This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

Online Algorithm for Robots to Learn Object Concepts and Language Model

Joe Nishihara, Tomoaki Nakamura, Takayuki Nagai

***Abstract*—Humans form concept of objects by classifying them into categories, and acquire language by simultaneously interacting with others. Thus, the meaning of a word can be learned by connecting a recognized word to its corresponding concept. We consider this ability important for robots to flexibly develop knowledge of language and concepts. In this paper, we propose an online algorithm for robots to acquire knowledge of natural language and learn object concepts. A robot learns the language model from word sequences, which are obtained by the segmentation of phoneme sequences provided by a user, by using unsupervised word segmentation each time it is provided with a new object. Moreover, the robot acquires object concepts using these word sequences as well as multimodal information obtained by observing objects. The crucial aspect of our model is the interdependence of words and concepts: there is a high probability that the same words will be uttered to describe objects in the same category. By taking this relationship into account, our proposed method enables robots to acquire a more accurate language model and object concepts online. Experimental results verify this.**

***Index Terms*—multimodal categorization, MLDA, object concepts, language acquisition, online learning, unsupervised learning,**

I. Introduction

IT is well known that the categorization of objects plays an important role in human intelligence [1]. Humans form concepts through categorization. An important aspect of concepts is that they allow us to predict unobserved information. For example, we can infer how hard an object is, how to use it, and so on by simply watching. This is possible because we form concepts while experiencing the world by categorizing multimodal information, which we call "multimodal categorization." Another important aspect of concepts is their role in humans' comprehension of the meanings of words. Words are phonetic labels of concepts, and we can understand their meanings by associating them with concepts. In other words, the problem of language acquisition at a preliminary stage can be defined as a combination of the segmentation of speech signals (word acquisition), multimodal categorization (concept acquisition), and the association between the two. The mechanism in humans of understanding the meanings of words can thus be explained as inference using concepts.

In order to design robots that form concepts and understand meanings of words in the same manner as humans, the ability to categorize experience is considered important [2]. To accomplish this, robots would need a model to infer unobserved information using the generated concepts, i.e., multimodal categories. Hence, in [3], we proposed a categorization model that uses Multimodal Latent Dirichlet Allocation (MLDA), which is an extension of LDA [4]. This model enables robots to form natural object categories using multimodal information.

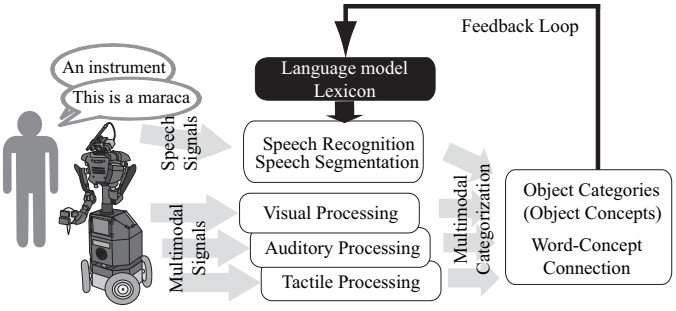


Fig. 1. Basic idea underlying the learning of concepts and words. Please note that the "Feedback Loop" represented by the thick allow in the figure is the main difference between our proposal and [6], [7]. This feedback loop significantly improves learning results, as shown in later experiments.

The idea underlying MLDA is as follows: In the learning process, the robot obtains speech signals uttered by a human partner as well as multimodal data regarding the target object. The speech signals are recognized as phoneme sequences followed by unsupervised word segmentation [5]. We use the bag-of-words model to represent multimodal data. At the recognition stage, the trained model can be used to infer unobserved information from observed information. Since the original MLDA encounters problems with batch learning, it was extended to an online version in [6], [7]. A batch-type algorithm normally assumes that the system can immediately use all available training data.

