Spatial Concept Acquisition for a Mobile Robot that Integrates Self-Localization and Unsupervised Word Discovery from Spoken Sentences

Akira Taniguchi, Tadahiro Taniguchi, *Member, IEEE,* and Tetsunari Inamura, *Member, IEEE,*

***Abstract*—In this paper, we propose a novel unsupervised learning method for the lexical acquisition of words related to places visited by robots, from human continuous speech signals. We address the problem of learning novel words by a robot that has no prior knowledge of these words except for a primitive acoustic model. Further, we propose a method that allows a robot to effectively use the learned words and their meanings for self-localization tasks. The proposed method is nonparametric Bayesian spatial concept acquisition method (SpCoA) that integrates the generative model for self-localization and the unsupervised word segmentation in uttered sentences via latent variables related to the spatial concept. We implemented the proposed method SpCoA on SIGVerse, which is a simulation environment, and TurtleBot 2, which is a mobile robot in a real environment. Further, we conducted experiments for evaluating the performance of SpCoA. The experimental results showed that SpCoA enabled the robot to acquire the names of places from speech sentences. They also revealed that the robot could effectively utilize the acquired spatial concepts and reduce the uncertainty in self-localization.**

***Index Terms*—Learning place names, lexical acquisition, self-localization, spatial concept**

I. Introduction

AUTONOMOUS robots, such as service robots, operating in the human living environment with humans have to be able to perform various tasks and language communication. To this end, robots are required to acquire novel concepts and vocabulary on the basis of the information obtained from their sensors, e.g., laser sensors, microphones, and cameras, and recognize a variety of objects, places, and situations in an ambient environment. Above all, we consider it important for the robot to learn the names that humans associate with places in the environment and the spatial areas corresponding to these names; i.e., the robot has to be able to understand words related to places. Therefore, it is important to deal with considerable uncertainty, such as the robot's movement errors, sensor noise, and speech recognition errors.

Several studies on language acquisition by robots have assumed that robots have no prior lexical knowledge. These studies differ from speech recognition studies based on a large vocabulary and natural language processing studies based on lexical, syntactic, and semantic knowledge [1], [2]. Studies on language acquisition by robots also constitute a constructive approach to the human developmental process and the emergence of symbols.

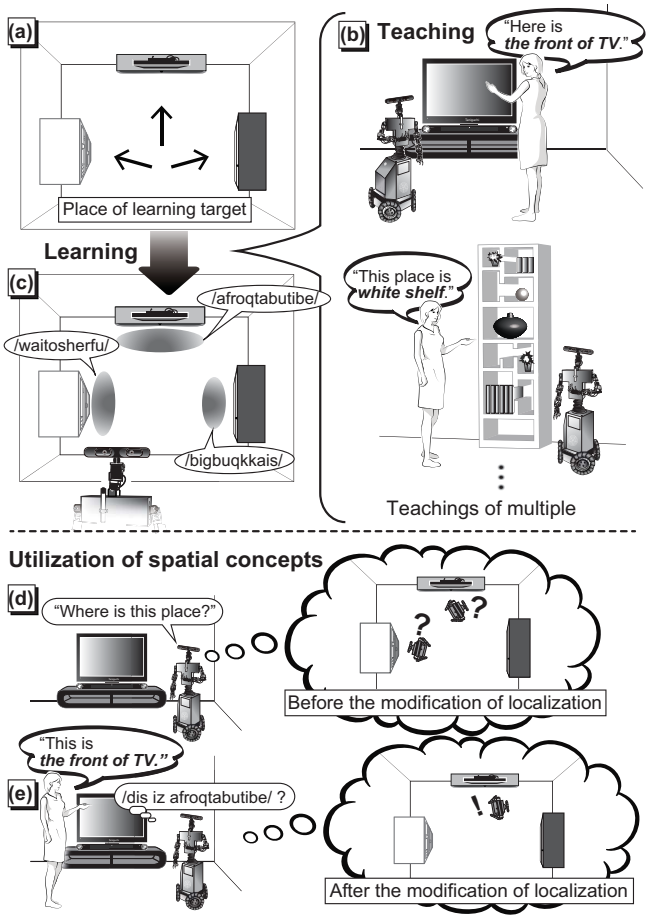


Fig. 1. Schematic representation of the target task: (a) Learning targets are three places near the objects. (b) When an utterer and a robot came in front of the TV, the utterer spoke *"Here is the front of TV."* The same holds true for the other places. (c) The robot performs word discovery from the uttered sentences and learns the related spatial concepts. Words are related to each place. (d) The robot is in front of TV actually. However, the hypothesis of the self-position of the robot is uncertain. Then, the robot asks a neighbor what the current place is. (e) By utilizing spatial concepts and an uttered sentence, the robot can narrow down the hypothesis of self-position.

The objectives of this study were to build a robot that learns words related to places and efficiently utilizes this learned vocabulary in self-localization. Lexical acquisition related to places is expected to enable a robot to improve its spatial

Akira Taniguchi and Tadahiro Taniguchi are with Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan (e-mail:a.taniguchi@em.ci.ritsumei.ac.jp; taniguchi@em.ci.ritsumei.ac.jp).

Tetsunari Inamura is with National Institute of Informatics/The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan (e-mail:inamura@nii.ac.jp).

cognition. A schematic representation depicting the target task of this study is shown in Fig. 1. This study assumes that a robot does not have any vocabularies in advance but can recognize syllables or phonemes. The robot then performs self-localization while moving around in the environment, as shown in Fig. 1 (a). An utterer speaks a sentence including the name of the place to the robot, as shown in Fig. 1 (b). For the purposes of this study, we need to consider the problems of self-localization and lexical acquisition simultaneously.

When a robot learns novel words from utterances, it is difficult to determine segmentation boundaries and the identity of different phoneme sequences from the speech recognition results, which can lead to errors. First, let us consider the case of the lexical acquisition of an isolated word. For example, if a robot obtains the speech recognition results *"aporu"*, *"epou"*, and *"aqpuru"* (incorrect phoneme recognition of *apple*), it is difficult for the robot to determine whether they denote the same referent without prior knowledge. Second, let us consider a case of the lexical acquisition of the utterance of a sentence. For example, a robot obtains a speech recognition result, such as *"thisizanaporu."* The robot has to necessarily segment a sentence into individual words, e.g., *"this"*, *"iz"*, *"an"*, and *"aporu"*. In addition, it is necessary for the robot to recognize words referring to the same referent, e.g., the fruit *apple*, from among the many segmented results that contain errors. In case of Fig. 1 (c), there is some possibility of learning names including phoneme errors, e.g., *"afroqtabutibe,"* because the robot does not have any lexical knowledge.

On the other hand, when a robot performs online probabilistic self-localization, we assume that the robot uses sensor data and control data, e.g., values obtained using a range sensor and odometry. If the position of the robot on the global map is unclear, the difficulties associated with the identification of the self-position by only using local sensor information become problematic. In the case of global localization using local information, e.g., a range sensor, the problem that the hypothesis of self-position is present in multiple remote locations, frequently occurs, as shown in Fig. 1 (d).

In order to solve the abovementioned problems, in this study, we adopted the following approach. An utterance is recognized as not a single phoneme sequence but a set of candidates of multiple phonemes. We attempt to suppress the variability in the speech recognition results by performing word discovery taking into account the multiple candidates of speech recognition. In addition, the names of places are learned by associating with words and positions. The lexical acquisition is complemented by using certain particular spatial information; i.e., this information is obtained by hearing utterances including the same word in the same place many times. Furthermore, in this study, we attempt to address the problem of the uncertainty of self-localization by improving the self-position errors by using a recognized utterance including the name of the current place and the acquired spatial concepts, as shown in Fig. 1 (e).

In this paper, we propose *nonparametric Bayesian spatial concept acquisition method* (SpCoA) on basis of unsupervised word segmentation and a nonparametric Bayesian generative model that integrates self-localization and a clustering in both words and places. The main contributions of this paper are as follows:

- We have proposed a learning method for spatial concepts that can perform the lexical acquisition related to places, i.e., the names of places, from a continuous speech signal in an unsupervised manner.
- We have achieved relatively accurate lexical acquisition that reduced the variability and errors in phonemes by performing word discovery using the multiple candidates of the speech recognition results, i.e., by using a lattice format.
- In addition to the general self-localization method of mobile robots, we showed that self-localization by the proposed method can reduce the uncertainty of self-position by utilizing the learned spatial concepts and an uttered sentence about the current position.

The remainder of this paper is organized as follows: In Section II, previous studies on language acquisition and lexical acquisition relevant to our study are described. In Section III, the proposed method SpCoA is presented. In Sections IV and V, we discuss the effectiveness of SpCoA in the simulation and in the real environment. Section VI concludes this paper.

II. Related Works

A. Lexical acquisition

Most studies on lexical acquisition typically focus on lexicons about objects [1], [3]–[11]. Many of these studies have not be able to address the lexical acquisition of words other than those related to objects, e.g., words about places.

Roy et al. proposed a computational model that enables a robot to learn the names of objects from an object image and spontaneous infant-directed speech [1]. Their results showed that the model performed speech segmentation, word discovery, and visual categorization. Iwahashi et al. reported that a robot properly understands the situation and acquires the relationship of object behaviors and sentences [3]–[5]. Qu & Chai focused on the conjunction between speech and eye gaze and the use of domain knowledge in lexical acquisition [7], [8]. They proposed an unsupervised learning method that automatically acquires novel words for an interactive system. Qu & Chai's method based on the IBM translation model [12] estimates the word-entity association probability.

