1}^{L_k} w_{k,lf} h_{k,lt} \right). \tag{5}$$

The estimation of the model parameters is performed with multiplication update rules [1] as in the original NMF.

Once the model parameters are estimated, the landmark source signals can be estimated by Wiener filtering as follows:

$$\begin{aligned} \hat{s}_{k,ft} &= \mathbb{E}[s_{k,ft} \mid y_{ft}, \mathbf{W}, \mathbf{H}] \\ &= \frac{\sum_{l=1}^{L_k} w_{k,lf} h_{k,lt}}{\sum_{k=0}^{K} \sum_{l=1}^{L_k} w_{k,lf} h_{k,lt}} y_{ft}. \end{aligned} \tag{6}$$

### *B. Spatial likelihood-based localization with GP regression*

The indoor sound propagation process depends on many factors, such as the direction of sources, reflections, and reverberations. Thus, sound distribution in the target area for localization was represented by using GP regression [23]. Note that conventional methods use regression (or classification) to estimate the source location directly. In contrast, the proposed method predicts sound distribution by GP regression and calculates the spatial likelihood.

The GP regression was trained by using mixture recordings $\mathbf{Y}_n$ sampled at locations $\mathbf{x}_n \in \mathbb{R}^2$ $(n = 1, \ldots, N)$. The mixture recordings were initially converted into NMF-based feature values $\mu_{n,k}$ for source $k$ at a sampled location $\mathbf{x}_n$. The Gaussian process is then trained to predict the feature values conditioned by the sampled locations $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$:

$$[\mu_{1,k}, \ldots, \mu_{N,k}]^{\mathsf{T}} \mid \mathbf{X} \sim \mathcal{N} \left( \mathbf{0}, \mathbf{K}(\mathbf{X}) + \sigma^2 \mathbf{I} \right), \tag{7}$$

where $\sigma^2 \in \mathbb{R}_+$ is a variance hyperparameter, and $\mathbf{K}_k(\mathbf{X}) \in \mathbb{S}_+^{N \times N}$ is a covariance matrix whose $i, j$-element $\kappa_k(\mathbf{x}_i, \mathbf{x}_j)$ is defined by a kernel function. In this study, we use the scaled radial basis function (RBF) kernel [23] for $\kappa_k(\mathbf{x}_i, \mathbf{x}_j)$:

$$\kappa_k(\mathbf{x}_i, \mathbf{x}_j) = \theta_k \exp \left( -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\gamma_k^2} \right), \tag{8}$$

where $\theta_k \in \mathbb{R}_+$ and $\gamma_k \in \mathbb{R}$ are the kernel parameters. The training optimizes these parameters to maximize the log-likelihood of Eq. (7) by using gradient descent.

Once the kernel parameters are trained, a likelihood function can be determined for an arbitrary location $\mathbf{x}$ with an unseen feature vector $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_K]$ by GP regression as follows:

$$\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D} \sim \prod_{k=1}^{K} \mathcal{N} \left( \mu_k \mid \hat{m}_k(\mathbf{x}), \hat{v}_k(\mathbf{x}) \right) \tag{9}$$

where $\mathcal{D} = \{\boldsymbol{\mu}_1, \mathbf{x}_1 \ldots, \boldsymbol{\mu}_N, \mathbf{x}_N\}$ are the training data, and $\hat{m}_k(\mathbf{x}) \in \mathbb{R}$ and $\hat{v}_k(\mathbf{x}) \in \mathbb{R}_+$ are the predictive mean and variance of a feature $\mu_k$ determined from the pretrained GP model [23]. The microphone location $\hat{\boldsymbol{x}}_{\mathrm{ML}}$ can be estimated by finding the location where the likelihood function for its feature vector is maximized:

$$\hat{\mathbf{x}}_{\mathrm{ML}} = \operatorname*{argmax}_{\mathbf{x}} p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}). \tag{10}$$

A simple way to find the maximum estimate is to perform a grid search in the likelihood space. In addition, since our localization is based on a spatial probabilistic (likelihood) model, it is easy to combine with other modalities by introducing a prior distribution. If a prior distribution $p(\mathbf{x})$ is available (e.g., inertial sensor estimate), we can calculate the posterior distribution $p(\mathbf{x} \mid \boldsymbol{\mu}, \mathcal{D}) \propto p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) p(\mathbf{x})$, and the location can be estimated using maximum-a-posteriori estimation as follows:

$$\hat{\mathbf{x}}_{\mathrm{MAP}} = \operatorname*{argmax}_{\mathbf{x}} p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) p(\mathbf{x}). \tag{11}$$

