Robotics and Autonomous Systems 99 (2018) 166–180

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Unsupervised spatial lexical acquisition by updating a language model with place clues☆

Akira Taniguchi [a,*], Tadahiro Taniguchi [a], Tetsunari Inamura [b]

[a] Ritsumeikan University, 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan
[b] National Institute of Informatics/The Graduate University for Advanced Studies, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

HIGHLIGHTS

- We improve the accuracy of lexical acquisition by updating a language model with place clues.
- The robot can learn spatial concepts with high accuracy as unsupervised place categorization.
- The mutual information is used to select words related to a place effectively.

ARTICLE INFO

Article history:
Received 31 March 2017
Received in revised form 22 September 2017
Accepted 19 October 2017
Available online 2 November 2017

Keywords:
Ambiguous speech recognition
Bayesian nonparametrics
Lexical acquisition
Place categorization
Spatial concept acquisition
Unsupervised word segmentation

ABSTRACT

This paper describes how to achieve highly accurate unsupervised spatial lexical acquisition from speech-recognition results including phoneme recognition errors. In most research into lexical acquisition, the robot has no pre-existing lexical knowledge. The robot acquires sequences of some phonemes as words from continuous speech signals. In a previous study, we proposed a nonparametric Bayesian spatial concept acquisition method (SpCoA) that integrates the robot's position and words obtained by unsupervised word segmentation from uncertain syllable recognition results. However, SpCoA has a very critical problem to be solved in lexical acquisition; the boundaries of word segmentation are incorrect in many cases because of many phoneme recognition errors. Therefore, we propose an unsupervised machine learning method (SpCoA++) for the robust lexical acquisition of novel words relating to places visited by the robot. The proposed SpCoA++ method performs an iterative estimation of learning spatial concepts and updating a language model using place information. SpCoA++ can select a candidate including many words that better represent places from multiple word-segmentation results by maximizing the mutual information between segmented words and spatial concepts. The experimental results demonstrate a significant improvement of the phoneme accuracy rate of learned words relating to place in the proposed method by word-segmentation results based on place information, in comparison to the conventional methods. We indicate that the proposed method enables the robot to acquire words from speech signals more accurately, and improves the estimation accuracy of the spatial concepts.

1. Introduction

Autonomous robots operating in a human living environment need to understand the spatial lexical knowledge of their ambient environment in order to facilitate interactions with people. For example, a robot might be required to recognize the name of its current position and those of other areas on its environmental map, such as *"kitchen"*, *"entrance-way"*, and certain proper nouns specific to various places. Therefore, we consider it to be important for robots to be able to learn the novel and various words that people associate with particular places in their environments, and the spatial areas corresponding to those names, i.e., robots must be able to acquire novel and various words relating to places. To do this, a robot could use speech signals recognized from microphones and the sensory-motor information obtained from odometry and laser sensors in the ambient environment.

Lexical acquisition means that a robot with no pre-existing lexicon learns phoneme sequences from the continuous speech signals of a person. In this case, the robot must be able to manage a considerable degree of uncertainty in the speech recognition, i.e., phoneme recognition errors. When the robot learns novel

☆ This research was partially supported by JST CREST, and a Grant-in-Aid for Scientific Research on Innovative Areas (16H06569) funded by the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

* Corresponding author.
E-mail addresses: a.taniguchi@em.ci.ritsumei.ac.jp (A. Taniguchi), taniguchi@em.ci.ritsumei.ac.jp (T. Taniguchi), inamura@nii.ac.jp (T. Inamura).

https://doi.org/10.1016/j.robot.2017.10.013
0921-8890/

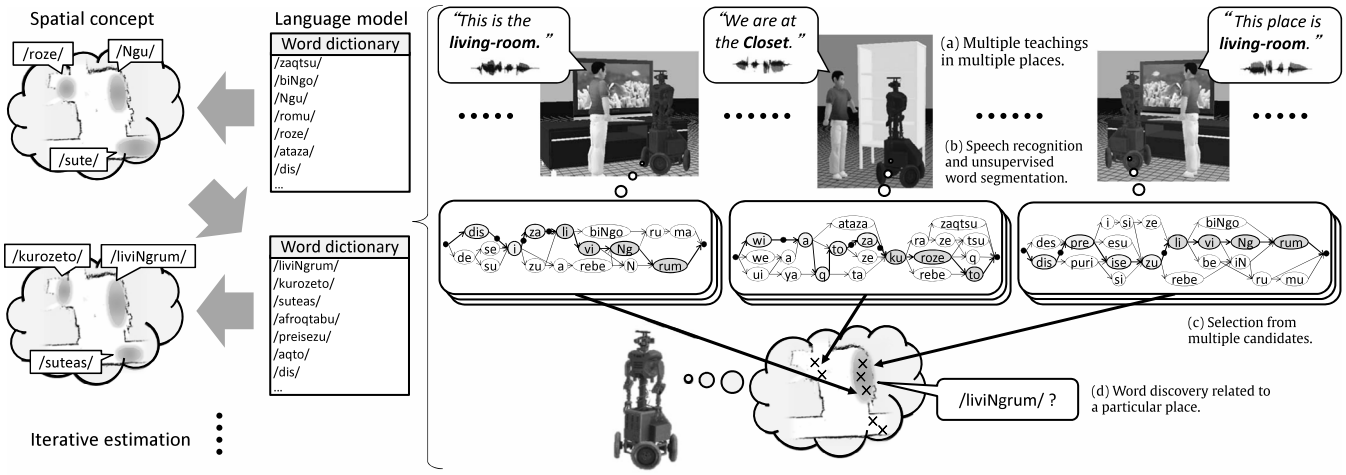

Fig. 1. Schematic representation of spatial lexical acquisition used in this study. Our approach involves the updating of both spatial concepts and a language model for highly accurate word discovery. (a) The user teaches the robot different sentences in different places. (b) The robot recognizes the user's speech signals and segments them into words. The thickest line in each weighted finite-state transducer (WFST) represents the selected path, and a black dot in the middle of an arrow represents a word boundary. For example, in the WFST on the right, the words that were obtained are */dis/*, */preisezu/*, and */liviNgrum/*. (c) The robot selects words from multiple candidates of segmented words by using place information. (d) The robot learns words related to a particular place.

words from speech signals, it is difficult to determine segmentation boundaries and the identity of words by using speech recognition, which can lead to various errors. Without pre-existing lexical knowledge, the robot makes speech-recognition mistakes more frequently than if it had the required lexical knowledge [1]. Let us consider a problem related to lexical acquisition from uttered sentences. For example, the robot obtains a speech-recognition result such as */heaiznyuyoak/* (an incorrect phoneme recognition of *"Here is New York"*.). The robot must segment the sentence into the true boundaries as individual words, e.g., */hea/*, */iz/*, and */nyuyoak/*. However, in many cases, this speech-recognition result is segmented into incorrect boundaries through either under- or over-segmentation [2], e.g., */he/*, */aiz/*, */nyuyo/*, and */ak/*. Furthermore, it is necessary for the robot to recognize words referring to the same thing from among these numerous segmented results that contain errors.

This study addresses the above lexical-acquisition problems by complementing ambiguous speech-recognition and word-segmentation results with place information. We assume that the robot has not acquired any vocabulary in advance, and can recognize only phonemes or syllables. We represent the spatial area of the environment in terms of a *position distribution*. Furthermore, we define a *spatial concept* as a place category that includes place names and the position distributions corresponding to those names.

Taniguchi et al. [3] proposed nonparametric Bayesian spatial concept acquisition method (SpCoA) based on an unsupervised word-segmentation method known as latticelm[1] [4]. This method enables word segmentation with consideration of phoneme errors in speech recognition more efficiently than does the nested Pitman–Yor language model (NPYLM) [5]. However, in many cases, the original word representing the name of the place is finely segmented into several words, i.e., over-segmentation. We consider this problem to be caused by a word-segmentation method that does not use place information, i.e., the words are segmented from syllable sequences only. In this paper, as a solution to this problem, we propose the SpCoA++ method that iteratively constructs spatial concepts and a language model, and that performs word segmentation and spatial concept acquisition more accurately. In addition, we propose a method that enables a word related to a particular place to be selected using the mutual information.

A schematic diagram depicting our study is shown in Fig. 1. The left parts of Fig. 1 represent the iterative estimation of spatial concepts and a language model. In Fig. 1(a), when the robot arrives at a place that is a designated learning target, the user speaks a sentence (including the name of the place) to the robot, which moves within the environment while performing self-localization. In Fig. 1(b), the robot performs speech recognition and unsupervised word segmentation from the human speech signals. In our approach, in order to cope with the uncertainty of speech recognition, we use speech recognition based on a weighted finite-state transducer (WFST), which is a word graph representing the speech-recognition results. In addition, we use the latticelm [4], which can segment the speech-recognition results in the WFST format. In Fig. 1(c), from the multiple candidates of word boundaries and paths in a WFST, the robot selects the one that includes words that best represent a particular place. In Fig. 1(d), the robot learns a word that is frequently obtained in only a specific place as the name of the place, e.g., */liviNgrum/*. The robot then uses the selected words to update a language model. We consider that our approach can greatly improve the performance of unsupervised word segmentation from speech-recognition results that contain phoneme-recognition errors in the lexical acquisition related to places.

The main contributions of this paper are as follows.

- We improve the accuracy of spatial lexical acquisition by updating a language model with place clues. In comparative experiments, the proposed method performs better than the conventional methods.
- We show that it is possible to learn spatial concepts with increased accuracy as unsupervised place categorization by obtaining highly accurate word segmentation.
- We show that it is possible to select word-segmentation results and words related to a particular place by using the mutual information. In addition, we show that the selection of word-segmentation candidates by using this mutual information is theoretically valid.

The remainder of this paper is organized as follows. In Section 2, we discuss previous studies on lexical acquisition and semantic mapping that are relevant to our study. In Sections 3 and 4, we

[1] latticelm is an unsupervised word-segmentation tool that [4] is implemented and is treated as the name of the method in this study.

present our conventional SpCoA and proposed SpCoA++ methods, respectively. In Sections 5 and 6, we discuss the effectiveness of the proposed method in a simulation and a real environment, respectively. Section 7 concludes the paper.