Nakamura et al. proposed a method to learn object concepts and word meanings from multimodal information and verbal information [10]. The method proposed in [10] is a categorization method based on multimodal latent Dirichlet allocation (MLDA) that enables the acquisition of object concepts from multimodal information, such as visual, auditory, and haptic information [13]. Araki et al. addressed the development of a method combining unsupervised word segmentation from uttered sentences by a nested Pitman-Yor language model (NPYLM) [14] and the learning of object concepts by MLDA [11]. However, the disadvantage of using NPYLM was that phoneme sequences with errors did not result in appropriate word segmentation.

These studies did not address the lexical acquisition of the space and place that can also tolerate the uncertainty of

phoneme recognition. However, for the introduction of robots into the human living environment, robots need to acquire a lexicon related to not only objects but also places. Our study focuses on the lexical acquisition related to places. Robots can adaptively learn the names of places in various human living environments by using SpCoA. We consider that the acquired names of places can be useful for various tasks, e.g., tasks with a movement of robots by the speech instruction.

B. Simultaneous learning of places and vocabulary

The following studies have addressed lexical acquisition related to places. However, these studies could not utilize the learned language knowledge in other estimations such as the self-localization of a robot.

Taguchi et al. proposed a method for the unsupervised learning of phoneme sequences and relationships between words and objects from various user utterances without any prior linguistic knowledge other than an acoustic model of phonemes [2], [15]. Further, they proposed a method for the simultaneous categorization of self-position coordinates and lexical learning [16]. These experimental results showed that it was possible to learn the name of a place from utterances in some cases and to output words corresponding to places in a location that was not used for learning.

Milford et al. proposed RatSLAM inspired by the biological knowledge of a pose cell of the hippocampus of rodents [17]. Milford et al. proposed a method that enables a robot to acquire spatial concepts by using RatSLAM [18]. Further, Lingodroids, mobile robots that learn a language through robot-to-robot communication, have been studied [19]–[21]. Here, a robot communicated the name of a place to other robots at various locations. Experimental results showed that two robots acquired the lexicon of places that they had in common. In [21], the researchers showed that it was possible to learn temporal concepts in a manner analogous to the acquisition of spatial concepts. These studies reported that the robots created their own vocabulary. However, these studies did not consider the acquisition of a lexicon by human-to-robot speech interactions.

Welke et al. proposed a method that acquires spatial representation by the integration of the representation of the continuous state space on the sensorimotor level and the discrete symbolic entities used in high-level reasoning [22]. This method estimates the probable spatial domain and word from the given objects by using the spatial lexical knowledge extracted from Google Corpus and the position information of the object. Their study is different from ours because their study did not consider lexicon learning from human speech.

In the case of global localization, the hypothesis of self-position often remains in multiple remote places. In this case, there is some possibility of performing an incorrect estimation and increasing the estimation error. This problem exists during teaching tasks and self-localization after the lexical acquisition. The abovementioned studies could not deal with this problem. In this paper, we have proposed a method that enables a robot to perform more accurate self-localization by reducing the estimation error of the teaching time by using a smoothing method in the teaching task and by utilizing words acquired through the lexical acquisition. The strengths of this study are that learning of spatial concept and self-localization represented as one generative model and robots are able to utilize acquired lexicon to self-localization autonomously.

III. Spatial Concept Acquisition

We propose *nonparametric Bayesian spatial concept acquisition method* (SpCoA) that integrates a nonparametric morphological analyzer for the lattice [23], i.e., latticelm[1], a spatial clustering method, and Monte Carlo localization (MCL) [24].

A. Generative model

In our study, we define a *position* as a specific coordinate or a local point in the environment, and the *position distribution* as the spatial area of the environment. Further, we define a *spatial concept* as the names of places and the position distributions corresponding to these names.

The model that was developed for spatial concept acquisition is a probabilistic generative model that integrates a self-localization with the simultaneous clustering of places and words. Fig. 2 shows the graphical model for spatial concept acquisition. Table I shows each variable of the graphical model. The number of words in a sentence at time t is denoted as B_t. The generative model of the proposed method is defined as equation (1-10).

$$\begin{aligned}
\pi &\sim \mathrm{GEM}(\gamma) && (1)\\
C_t &\sim \mathrm{Mult}(\pi) && (2)\\
W &\sim \mathrm{Dir}(\beta_0) && (3)\\
O_{t,b} &\sim \mathrm{Mult}(W_{C_t}) && (4)\\
\phi_l &\sim \mathrm{GEM}(\alpha) && (5)\\
i_t &\sim p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t) && (6)\\
\Sigma &\sim \mathcal{IW}(\Sigma \mid V_0, \nu_0) && (7)\\
\mu &\sim \mathcal{N}(\mu \mid m_0, (\Sigma/\kappa_0)) && (8)\\
x_t &\sim p(x_t \mid x_{t-1}, u_t) && (9)\\
z_t &\sim p(z_t \mid x_t) && (10)
\end{aligned}$$

Then, the probability distribution for equation (6) can be defined as follows:

$$\begin{aligned}
&p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t)\\
&\quad = \frac{\mathcal{N}(x_t \mid \mu_{i_t}, \Sigma_{i_t})\mathrm{Mult}(i_t \mid \phi_{C_t})}{\sum_{i_t=j}\mathcal{N}(x_t \mid \mu_j, \Sigma_j)\mathrm{Mult}(j \mid \phi_{C_t})}. \qquad (11)
\end{aligned}$$

The prior distribution configured by using the stick breaking process (SBP) [25] is denoted as $\mathrm{GEM}(\cdot)$, the multinomial distribution as $\mathrm{Mult}(\cdot)$, the Dirichlet distribution as $\mathrm{Dir}(\cdot)$, the inverse–Wishart distribution as $\mathcal{IW}(\cdot)$, and the multivariate Gaussian (normal) distribution as $\mathcal{N}(\cdot)$. The motion model and the sensor model of self-localization are denoted as $p(x_t \mid x_{t-1}, u_t)$ and $p(z_t \mid x_t)$ in equations (9) and (10), respectively.

[1]latticelm is the name of the tool that [23] is implemented and is treated as the name of the method in this study. http://www.phontron.com/latticelm/

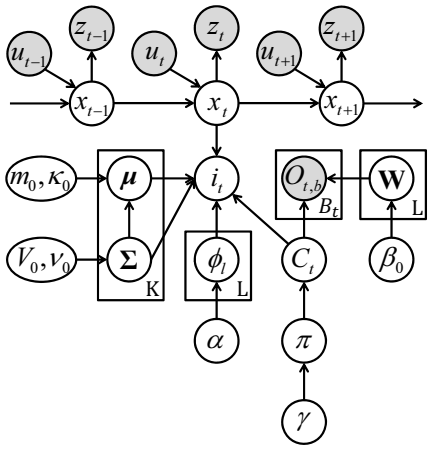


Fig. 2. Graphical model of the proposed method SpCoA

TABLE I
EACH ELEMENT OF THE GRAPHICAL MODEL

x_t	Self-position of a robot
u_t	Control data
z_t	Sensor data
C_t	Index of spatial concepts
$O_{t,b}$	Segmented word in a uttered sentence
$\mathbf{W}$	Multinomial distribution as the word probability of the names of places
$\boldsymbol{\mu}, \boldsymbol{\Sigma}$	Gaussian distribution as a position distribution (mean vector, covariance matrix)
i_t	Index of a position distribution
ϕ_l	Multinomial distribution of index i_t of Gaussian distribution
π	Multinomial distribution of index C_t of spatial concepts
α	Hyperparameter of multinomial distributions ϕ_l
γ	Hyperparameter of multinomial distribution π
β_0	Hyperparameter of Dirichlet prior distribution
$m_0, \kappa_0,$ V_0, ν_0	Hyperparameters of Gaussian–inverse–Wishart prior distribution

This model can learn an appropriate number of spatial concepts, depending on the data, by using a nonparametric Bayesian approach. We use the SBP, which is one of the methods based on the Dirichlet process. In particular, this model can consider a theoretically infinite number of spatial concepts $L \to \infty$ and position distributions $K \to \infty$. SBP computations are difficult because they generate an infinite number of parameters. In this study, we approximate a number of parameters by setting sufficiently large values, i.e., a weak-limit approximation [26].

It is possible to correlate a name with multiple places, e.g., *"staircase"* is in two different places, and a place with multiple names, e.g., *"toilet"* and *"restroom"* refer to the same place. Spatial concepts are represented by a word distribution of the names of the place W_l and several position distributions (μ_k, Σ_k) indicated by a multinomial distribution ϕ_l. In other words, this model is capable of relating the mixture of Gaussian distributions to a multinomial distribution of the names of places. It should be noted that the arrows connecting i_t to the surrounding nodes of the proposed graphical model differ from those of ordinal Gaussian mixture model (GMM). We assume that words obtained by the robot do not change its position, but that the position of the robot affects the distribution of words. Therefore, the proposed generative process assumes that the index of position distribution i_t, i.e., the category of the place, is generated from the position of the robot x_t. This change can be naturally introduced without any troubles by introducing equation (11).