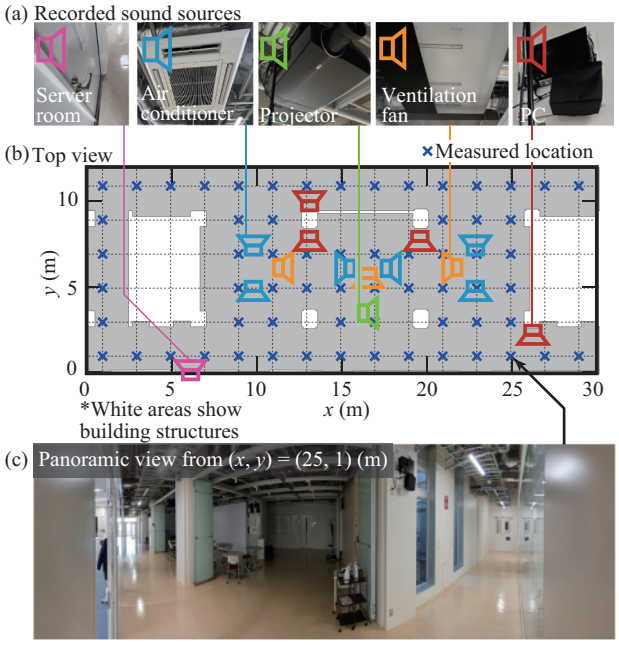



Fig. 3: Target area for localization. (a) lists landmark sound sources used in our experiment, (b) shows the top view of the area, and (c) shows a panoramic view of the area.

## IV. Experimental Evaluation

### A. Experimental setting

The proposed method was evaluated in the indoor environment, as shown in Fig. 3. The shape of the environment was approximately $(x, y) = (30, 12)$ (m), with pillars and mechanical voids. This environment has $K = 5$ kinds of significant sound sources. The approximate locations and representative pictures of these sounds are shown in Fig. 3(a). The sound signals were recorded with a linear PCM recorder (PCM-D100, Sony) and a microphone (M50, EarthWorks) at a sampling frequency of 48 kHz. To train the GP model and evaluate the proposed method, mixture recordings were sampled on a grid with 2 m spacing and a height of $z = 1$ m, as shown in Fig. 3(b).Each mixture recording lasted more than 30 sec. and was windowed with 1 sec to make 30 samples for each location. The evaluation was performed by leave-one-out cross-validation with the following procedure. A mixture of samples at one location was chosen for evaluation and used the remaining samples for training.

The hyperparameters of NMF and GP were heuristically determined as follows. The numbers of the basis vector for NMF were set to $L_k = 5$ $(k = 1, \ldots, K)$ and $L_0 = 4$. The multiplication updates of NMF were iterated 100 times for both the training and localization. The model parameters of GP regression were optimized by an Adam optimizer with a learning rate of 0.1 for 100 iterations. The kernel parameter $\gamma_k$ was constrained to be greater than 3 m to represent the smooth spatial distribution of sound energy. The localization was performed by a grid search on the likelihood function or posterior distribution. The grid size was 0.1 m for $x$ and $y$ in the room, whereas a height of $z$ = 1 m was given; this is because a microphone attached to a human is usually at a consistent height.

The proposed method was evaluated under different signal-to-noise ratios (SNR). A sound from the MIMII dataset [24] was added as out-of-domain background noise. SNR was evaluated from -60 to +18 dB at every 3 dB.

The advantage of the proposed method is a straightforward integration with other Bayesian localization methods. To show the integration example, an integration with IMU-based localization was simulated. IMU-based localization suffers from drift over time, while it works without any environmental setup. The following probability distribution with location drift was chosen as an IMU-like prior:

$$p(\mathbf{x}) = \mathcal{N}\left((x, y) \mid (x_{\mathrm{gt}}, y_{\mathrm{gt}}) + (\epsilon_x, \epsilon_y), 5^2\mathbf{I}\right) \tag{12}$$

where $\epsilon_x, \epsilon_y = (5, 5)$ (m) are the drift components to simulate the drifted prior of IMU. The expected localization error of $p(\mathbf{x})$ is approximately 7 m.

The proposed method was compared with naïve implementation of feature extraction and regression. A feature extraction baseline with the Mel-frequency cepstrum coefficient (MFCC) with 20 components and a GP regression baseline were chosen to directly predict $x$ and $y$ coordinates of the microphone. Localization results were evaluated with circular error (CE) and its empirical cumulative distribution function (eCDF). A CE is a Euclidean distance between the true location and the estimated location in two dimensions. These metrics are defined in ISO/IEC 18305:2016 and have been used for various indoor localization evaluations [25]. The 50 percentiles (circular error probable: CEP), mean, and 95 percentiles (CE95) of CEs were used as representative values.