2. Related work

2.1. Iterative update of a language model for lexical acquisition

Heymann et al. [6] proposed an iterative optimization method that alternately updates the phoneme recognition results and the language model by using the method of unsupervised word segmentation. As a result, they showed that it is possible to improve the accuracy of word segmentation. However, their study used only continuous speech signals rather than other information with high co-occurrence (e.g., place or object). We reason that accuracy of the lexical acquisition would be improved by supplementing speech information with other information. Araki et al. [1] proposed a method for learning object concepts and word meanings from multimodal information and spoken sentences, which were segmented using an unsupervised morphological analyzer based on the NPYLM [5]. However, the disadvantage of using the NPYLM is that phoneme sequences with errors do not result in appropriate word segmentation. In this case, the speech-recognition results include many phoneme-recognition errors, and the word-segmentation results include many incorrect boundaries due to phoneme recognition errors. Therefore, as a similar approach, Nakamura et al. [7,8] proposed a mutual learning method based on integrating the learning of object concepts with a language model. However, they used the NPYLM, which does not consider phoneme recognition errors, for the word segmentation of N-best speech-recognition results. Taniguchi et al. [3] addressed these NPYLM-related problems, and reduced the errors and variability of speech-recognition results by using latticelm [4]. In this paper, we propose the SpCoA++ method that iteratively constructs spatial concepts and a language model, and that performs word segmentation and spatial concept acquisition more accurately.

2.2. Lexical acquisition of objects

Studies on lexical acquisition typically focus on learning lexicons regarding objects [1,7–17]. Roy et al. [9] proposed a computational model by which a robot could learn the names of objects from object images and natural infant-directed speeches. Their results showed that the model could perform speech segmentation, lexical acquisition, and visual categorization. Iwahashi et al. [11,12] developed the LCore on-line machine learning system that enables a robot to learn communicative capabilities through speech and behavioral interactions with a person in a probabilistic framework. Qu and Chai [13,14] proposed an unsupervised learning method that automatically acquires novel words for an interactive system. They focused on the co-occurrence between speech and eye gaze and the use of domain knowledge in lexical acquisition. Hornstein et al. [15] proposed a methodology that mimics a human infant for the purposes of language acquisition in humanoid robots. Their model acquires the language from visual and auditory information by interaction with a person. It acquires words and phonemes by pattern recognition and hierarchical clustering without any pre-programmed linguistic knowledge. Attamimi et al. [16] proposed a method for learning novel objects using word segmentation of out-of-vocabulary phoneme sequences, as well as a method for object detection. Nakamura et al. [17] proposed a categorization method based on multimodal latent Dirichlet allocation (MLDA) that enables the acquisition of object concepts from multimodal information, i.e., visual, auditory, haptic, and verbal information. Many of these studies have been unable to address the lexical acquisition of words other than those related to objects, e.g., words related to space and places. However, to incorporate robots into the human living environment, the robots must be able to learn a lexicon that is related to not only objects but also places. In general, places are different from objects in that the boundary of a place is ambiguous. In addition, the names of places, the areas of places, and the number of places can all differ according to the user and the home environment. Therefore, the robot has to acquire the spatial concepts adaptively depending on the environment. In this study, we focus on robust word discovery related to places. The proposed method enables a robot to learn the names of places with high accuracy in various human living environments. We conclude that a robot can obey a spoken user instruction that involves spatial concepts and movements more effectively if it learns a more accurate place-related lexicon.

2.3. Semantic mapping, place categorization, and spatial lexical acquisition

The related research fields of semantic mapping and place categorization [18] have attracted considerable interest in recent years. Studies on semantic mapping and place categorization typically focus on grounding concepts, labels, or symbols to a geometric or topological map constructed by using simultaneous localization and mapping (SLAM) [19] in a form that can be understood by a robot. In most of these studies, the robot can create a semantic map that associates only a preset vocabulary that is given in advance. However, the robot cannot acquire novel words from human speech signals. We consider our study to be in the field of semantic mapping in the sense that the spoken language is grounded to a map. Therefore, our study is a challenging research problem based on the unsupervised manner in which a robot autonomously learns place categories including novel words from its own experience of the surrounding space (i.e., control data, sensor data, and speech information). The following previous studies have addressed semantic mapping, place categorization, and language acquisition related to places.

Cummins and Newman [20] proposed the FAB-MAP visual SLAM system that uses image features converted into a bag-of-visual-words representation in the space of appearance. Their method was able to detect loop closure by matching appearance information alone; it did not use linguistic knowledge such as speech or word information. Therefore, the robot could not understand words related to places in human–robot interaction.

Walter et al. [21] proposed an algorithm that enables robots to efficiently learn semantic graphs that integrate semantic representations from natural language descriptions such as labels and spatial relationships with a metric map. Welke et al. [22] proposed a method for learning a spatial representation by integrating the representation between the discrete symbolic entities used in high-level reasoning and the continuous state space of the sensorimotor level. Their method can estimate a probable spatial area and a word of an object by using the spatial lexical knowledge and the position of the object. Bastianelli et al. [23] developed a system for on-line semantic mapping, based on multimodal human–robot interaction. Their system allows to acquire the names and positions of new objects in the map. In addition, Bastianelli et al. [24] proposed a discriminative approach using semantic mapping and natural language processing. They improved the lexical generalization of spoken commands by adopting distributional models of lexical semantics, i.e., word2vec [25]. However, these studies [21–24] did not consider the acquisition of novel words from human speech signals. To have a robot learn new words accurately without any pre-existing vocabulary can be regarded as a very difficult problem. Our study focuses on robust word discovery by using the spatial information from speech signals.

Milford et al. [26] implemented a method that enables a robot to learn spatial concepts using RatSLAM [27], which is inspired by biological knowledge. In addition, they implemented mobile robots that learn lexical knowledge through robot-to-robot communication. These robots are called Lingodroids [28–31]. In [29,31], two robots that had different sensors and SLAM algorithms were used. These studies reported that the robots created their own vocabulary. Spranger and Steels [32] proposed a model for co-acquisition of the syntax and semantics of the spatial language. Experimental results showed that the robot could learn spatial grammar and spatial categories related to direction. However, these studies did not consider lexical acquisition by human-to-robot speech interactions.

Taguchi et al. [33,34] proposed an unsupervised method for learning words and relationships between objects and phoneme sequences from various user speeches without any prior linguistic knowledge other than an acoustic model. In addition, they implemented a method for simultaneously categorizing self-positions and lexicon [35]. These experimental results showed that the method was able to learn the names of places from user speech and to understand words related to places in different positions that were not used for learning. However, these studies were unable to use the learned words for self-localization of the robot. In our study, the proposed method can use words related to places for self-localization. The strengths of our study are that learning spatial concepts and self-localization are represented as one generative model, and that robots can use the learned words to self-localize autonomously. We have previously reported experimental results for self-localization using spatial concepts in SpCoA [3].

Taniguchi et al. [36] proposed a method for simultaneously estimating self-position and words from noisy sensory information and utterances. Their method integrated ambiguous speech-recognition results with the self-localization method for learning spatial concepts. However, the method of [36] assumed that a name of a place would be learned from utterances of an isolated word. Our proposed method using SpCoA can learn names of places from uttered sentences including multiple words. As a similar approach to that of [3,36], Ishibushi et al. [37] proposed a self-localization method that exploits image features using a convolutional neural network (CNN) [38]. The experimental results showed that the method was able to converge particles for self-localization and to reduce estimation errors in the global self-localization. Hagiwara et al. [39] implemented a method based on hierarchical MLDA (hMLDA) [40] for learning place concepts using position and visual information. The experimental results demonstrated the formation of hierarchical place concepts by hMLDA. We consider that [37,39] can be explained as unsupervised place categorization. However, they did not perform the lexical acquisition using speech signals or word information. We believe that a robot can learn multimodal spatial concepts from positions, visual images, and words by integrating image features using our proposed method.

3. Nonparametric Bayesian spatial concept acquisition method (SpCoA)

This model enables robots to learn the relationship between words and places [3]. The model developed for spatial concept acquisition is based on unsupervised word segmentation and a nonparametric Bayesian generative model that integrates self-localization and clustering in both words and places. The self-localization method adopts Monte Carlo localization (MCL) [41], a method that is used for the localization of mobile robots. More details are given in [3].

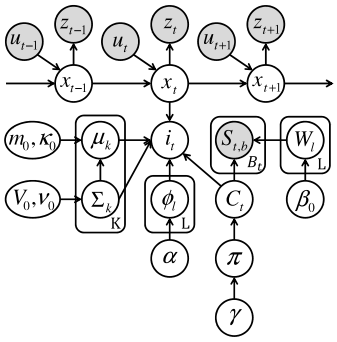

Fig. 2. Graphical model of SpCoA for spatial concept acquisition [3].

3.1. Overview and main features of SpCoA

Fig. 2 shows the graphical model for the acquisition of spatial concepts, and Table 1 lists each variable of the graphical model. The number of words of the sentence uttered at time t is denoted as B_t. The main features of this model and the model structure are as follows.

(1) This model can learn many-to-many correspondences between names and places by relating several places to several names via spatial concepts. Specifically, spatial concepts are represented by the word distribution W_l of the place names 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 place names.
(2) It can learn an appropriate number of spatial concepts, depending on the data, by using a nonparametric Bayesian approach. Specifically, this method uses the stick-breaking process (SBP) [42], a method based on the Dirichlet process. Therefore, this method can consider theoretically infinite numbers L of spatial concepts and K of position distributions. In this paper, we approximate a number of parameters by setting sufficiently large values, i.e., a weak-limit approximation [43].
(3) This model can learn words related to places from continuous speech signals. This method uses an unsupervised word-segmentation method known as latticelm [4], which can use speech-recognition results in the WFST format to segment continuous speech signals directly. This method reduces the variability of speech recognition.

3.2. Spatial concept acquisition by Gibbs sampling

This model learns by using batch estimation based on multiple training data elements. In this case, a set of taught times is denoted by $T_o = \{t_1, t_2, \ldots, t_N\}$, where N is the number of training data elements. An uttered sentence is converted into a bag-of-words (BoW) representation. The model parameters are denoted as $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}, \pi\}$. Then, the sets of variables are denoted as $\mathbf{W} = \{W_1, \ldots, W_L\}$, $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_L\}$, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$, and $\boldsymbol{\Sigma} = \{\Sigma_1, \ldots, \Sigma_K\}$. Furthermore, the sampling values of the model parameters from the following joint posterior distribution are obtained by performing Gibbs sampling:

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

where the hyperparameters of the model are denoted as $\mathbf{h} = \{\alpha, \gamma, \beta_0, m_0, \kappa_0, V_0, \nu_0\}$. The initial values of the model parameters can be set arbitrarily in accordance with a condition. More details are given in [3].