B. Overview of the proposed method SpCoA

We assume that a robot performs self-localization by using control data and sensor data at all times. The procedure for the learning of spatial concepts is as follows:

1) An utterer teaches a robot the names of places, as shown in Fig. 1 (b). Every time the robot arrives at a place that was a designated learning target, the utterer says a sentence, including the name of the current place.
2) The robot performs speech recognition from the uttered speech signal data. Thus, the speech recognition system includes a word dictionary of only Japanese syllables. The speech recognition results are obtained in a lattice format.
3) Word segmentation is performed by using the lattices of the speech recognition results.
4) The robot learns spatial concepts from words obtained by word segmentation and robot positions obtained by self-localization for all teaching times. The details of the learning are given in III-C.

The procedure for self-localization utilizing spatial concepts is as follows:

1) The words of the learned spatial concepts are registered to the word dictionary of the speech recognition system.
2) When a robot obtains a speech signal, speech recognition is performed. Then, a word sequence as the 1-best speech recognition result is obtained.
3) The robot modifies the self-localization from words obtained by speech recognition and the position likelihood obtained by spatial concepts. The details of self-localization are provided in III-D.

The proposed method can learn words related to places from the utterances of sentences. We use an unsupervised word segmentation method latticelm that can directly segment words from the lattices of the speech recognition results of the uttered sentences [23]. The *lattice* can represent to a compact the set of more promising hypotheses of a speech recognition result, such as N-best, in a directed graph format. Unsupervised word segmentation using the lattices of syllable recognition is expected to be able to reduce the variability and errors in phonemes as compared to NPYLM [14], i.e., word segmentation using the 1-best speech recognition results.

The self-localization method adopts MCL [24], a method that is generally used as the localization of mobile robots for simultaneous localization and mapping (SLAM) [27]. We

assume that a robot generates an environment map by using MCL-based SLAM such as FastSLAM [28], [29] in advance, and then, performs localization by using the generated map. Then, the environment map of both an occupancy grid map and a landmark map is acceptable.

C. Learning of spatial concept

Spatial concepts are learned from multiple teaching data, control data, and sensor data. The teaching data are a set of uttered sentences for all teaching times. Segmented words of an uttered sentence are converted into a bag-of-words (BoW) representation as a vector of the occurrence counts of words $O_{t,\mathbf{B}}$. The set of the teaching times is denoted as $T_o = \{t_1, t_2, \ldots, t_N\}$, and the number of teaching data items is denoted as N. The model parameters are denoted as $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, \pi\}$. The initial values of the model parameters can be set arbitrarily in accordance with a condition. Further, the sampling values of the model parameters from the following joint posterior distribution are obtained by performing Gibbs sampling.

$$p(x_{0:T}, i_{T_o}, C_{T_o}, \Theta \mid O_{T_o,\mathbf{B}}, u_{1:T}, z_{1:T}, \mathbf{h}) \tag{12}$$

where the hyperparameters of the model are denoted as $\mathbf{h} = \{\alpha, \gamma, \beta_0, m_0, \kappa_0, V_0, \nu_0\}$. The algorithm of the learning of spatial concepts is shown in Algorithm 1.

The conditional posterior distribution of each element used for performing Gibbs sampling can be expressed as follows: An index i_t of the position distribution is sampled for each data $t \in T_o$ from a posterior distribution as follows:

$$\begin{aligned} i_t &\sim p(i_t = k \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t) \\ &\propto \mathcal{N}(x_t \mid \mu_{k=i_t}, \Sigma_{k=i_t}) \mathrm{Mult}(i_t = k \mid \phi_{l=C_t}). \end{aligned} \tag{13}$$

An index C_t of the spatial concepts is sampled for each data item $t \in T_o$ from a posterior distribution as follows:

$$\begin{aligned} C_t &\sim p(C_t = l \mid x_t, i_t, O_{t,\mathbf{B}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, \pi) \\ &\propto \mathrm{Mult}(O_{t,\mathbf{B}} \mid W_{l=C_t}) \mathrm{Mult}(i_t = k \mid \phi_{l=C_t}) \\ &\qquad \mathrm{Mult}(C_t = l \mid \pi) \end{aligned} \tag{14}$$

where $O_{t,\mathbf{B}}$ denotes a vector of the occurrence counts of words in the sentence at time t. A posterior distribution representing word probabilities of the name of place $\mathbf{W}$ is calculated as follows:

$$\begin{aligned} &p(\mathbf{W} \mid C_{T_o}, O_{T_o,\mathbf{B}}) \\ &\quad \propto \prod_{l \in L} \Bigg[\prod_{\substack{l=C_t \\ t \in T_o}} p(O_{t,\mathbf{B}} \mid W_{l=C_t}) \Bigg] p(W_l) \end{aligned} \tag{15}$$

where variables with the subscript T_o denote the set of all teaching times. A word probability of the name of place W_l is sampled for each $l \in L$ as follows:

$$W_l \sim \mathrm{Mult}(\mathbf{O}_l \mid W_l) \mathrm{Dir}(W_l \mid \beta_0) \propto \mathrm{Dir}(W_l \mid \beta_{n_l}) \tag{16}$$

where β_{n_l} represents the posterior parameter and $\mathbf{O}_l$ denotes the BoW representation of all sentences of $C_t = l$ in $t \in T_o$. A posterior distribution representing the position distribution $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ is calculated as follows:

$$\begin{aligned} &p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid i_{T_o}, x_{T_o}, C_{T_o}, \phi_l) \\ &\quad \propto \prod_{k \in K} \Bigg[\prod_{\substack{k=i_t \\ t \in T_o}} p(x_t \mid \mu_{k=i_t}, \Sigma_{k=i_t}) \Bigg] p(\mu_k, \Sigma_k). \end{aligned} \tag{17}$$

A position distribution μ_k, Σ_k is sampled for each $k \in K$ as follows:

$$\begin{aligned} \mu_k, \Sigma_k &\sim \mathcal{N}(\mathbf{x}_k \mid \mu_k, \Sigma_k) \mathcal{NIW}(\mu_k, \Sigma_k \mid m_0, \kappa_0, V_0, \nu_0) \\ &\propto \mathcal{NIW}(\mu_k, \Sigma_k \mid m_{n_k}, \kappa_{n_k}, V_{n_k}, \nu_{n_k}) \end{aligned} \tag{18}$$

where $\mathcal{NIW}(\cdot)$ denotes the Gaussian–inverse–Wishart distribution; $m_{n_k}, \kappa_{n_k}, V_{n_k}$, and ν_{n_k} represent the posterior parameters; and $\mathbf{x}_k$ indicates the set of the teaching positions of $i_t = k$ in $t \in T_o$. A topic probability distribution π of spatial concepts is sampled as follows:

$$\pi \sim \mathrm{Mult}(C_{T_o} \mid \pi) \mathrm{Dir}(\pi \mid \gamma) \propto \mathrm{Dir}(\pi \mid C_{T_o}, \gamma). \tag{19}$$

A posterior distribution representing the mixed weights ϕ_l of the position distributions is calculated as follows:

$$\begin{aligned} &p(\phi_l \mid x_{T_o}, i_{T_o}, C_{T_o}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &\quad \propto \prod_{l \in L} \Bigg[\prod_{\substack{l=C_t \\ t \in T_o}} p(i_t \mid \phi_{l=C_t}) \Bigg] p(\phi_l). \end{aligned} \tag{20}$$

A mixed weight ϕ_l of the position distributions is sampled for each $l \in L$ as follows:

$$\phi_l \sim \mathrm{Mult}(\mathbf{i}_l \mid \phi_l) \mathrm{Dir}(\phi_l \mid \alpha) \propto \mathrm{Dir}(\phi_l \mid \mathbf{i}_l, \alpha) \tag{21}$$

where $\mathbf{i}_l$ denotes a vector counting all the indices of the Gaussian distribution of $C_t = l$ in $t \in T_o$.

Self-positions $x_{0:T}$ are sampled by using a Monte Carlo fixed-lag smoother [30] in the learning phase. The smoother can estimate self-position $x_{0:t}$ and not $p(x_{0:t} \mid u_{1:t}, z_{1:t})$, i.e., a sequential estimation from the given data $u_{1:t}, z_{1:t}$ until time t, but it can estimate $p(x_{0:t} \mid u_{1:T}, z_{1:T})$, i.e., an estimation from the given data $u_{1:T}, z_{1:T}$ until time T later than t $(t < T)$. In general, the smoothing method can provide a more accurate estimation than the MCL of online estimation. In contrast, if the self-position of a robot x_t is sampled like direct assignment sampling for each time t, the sampling of x_t is divided in the case with the teaching time $t \in T_o$ and another time $t \notin T_o$ as follows:

$$x_t \sim \begin{cases} p(x_t \mid x_{t-1}, x_{t+1}, u_t, u_{t+1}, z_t) \\ \propto p(x_{t+1} \mid x_t, u_{t+1}) p(z_t \mid x_t) p(x_t \mid x_{t-1}, u_t) \\ \qquad\qquad (t \notin T_o), \\ p(x_t \mid x_{t-1}, x_{t+1}, u_t, u_{t+1}, z_t, i_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t) \\ \propto p(x_{t+1} \mid x_t, u_{t+1}) p(z_t \mid x_t) p(x_t \mid x_{t-1}, u_t) \\ \quad p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t) \\ \qquad\qquad (t \in T_o). \end{cases} \tag{22}$$