This is not the case for the type of concept learning discussed in this paper, because our target is a robot that learns concepts and language in the same manner as humans. Humans gradually acquire training data through experience and learn incrementally. Moreover, humans learn concepts and language interactively, through communication with others, and the level of communication changes dynamically depending on the linguistic knowledge of the learner [8]. To realize such interactive learning, a batch-type algorithm is insufficient, because it requires too much time to learn objects, and because the robot cannot quickly respond to the teacher. This occurs because the batch-type algorithm uses all of the data for training, including new data. This requires considerable computation and memory capacity. This results in significant delays when the robot

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

responds to the partner, hindering the natural interaction between the robot and the partner. Moreover, the situation is exacerbated as the learning progresses. Therefore, an algorithm that can learn more quickly is desired. To tackle this problem, we have in past work proposed a "Particle Filter-based online MLDA" (PFoMLDA), an online extension of MLDA [6], [7]. PFoMLDA allows robots to incrementally form object concepts in real environments, and is the baseline algorithm used in this paper.

Although the PFoMLDA yields results that are comparable with those of batch-type MLDA [6], we address the following issue in this paper: The most important assumption in this paper is that robots do not have a priori knowledge of natural language, except for phonetic knowledge. This is because we are interested in the language acquisition process in robots from the very beginning. It is difficult for robots to correctly recognize phonemes uttered by a human without a lexicon or a language model. This causes phoneme recognition errors that significantly affect the learning of concepts. Reference [7] reported a degradation in learning performance due to phoneme recognition errors. As is well known, the performance of recent speech recognition technology heavily depends on the language model. Therefore, the robot should incrementally learn the language model as the lexicon grows. However, this is obviously a chicken-and-egg problem, since the learning of a language model requires correct speech recognition results, and vice versa. The idea is that object concepts formed by multimodal perception can be the other source of information to learn the language model. Unfortunately, this raises another chicken-and-egg problem, since object concepts are formed using a lexicon.

To solve the above problems, we propose an online method by which robots can learn object concepts and a language model iteratively. In this paper, "language model" refers to a set of phoneme sequences as a lexicon and bigram counts of the words in the lexicon. In our method, the robot learns the language model and the category to which a given object belongs jointly and online by using human utterances and multimodal information acquired from the target object, respectively. Speech signals uttered by a human partner are converted into phoneme sequences by a speech recognizer. Since co-occurrence is the key to learning the meanings of words, these phoneme sequences are expected to represent features of the target object. Therefore, object categories, which are formed using multimodal information, are intimately related to co-occurring phoneme sequences. The critical idea underlying our proposal is a recognition of the interdependence of words and concepts: there is a high probability that the same words are used to describe objects in the same category, and objects referred to by identical words are highly likely to have identical features. Using this relation, the accuracy of both phoneme recognition and object classification can be improved. The idea of joint learning was first proposed in [9] by the co-authors of this paper. However, the algorithm proposed in [9] is a batch-type algorithm, subject to the problems described above. Therefore, in this paper, the PFoMLDA is extended to a model that can incrementally acquire a language model and the concepts of different objects.

The contributions of this paper are twofold. First, in a rigorous manner, we formulate the online joint learning of concepts and a language model, based on a generative model. In contrast to this paper, [6] and [7] provide no theoretical formulation of the joint learning problem, and the online algorithm is not involved in [9]. Hence, this is the first attempt to propose an online joint learning algorithm that achieves the aforementioned learning objectives, as shown in Fig. 1. The most important claim is that the feedback loop in Fig. 1 significantly improves performance when learning both concepts and words online. Second, we analyze the online joint learning process in detail to determine what is happening during the learning process. The dynamics of learning the language model and concepts are revealed in the results of an experiment. We strongly believe that these results are significant, because they provide insight into the cognitive model of language learning.

This paper is organized as follows: Related work is described in the next section. Our proposed model is explained in Section III, and an overview of parameter estimation of this model is explained in Section IV. The methods for learning the language model and object concepts are described in Sections V and VI, respectively, and the algorithm for learning the model is presented in Section VII. Experiments are described in Section VIII, and the results are discussed in Section IX. Finally, Section X concludes this study.

II. Related Work

A considerable amount of research has been conducted on designing robots that can simultaneously acquire concepts and words [10], [11]. However, no study to date has simultaneously considered the acquisition of the concepts of objects as well as language based on various modalities, e.g., visual, audio, haptic, and word information, as in this paper. Roy and Pentland [10] proposed a method for robots to learn objects and their names using audio cues and co-occurrence relations among objects. However, the accuracy of word segmentation obtained from continuous speech using their method was approximately 30%; nor did they consider a language model. Reference [11] proposed a method for a robot to learn the names of locations. However, the pattern of teacher utterance was fixed in this experiment. On the contrary, we propose for robots to acquire the concepts of objects as well as a language model using words segmented from free expressions and multimodal information. Online learning is another important feature of our method.