Table 1
Each element of the graphical model of SpCoA.

x_t	Self-position of robot
u_t	Control data
z_t	Sensor data (range data)
C_t	Index of spatial concepts
i_t	Index of position distributions
$S_{t,b}$	bth segmented word in time t
π	Multinomial distribution of index C_t of spatial concepts
ϕ_l	Multinomial distribution of index i_t of position distribution
W_l	Multinomial distribution of place names
μ_k, Σ_k	Position distribution Gaussian (mean vector, covariance matrix)
α, γ, β_0, m_0, κ_0, V_0, ν_0	Hyperparameters

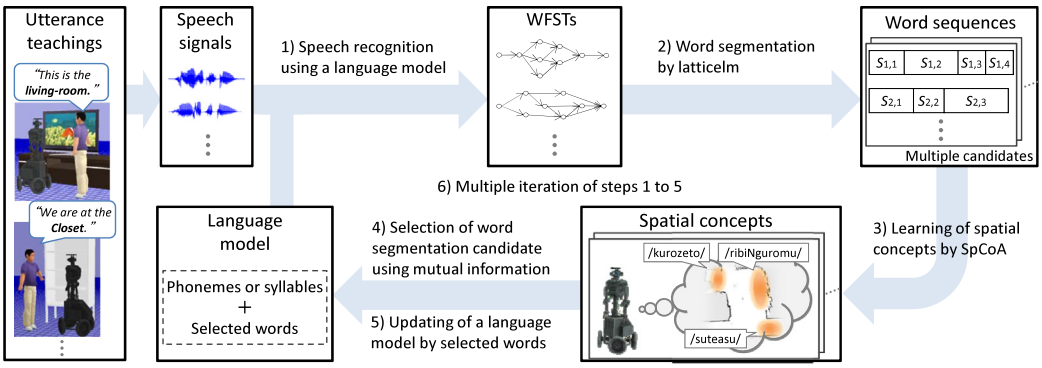

Fig. 3. Flow of iterative estimation of SpCoA++. This figure corresponds to steps 1–6 of the procedure in 4.1: (1) the robot gets WFSTs from speech signals of utterance teachings by speech recognition using a language model; (2) the robot gets various candidate word-segmentation results by unsupervised word segmentation using WFSTs; (3) the robot learns spatial concepts by SpCoA for each word-segmentation candidate; (4) the robot selects a word-segmentation candidate using mutual information; (5) the robot registers all the selected words in a word dictionary of a language model; (6) the robot performs multiple iterations of steps 1–5.

4. Iterative estimation of spatial concepts and language model (SpCoA++)

The proposed SpCoA++ method uses SpCoA to acquire the spatial concepts and then uses these to update the language model iteratively. We describe a method that mutually learns spatial concepts, which the method then uses to update the language model. SpCoA++ is capable of taking into account spatial concepts obtained from segmented results based on unsupervised word segmentation using the mutual information. In addition, we propose a method for selecting a word related to a particular place from segmented words. In this study, our approach is to select word-segmentation results, including words related to specific places, and to improve lexical acquisition by updating the language model. This is done by registering all the selected words in the language model's word dictionary, which contains only phonemes or syllables at this stage.

The remainder of this section is organized as follows. In Section 4.1, we describe the procedure of the proposed SpCoA++ method. In Section 4.2, we describe the graphical and generative models of the proposed method. In Section 4.3, we describe the formulation of speech recognition and unsupervised word segmentation. The mutual information for the selection of word-segmentation results is described in Section 4.4. In Section 4.5, we describe the mutual information by binarized variables (MIB) for the selection of words related to places. In Section 4.6, we describe the learning algorithm of the iterative estimation method based on Markov-chain Monte Carlo (MCMC). In Section 4.7, we describe the validity of the mutual-information maximization criterion.

4.1. Iterative estimation procedure of SpCoA++

This section describes the procedure of SpCoA++ based on the learning of spatial concepts and the language model. The schematic diagram of the flow of iterative estimation is shown in Fig. 3. The procedure of SpCoA++ is discussed as follows.

(1) **Speech recognition of utterances using a language model**
A robot performs speech recognition of continuous speech signals using a current language model. speech-recognition results are obtained in the WFST format.
(2) **Unsupervised word segmentation of WFSTs**
The robot performs unsupervised word segmentation using the speech-recognition results of the WFSTs produced by latticelm [4]. The robot then gets several word-segmentation results that represent various candidates of words.
(3) **Learning of spatial concepts**
The robot learns spatial concepts by using SpCoA for each word-segmentation candidate.
(4) **Selection of word-segmentation results**
A word-segmentation candidate is selected by using the maximum value of the mutual information between words $S_{t,b}$ and the indices of the spatial concepts C_t. The calculation of this mutual information is described in detail in Sections 4.4 and 4.5.
(5) **Updating of language model**
The robot registers all the words of a selected word-segmentation result with the word dictionary.
(6) **Iteration**
Multiple iterations are performed for the process described in steps 1–5.
(7) **After completion of the iteration**
The final learning result of the spatial concept is based on the final word-segmentation result obtained at the end of the iteration process.

A pseudo-code for the above procedure is given in Algorithm 1.

4.2. Graphical model and generative model

Fig. 4 shows the graphical model of SoCoA++. Table 2 lists the new variables of the proposed method extended from SpCoA.

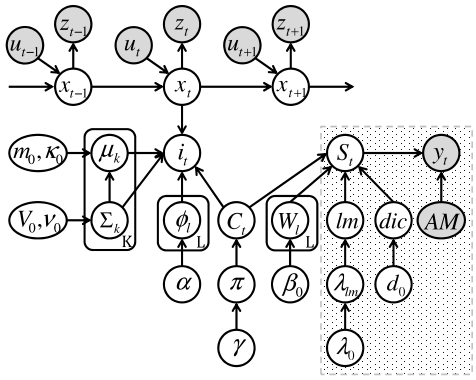

Fig. 4. Graphical model of the proposed SpCoA++ method. The rectangular shaded area represents the extension from SpCoA. The extended parts provide the language model and speech recognition.

Table 2
New variables of the proposed graphical model.

S_t	Sequence of words (word-segmentation result of the uttered sentence)
y_t	Speech data
AM	Acoustic model for speech recognition
lm	N-gram language model
λ_{lm}	Parameter of n-gram length for language model (the value of n for the n-gram)
λ_0	Parameter of prior distribution for n-gram length
dic	Word dictionary of a language model
d_0	Parameter of prior distribution for word dictionary (phonemes or syllables)

The proposed graphical model of SpCoA++ is made by adding a language model, an acoustic model, and an speech signal to the model of SpCoA. By simultaneous estimation of spatial concept, language model and word dictionary, SpCoA++ achieves more accurate spatial lexical acquisition than SpCoA. We describe the generative model newly defined by changing the graphical model as follows:

$$\lambda_{lm} \sim p(\lambda_{lm} \mid \lambda_0), \tag{2}$$

$$lm \sim p(lm \mid \lambda_{lm}), \tag{3}$$

$$dic \sim p(dic \mid d_0), \tag{4}$$

$$S_t \sim p(S_t \mid \mathbf{W}, C_t, lm, dic), \tag{5}$$

$$\overset{\text{UR}}{\approx} p(S_t \mid lm, dic) \prod_{B_t} \frac{p(S_{t,b} \mid W_{C_t})}{p(S_{t,b})}, \tag{6}$$

$$y_t \sim p(y_t \mid S_t, AM). \tag{7}$$

We assumes that Eq. (7) is corresponding to a process of generating speech signal from a sequence of words S_t and an acoustic model AM. In addition, S_t is generated by a language model lm rescaled by the word distribution W_{C_t} and the word dictionary dic in Eq. (6). Eq. (5) is approximated by using unigram rescaling [44] as shown in Eq. (6), were $\overset{\text{UR}}{\approx}$ represents the approximation by unigram rescaling.

4.3. Probabilistic formulation of speech recognition and unsupervised word segmentation

The 1-best speech recognition and the WFST speech recognition are represented respectively as follows:

$$S_t \sim p(S_t \mid y_t, AM, dic), \tag{8}$$

$$Lattice_t \sim p(Lattice_t \mid y_t, AM, dic), \tag{9}$$

where $Lattice_t$ denotes a speech-recognition result in the WFST format.

The unsupervised WFST word segmentation by latticelm is represented as follows:

$$S_{T_o}, lm \sim p(S_{T_o}, lm \mid Lattice_{T_o}, \lambda_{lm}). \tag{10}$$

The probability distribution of Eq. (11) can be calculated as Eq. (12) via $Lattice_{T_o}$. The speech-recognition result of the WFST-format $Lattice_{T_o}$ represents multiple candidates of n-best speech-recognition results as a compact graph format. Therefore, a WFST speech-recognition result is the approximate representation of all speech-recognition results:

$$p(S_{T_o}, lm \mid \lambda_{lm}, y_{T_o}, AM, dic) \tag{11}$$

$$\approx p(S_{T_o}, lm \mid Lattice_{T_o}, \lambda_{lm}) \; p(Lattice_{T_o} \mid y_{T_o}, AM, dic). \tag{12}$$

4.4. Selection of word-segmentation results using mutual information

The mutual information between words $S_{t,b}$ and indices of spatial concept C_t is calculated as follows:

$$\mathrm{MI}(S_{t,b}; C_t \mid \Theta) = \sum_{S_{t,b}} \sum_{C_t} p(S_{t,b}, C_t \mid \Theta) \log \frac{p(S_{t,b}, C_t \mid \Theta)}{p(S_{t,b} \mid \Theta) p(C_t \mid \Theta)}. \tag{13}$$

The iterative selection using the mutual information is designed to select the word-segmentation result that includes many words that more appropriately segment the names of places. The updated language model can be used effectively for the next round of speech recognition.

4.5. Selection of words using mutual information by binarized variables

In Section 4.4, the iterative estimation is calculated by using all learned words. In this case, words that do not relate to places can affect the mutual information of Eq. (13) as noise, e.g., *"is"*, *"this"*, and *"a"*. Therefore, we described the procedure for selecting word-segmentation results considering the effects of only words related to places. We proposed a method that extracts words related to places by generating the MIB. Basically, the MIB indicates a high value with respect to a pair having a high co-occurrence between a specific place and a specific word. The robot can automatically distinguish words that do not relate to places as stop-words by using the MIB. The set of all of words is denoted as S_{all}. This mutual information between a word $s \in S_{all}$ and an index of spatial concept $c \in \{1, 2, \ldots, L\}$ is calculated in Eq. (14). Let $S \in (s, \bar{s})$ and $C \in (c, \bar{c})$. Then,

$$\mathrm{MIB}(s; c \mid \Theta) = \sum_{S,C} p(S, C \mid \Theta) \log \frac{p(S, C \mid \Theta)}{p(S \mid \Theta) p(C \mid \Theta)}, \tag{14}$$

where an index of the spatial concept c is an index of the calculation object, and $\bar{c}$ represents all indices excluding index c. A word s is a word of the calculation object, and $\bar{s}$ represents all words excluding the word s.