### B. Sound source separation and Gaussian regression

Fig. 4 shows examples of the predicted RMS (energy) of Gaussian process models. The predicted RMS values have high values around the sound sources. Slight directivity is also shown in Fig. 4(a). The panel of the projector emitted environmental sound, which faced the positive $x$ direction. The slight directivity is also observed in Fig. 4(c). They may have vertical directivity since the RMS distribution is relatively narrow. However, sounds from the PCs and ventilation fans in Figs. 4(b) and (e) diffract around the mechanical void at left and right. One advantage of using supervised NMF is that it is easier to understand the features corresponding to the real world than naïve spectral features. From these results, we can confirm that these GP models provide reasonable RMS predictions.

### C. Localization performance

TABLE I shows localization performance for all the combinations of feature extraction methods and regression methods. The "SNMF" feature was the average temporal activation for each sound source $h_{k,lt}$ estimated by the supervised NMF. The "SNMF-WF" feature (proposed) was the RMS of the extracted sound source (Eq. (2)). We first compare the performance differences of the direct regression and likelihood-based localization (top three rows vs. middle

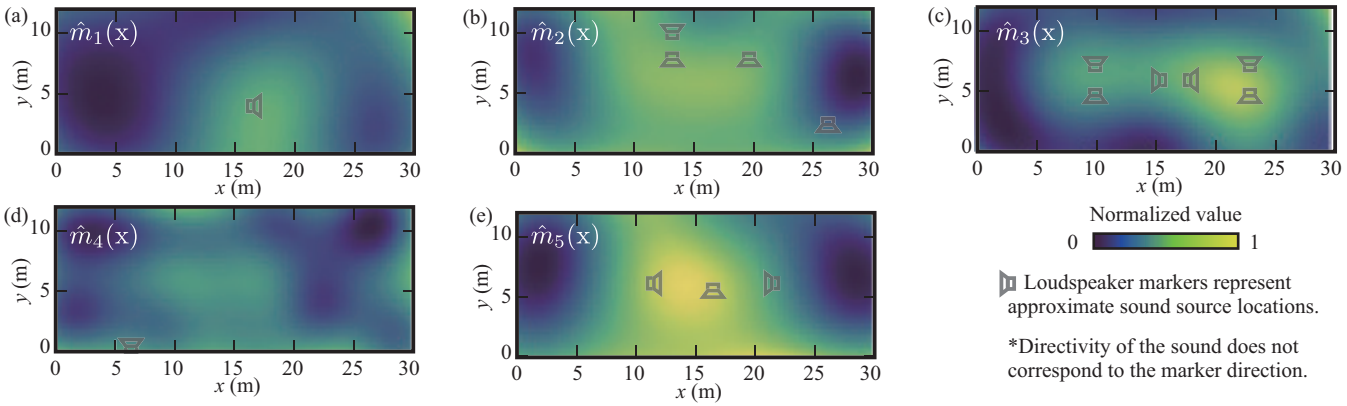



Fig. 4: Examples of GP regression models with separated source RMS for each location. Figures correspond to the sound sources described in Fig. 3: (a) a projector, (b) PCs, (c) air conditioners, (d) server room, and (e) ventilation fans, respectively.

TABLE I: Localization errors for each combination of the methods.

| Method | | Localization error (m) | | |
|---|---|---|---|---|
| Feature | Localization | CEP | Mean | CE95 |
| MFCC | Regression | 6.8 | 7.7 | 20.6 |
| SNMF | Regression | 4.5 | 5.4 | 12.8 |
| SNMF + WF | Regression | 5.3 | 6.3 | 14.0 |
| MFCC | Likelihood | 3.3 | 5.7 | 16.6 |
| SNMF | Likelihood | 3.9 | 6.1 | 16.9 |
| SNMF + WF | Likelihood | 3.4 | 6.1 | 18.6 |
| MFCC | Likelihood + Prior | 3.0 | 5.1 | 15.2 |
| SNMF | Likelihood + Prior | 3.2 | 4.9 | 13.5 |
| SNMF + WF | Likelihood + Prior | 2.7 | 3.3 | 9.2 |