Algorithm 1 Learning of spatial concepts

$\mathcal{L} = \emptyset$, $T_o = \emptyset$
// Localization and speech recognition
for $t = 0$ to T **do**
  $x_{0:t} \sim$ Monte_Carlo_smoother$(x_{0:t-1}, u_{1:t}, z_{1:t})$ [30]
  if the speech signal is observed **then**
    $lattice_t =$ speech_recognition$(speech_signal)$
    add $lattice_t$ to $\mathcal{L}$ // Registering the lattice
    add t to T_o // Registering the teaching time
  end if
end for
// Word segmentation using lattices
$O_{T_o,\mathbf{B}} \sim$ latticelm$(\mathcal{L})$ [23]
// Gibbs sampling
Initialize parameters i_{T_o}, C_{T_o}, $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, \pi\}$
for $j = 1$ to $iteration\_number$ **do**
  $i_{T_o} \sim p(i_{T_o} \mid x_{T_o}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_{T_o})$ (13)
  $C_{T_o} \sim p(C_{T_o} \mid x_{T_o}, i_{T_o}, O_{T_o,\mathbf{B}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, \pi)$ (14)
  $\mathbf{W} \sim p(\mathbf{W} \mid C_{T_o}, O_{T_o,\mathbf{B}})$ (16)
  $\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim p(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid i_{T_o}, x_{T_o}, C_{T_o}, \phi_l)$ (18)
  $\pi \sim p(\pi \mid C_{T_o})$ (19)
  $\phi_l \sim p(\phi_l \mid x_{T_o}, i_{T_o}, C_{T_o}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (21)
  for $t = 0$ to T **do**

$$x_t \sim \begin{cases} p(x_t \mid x_{t-1}, x_{t+1}, u_t, u_{t+1}, z_t) \\ \qquad\qquad (t \notin T_o) \\ p(x_t \mid x_{t-1}, x_{t+1}, u_t, u_{t+1}, z_t) \\ p(i_t \mid x_t, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, C_t) \\ \qquad\qquad (t \in T_o) \end{cases} \quad (22)$$

  end for
end for
return Θ

D. Self-localization of after learning spatial concepts

A robot that acquires spatial concepts can leverage spatial concepts to self-localization. The estimated model parameters $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi_l, \pi\}$ and a speech recognition sentence $O_{t,\mathbf{B}}$ at time t are given to the condition part of the probability formula of MCL as follows:

$$\begin{aligned} &p(x_{0:t} \mid z_{1:t}, u_{1:t}, O_{1:t,\mathbf{B}}, \Theta) \\ &\quad \propto p(z_t \mid x_t) p(O_{t,\mathbf{B}} \mid x_t, \Theta) p(x_t \mid x_{t-1}, u_t) \\ &\qquad p(x_{0:t-1} \mid z_{1:t-1}, u_{1:t-1}, O_{1:t-1,\mathbf{B}}, \Theta). \end{aligned} \quad (23)$$

When the robot hears the name of a place spoken by the utterer, in addition to the likelihood of the sensor model of MCL, the likelihood of x_t with respect to a speech recognition sentence is calculated as follows:

$$\begin{aligned} &p(O_{t,\mathbf{B}} \mid x_t, \Theta) \\ &\propto \sum_{C_t} \Big[p(O_{t,\mathbf{B}} | W_{C_t}) \sum_{i_t} \Big[p(x_t | \mu_{i_t}, \Sigma_{i_t}) p(i_t | \phi_{C_t}) \Big] p(C_t | \pi) \Big]. \end{aligned} \quad (24)$$

The algorithm of self-localization utilizing spatial concepts is shown in Algorithm 2. The set of particles is denoted as X_t, the temporary set that stores the pairs of the particle $x_t^{[m]}$ and the weight $w_t^{[m]}$, i.e., $\langle x_t^{[m]}, w_t^{[m]} \rangle$, is denoted as $\bar{X}_t$. The number of particles is M. The function sample_motion_model is a function that moves each particle from its previous state x_{t-1} to its current state x_t by using control data. The function sensor_model calculates the likelihood of each particle $x_t^{[m]}$ using sensor data z_t. These functions are normally used in MCL. For further details, please refer to [27]. In this case, a speech recognition sentence $O_{t,\mathbf{B}}$ is obtained by the speech recognition system using a word dictionary containing all the learned words.

Algorithm 2 Self-localization utilizing spatial concepts

procedure Localization$(X_{t-1}, u_t, z_t, O_{t,\mathbf{B}}, \Theta)$
  $\bar{X}_t = X_t = \emptyset$
  for $m = 1$ to M **do**
    $x_t^{[m]} =$ sample_motion_model$(u_t, x_{t-1}^{[m]})$ (9)
    $w_t^{[m]} =$ sensor_model$(z_t, x_t^{[m]})$ (10)
    if the speech signal is observed **then**
      $w_t^{[m]} = w_t^{[m]} \times p(O_{t,\mathbf{B}} \mid x_t, \Theta)$
    end if
    add $\langle x_t^{[m]}, w_t^{[m]} \rangle$ to $\bar{X}_t$
  end for
  for $m = 1$ to M **do**
    draw i with probability $\propto w_t^{[i]}$
    add $x_t^{[i]}$ to X_t
  end for
  return X_t
end procedure

IV. Experiment I

In this experiment, we validate the evidence of the proposed method (SpCoA) in an environment simulated on the simulator platform SIGVerse[2] [31], which enables the simulation of social interactions. The speech recognition is performed using the Japanese continuous speech recognition system Julius[3] [32], [33]. The set of 43 Japanese phonemes defined by Acoustical Society of Japan (ASJ)'s speech database committee is adopted by Julius [32]. The representation of these phonemes is also adopted in this study. The Julius system uses a word dictionary containing 115 Japanese syllables. The microphone attached on the robot is SHURE's PG27-USB. Further, an unsupervised morphological analyzer, a latticelm 0.4, is implemented [23].

In the experiment, we compare the following three types of word segmentation methods. A set of syllable sequences is given to the graphical model of SpCoA by each method. This set is used for the learning of spatial concepts as recognized uttered sentences $O_{T_o,\mathbf{B}}$.

(A) latticelm (proposed method)
  Syllable recognition results in the lattice format are segmented by using latticelm.
(B) 1-best NPYLM
  Syllable recognition results of the 1-best method are segmented by using latticelm. In this case, latticelm [23] is almost equivalent to NPYLM [14].

[2]SIGServer-2.2.2, SIGViewer-2.2.0, http://www.sigverse.com/wiki/
[3]Julius dictation-kit-v4.3.1-linux, GMM-HMM decoding, http://julius.sourceforge.jp/

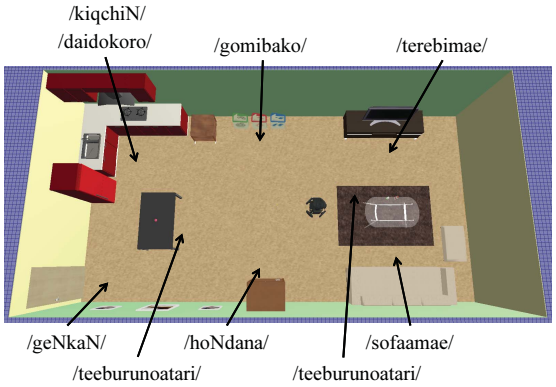


Fig. 3. Environment to be used for learning and localization on SIGVerse: This is a pseudo-room in the simulated real world. There is a robot in the center of the room. The size of the room is 500 cm × 1,000 cm, and the size of the robot is 50 cm × 50 cm.

TABLE II
VARIOUS PHRASES OF EACH JAPANESE SENTENCE: "**" IS USED AS A PLACEHOLDER FOR THE NAME OF EACH PLACE. EXAMPLES OF THESE PHRASES ARE *"** is here."*, *"This place is **."*, *"This place's name is **."*, AND *"Came to **."* IN ENGLISH.

** da yo	** wa kochira desu
** desu	kochira ga ** ni nari masu
koko ga **	kono basho ga ** da yo
koko wa ** desu	kono basho no namae wa **
** ni ki mashi ta	koko no namae wa ** da yo

(C) Bag-of-syllables (BoS)

Syllable recognition results of the 1-best method are segmented by each syllable. In other words, this method is used for segmenting not words but elements of the recognized syllable sequences directly in the BoS (bag-of-letters) representation.

The remainder of this section is organized as follows: In Section IV-A, the conditions and results of learning spatial concepts are described. The experiments performed using the learned spatial concepts are described in Section IV-B to IV-E. In Section IV-B, we evaluate the accuracy of the phoneme recognition and word segmentation for uttered sentences. In Section IV-C, we evaluate the clustering accuracy of the estimation results of index C_t of spatial concepts for each teaching utterance. In Section IV-D, we evaluate the accuracy of the acquisition of names of places. In Section IV-E, we show that spatial concepts can be utilized for effective self-localization.