Many researchers have investigated methods for robots to learn relationships among multimodal sensory data [12]–[16]. The purpose of these studies was to predict one modality from another by using models. This mapping is very useful for robots because it allows them to select suitable actions with regard to the target object simply by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

watching it, for instance. Although we share the objectives of past work in the area, the problem of the acquisition of the meanings of words is fundamentally different from the prediction of modalities.

Several studies have been carried out on the categorization of objects using images in an unsupervised manner [17]–[20]. Moreover, unsupervised object categorization has been effected by using point clouds acquired from a laser range finder or a time-of-flight camera [21]. However, object categories in these cases were not solely determined from visual information. Sinapov and Stoytchev showed that object categorization can be accomplished by utilizing sounds made when a robot touches an object [22]. However, they focused on categorization to the exclusion of the acquisition of the meanings of words. Studies using tactile sensors have also been conducted [23]–[25], but with the aim of object recognition (using their shapes) rather than categorization. We believe that humans categorize objects using multimodal information, and this is thus needed for robots to form categories.

Several researchers have studied scene and action recognition using multiple information such as objects and language [26]–[30]. In [26]–[28], human actions are recognized by considering manipulated objects. Moreover, in [29], actions are recognized with a language model trained with a large corpus. The model extracts the relationships between actions (verbs) and tools (nouns). In [30], a method was proposed for segmenting and recognizing images using verbal information. These studies have shown that performance can be improved by using multiple information. However, scene and action recognition is a focused research direction, and language acquisition was not considered in these studies.

Online algorithms for LDA have been proposed, and there are two types of inference algorithms. One is based on variational Bayes (VB) inference [31], [32], [33], and the other is based on a sampling approach [34]. We employed the sampling-based approach, because [35] reported that this approach provides relatively better performance. Moreover, an implementation of the VB-based approach is generally more complex than the sampling-based approach. Canini et al. proposed a sampling-based online algorithm for LDA that is similar to our use of PFoMLDA in this paper. However, old data is used during resampling step in their algorithm. This is not applicable to our method, because we assume that the robot learns incrementally using only recent data. Our PFoMLDA updates its parameters from a single object exclusively, and it does not use old data.

III. Integrated Language and Object Concept Acquisition Model

A. Multimodal Categorization by MLDA [3]

We have in past work proposed a series of methods to form object concepts using multimodal data and user utterances [3], [36]–[38]. The basic idea underlying these methods is "multimodal categorization," which assumes

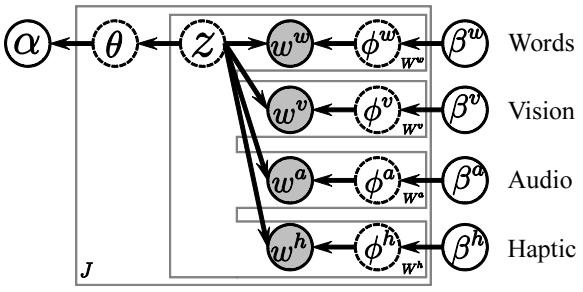


Fig. 2. Graphical model of MLDA.

no prior linguistic knowledge. The robot is assumed to have only acoustic models that enable it to recognize the phoneme sequences of the utterances of a human teacher. The robot also obtains multimodal information, such as visual, audio, and haptic, by using its embodiment and word information by segmenting the phoneme sequences into words. The robot can form the concepts of objects like humans do by categorizing such multimodal information. Fig.2 shows the graphical model of MLDA. w^* in the figure denotes each item of multimodal information, which are generated by a multinomial distribution parameterized by ϕ^*. These parameters of multinomial distributions are in turn generated from a Dirichlet distribution parameterized by β^*. z represents a category generated from a multinomial distribution parameterized by θ, and θ is generated in turn from a Dirichlet distribution parameterized by α. In this figure, gray nodes represent observable information, and dashed nodes represent latent parameters. The problem of object categorization is thus equivalent to estimating these latent parameters from the observable information w^*.