Firstly, a set of words S_{select} of which the MIB is equal to or greater than the threshold ϵ is determined as follows:

$$S_{select} = \{s \in S_{all} \mid \mathrm{MIB}(s; c \mid \Theta) \geq \epsilon\}. \tag{15}$$

The threshold ϵ can be arbitrarily set to an empirically lead value to include several words for each spatial concept. Multinomial distributions of the names of places $\mathbf{W}$ were recreated by using the set

Algorithm 1 Algorithm for learning of SpCoA++

Observe N speech signal data y_{T_o}
Setting of hyperparameters $\mathbf{h} = \{\alpha, \gamma, \beta_0, m_0, \kappa_0, V_0, \nu_0, \lambda_0, d_0\}$
Initialize parameters $i_{T_o}, C_{T_o}, \Theta = \{\mathbf{W}, \boldsymbol{\mu}, \Sigma, \boldsymbol{\phi}, \pi\}$
for $j = 1$ to J **do** // Iteration of sampling of model parameters of SpCoA++
  for $m = 1$ to M **do** // Calculating sample for each multiple candidate
    $\lambda_{lm}^{[m]} \sim p(\lambda_{lm} \mid \lambda_0)$ // Preparing m hyperparameter candidates of latticelm
    $S_{T_o}^{[m]}, lm^{[m]} \sim p(S_{T_o}, lm \mid \lambda_{lm}^{[m]}, y_{T_o}, AM, dic)$ // Speech recognition and unsupervised word segmentation
    $\Theta^{[m]} \sim p(x_{0:T}, i_{T_o}, C_{T_o}, \Theta \mid S_{T_o}^{[m]}, u_{1:T}, z_{1:T}, \mathbf{h})$ // Learning spatial concepts by SpCoA
    $dic^{[m]} \sim p(dic \mid C_{T_o}^{[m]}, S_{T_o}^{[m]}, \mathbf{W}^{[m]}, lm^{[m]}, d_0)$ // Registering all words in sample m to word dictionary
  end for
  $m^* = \arg\max_m \mathrm{MI}(S_{t,b}; C_t \mid \Theta^{[m]})$ // Selecting sample by maximum-mutual-information criterion
  $\Theta, lm, dic = \Theta^{[m^*]}, lm^{[m^*]}, dic^{[m^*]}$
end for
return Θ, lm, dic

of words S_{select} as W_{select}, which is the normalized probability value. This operation means replacing S_{all} and $\mathbf{W}$ with S_{select} and W_{select} respectively. Finally, the mutual information for the selection of word-segmentation results is calculated using S_{select} and W_{select} in Eq. (13) of Section 4.4.

4.6. Learning algorithm of SpCoA++

Algorithm 1 is the batch learning algorithm of SpCoA++. The number of iterations for the estimation of model parameters is denoted as J. The number of candidates for the word-segmentation results is denoted as M. We describe the model parameter estimation of the graphical model of SpCoA++ by using sampling. The set of parameters of spatial concepts is $\Theta = \{\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}, \pi\}$, and the set of hyperparameters is $\mathbf{h} = \{\alpha, \gamma, \beta_0, m_0, \kappa_0, V_0, \nu_0, \lambda_0, d_0\}$. The initial values of the parameters can be set arbitrarily in accordance with a environmental condition. The joint posterior distribution of the model parameters is

$$p(x_{0:T}, i_{T_o}, C_{T_o}, S_{T_o}, \Theta, lm, \lambda_{lm}, dic \mid u_{1:T}, z_{1:T}, y_{T_o}, AM, \mathbf{h}). \tag{16}$$

The model parameters of Eq. (16) are estimated using an MCMC framework similar to that of Gibbs sampling. In one iteration of the SpCoA++ sampling algorithm, a more accurate sampling value can be obtained by applying weighted sampling to multiple sampling values, where the weighting is that of the mutual information. Multiple sampling values are obtained from each conditional distribution for the unsupervised word segmentation from speech signals, spatial concept acquisition, and updating the word dictionary. Therefore, the proposal distribution in the weighted sampling method is approximated by Monte Carlo sampling. As an approximation of weighted sampling, this sampling method selects the sample with the maximum weighted value. When the proposed method estimates the model parameters from the joint distribution of Eq. (16) using Gibbs sampling, it needs to obtain a sampling value alternately from each conditional distribution shown as follows.

(I) We describe the conditional probability distribution for speech recognition and unsupervised word segmentation:

$$\begin{aligned} &p(S_{T_o}, lm, \lambda_{lm} \mid C_{T_o}, \mathbf{W}, y_{T_o}, AM, dic, \lambda_0) \\ &\overset{\mathrm{UR}}{\approx} \prod_{T_o} \prod_{B_t} \frac{p(S_{t,b} \mid \mathbf{W}, C_t)}{p(S_{t,b})} \\ &\qquad p(S_{T_o}, lm \mid \lambda_{lm}, y_{T_o}, AM, dic) p(\lambda_{lm} \mid \lambda_0). \end{aligned} \tag{17}$$

The first term of Eq. (17) corresponds to the mutual information. The second term is the probability distribution for the speech recognition and unsupervised word segmentation. The third term is the probability distribution regarding the n-gram length. However, it is difficult to directly sample the parameters S_{T_o}, lm, and λ_{lm} from Eq. (17). Therefore, we consider sampling the parameters from Eq. (18) as follows:

$$\begin{aligned} &p(S_{T_o}, lm, \lambda_{lm} \mid y_{T_o}, AM, dic, \lambda_0) \\ &\quad \approx p(S_{T_o}, lm \mid \lambda_{lm}, y_{T_o}, AM, dic) p(\lambda_{lm} \mid \lambda_0). \end{aligned} \tag{18}$$

Then, we assume $p(\lambda_{lm} \mid \lambda_0)$ as a discrete uniform distribution. We samples the m number of λ_{lm} from $p(\lambda_{lm} \mid \lambda_0)$ as candidates of n-gram length for latticelm. The language model has word n-gram and character n-gram because latticelm is based on NPYLM. The parameter of n-gram length λ_{lm} is represented by two n-gram length pairs of word and character, i.e., $\lambda_{lm} = (\lambda_{wlm}, \lambda_{ulm})$. The hyperparameter λ_0 is a pair of a set of possible values of λ_{wlm} and λ_{ulm} for the uniform distribution.

(II) We describe the conditional probability distribution for learning spatial concepts:

$$\begin{aligned} &p(x_{0:T}, i_{T_o}, C_{T_o}, \Theta \mid S_{T_o}, lm, u_{1:T}, z_{1:T}, \mathbf{h}, dic) \\ &\quad = \frac{p(lm \mid \mathbf{W}, C_{T_o}, S_{T_o}, \mathbf{h}, dic)}{p(lm \mid S_{T_o}, \mathbf{h}, dic)} \\ &\qquad p(x_{0:T}, i_{T_o}, C_{T_o}, \Theta \mid S_{T_o}, u_{1:T}, z_{1:T}, \mathbf{h}, dic). \end{aligned} \tag{19}$$

Then, the probability distribution regarding lm is

$$\begin{aligned} &\frac{p(lm \mid \mathbf{W}, C_{T_o}, S_{T_o}, \mathbf{h}, dic)}{p(lm \mid S_{T_o}, \mathbf{h}, dic)} \\ &\quad = \frac{p(S_{T_o} \mid \mathbf{W}, C_{T_o}, lm, dic) p(S_{T_o})}{p(S_{T_o} \mid \mathbf{W}, C_{T_o}) p(S_{T_o} \mid lm, dic)} \overset{\mathrm{UR}}{\approx} 1. \end{aligned} \tag{20}$$

Therefore, Eq. (19) can represent only the posterior distribution of the parameters for learning spatial concepts. We can estimate these parameters using Gibbs sampling in the same way as SpCoA [3].

(III) We describe the conditional probability distribution for updating the word dictionary as follows:

$$p(dic \mid C_{T_o}, S_{T_o}, \mathbf{W}, lm, d_0). \tag{21}$$

Then, an estimated result from the posterior probability of Eq. (21) is

$$dic = \{d_0 \cup O_{\mathbf{B}}\}, \tag{22}$$

where $O_{\mathbf{B}}$ denotes all of the words of a selected word-segmentation result. Eq. (22) represents adding $O_{\mathbf{B}}$ to the initial word dictionary d_0, e.g., Japanese syllables.

We expect to find an estimated value of the model parameters by repeating the parameter sampling of (I), (II), and (III). However, we obtain an estimated value from Eq. (18) instead of Eq. (17) because it is difficult to directly sample the parameters from Eq. (17) in (I). Therefore, we perform weighted sampling using the target distribution $P = (16)$ and the proposal distribution Q. We define the values obtained by performing Gibbs sampling from Eqs. (18), (19) and (21) as sampling values from the proposal distribution Q. The weight is denoted as $\omega = P/Q$:

$$P = \frac{P}{Q}Q = \omega Q. \tag{23}$$

Then, the weight ω is

$$\omega = \prod_{T_o}\prod_{B_t}\frac{p(S_{t,b} \mid \mathbf{W}, C_t)}{p(S_{t,b})}. \tag{24}$$

Eq. (24) is the second term of the unigram rescaling and can be represented as the mutual information between $S_{t,b}$ and C_t. In Section 4.7, we describe a relationship between the weight ω and the mutual information.

We approximate the proposal distribution Q by Monte Carlo sampling. We perform the weighted sampling by selecting a sampling candidate with a maximum value of the weight ω as the approximation of one iteration of Gibbs sampling:

$$P \approx \frac{1}{M}\sum_{m=1}^{M}\omega^{[m]}Q^{[m]}. \tag{25}$$

4.7. Relationship between weight ω and mutual information

We describe the validity of using the criterion of maximizing the mutual information. The proposed method selects the candidate with the maximum value of $\mathrm{MI}(S_{t,b}; C_t \mid \Theta)$ from multiple candidates. In the sampling procedure, Eq. (24) is obtained by performing unigram rescaling. We want to select the candidate with the maximum value in Eq. (24).