A. Learning of spatial concepts

1) Conditions: We conduct this experiment of spatial concept acquisition in the environment prepared on SIGVerse. The experimental environment is shown in Fig. 3. A mobile robot can move by performing forward, backward, right rotation, or left rotation movements on a two-dimensional plane. In this experiment, the robot can use an approximately correct map of the considered environment. The robot has a range

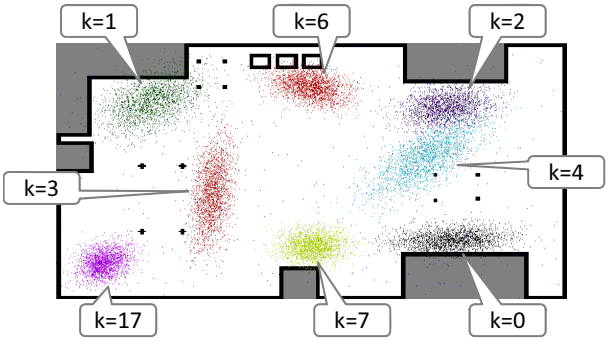


Fig. 4. Learning result of the position distribution: A point group of each color to represent each position distribution is drawn on an map of the considered environment. The colors of the point groups are determined randomly. Each balloon shows the index number for each position distribution.

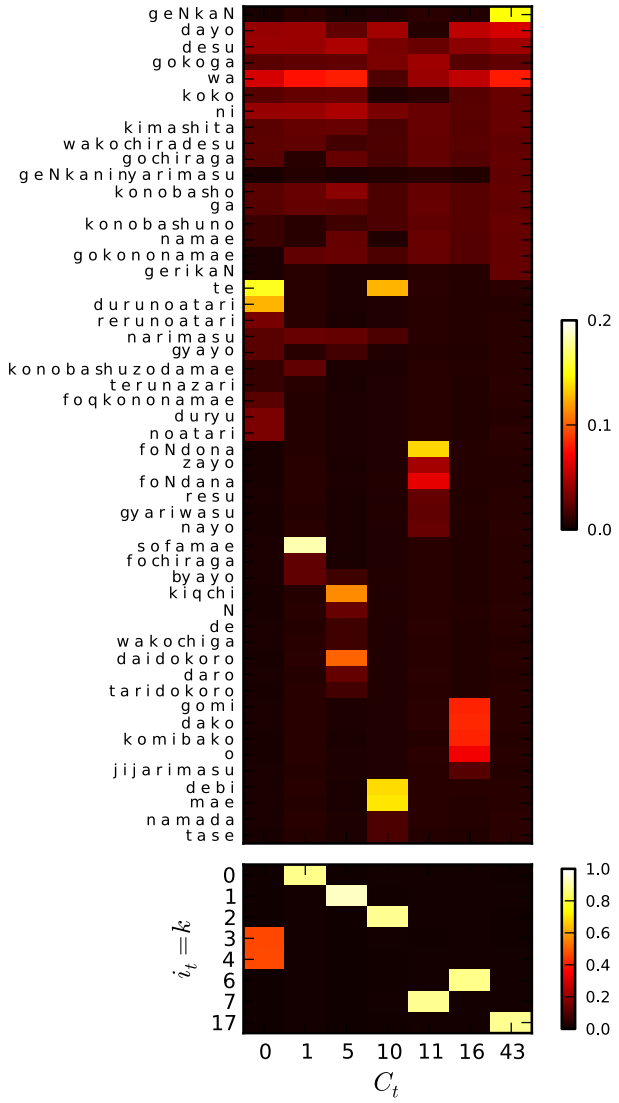


Fig. 5. Learning result of the multinomial distributions of the names of places W (top); multinomial distributions of the index of the position distribution ϕ_l (bottom): All the words obtained during the experiment are shown.

sensor in front and performs self-localization on the basis of an occupancy grid map. The initial particles are defined by the true initial position of the robot. The number of particles is $M = 1000$.

The lag value of the Monte Carlo fixed-lag smoothing is fixed at 100. The other parameters of this experiment are as follows: $L = 50$, $K = 50$, $\alpha = 1.5$, $\gamma = 8$, $\beta_0 = 0.5$, $m_0 = [0, 0]^{\mathrm{T}}$, $\kappa_0 = 0.001$, $V_0 = \mathrm{diag}(1000, 1000)$, and $\nu_0 = 2$. The number of iterations used for Gibbs sampling is 100. This experiment does not include the direct assignment sampling of x_t in equation (22), i.e., lines 22–24 of Algorithm 1 are omitted, because we consider that the self-position can be obtained with sufficiently good accuracy by using the Monte Carlo smoothing. Eight places are selected as the learning targets, and eight types of place names are considered. Each uttered place name is shown in Fig. 3. These utterances include the same name in different places, i.e., *"teeburunoatari"* (which means *near the table* in English), and different names in the same place, i.e., *"kiqchiN"* and *"daidokoro"* (which mean *a kitchen* in English). The other teaching names are *"geNkaN"* (which means *an entrance* or *a doorway* in English); *"terebimae"* (which means *the front of the TV* in English); *"gomibako"* (which means *a trash box* in English); *"hoNdana"* (which means *a bookshelf* in English); and *"sofaamae"* (which means *the front of the sofa* in English). The teaching utterances, including the 10 types of phrases, are spoken for a total of 90 times. The phrases in each uttered sentence are listed in Table II.

2) Results: The learning results of spatial concepts obtained by using the proposed method are presented here. Fig. 4 shows the position distributions learned in the experimental environment. Fig. 5 (top) shows the word distributions of the names of places for each spatial concept, and Fig. 5 (bottom) shows the multinomial distributions of the indices of the position distributions. Consequently, the proposed method can learn the names of places corresponding to each place of the learning target. In the spatial concept of index $C_t = 1$, the highest probability of words was *"sofamae"*, and the highest probability of the indices of the position distribution was $k = 0$; therefore, the name of a place *"sofamae"* was learned to correspond to the position distribution of $k = 0$. In the spatial concept of index $C_t = 5$, *"kiqchi"* and *"daidokoro"* were learned to correspond to the position distribution of $k = 1$. Therefore, this result shows that multiple names can be learned for the same place. In the spatial concept of index $C_t = 0$, *"te"* and *"durunoatari"* (one word in a normal situation) were learned to correspond to the position distributions of $k = 3$ and $k = 4$. Therefore, this result shows that the same name can be learned for multiple places.

B. Phoneme recognition accuracy of uttered sentences

1) Conditions: We compared the performance of three types of word segmentation methods for all the considered uttered sentences. It was difficult to weigh the ambiguous syllable recognition and the unsupervised word segmentation separately. Therefore, this experiment considered the positions of a delimiter as a single letter. We calculated the matching rate of a phoneme string of a recognition result of each uttered sentence and the correct phoneme string of the teaching data that was suitably segmented into Japanese morphemes using MeCab[4], which is an off-the-shelf Japanese morphological analyzer that is widely used for natural language processing. The matching rate of the phoneme string was calculated by using the phoneme accuracy rate (PAR) as follows:

$$\mathrm{PAR} = 1 - \frac{S + D + I}{N}. \tag{25}$$

The numerator of equation (25) is calculated by using the Levenshtein distance between the correct phoneme string and the recognition phoneme string. S denotes the number of substitutions; D, the number of deletions; and I, the number of insertions. N represents the number of phonemes of the correct phoneme string.

TABLE III
COMPARISON OF THE PHONEME ACCURACY RATES OF UTTERED SENTENCES FOR DIFFERENT WORD SEGMENTATION METHODS

	latticelem (used in SpCoA)	1-best NPYLM	BoS
PAR	<u>0.82</u>	0.71	0.67

2) Results: Table III shows the results of PAR. Table IV presents examples of the word segmentation results of the three considered methods. We found that the unsupervised morphological analyzer capable of using lattices improved the accuracy of phoneme recognition and word segmentation. Consequently, this result suggests that this word segmentation method considers the multiple hypothesis of speech recognition as a whole and reduces uncertainty such as variability in recognition by using the syllable recognition results in the lattice format.

C. Estimation accuracy of spatial concepts

1) Conditions: We compared the matching rate with the estimation results of index C_t of the spatial concepts of each teaching utterance and the classification results of the correct answer given by humans. The evaluation of this experiment used the adjusted Rand index (ARI) [34]. ARI is a measure of the degree of similarity between two clustering results.

Further, we compared the proposed method with a method of word clustering without location information for the investigation of the effect of lexical acquisition using location information. In particular, a method of word clustering without location information used the Dirichlet process mixture (DPM) of the unigram model of an SBP representation. The parameters corresponding to those of the proposed method were the same as the parameters of the proposed method and were estimated using Gibbs sampling.

2) Results: Fig. 6 shows the results of the average of the ARI values of 10 trials of learning by Gibbs sampling. Here, we found that the proposed method showed the best score. These results and the results reported in Section IV-B suggest that learning by uttered sentences obtained by better phoneme

[4]MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

TABLE IV
EXAMPLES OF WORD SEGMENTATION RESULTS OF UTTERED SENTENCES. "|" DENOTES A WORD SEGMENT POINT.