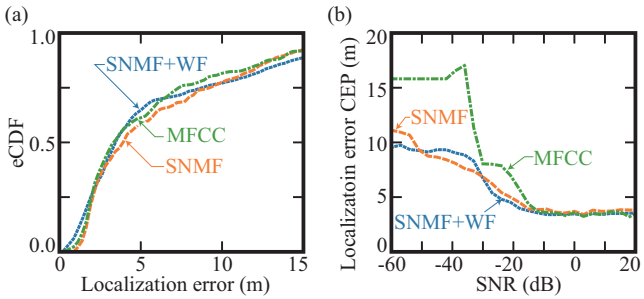



Fig. 5: Localization performance. (a) Localization with proposed spatial likelihood in the experiment, (b) effect of SNR on localization error (CEP).

three rows). Compared with the direct regression, the proposed spatial likelihood significantly reduced the localization errors in CEP. The mean error and CE95, however, had no significant difference. Since the proposed method first predicts the feature value at a location candidate and then compares it with the measurement, this process may create high peaks in the likelihood in completely different locations, which seem to have similar feature values (source energies). Since environmental sounds may not be unique at several locations, the likelihood inherently becomes multi-modal distribution, and the microphone could not be uniquely localized from only audio data. Fig. 5(a) show the localization performance using the proposed spatial-likelihood. These results show that the proposed likelihood estimation is a reasonable extension of regression, with almost the same localization error.

The effect of SNR on localization error is shown in Fig. 5(b). There is a certain threshold according to SNR, beyond which the accuracy deteriorates rapidly. The localization error of MFCC has a sharp change in accuracy starting at about -15 dB and saturating at around -24 dB. The relevant fingerprint sounds were mistaken for fingerprints from another location. The localization error of SNMF changes from a similar SNR, but does not have as drastic an effect on positional accuracy as MFCC. This might be due to the fact that MFCC directly uses mixture signals, which drastically change according to SNR, whereas SNMF and SNMF-WF first decompose the input mixture into individual source signals, which might not change significantly. The localization error of the SNMF-based method saturates at around -33 dB, which is a 9 dB improvement from MFCC.

### *D. Integration with other localization method*

Finally, the effectiveness of the proposed likelihood-based localization is demonstrated by comparing it with the naïve direct regression. Simulating an IMU-based localization with large errors in its estimates, the posterior distribution was derived from our audio likelihood and the IMU-based prior to integrate these two modalities. Fig. 6 shows an example of the integration results. As mentioned in section IV-C, the spatial likelihood can have multiple peaks at different locations. Fig. 6(a) shows such a case, where the likelihood has three sharp peaks. Fig. 6(b) shows a drifted prior distribution. The combination of these distributions narrows the choice of location, as shown in Fig. 6(c). In this case, the localization error was reduced even with the biased prior, whose expected error is approximately 7 m, as shown in the bottom three rows of TABLE I. Note that prior distribution does not always improve the accuracy, such as when the prior picks up a different peak of likelihood. From the above example, we can confirm the straightforward Bayesian integration of the

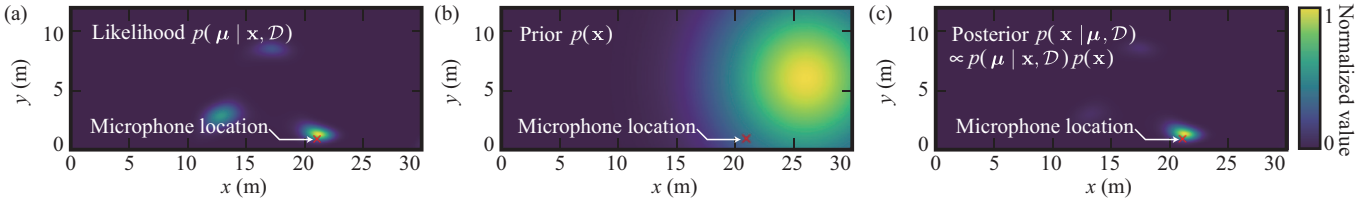



Fig. 6: Example of integration with prior distribution. (a) spatial likelihood $p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D})$, (b) IMU-based prior distribution $p(\mathbf{x}) = \mathcal{N}\left((x, y) \mid (26, 6), 5^2 I\right)$, and (c) posterior distribution $p(\mathbf{x} \mid \boldsymbol{\mu}, \mathcal{D}) \propto p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) p(\mathbf{x})$ calculated from (a) and (b).

proposed method enables collaboration with existing methods. Time-series tracking by particle filtering is also possible, which is in future work.

## V. Conclusions

In this study, a method for infrastructure-free indoor microphone localization from indoor environmental sounds was proposed. The proposed method first extracts landmark sounds from an observed a mixture signal using supervised NMF and Wiener filtering. Then, the microphone is localized by maximizing the spatial likelihood from the GP regression. The proposed method was evaluated in an approximately 12 m $\times$ 30 m room with five types of environmental sounds. The results showed that NMF-based feature extraction improves the localization performance, specifically against noise from the naïve MFCC feature. The likelihood-based localization was demonstrated to be easily integrated with a prior distribution. This characteristic enables the further extension of the proposed method with sensor fusion.