We take the log of both sides in Eq. (24). In this case, the candidate with the maximum parameter values do not change because log is a monotonically increasing function:

$$\log\omega = \sum_{T_o}\sum_{B_t}\log\frac{p(S_{t,b} \mid \mathbf{W}, C_t)}{p(S_{t,b})}. \tag{26}$$

Eq. (26) means that the logarithmic probabilities of data estimated as $S_{t,b} = s$ and $C_t = c$ are added for all $S_{t,b}$ and C_t in all of the training data. The ratio of data estimated as $S_{t,b} = s$ and $C_t = c$ in all of the data is

$$p(S_{t,b} = s, C_t = c \mid \Theta) \approx \frac{n_t^{(s,c)}}{N_w}, \tag{27}$$

where $n_t^{(s,c)}$ denotes the number of data estimated as $S_{t,b} = s$ and $C_t = c$ in all of the data, and N_w denotes the number of words in all of the data. In practice, the parameter Θ is affected by the probability smoothing of the hyperparameters γ and β_0. However, if we set γ and β_0 to sufficiently small values, Eq. (27) is a better approximation of joint probability distribution of s and c.

In addition, the probability of the occurrence of words in all of the training data is

$$\sum_{T_o}\sum_{B_t}\log p(S_{t,b}) \approx \sum_{s\in S_{t,b}}\log p(S_{t,b} = s \mid \Theta). \tag{28}$$

Therefore, Eq. (26) becomes

$$\begin{aligned}
&\sum_{T_o}\sum_{B_t}\log\frac{p(S_{t,b} \mid \mathbf{W}, C_t)}{p(S_{t,b})} \\
&\approx \sum_{s\in S_{t,b}}\sum_{c\in C_t} n_t^{(s,c)}\log\frac{p(S_{t,b} = s \mid \mathbf{W}, C_t = c)}{p(S_{t,b} = s \mid \Theta)} \\
&\propto \sum_{s\in S_{t,b}}\sum_{c\in C_t} p(S_{t,b} = s, C_t = c \mid \Theta) \\
&\qquad \log\frac{p(S_{t,b} = s \mid \mathbf{W}, C_t = c)}{p(S_{t,b} = s \mid \Theta)}.
\end{aligned} \tag{29}$$

Therefore, the sample candidate m^* selected by weighted sampling is

$$\begin{aligned}
m^* &= \operatorname*{argmax}_m \prod_{T_o}\prod_{B_t}\frac{p(S_{t,b}^{[m]} \mid \mathbf{W}^{[m]}, C_t^{[m]})}{p(S_{t,b}^{[m]})} \\
&\approx \operatorname*{argmax}_m \mathrm{MI}(S_{t,b}; C_t \mid \Theta^{[m]}).
\end{aligned} \tag{30}$$

5. Experiments I: Simulator

5.1. Conditions

In this experiment, we validate the evidence of the proposed SpCoA++ method in an environment simulated on the simulator platform SIGVerse[2] [45], which enables the simulation of social interactions. The speech recognition is performed using the Japanese continuous-speech-recognition system Julius[3] [46,47]. The set of 43 Japanese phonemes defined by the speech database committee of the Acoustical Society of Japan (ASJ) is adopted by Julius [46]. The representation of these phonemes is also adopted in this study. The Julius system uses a word dictionary containing 115 Japanese syllables. A microphone (SHURE PG27-USB) is attached to the head of the robot. Furthermore, we use an unsupervised morphological analyzer[4] (NPYLM [5] and latticelm [4]). The mobile robot model consists of an omni-directional mobile base and an upper body humanoid. A range sensor is attached to the mobile base. The robot model and the room environment are shown in Fig. 5. The robot can move by performing forward, backward, right or left rotation movements on a two-dimensional plane. In this experiment, the robot can use an approximately correct map of the considered environment.

5.2. Comparison methods

For comparison purposes, we compare the performances of eight methods.

(A) SpCoA [3]
SpCoA performs the learning of spatial concepts. This method estimates the model parameters in Eq. (1) from the word-segmentation results without the iterative estimation. The parameters of the latticelm tool are the initial settings.

(B) Iterative optimization [6] (using NPYLM)
Iterative optimization performs speech recognition repeatedly using the updated language model, and unsupervised word segmentation using the recognized phoneme sequences. This method explains the iterative estimation

[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/.

[4] latticelm 0.4, http://www.phontron.com/latticelm/.

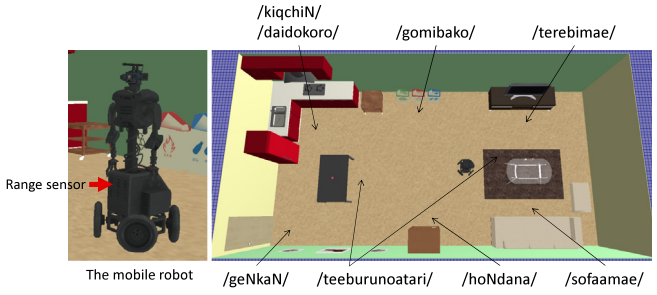

Fig. 5. (left) The robot model: the robot has a range sensor in front and performs self-localization by an occupancy grid map. The red arrow shows the position of the range sensor. (right) Environment to be used for learning and localization on SIGVerse: this is a pseudo-room in the simulated real world. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

method without the selection of word-segmentation results. We register a word-segmentation result directly to the language model without multiple candidates for word-segmentation results. We use 1-best speech-recognition results, which are segmented using NPYLM [5].

(C) Iterative optimization [6] (using latticelm)
As (B), except we use speech-recognition results of the WFST format, which are segmented using latticelm.

(D) Nakamura et al. [7,8]
Nakamura's method [7,8] is applied for learning spatial concepts. In this case, the MLDA for learning object concepts correspond to SpCoA. This method uses N-best speech-recognition results, which are segmented using NPYLM.

(E) SpCoA++ (without the selection of words)
The proposed method performs iterative estimation in order to learn spatial concepts and update the language model.

(F) SpCoA++
As (E), except this method selects word-segmentation results from a set of words related to spatial concepts, as discussed in Section 4.5.

(G) SpCoA (using Japanese syllables and minimum word dictionary)
This method learns spatial concepts using speech recognition with word information. We register Japanese syllables and teaching words (correct phoneme sequences) to a word dictionary of Julius. The registered words include all of the words described in Table 3. The language model assumes that all words are obtained with equal probability. The robot learns spatial concepts from speech-recognition results using a minimum word dictionary that includes Japanese syllables and all the teaching words.

(H) SpCoA (using the existing large-vocabulary word dictionary of Julius)
As (G), except this method uses an existing large-vocabulary dictionary of Julius for speech recognition and learning spatial concepts. The language model included in this system has been developed by using the Balanced Corpus of Contemporary Written Japanese (BCCWJ) compiled and owned by the National Institute for Japanese Language and Linguistics. The word dictionary includes 64,271 words.

5.3. Learning spatial concepts

5.3.1. Conditions

We conducted an experiment to test the spatial concept acquisition in a simulator environment. The environmental map was a two-dimensional occupancy grid map. We used self-localization results obtained by using Monte Carlo smoothing [48] as self-position data $x_{0:T}$ in the training data. The initial particles for self-localization were set according to the true initial position of the robot. The number of particles was $M = 1{,}000$. The teaching utterances were performed a total of 90 times including 10 types of various phrases. The phrases in each uttered sentence are listed in Table 3. The number of teaching places was eight, and the number of teaching names was eight. Each uttered place name is shown in Fig. 5. These utterances include the same name for different places, i.e., */teeburunoatari/* (which means *near the table* in English), and different names for the same place, i.e., */kiqchiN/* and */daidokoro/* (both of which mean *kitchen*). The other teaching names are */geNkaN/* (*entrance* or *doorway*); */terebimae/* (*the front of the TV*); */gomibako/* (*trash box*); */hoNdana/* (*bookshelf*); and */sofaamae/* (*the front of the sofa*). The iterative estimation procedure involved 10 iterations. The number of Gibbs sampling iterations was 100. The hyperparameters were set as follows: $L = 50$, $K = 50$, $\alpha = 1.5$, $\gamma = 8$, $\beta_0 = 0.5$, $m_0 = [0, 0]^{\mathrm{T}}$, $\kappa_0 = 0.001$, $V_0 = \mathrm{diag}(1000, 1000)$, $\nu_0 = 2$, and $\lambda_0 = (\{2, 3, 4\}, \{3, 4\})$. The above hyperparameters were set so that all methods in the comparison were tested under the same conditions. The number of candidates for word-segmentation results was six. The threshold of the mutual information for word selection was specified as $\epsilon = 0.01$. The hyperparameters have been determined manually, i.e., empirically.

Table 3
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 *"We are at the **."* 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

5.3.2. Results

The learning results of spatial concepts obtained using the proposed method are presented here. The number of estimated spatial concepts was seven. The number of estimated position distributions was nine. Fig. 6 shows the position distributions learned on the map of the experimental environment. Table 4 lists the five best elements in terms of the MIB value 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. The proposed method learned almost the exact place name(s) corresponding to each learning target. In addition, the word segmentation of place names was mostly accurate, whereas the acquired words included a few phoneme-recognition errors. Moreover, the experimental results show that words relating to spatial concepts can be determined by selecting words using MIB.

5.4. Comparative evaluation of methods

5.4.1. Conditions

We evaluated the methods according to the following three metrics.

- Estimation accuracy rate of spatial concepts (ARI)
  We compare the matching rate for an estimated index C_t of the spatial concept of each teaching utterance and the classification results of correct answers by a person. The evaluation of this experiment uses the adjusted Rand index (ARI) [49], which is a measure of the degree of similarity between two clustering results.

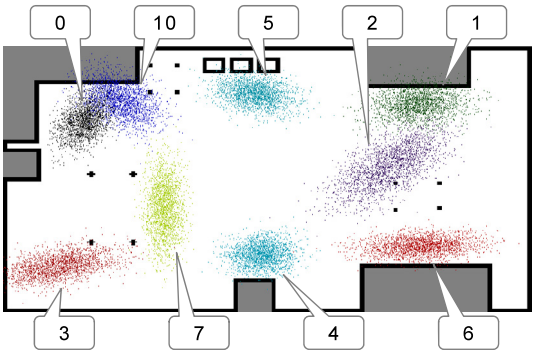

Fig. 6. Learning results of each position distribution. A point group of each color denoting each position distribution was drawn on the map. The colors of the point groups were determined randomly. Furthermore, each index number is denoted as $i_t = k$.