Correct word sequence	geNkaN|wa|kochira|desu	gomibako|ni|ki|mashi|ta	kono|basho|no|namae|wa|hoNdana
latticelm	geNkaN|wakochiradesu	komibako|o|ni|kimashita	konobashuno|namae|wa|foNdana
1-best NPYLM	ki|nika|N|wa|kochira|de|su	go|mibako|niki|na|shita	kono|bo|shu|no|namae|wa|fo|N|da|na
BoS	ki|ni|ka|N|wa|ko|chi|ra|de|su	go|mi|ba|ko|ni|ki|na|shi|ta	ko|no|bo|shu|no|na|ma|e|wa|fo|N|da|na

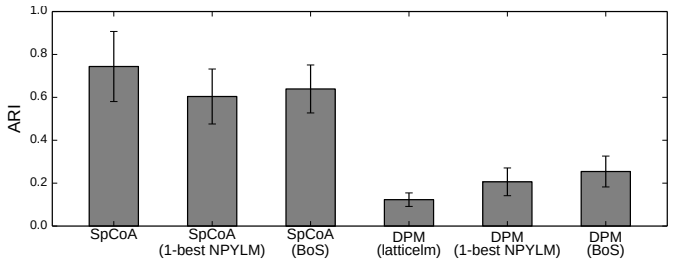


Fig. 6. Comparison of the accuracy rates of the estimation results of spatial concepts

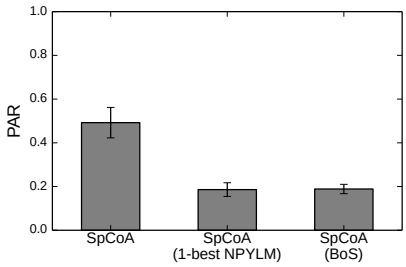


Fig. 7. PAR scores for the word considered the name of a place

recognition and better word segmentation produces a good result for the acquisition of spatial concepts. Furthermore, in a comparison of two clustering methods, we found that SpCoA was considerably better than DPM, a word clustering method without location information, irrespective of the word segmentation method used. The experimental results showed that it is possible to improve the estimation accuracy of spatial concepts and vocabulary by performing word clustering that considered location information.

D. Accuracy of acquired phoneme sequences representing the names of places

1) Conditions: We evaluated whether the names of places were properly learned for the considered teaching places. This experiment assumes a request for the best phoneme sequence $O_{t,\mathrm{best}}$ representing the self-position x_t for a robot. The robot moves close to each teaching place. The probability of a word $O_{t,\mathrm{best}}$ when the self-position x_t of the robot is given, $p(O_{t,\mathrm{best}} \mid x_t)$, can be obtained by using equation (24). The word having the best probability was selected. We compared the PAR with the correct phoneme sequence and a selected name of the place. Because *"kiqchiN"* and *"daidokoro"* were taught for the same place, the word whose PAR was the higher score was adopted.

2) Results: Fig. 7 shows the results of PAR for the word considered the name of a place. SpCoA (latticelm), the proposed method using the results of unsupervised word segmentation on the basis of the speech recognition results in the lattice format, showed the best PAR score. In the 1-best and BoS methods, a part syllable sequence of the name of a place was more minutely segmented as shown in Table IV. Therefore, the robot could not learn the name of the teaching place as a coherent phoneme sequence. In contrast, the robot could learn the names of teaching places more accurately by using the proposed method.

E. Self-localization that utilizes acquired spatial concepts

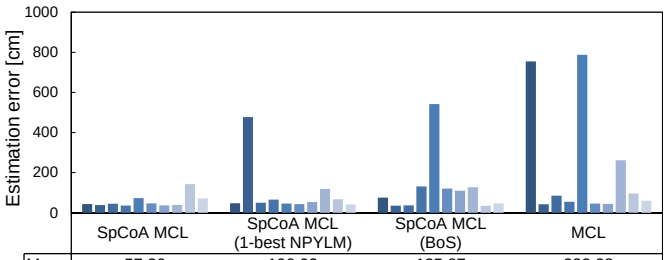
1) Conditions: In this experiment, we validate that the robot can make efficient use of the acquired spatial concepts. We compare the estimation accuracy of localization for the proposed method (SpCoA MCL) and the conventional MCL. When a robot comes to the learning target, the utterer speaks out the sentence containing the name of the place once again for the robot. The moving trajectory of the robot and the uttered positions are the same in all the trials. In particular, the uttered sentence is *"kokowa ** dayo"*. When learning a task, this phrase is not used. The number of particles is $M = 1000$, and the initial particles are uniformly distributed in the considered environment. The robot performs a control operation for each time step.

The estimation error in the localization is evaluated as follows: While running localization, we record the estimation error (equation (26)) on the $\mathbf{xy}$ plane of the floor for each time step.

$$e_t = \sqrt{(\bar{\mathbf{x}}_t - \mathbf{x}_t^*)^2 + (\bar{\mathbf{y}}_t - \mathbf{y}_t^*)^2} \tag{26}$$

where $\mathbf{x}_t^*, \mathbf{y}_t^*$ denote the true position coordinates of the robot as obtained from the simulator, and $\bar{\mathbf{x}}_t = \sum_{i=1}^{M} w_t^{(i)} \mathbf{x}_t^{(i)}$, $\bar{\mathbf{y}}_t = \sum_{i=1}^{M} w_t^{(i)} \mathbf{y}_t^{(i)}$ represent the weighted mean values of localization coordinates. The normalized weight $w_t^{(i)}$ is obtained from the sensor model in MCL as a likelihood. In the utterance time, this likelihood is multiplied by the value calculated using equation (24). $\mathbf{x}_t^{(i)}$, $\mathbf{y}_t^{(i)}$ denote the $\mathbf{x}$-coordinate and the $\mathbf{y}$-coordinate of index i of each particle at time t. After running the localization, we calculated the average of e_t.

Further, we compared the estimation accuracy rate (EAR) of the global localization. In each trial, we calculated the proportion of time step in which the estimation error was less than 50 cm.



	SpCoA MCL	SpCoA MCL (1-best NPYLM)	SpCoA MCL (BoS)	MCL
Mean	57.20	100.92	125.87	223.38
S.E.	9.89	40.23	45.49	88.86
EAR	0.81	0.70	0.65	0.56

Fig. 8. Results of estimation errors and EARs of self-localization

Fig. 9. Autonomous mobile robot TurtleBot 2: The robot is based on Yujin Robot Kobuki and Microsoft Kinect for its use as a range sensor.

2) Results: Fig. 8 shows the results of the estimation error and the EAR for 10 trials of each method. All trials of SpCoA MCL (latticelm) and almost all trials of the method using 1-best NPYLM and BoS showed relatively small estimation errors. Results of the second trial of 1-best NPYLM and the fifth trial of BoS showed higher estimation errors. In these trials, many particles converged to other places instead of the place where the robot was, based on utterance information. Nevertheless, compared with those of the conventional MCL, the results obtained using spatial concepts showed an obvious improvement in the estimation accuracy. Consequently, spatial concepts acquired by using the proposed method proved to be very helpful in improving the localization accuracy.

V. Experiment II

In this experiment, the effectiveness of the proposed method was tested by using an autonomous mobile robot TurtleBot 2[5] in a real environment. Fig. 9 shows TurtleBot 2 used in the experiments. Mapping and self-localization are performed by the robot operating system (ROS). The speech recognition system, the microphone, and the unsupervised morphological analyzer were the same as those described in Section IV.

A. Learning of spatial concepts in the real environment

1) Conditions: We conducted an experiment of the spatial concept acquisition in a real environment of an entire floor of a building. In this experiment, self-localization was performed using a map generated by SLAM. The initial particles are defined by the true initial position of the robot. The generated map in the real environment and the names of teaching places are shown in Fig. 10. The number of teaching places was

[5]TurtleBot 2, http://turtlebot.com/

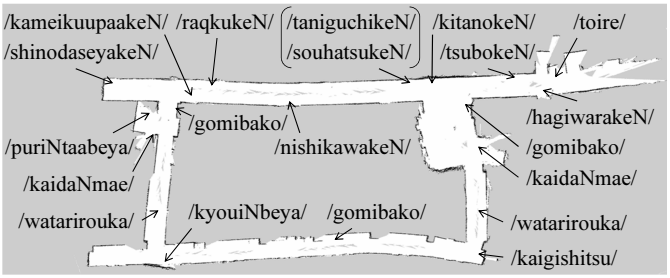


Fig. 10. Teaching places and the names of places shown on the generated map. The teaching places included places having two names each and multiple places having the same names.

19, and the number of teaching names was 16. The teaching utterances were performed for a total of 100 times.

2) Results: Fig. 11 shows the position distributions learned on the map. Table V shows the five best elements of the multinomial distributions of the name of place W_{C_t} and the multinomial distributions of the indices of the position distribution ϕ_{C_t} for each index of spatial concept C_t. Thus, we found that the proposed method can learn the names of places corresponding to the considered teaching places in the real environment. For example, in the spatial concept of index $C_t = 10$, *"torire"* was learned to correspond to a position distribution of $k = 42$. Similarly, *"kidanokeN"* corresponded to $k = 8$ in $C_t = 29$, and *"kaigihitsu"* was corresponded to $k = 60$ in $C_t = 32$. In the spatial concept of index $C_t = 27$, a part of the syllable sequences was minutely segmented as *"sohatsuke"*, *"N"*, and *"tani"*, *"guchi"*. In this case, the robot was taught two types of names. These words were learned to correspond to the same position distribution of $k = 55$. In $C_t = 8$, *"gomibako"* showed a high probability, and it corresponded to three distributions of the position of $k = 0, 36, 59$. The position distribution of $k = 13$ had the fourth highest probability in the spatial concept $C_t = 8$. Therefore, *"raqkukeN,"* which had the fifth highest probability in the spatial concept $C_t = 8$ (and was expected to relate to the spatial concept $C_t = 74$), can be estimated as the word drawn from spatial concept $C_t = 8$. However, in practice, this situation did not cause any severe problems because the spatial concept of the index $C_t = 74$ had the highest probabilities for the word *"rapukeN"* and the position distribution $k = 13$ than $C_t = 8$. In the probabilistic model, the relative probability and the integrative information are important. When the robot listened to an utterance related to *"raqkukeN,"* it could make use of the spatial concept of index $C_t = 74$ for self-localization with a high probability, and appropriately updated its estimated self-location. We expected that the spatial concept of index $C_t = 2$ was learned as two separate spatial concepts. However, *"watarirooka"* and *"kaidaNmae"* were learned as the same spatial concept. Therefore, the multinomial distribution ϕ_2 showed a higher probability for the indices of the position distribution corresponding to the teaching places of both *"watarirooka"* and *"kaidaNmae"*.