Future work should involve time-series estimation and fusion with a magnetic or IMU-based method. Since the proposed method uses the RMS of each sound, sound propagation simulation can generate the fingerprint.

## Acknowledgments

This work was supported in part by JSPS KAKENHI under Grant JP22K17922 and JST ACT-X No. JPMJAX200N.

## References

[1] C. Févotte *et al.*, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural computation*, vol. 21, pp. 793–830, 2008.

[2] P. Bahl *et al.*, "RADAR: an in-building RF-based user location and tracking system," in *Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, 2000, pp. 775–784.

[3] M. Murata *et al.*, "Smartphone-based Indoor Localization for Blind Navigation across Building Complexes," in *Proceedings of the 2018 IEEE International Conference on Pervasive Computing and Communications*, 2018, pp. 1–10.

[4] K. Urano *et al.*, "An End-to-End BLE Indoor Localization Method Using LSTM," *Journal of Information Processing*, vol. 29, pp. 58–69, 2021.

[5] M. Murata *et al.*, "Smartphone-based localization for blind navigation in building-scale indoor environments," *Pervasive and Mobile Computing*, vol. 57, pp. 14–32, 2019.

[6] J. Urena *et al.*, "Acoustic Local Positioning With Encoded Emission Beacons," *Proceedings of the IEEE*, vol. 106, no. 6, pp. 1042–1062, 2018.

[7] S. Ogiso, "Robust acoustic localization in a reverberant environment for synchronous and asynchronous beacons," in *Proceedings of the 2021 International Conference on Indoor Positioning and Indoor Navigation*, 2021, pp. 1–29.

[8] C. Campos *et al.*, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[9] M. Labbé *et al.*, "RTAB-Map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019.

[10] H. Xie *et al.*, "MaLoc: a practical magnetic fingerprinting approach to indoor localization using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014, pp. 243–253.

[11] M. Kourogi *et al.*, "Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera," in *Proceedings of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, 2003, pp. 103–112.

[12] E. Foxlin, "Pedestrian tracking with shoe-mounted inertial sensors," *IEEE Computer Graphics and Applications*, vol. 25, no. 6, pp. 38–46, 2005.

[13] T. Yang *et al.*, "A Survey of Recent Indoor Localization Scenarios and Methodologies," *Sensors*, vol. 21, no. 23, p. 8086, 2021.

[14] H.-S. Kim *et al.*, "Indoor Positioning System Using Magnetic Field Map Navigation and an Encoder System," *Sensors*, vol. 17, no. 3, pp. 1–16, 2017.

[15] C. Evers *et al.*, "Acoustic SLAM," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 26, no. 9, pp. 1484–1498, 2018.

[16] N. R. Rypkema *et al.*, "Passive Inverted Ultra-Short Baseline (piUSBL) Localization: An Experimental Evaluation of Accuracy," in *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 7197–7204.

[17] S. Ogiso *et al.*, "Robust Indoor Localization in a Reverberant Environment Using Microphone Pairs and Asynchronous Acoustic Beacons," *IEEE Access*, vol. 7, pp. 123 116–123 127, 2019.

[18] M. Azizyan *et al.*, "SurroundSense: mobile phone localization using ambient sound and light," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 13, no. 1, pp. 69–72, 2009.

[19] S. P. Tarzia *et al.*, "Indoor localization without infrastructure using the acoustic background spectrum," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, 2011, pp. 155–168.

[20] K. Aono *et al.*, "Infrasonic scene fingerprinting for authenticating speaker location," in *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 361–365.

[21] R. Leonardo *et al.*, "A Framework for Infrastructure-Free Indoor Localization Based on Pervasive Sound Analysis," *IEEE Sensors Journal*, vol. 18, no. 10, pp. 4136–4144, 2018.

[22] V. Bisot *et al.*, "Supervised nonnegative matrix factorization for acoustic scene classification," in *DCASE*, 2016, pp. 1–5.

[23] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer school on machine learning*. Springer, 2003, pp. 63–71.

[24] H. Purohit *et al.*, "MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection," in *4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019, pp. 209–213.

[25] R. Ichikari *et al.*, "Off-Site Indoor Localization Competitions Based on Measured Data in a Warehouse," *Sensors*, vol. 19, no. 4, pp. 1–29, 2019.