- Place recognition rate using speech signal (PRR)

  When the robot hears user speech y_t including the name of a particular place, the robot estimates a position x_{best} indicated by the uttered sentence. The user says */** ni iqte/* (which means *"Go to **."* in English). The number of utterances is eight. The speech recognition performs 1-best speech recognition using the learned word dictionary. The 1-best speech recognition and the estimation of x_{best} are calculated as follows:

$$S_t \sim p(S_t \mid y_t, AM, dic), \tag{31}$$

$$x_{\text{best}} = \underset{x_t}{\operatorname{argmax}}\, p(x_t \mid S_t, \Theta). \tag{32}$$

  The operations of Eq. (31) and (32) are an approximation of $p(x_t \mid y_t, AM, dic, \Theta)$. In reality, it is difficult to calculate Eq. (32) for all of the position coordinates. Therefore, we use the value of the mean vector μ_{C_t} for each learned position distribution and the 10 position coordinates sampled for each position distribution as candidates for x_{best}. As a justification for this, we consider that a position near the maximum value of a Gaussian distribution becomes a possible candidate for calculating Eq. (32).

  In this experiment, we decided to correct the position within the rectangular area surrounding the position coordinates taught as the same place (including 10-cm margins up, down, left, and right). If the user speaks about */kiqchiN/* or */daidokoro/*, the positions within the area surrounding both teaching positions are correct. If the user speaks about */teeburunoatari/*, the positions within the areas of two places are correct. The place recognition rate (PRR) is calculated as follows:

$$\text{PRR} = \frac{n_{\text{c}}}{n_{\text{u}}}, \tag{33}$$

  where n_{u} denotes the number of utterances and n_{c} denotes the number of correct positions.
- Phoneme accuracy rate of acquired words (PARw)

  We evaluate whether a phoneme sequence learned as the name of a place is properly segmented. This experiment assumes a request for the best phoneme sequence S_{best} representing the self-position x_t for a robot. We compare the phoneme accuracy rate (PAR) with the correct phoneme sequence and a selected word for each teaching place. The matching rate of the phoneme string was calculated by using the PAR as follows:

$$\text{PAR} = 1 - \frac{n_{\text{s}} + n_{\text{d}} + n_{\text{i}}}{n_{\text{p}}}. \tag{34}$$

Table 4
Learning results of high-probability words and indices of the position distribution for each spatial concept.

Index	W_{C_t}		ϕ_{C_t}	
C_t	Word	(MIB)	Index i_t	(Probability)
0	**gomibakoo**	(**0.044**)	**5**	(**0.874**)
	a	(0.014)	11	(0.004)
	tereburunoatari	(0.006)	4	(0.004)
	daidokoro	(0.003)	33	(0.003)
	kiqchiN	(0.003)	28	(0.003)
1	**qgeNkaN**	(**0.063**)	**3**	(**0.873**)
	tereburunoatari	(0.005)	25	(0.004)
	i	(0.003)	4	(0.004)
	kiqchiN	(0.002)	48	(0.004)
	daidokoro	(0.002)	6	(0.003)
2	**kiqchiN**	(**0.043**)	**10**	(**0.509**)
	daidokoro	(**0.041**)	**0**	(**0.422**)
	tereburunoatari	(0.014)	5	(0.002)
	i	(0.008)	3	(0.002)
	sofamae	(0.007)	33	(0.002)
3	**terebimae**	(**0.062**)	**1**	(**0.872**)
	tereburunoatari	(0.005)	49	(0.004)
	i	(0.003)	22	(0.003)
	kiqchiN	(0.003)	7	(0.003)
	daidokoro	(0.002)	3	(0.003)
4	**tereburunoatari**	(**0.091**)	**2**	(**0.493**)
	i	(0.040)	**7**	(**0.436**)
	a	(0.011)	33	(0.002)
	kiqchiN	(0.008)	47	(0.002)
	daidokoro	(0.007)	36	(0.002)
12	**sofamae**	(**0.066**)	**6**	(**0.874**)
	tereburunoatari	(0.005)	3	(0.004)
	i	(0.003)	31	(0.003)
	kiqchiN	(0.002)	16	(0.003)
	a	(0.002)	34	(0.003)
21	**hoNdana**	(**0.061**)	**4**	(**0.875**)
	tereburunoatari	(0.005)	9	(0.003)
	i	(0.003)	37	(0.003)
	daidokoro	(0.002)	43	(0.003)
	kiqchiN	(0.002)	25	(0.003)

The numerator of Eq. (34) is calculated by using the Levenshtein distance between the correct phoneme string and the recognition phoneme string. Here, n_{s} is the number of substitutions, n_{d} is the number of deletions, n_{i} is the number of insertions, and n_{p} is the number of phonemes of the correct phoneme string. The selection of S_{best} is calculated as follows:

$$S_{\text{best}} = \underset{S_{t,b}}{\operatorname{argmax}}\, p(S_{t,b} \mid x_t, \Theta). \tag{35}$$

In this experiments, the self-position x_t used in the evaluation of PARw was not contained in the training dataset. If the method performs more accurate recognition of words and more accurate acquisition of spatial concepts, the PARw indicates a higher value. We consider this evaluation index as an overall measure for evaluating the proposed method.

5.4.2. Results

In these experiments, 10 trials were performed for each method. Table 5 lists the averages of the evaluation values for each method. Bold and underscore indicate the highest evaluation values, and underscore indicates the second highest evaluation values. The proposed methods SpCoA++ (E, F) generated considerably higher evaluation values than those of the others methods. In particular, the proposed method (F) shows the highest values. In the PARw results, SpCoA (A) using latticelm and iterative optimization method (B) using NPYLM have almost the same low

Table 5
Evaluation value of each method.

Methods	ARI	PRR	PARw
(A) SpCoA [3]	0.642	0.867	0.206*
(B) Iterative optimization [6] (using NPYLM)	0.615	0.678	0.250*
(C) Iterative optimization [6] (using latticelm)	0.741	0.889	0.651*
(D) Nakamura et al. [7,8]	0.656	0.578	0.461*
(E) SpCoA++ (w/o the selection of words)	0.853	0.922	0.784*
(F) SpCoA++	**0.935**	**0.978**	**0.863**
(G) SpCoA (using Japanese syllables and minimum word dictionary)	0.679	0.844	0.843*
(H) SpCoA (using the existing large-vocabulary word dictionary of Julius)	0.612	0.689	0.566*

* Significant at 0.01 level in comparison between (F) and each method.

values because the part phoneme sequences learned by (A, B) were minutely segmented. In the PARw values of the two methods (B, D) using NPYLM, method (D) with the selection of word-segmentation results using N-best speech-recognition results shows higher values than those of method (B) using 1-best speech-recognition results. In the comparison of methods (B, D) using NPYLM and methods (A, C, E, F) using latticelm, the methods using latticelm tend to have better evaluation values than those of the methods using NPYLM, particularly in the PRR. In all of the evaluation values of (C, D), the iterative optimization method (C) using latticelm showed higher values than those of method (D) with the selection of word-segmentation results by NPYLM using N-best speech-recognition results. In the methods using latticelm (WFST), all of the evaluation values showed high values in the descending order of (F, E, C, A). Specifically, the proposed methods (E, F) of selecting a candidate from multiple samples have significant advantages over conventional methods. As a result, the proposed method was able to select better word-segmentation candidates by using the mutual information. In addition, the proposed method (F) showed higher values than those of methods (G, H) using speech recognition with word knowledge. Furthermore, we performed a matched-pairs test [50] on the PARw values. We investigated the statistical significance between the proposed method (F) SpCoA++ and other methods. As a result, the statistical significance at the 0.01 level was shown in all comparison targets. We consider that the robot could learn more suitable phoneme sequences for learning spatial concepts from speech signal data by the proposed method.

Fig. 7 shows the PARw values for each iteration of the five iterative methods (B)–(F) for comparison. The PARw value tends to increase throughout each iteration. In particular, the proposed methods (E, F) show higher values from the outset. In the first half of each iteration, the evaluation values of each method are roughly similar. However, method (F) shows higher values than those of (E) in the second half of the iteration. We consider that the selection of words using MIB works better for the discovery of words related to places. In method (C), the PARw values increase significantly. However, the PARw of (C) does not reach those of the proposed methods (E, F).

Hence, it is confirmed that the robot can learn the names of places effectively using place information for robust lexical acquisition. The methods (B, D) using NPYLM did not increase sufficiently during the iteration. We consider that the different phoneme sequences (including various errors) existed in the speech-recognition results in the progress of iteration, i.e., methods (B, D) could not reduce the variability of speech recognition effectively. This result suggests that the iterative estimation methods using latticelm can reduce variability in recognition by using phoneme recognition results in the WFST format.

Fig. 8 shows the ARI values for each iteration of the iterative estimation by the proposed method. The black points are for a candidate selected by the proposed method, i.e., the candidate with maximum mutual information. The mutual information is then a

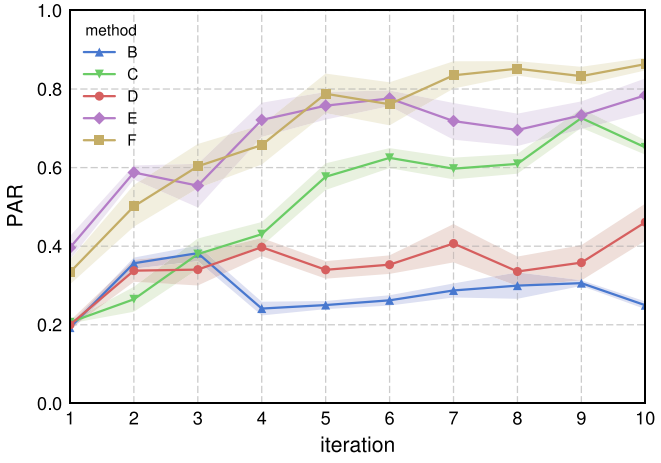

Fig. 7. PARw values for each iteration of the five iterative methods for comparison. The broken line shows the average of PARw on 10 trials for each method. The error bars show the standard errors of PARw on 10 trials for each method. The blue line (B) is iterative optimization [6] (using NPYLM); the green line (C) is iterative optimization [6] (using latticelm); the red line (D) is Nakamura's method [7,8]; the purple line (E) is the proposed SpCoA++ method (w/o the selection of words); the yellow line (F) is SpCoA++. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

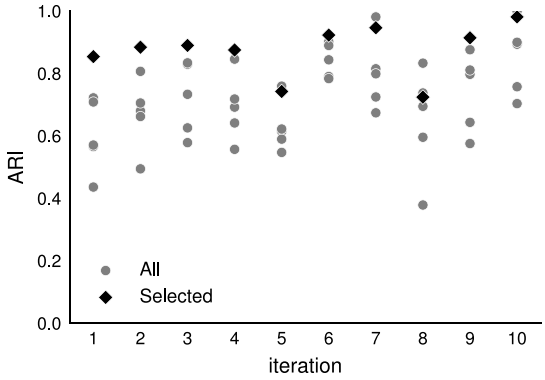

Fig. 8. ARI values for each iteration of the iterative estimation by the proposed method. Plotted gray points (All) are all of the candidates for samples. A black point (Selected) is a candidate selected by the maximization of mutual information.