The proposed method adopts a nonparametric Bayesian method in which it is possible to form spatial concepts that allow many-to-many correspondences between names and

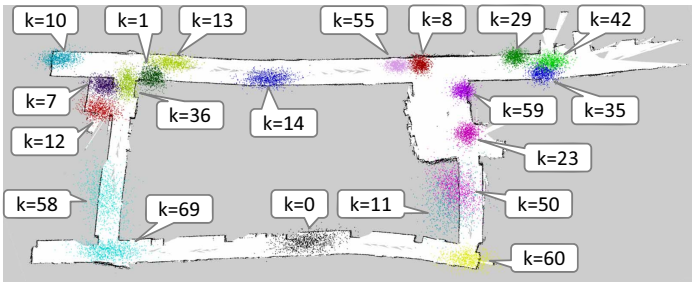


Fig. 11. Learning result of each position distribution: A point group of each color denoting each position distribution was drawn on the map. The colors of the point groups were determined randomly. Further, each index number is denoted as $i_t = k$.

TABLE V
LEARNING RESULT OF HIGH-PROBABILITY WORDS AND INDICES OF THE POSITION DISTRIBUTION FOR EACH SPATIAL CONCEPT

Index C_t	W_{C_t} Word	(Probability)	ϕ_{C_t} Index i_t	(Probability)
2	**watarirooka**	(**0.165**)	**23**	(**0.175**)
	kaidaNmae	(**0.151**)	**58**	(**0.174**)
	desu	(0.117)	**50**	(**0.142**)
	nikimashita	(0.069)	**12**	(**0.141**)
	ewa	(0.068)	**11**	(**0.039**)
3	**nokeN**	(**0.189**)	**29**	(**0.342**)
	tsu	(**0.185**)	64	(0.008)
	a	(0.046)	49	(0.008)
	nayo	(0.045)	82	(0.008)
	de	(0.045)	94	(0.008)
6	**shinozaseya**	(**0.178**)	**10**	(**0.339**)
	keN	(**0.174**)	37	(0.008)
	desu	(0.075)	5	(0.008)
	koko	(0.042)	94	(0.008)
	wa	(0.041)	29	(0.008)
8	desu	(0.195)	**36**	(**0.174**)
	gomibako	(**0.180**)	**59**	(**0.136**)
	koko	(0.102)	**0**	(**0.136**)
	a	(0.081)	13	(0.135)
	rapukeN	(0.043)	9	(0.006)
10	**torire**	(**0.154**)	**42**	(**0.340**)
	wa	(0.118)	83	(0.008)
	kokoga	(0.080)	9	(0.008)
	nikimashita	(0.045)	46	(0.008)
	byayo	(0.043)	11	(0.008)
27	**tani**	(**0.098**)	**55**	(**0.507**)
	sohatsuke	(**0.098**)	84	(0.006)
	N	(**0.098**)	21	(0.006)
	guchi	(**0.096**)	42	(0.006)
	desu	(0.078)	37	(0.006)
29	**kidanokeN**	(**0.209**)	**8**	(**0.336**)
	desu	(0.091)	60	(0.009)
	dayo	(0.090)	70	(0.008)
	a	(0.050)	17	(0.008)
	konobashoga	(0.050)	2	(0.008)

Index C_t	W_{C_t} Word	(Probability)	ϕ_{C_t} Index i_t	(Probability)
32	**kaigihitsu**	(**0.181**)	**60**	(**0.301**)
	nikimashita	(0.113)	0	(0.065)
	gomirako	(0.079)	7	(0.064)
	a	(0.078)	36	(0.007)
	dayo	(0.076)	33	(0.007)
34	**N**	(**0.113**)	**1**	(**0.197**)
	kameikukache	(**0.112**)	36	(0.075)
	ewa	(0.109)	42	(0.075)
	konobashunonama	(0.108)	29	(0.009)
	ninarimasu	(0.043)	61	(0.008)
44	**rakeN**	(**0.159**)	**35**	(**0.296**)
	hagiwa	(**0.157**)	30	(0.009)
	desu	(0.080)	48	(0.008)
	wakochira	(0.046)	32	(0.008)
	ewa	(0.045)	68	(0.008)
47	**buriN**	(**0.132**)	**7**	(**0.321**)
	bea	(**0.130**)	0	(0.067)
	pa	(**0.107**)	22	(0.008)
	ewa	(0.083)	5	(0.008)
	dayo	(0.078)	96	(0.008)
66	**wakeN**	(**0.133**)	**14**	(**0.332**)
	nishikya	(**0.132**)	90	(0.008)
	desu	(0.103)	25	(0.008)
	a	(0.071)	69	(0.008)
	nishi	(**0.040**)	87	(0.008)
74	**rapukeN**	(**0.145**)	**13**	(**0.173**)
	nikimashita	(0.081)	56	(0.011)
	nayo	(0.080)	2	(0.010)
	nokeN	(0.017)	91	(0.010)
	wakeN	(0.016)	58	(0.010)
75	**bea**	(**0.153**)	**69**	(**0.343**)
	N	(**0.151**)	34	(0.008)
	kyoi	(**0.149**)	15	(0.008)
	dayo	(0.064)	75	(0.008)
	desu	(0.062)	27	(0.008)

places. In contrast, this can create ambiguity that classifies originally different spatial concepts into one spatial concept as a side effect. There is a possibility that the ambiguity of concepts such as $C_t = 2$ will have a negative effect on self-localization, even though the self-localization performance was (overall) clearly increased by employing the proposed method. The solution of this problem will be considered in future work.

In terms of the PAR of uttered sentences, the evaluation value from the evaluation method used in Section IV-B is 0.83; this value is comparable to the result in Section IV-B. However, in terms of the PAR of the name of the place, the evaluation value from the evaluation method used in Section IV-D is 0.35, which is lower than that in Section IV-D. We consider that the increase in uncertainty in the real environment and the increase in the number of teaching words reduced the performance. We expect that this problem could be improved using further experience related to places, e.g., if the number of utterances per place is increased, and additional sensory information is provided.

B. Modification of localization by the acquired spatial concepts

1) Conditions: In this experiment, we verified the modification results of self-localization by using spatial concepts in global self-localization. This experiment used the learning results of spatial concepts presented in Section V-A. The experimental procedures are shown below. The initial particles were uniformly distributed on the entire floor. The robot begins to move from a little distance away to the target place. When the robot reached the target place, the utterer spoke the sentence containing the name of the place for the robot. Upon obtaining the speech information, the robot modifies the self-localization on the basis of the acquired spatial concepts. The number of particles was the same as that mentioned in Section V-A.

2) Results: Fig. 12 shows the results of the self-localization before (the top part of the figure) and after (the bottom part of the figure) the utterance for three places. The particle states are denoted by red arrows. The moving trajectory of the robot is indicated by a green dotted arrow. Figs. 12 (a), (b), and (c) show the results for the names of places *"toire"*, *"souhatsukeN"*, and *"gomibako"*. Further, three spatial concepts, i.e., those at $k = 0, 36, 59$, were learned as *"gomibako"*. In this experiment, the utterer uttered to the robot when the robot came close to the place of $k = 36$. In all the examples shown in the top part of the figure, the particles were dispersed in several places. In contrast, the number of particles near the true position of the robot showed an almost accurate increase in all the examples shown in the bottom part of the figure. Thus, we can conclude that the proposed method can modify self-localization by using spatial concepts.

VI. Conclusion and Future Work

In this paper, we discussed the spatial concept acquisition, lexical acquisition related to places, and self-localization using acquired spatial concepts. We proposed *nonparametric Bayesian spatial concept acquisition method* SpCoA that integrates latticelm [23], a spatial clustering method, and MCL. We conducted experiments for evaluating the performance of SpCoA in a simulation and a real environment. SpCoA showed good results in all the experiments. In experiments of the learning of spatial concepts, the robot could form spatial concepts for the places of the learning targets from human continuous speech signals in both the room of the simulation environment and the entire floor of the real environment. Further, the unsupervised word segmentation method latticelm could reduce the variability and errors in the recognition of phonemes in all the utterances. SpCoA achieved more accurate lexical acquisition by performing word segmentation using the lattices of the speech recognition results. In the self-localization experiments, the robot could effectively utilize the acquired spatial concepts for recognizing self-position and reducing the estimation errors in self-localization.