value estimated by a proposed unsupervised method from learning data. The ARI is a value calculated by comparing an estimated result with the correct classification result. The ARI of the selected candidate was generally the highest value. In addition, we calculated the correlation between ARI values and MI values in 10 trials. The Pearson correlation coefficient showed a comparatively strong correlation ($r = 0.695$). Therefore, the experimental results show

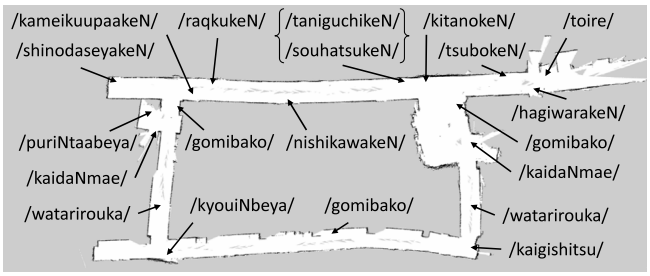

Fig. 9. Map of the experimental environment and teaching words. These teaching words include the same name for different places, i.e., */gomibako/* (*trash box*), */kaidaNmae/* (*the front of staircase*) and */watarirouka/* (*a connecting corridor*); and different names for the same place, i.e., */taniguchikeN/* and */souhatsukeN/* (*emergent systems lab.*). Others are */priNtaabeya/* (*a printing room*); */kyouiNbeya/* (*a teacher's room*); */toire/* (*a rest room*); and */kaigishitsu/* (*a meeting room*). The others are names of the laboratories.

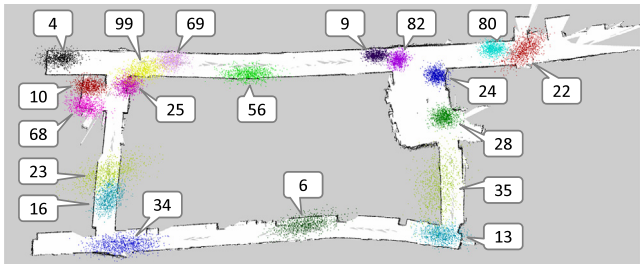

Fig. 10. Learning result of position distributions. A point group of each color denoting each position distribution was drawn on the map. The colors of the point groups were determined randomly. Furthermore, each index number is denoted as $i_t = k$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that by selecting a candidate using the maximization of mutual information, it is possible to select a candidate that is categorizing places accurately.

6. Experiments II: Real mobile robot

In this experiment, the effectiveness of the proposed method was tested by using an autonomous mobile robot TurtleBot 2[5] in a real-world environment. The robot is based on Yujin Robot's Kobuki; a Microsoft Kinect sensor was used as a range sensor. A microphone (SHURE PG27-USB) was attached to the robot. The speech recognition system was Julius. The unsupervised word segmentation system was latticelm.

6.1. Comparison methods

For comparison purposes, we compare the performance of four methods.

(A) SpCoA [3]
(B) Iterative optimization [6] (using latticelm)
(C) SpCoA++ (w/o the selection of words)
(D) SpCoA++.

6.2. Learning of spatial concepts by iterative estimation

6.2.1. Condition

We conducted an experiment to test spatial concept acquisition in an actual environment (an entire floor of a building). In this experiment, the robot performed self-localization using an environment map generated by using the SLAM package of the robot operating system (ROS), i.e., AMCL and gmapping. The generated map of the real environment and true phoneme sequences of the names of places are shown in Fig. 9. The initial particles were set according to the true initial position of the robot. The number of particles was $M = 2{,}000$. The number of teaching places was 19, and the number of teaching names was 16. The teaching utterances were performed a total of 100 times. The phrases in each uttered sentence are listed in Table 3. The iterative estimation procedure involved 10 iterations. The number of Gibbs sampling iterations was 100. The hyperparameters were set as follows: $L = 100$, $K = 100$, $\alpha = 10$, $\gamma = 20$, $\beta_0 = 0.2$, $m_0 = [0, 0]^{\mathrm{T}}$, $\kappa_0 = 0.001$, $V_0 = \mathrm{diag}(1, 1)$, $\nu_0 = 2$, and $\lambda_0 = (\{2, 3, 4\}, \{3, 4\})$. The above parameters were set so that all methods in the comparison were tested under the same conditions. The number of candidates for the word-segmentation results was six. The threshold of the mutual information for word selection was specified as $\epsilon = 0.01$. The hyperparameters have been determined manually, i.e., empirically.

6.2.2. Results

The learning results achieved with the proposed method are as follows. Fig. 10 shows the position distributions learned on the map. Each balloon shows the index number for each position distribution. Table 6 lists the five best elements in terms of the MIB of words of 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. As a result, we found that the proposed method can learn the names of places corresponding to the considered teaching places in the real environment. In addition, the word segmentation of names of places was mostly accurate, whereas the acquired words included a few phoneme-recognition errors.

In the index $C_t = 8$ and 30, several places indicating different place's names are merged into one spatial concept. 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 places. However, there is possibility that classifies different spatial concepts into one spatial concept as a side-effect. We consider that this problem can be improved by integrating visual scene information into the proposed method. Since this paper focused on lexical acquisition, the solution of this problem will be considered in future work.

6.3. Comparative evaluation of methods

6.3.1. Condition

We evaluated the methods according to the following three metrics.

- Estimation accuracy rate of spatial concepts (ARI)
  We compare the matching rate for an estimated index C_t of the spatial concept of each teaching utterance and the classification results of correct answers by a person. The evaluation of this experiment uses ARI.
- Phoneme accuracy rate of uttered sentences (PARs)
  We compare the accuracy rate of phoneme recognition and word segmentation for all recognized uttered sentences. We calculate the matching rate of a phoneme string of a recognition result of each uttered sentence and the correct phoneme string of the training data. This experiment considered the positions of a delimiter as a single letter. The PAR is calculated by Eq. (34).

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

Table 6
Learning result of words and indices of the position distribution for spatial concepts.

C_t	W_{C_t}		ϕ_{C_t}		C_t	W_{C_t}		ϕ_{C_t}	
	Word	(MIB)	i_t	(Prob.)		Word	(MIB)	i_t	(Prob.)
	qkidanokeN	(**0.017**)	**82**	(**0.264**)		**kaigihitsu**	(**0.016**)	**13**	(**0.305**)
	q	(0.012)	34	(0.107)		desu	(0.002)	9	(0.010)
1	dayo	(0.005)	35	(0.103)	21	u	(0.002)	29	(0.010)
	gonomasyogawa	(0.005)	13	(0.053)		gu	(0.002)	58	(0.009)
	a	(0.003)	80	(0.006)		a	(0.001)	91	(0.008)
	nishikyowakeN	(**0.027**)	**56**	(**0.340**)		**kyoibea**	(**0.013**)	**34**	(**0.235**)
	wa	(0.002)	61	(0.008)		u	(0.004)	86	(0.009)
2	q	(0.002)	46	(0.008)	23	ninarimasuq	(0.002)	77	(0.009)
	gokowawa	(0.002)	67	(0.008)		a	(0.002)	82	(0.009)
	gokononamaewa	(0.001)	57	(0.008)		kochigagaa	(0.001)	36	(0.008)
	hagiewarakeN	(**0.022**)	**22**	(**0.503**)		**gomibako**	(**0.041**)	**24**	(**0.143**)
	toire	(**0.022**)	3	(0.007)		**uwatarirooka**	(**0.020**)	**6**	(**0.142**)
4	a	(0.004)	62	(0.006)	30	u	(0.019)	**35**	(**0.094**)
	kaidaNmae	(0.003)	28	(0.006)		nikimajita	(0.008)	**23**	(**0.090**)
	u	(0.002)	48	(0.006)		a	(0.008)	10	(0.048)
	kaidaNmae	(**0.060**)	**28**	(**0.213**)		**qpuriNpabea**	(**0.023**)	**10**	(**0.298**)
	raqguqkeN	(**0.016**)	**68**	(**0.212**)		a	(0.011)	25	(0.125)
8	e	(0.007)	**69**	(**0.166**)	54	qgomibakoo	(0.006)	0	(0.007)
	nayo	(0.004)	46	(0.005)		desu	(0.004)	64	(0.007)
	a	(0.003)	94	(0.005)		kokoga	(0.002)	3	(0.007)
	ashinodaseyaqkeN	(**0.026**)	**4**	(**0.319**)		**paniguchikeN**	(**0.022**)	**9**	(**0.503**)
	e	(0.004)	6	(0.068)		**usohatsukeN**	(**0.016**)	70	(0.007)
12	gokowa	(0.002)	95	(0.008)	60	nikimajita	(0.003)	87	(0.006)
	kaidaNmae	(0.001)	97	(0.008)		gonomashogawa	(0.003)	82	(0.006)
	desu	(0.001)	52	(0.008)		kaidaNmae	(0.003)	18	(0.006)
	kameikupaaqkeN	(**0.026**)	**99**	(**0.320**)		**atsuhokeN**	(**0.028**)	**80**	(**0.291**)
	gokowawa	(0.012)	25	(0.108)		a	(0.002)	21	(0.010)
20	qgomibako	(0.010)	24	(0.057)	65	nayo	(0.001)	71	(0.009)
	kochigagaa	(0.004)	31	(0.007)		wakochiradesuq	(0.001)	95	(0.009)
	q	(0.003)	15	(0.007)		q	(0.001)	38	(0.009)

- Phoneme accuracy rate of acquired words (PARw)
  We evaluate whether a phoneme sequence learned as the name of a place is properly segmented. This experiment assumes a request for the best phoneme sequence $S_{t,\mathrm{best}}$ representing the self-position x_t for a robot. We compare the PAR with the correct phoneme sequence and a selected word for each teaching place. The selection of S_{best} is calculated as Eq. (35).

Table 7
Evaluation value of each method.

Methods	ARI	PARs	PARw
(A) SpCoA [3]	0.570	0.667	0.142
(B) Iterative optimization [6]	0.516	0.700	0.309
(C) SpCoA++ (w/o the selection of words)	0.614	0.755	0.492
(D) SpCoA++	**0.626**	**0.761**	**0.511**

6.3.2. Results

In these experiments, 10 trials were performed for each method. Table 7 lists the evaluation-value averages calculated using the above metrics. The proposed methods (C) and (D) generated higher values than those of the conventional methods (A) and (B) for all metrics. In particular, the proposed method (D) with the selection of words showed the highest values. In the ARI results, SpCoA++ proved to be effective for learning spatial concepts. In the PARs results, we showed that the proposed method was able to perform more accurate phoneme recognition and word segmentation. The PARw results indicated that the proposed method enables the robot to acquire words from speech with more accuracy, and improves the estimation accuracy of the spatial lexical acquisition. As a result, the proposed method was able to select better word-segmentation candidates by using the mutual information. Moreover, the experimental results showed that words relating to the spatial concept could be determined by selecting words using MIB.