As a method that further improves the performance of the lexical acquisition, a mutual learning method was proposed by Nakamura et al. on the basis of the integration of the learning of object concepts with a language model [35], [36]. Following a similar approach, Heymann et al. proposed a method that alternately and repeatedly updates phoneme recognition results and the language model by using unsupervised word segmentation [37]. As a result, they achieved robust lexical acquisition. In our study, we can expect to improve the accuracy of lexical acquisition for spatial concepts by estimating both the spatial concepts and the language model.

Furthermore, as a future work, we consider it necessary for robots to learn spatial concepts online and to recognize whether the uttered word indicates the current place or destination. Furthermore, developing a method that simultaneously acquires spatial concepts and builds a map is one of our future objectives. We believe that the spatial concepts will have a positive effect on the mapping. We also intend to examine a method that associates the image and the landscape with spatial concepts and a method that estimates both spatial concepts and object concepts.

References

[1] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive science*, vol. 26, no. 1, pp. 113–146, 2002.

[2] R. Taguchi, N. Iwahashi, T. Nose, K. Funakoshi, and M. Nakano, "Learning lexicons from spoken utterances based on statistical model selection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2009, pp. 2731–2734.

[3] N. Iwahashi, "Language acquisition through a human–robot interface by combining speech, visual, and behavioral information," *Information Sciences*, vol. 156, no. 1, pp. 109–121, 2003.

[4] ——, "Robots that learn language: A developmental approach to situated human-robot conversations," in *Human Robot Interaction*. InTech, 2007, pp. 95–118.

[5] N. Iwahashi, R. Taguchi, K. Sugiura, K. Funakoshi, and M. Nakano, "Robots that learn to converse: Developmental approach to situated language processing," in *Proceedings of International Symposium on Speech and Language Processing*, 2009, pp. 532–537.

[6] P. Gorniak and D. Roy, "Probabilistic grounding of situated speech using plan recognition and reference resolution," in *Proceedings of the 7th international conference on Multimodal interfaces*. ACM, 2005, pp. 138–143.

[7] S. Qu and J. Y. Chai, "Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 244–253.

[8] ——, "Context-based word acquisition for situated dialogue in a virtual world," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 247–278, 2010.

[9] J. Hörnstein, L. Gustavsson, J. Santos-Victor, and F. Lacerda, "Multimodal language acquisition based on motor learning and interaction," in *From Motor Learning to Interaction Learning in Robots*. Springer, 2010, pp. 467–489.

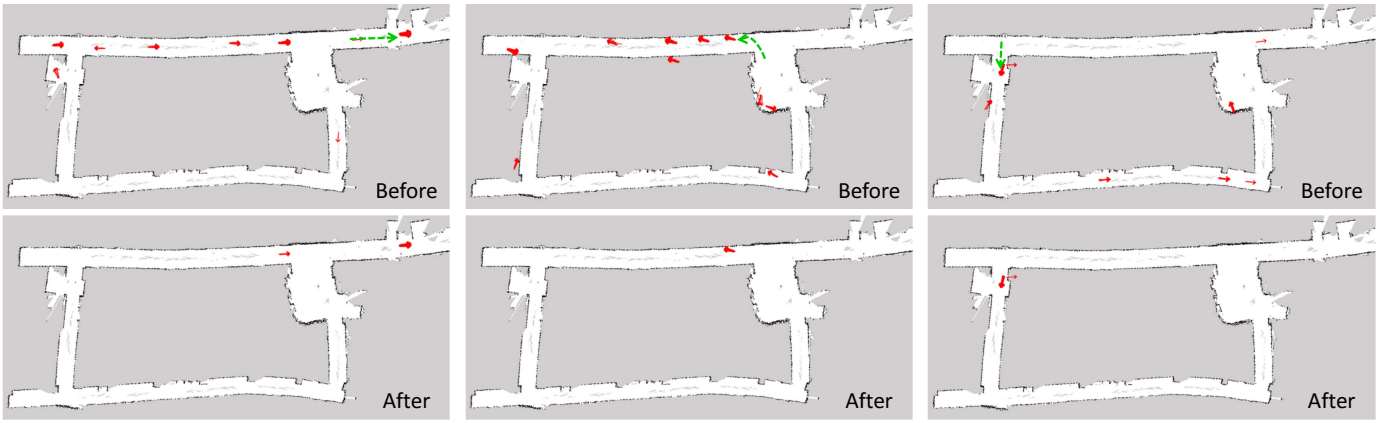
[10] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts," *Advanced Robotics*, vol. 25, no. 17, pp. 2189–2206, 2011.

[11] T. Araki, T. Nakamura, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Online learning of concepts and words using multimodal LDA and hierarchical Pitman-Yor Language Model," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2012, pp. 1623–1630.

[12] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[13] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal categorization by hierarchical Dirichlet process," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2011, pp. 1520–1525.

[14] D. Mochihashi, T. Yamada, and N. Ueda, "Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, 2009, pp. 100–108.



(a) The name of the place was *"toire."* (b) The name of the place was *"souhatsukeN."* (c) The name of the place was *"gomibako."*

Fig. 12. States of particles: before the teaching utterance (top); after the teaching utterance (bottom). The uttered sentence is *"kokowa ** dayo,"* (which means *"Here is **."*) *"**"* is the name of the place.

[15] R. Taguchi, N. Iwahashi, K. Funakoshi, M. Nakano, T. Nose, and T. Nitta, "Learning physically grounded lexicons from spoken utterances," *Human Machine Interaction–Getting Closer*, pp. 69–84, 2012.

[16] R. Taguchi, Y. Yamada, K. Hattori, T. Umezaki, M. Hoguro, N. Iwahashi, K. Funakoshi, and M. Nakano, "Learning place-names from spoken utterances and localization results by mobile robot," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 1325–1328.

[17] M. Milford, G. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2004, pp. 403–408.

[18] M. Milford, R. Schulz, D. Prasser, G. Wyeth, and J. Wiles, "Learning spatial concepts from RatSLAM representations," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 403–410, 2007.

[19] R. Schulz, G. Wyeth, and J. Wiles, "Lingodroids: socially grounding place names in privately grounded cognitive maps," *Adaptive Behavior*, vol. 19, no. 6, pp. 409–424, 2011.

[20] S. Heath, D. Ball, R. Schulz, and J. Wiles, "Communication between Lingodroids with different cognitive capabilities," in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2013, pp. 490–495.

[21] R. Schulz, G. Wyeth, and J. Wiles, "Are we there yet? grounding temporal concepts in shared journeys," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 2, pp. 163–175, 2011.

[22] K. Welke, P. Kaiser, A. Kozlov, N. Adermann, T. Asfour, M. Lewis, and M. Steedman, "Grounded spatial symbols for task planning based on experience," in *13th International Conference on Humanoid Robots (Humanoids). IEEE/RAS*, 2013.

[23] G. Neubig, M. Mimura, and T. Kawahara, "Bayesian learning of a language model from continuous speech," *IEICE TRANSACTIONS on Information and Systems*, vol. 95, no. 2, pp. 614–625, 2012.

[24] F. Dellaert, D. Fox, W. Burgard, and S. Thrun, "Monte carlo localization for mobile robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2. IEEE, 1999, pp. 1322–1328.

[25] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[26] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "A sticky HDP-HMM with application to speaker diarization," *The Annals of Applied Statistics*, pp. 1020–1056, 2011.

[27] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. MIT Press, 2005.

[28] M. Montemerlo, S. Thrun, D. Koller, and B. Wegbreit, "FastSLAM: A factored solution to the simultaneous localization and mapping problem," in *In Proceedings of the AAAI National Conference on Artificial Intelligence*. American Association for Artificial Intelligence, 2002, pp. 593–598.

[29] D. Hahnel, W. Burgard, D. Fox, and S. Thrun, "An efficient FastSLAM algorithm for generating maps of large-scale cyclic environments from raw laser range measurements," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2003, pp. 206–211.

[30] G. Kitagawa, "Computational aspects of sequential Monte Carlo filter and smoother," *Annals of the Institute of Statistical Mathematics*, vol. 66, no. 3, pp. 443–471, 2014.

[31] T. Inamura, T. Shibata, H. Sena, T. Hashimoto, N. Kawai, T. Miyashita, Y. Sakurai, M. Shimizu, M. Otake, K. Hosoda, *et al.*, "Simulator platform that enables social interaction simulation –SIGVerse: SocioIntelliGenesis simulator–," in *IEEE/SICE International Symposium on System Integration*, 2010, pp. 212–217.

[32] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Sharable software repository for Japanese large vocabulary continuous speech recognition," in *Fifth International Conference on Spoken Language Processing*, 1998.

[33] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," in *European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001.

[34] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[35] T. Nakamura, T. Araki, T. Nagai, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Multimodal concept and word learning using phoneme sequences with errors," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 157–162.

[36] T. Nakamura, T. Nagai, K. Funakoshi, S. Nagasaka, T. Taniguchi, and N. Iwahashi, "Mutual learning of an object concept and language model based on MLDA and NPYLM," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 600–607.

[37] J. Heymann, O. Walter, R. Haeb-Umbach, and B. Raj, "Iterative Bayesian Word Segmentation for Unsupervised Vocabulary Discovery from Phoneme Lattices," in *39th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.