7. Conclusions and future work

We proposed the unsupervised machine learning method SpCoA++ for robust spatial lexical acquisition. The proposed method can improve the performance of unsupervised word segmentation from continuous speech signals by using place information. Experimental results demonstrated that the proposed method is capable of estimating the spatial concept with high accuracy, and segmenting the name of a place resulting from an uttered sentence with improved accuracy. As a result, we achieved highly accurate lexical acquisition compared with the conventional methods. In addition, we showed that the candidate with high mutual information could perform not only word segmentation with high accuracy but also clustering of places with high accuracy. Furthermore, we confirmed that it is possible to select a word related to a particular place by using MIB. This result showed that a robot could recognize a word representing a place by extracting words from continuous speech signals. By using the proposed method, robots can adaptively learn the names of places in various human living environments. We consider that the acquired words related to places can be useful for various tasks as applications of the proposed method, e.g., navigation, guidance, and speech interaction.

In this study, we constructed the spatial concepts from position information and word information. However, the proposed method cannot infer a place from a landscape and understand which type of object is in a certain place. As further challenges, we consider that the robot can learn multimodal spatial concepts from positions, visual images, object information, and words by integrating image features [37] and object concepts [7] with the proposed method. By using multimodal spatial concepts, the robot will become able to estimate a name of a place from a visual image and an object existing in a place, e.g., there are cups in the *"kitchen"*. For example, the robot will be able to bring a cup from the kitchen

by the voice command *"Bring a cup"*. Furthermore, in future work, we will implement an online learning method for simultaneous estimation of the spatial concepts and environmental map. By online learning of spatial concepts, we aim to be able to perform additional learning from environmental changes and unknown environments without a map in advance.

References

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

[2] S. Goldwater, T.L. Griffiths, M. Johnson, A bayesian framework for word segmentation: Exploring the effects of context, Cognition 112 (1) (2009) 21–54.

[3] A. Taniguchi, T. Taniguchi, T. Inamura, Spatial concept acquisition for a mobile robot that integrates self-Localization and Unsupervised Word Discovery from Spoken Sentences, IEEE Trans. Cogn. Dev. Syst. 8 (4) (2016) 285–297. http://dx.doi.org/10.1109/TCDS.2016.2565542.

[4] G. Neubig, M. Mimura, T. Kawahara, Bayesian learning of a language model from continuous speech, IEICE Trans. Inf. Syst. 95 (2) (2012) 614–625.

[5] D. Mochihashi, T. Yamada, 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.

[6] J. Heymann, O. Walter, R. Haeb-Umbach, B. Raj, Iterative Bayesian word segmentation for unsupervised vocabulary discovery from phoneme lattices, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.

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

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

[9] D. Roy, A. Pentland, Learning words from sights and sounds: A computational model, Cogn. Sci. 26 (1) (2002) 113–146.

[10] N. Iwahashi, Language acquisition through a human–robot interface by combining speech, visual, and behavioral information, Inform. Sci. 156 (1) (2003) 109–121.

[11] N. Iwahashi, Robots that learn language: A developmental approach to situated human–robot conversations, in: N. Sarkar (Ed.), Human Robot Interaction, InTech, 2007, pp. 95–118.

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

[13] S. Qu, 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.

[14] S. Qu, J.Y. Chai, Context-based word acquisition for situated dialogue in a virtual world, J. Artificial Intelligence Res. 37 (1) (2010) 247–278.

[15] J. Hörnstein, L. Gustavsson, J. Santos-Victor, 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.

[16] M. Attamimi, A. Mizutani, T. Nakamura, K. Sugiura, T. Nagai, N. Iwahashi, H. Okada, T. Omori, Learning novel objects using out-of-vocabulary word segmentation and object extraction for home assistant robots, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2010, pp. 745–750.

[17] T. Nakamura, T. Araki, T. Nagai, N. Iwahashi, Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts, Adv. Robot. 25 (17) (2011) 2189–2206.

[18] I. Kostavelis, A. Gasteratos, Semantic mapping for mobile robotics tasks: a survey, Robot. Auton. Syst. 66 (2015) 86–103.

[19] S. Thrun, W. Burgard, D. Fox, Probabilistic Robotics, MIT Press, 2005.

[20] M. Cummins, P. Newman, FAB-MAP: Probabilistic localization and mapping in the space of appearance, Int. J. Robot. Res. 27 (6) (2008) 647–665.

[21] M.R. Walter, S. Hemachandra, B. Homberg, S. Tellex, S. Teller, Learning semantic maps from natural language descriptions, in: Proceedings of Robotics: Science and Systems (RSS), 2013.

[22] K. Welke, P. Kaiser, A. Kozlov, N. Adermann, T. Asfour, M. Lewis, M. Steedman, Grounded spatial symbols for task planning based on experience, in: Proceedings of the 13th IEEE-RAS International Conference on International Conference on Humanoid Robots (Humanoids), 2013, pp. 484–491.

[23] E. Bastianelli, D.D. Bloisi, R. Capobianco, F. Cossu, G. Gemignani, L. Iocchi, D. Nardi, On-line semantic mapping, in: Proceedings of the 16th International Conference on Advanced Robotics (ICAR), IEEE, 2013.

[24] E. Bastianelli, D. Croce, A. Vanzo, R. Basili, D. Nardi, A discriminative approach to grounded spoken language understanding in interactive robotics, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 2016, pp. 2747–2753.

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013. arXiv preprint arXiv:1301.3781.

[26] M. Milford, R. Schulz, D. Prasser, G. Wyeth, J. Wiles, Learning spatial concepts from ratslam representations, Robot. Auton. Syst. 55 (5) (2007) 403–410.

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

[28] R. Schulz, G. Wyeth, J. Wiles, Lingodroids: socially grounding place names in privately grounded cognitive maps, Adapt. Behav. 19 (6) (2011) 409–424.

[29] S. Heath, D. Ball, R. Schulz, J. Wiles, Communication between lingodroids with different cognitive capabilities, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2013, pp. 490–495.

[30] R. Schulz, G. Wyeth, J. Wiles, Are we there yet? grounding temporal concepts in shared journeys, IEEE Trans. Auton. Mental Dev. 3 (2) (2011) 163–175.

[31] S. Heath, D. Ball, J. Wiles, Lingodroids: cross-situational learning for episodic elements, IEEE Trans. Cogn. Dev. Syst. (ISSN: 2379-8920) 8 (1) (2016) 3–14. http://dx.doi.org/10.1109/TAMD.2015.2442619.

[32] M. Spranger, L. Steels, Co-acquisition of syntax and semantics: An investigation in spatial language, in: Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015, pp. 1909–1915.

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

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

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

[36] A. Taniguchi, T. Taniguchi, T. Inamura, Simultaneous estimation of self-position and word from noisy utterances and sensory information, in: Proceedings of the 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC HMS), 2016.

[37] S. Ishibushi, A. Taniguchi, T. Takano, Y. Hagiwara, T. Taniguchi, Statistical localization exploiting convolutional neural network for an autonomous vehicle, in: Proceedings of the 41st IEEE Annual Conference of Industrial Electronics Society (IECON), IEEE, 2015, pp. 1369–1375.

[38] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[39] Y. Hagiwara, I. Masakazu, T. Taniguchi, Place concept learning by hmlda based on position and vision information, in: Proceedings of the 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems (IFAC HMS), 2016.

[40] Y. Ando, T. Nakamura, T. Araki, T. Nagai, Formation of hierarchical object concept using hierarchical latent dirichlet allocation, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2013, pp. 2272–2279.

[41] F. Dellaert, D. Fox, W. Burgard, S. Thrun, Monte carlo localization for mobile robots, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Vol. 2, IEEE, 1999, pp. 1322–1328.

[42] J. Sethuraman, A constructive definition of Dirichlet priors, Statist. Sinica 4 (1994) 639–650.

[43] E.B. Fox, E.B. Sudderth, M.I. Jordan, A.S. Willsky, A sticky HDP-HMM with application to speaker diarization, Ann. Appl. Stat. (2011) 1020–1056.

[44] D. Gildea, T. Hofmann, Topic-based language models using EM, in: In Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH), 1999.

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

[46] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, K. Shikano, Sharable software repository for Japanese large vocabulary continuous speech recognition, in: Proceedings of 5th International Conference on Spoken Language Processing, 1998.

[47] A. Lee, T. Kawahara, K. Shikano, Julius—an open source real-time large vocabulary recognition engine, in: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), 2001.
[48] G. Kitagawa, Computational aspects of sequential Monte Carlo filter and smoother, Ann. Inst. Statist. Math. 66 (3) (2014) 443–471.
[49] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1) (1985) 193–218.
[50] L. Gillick, S.J. Cox, Some statistical issues in the comparison of speech recognition algorithms, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 1989, pp. 532–535.

Akira Taniguchi received his B.E. degree from Ritsumeikan University in 2013 and his M.E. degree from the Graduate School of Information Science and Engineering, Ritsumeikan University, in 2015. He is currently working toward his Ph.D. degree in the Emergent Systems Lab, Ritsumeikan University, Japan. From April 2017, he is a Japan Society for the Promotion of Science (JSPS) research fellow (DC2). His research interests include language acquisition, concept acquisition, and symbol emergence in robotics.

Tadahiro Taniguchi received M.E. and Ph.D. degrees from Kyoto University in 2003 and 2006, respectively. From April 2005 to March 2008, he was a research fellow of Japan Society for the Promotion of Science. He was an assistant professor from April 2008 to March 2010, an associate professor from April 2010 to March 2017, and a professor in the College of Information Science and Engineering, Ritsumeikan University. From September 2015 to September 2016, he was a visiting associate professor at Imperial College London. He is currently engaged in research on artificial intelligence and emergent systems.

Tetsunari Inamura received B.E., M.S., and Ph.D. degrees from the University of Tokyo in 1995, 1997, and 2000, respectively. He was a Researcher of the JST/CREST program from 2000 to 2003, and then joined the Department of Mechano-Informatics, School of Information Science and Technology, the University of Tokyo as a Lecturer till 2006. He is now an Associate Professor in National Institute of Informatics and the Department of Informatics, SOKENDAI (The Graduate University for Advanced Studies). His research interests include imitation learning, human motion analysis, and development of interactive robots through virtual reality.