B. Extension to Integrated Model

As mentioned earlier, phoneme recognition results include error because robots do not have prior linguistic knowledge. Words are obtained by segmenting recognized phoneme sequences in an unsupervised manner. However, it is difficult to segment phoneme sequences containing errors correctly into words, and appropriate concepts of objects are thus not formed. In order to represent the generative process of words more accurately, we need an integrated model that consists of both a language model and the concepts of objects. To this end, the model of MLDA shown in Fig.2 is extended to a more complex one to enable it to represent the generative process of uttered speech signals.

Fig.3 shows the extended model, which represents the generative process of speech signals, words, and concepts of objects. This model assumes that each observation is generated as follows:

1) Parameter θ of the multinomial distribution for defining object categories and parameter ϕ^* of the multinomial distribution for generating object information

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

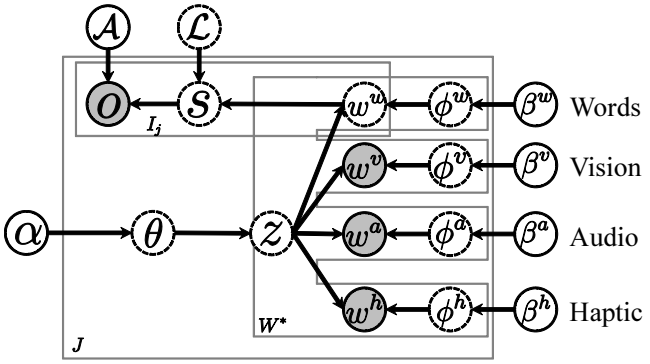


Fig. 3. Integrated model of language and concepts of objects.

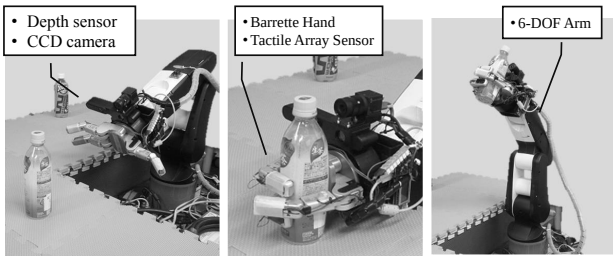


Fig. 4. Capturing (a) visual information, (b) haptic information, and (c) auditory information.

(observations) are generated as

$$\theta \sim \mathcal{D}(\theta|\alpha), \quad (1)$$

$$\phi^* \sim \mathcal{D}(\phi^*|\beta^*), \quad (2)$$

where $\mathcal{D}$ represents a Dirichlet distribution, and α and β^* are hyperparameters.

2) The following procedures are iterated for all objects $j \in \{1, 2, \cdots, J\}$:
   a) The followings are iterated W^* times for each modality $* \in \{w, v, a, h\}$:
      i) The object category z is drawn as

$$z \sim \mathcal{M}(z|\theta), \quad (3)$$

      where $\mathcal{M}$ represents a multinomial distribution.
      ii) Each observation w^* is drawn as

$$w^* \sim \mathcal{M}(w^*|\phi^*). \quad (4)$$

   b) The following procedures are iterated I_j times in order to generate user utterances that represent object features.
      i) A phoneme sequence is generated based on the above generated words $\boldsymbol{w}^w$ and the language model parameterized by $\mathcal{L}$.

$$s \sim P(s|\mathcal{L}, \boldsymbol{w}^w) \quad (5)$$

      ii) A speech signal o is generated based on the phoneme sequence s using the acoustic model parameterized by $\mathcal{A}$.

$$o_j \sim P(o|\mathcal{A}, s) \quad (6)$$

It is worth noting that the model integrates both object concepts and the language model. This means that the object category z affects the process of the generation of uttered speech signal o. The most important upshot is that object categories and phoneme sequences can be estimated by inferring the latent variables, which are represented by dashed nodes in the graphical model in Fig.3 by using the observable information. Hence, the learning problem here corresponds to the inference of the latent variables $\mathcal{L}$, ϕ^*, θ, s, and w^w based on observations w^v, w^a, w^h, and user utterances o for a fixed acoustic model $\mathcal{A}$.

The multimodal information used to learn the integrated model was obtained by using the robot shown in Fig.4. The details are as follows:

Visual information : w^v

A charge-coupled device (CCD) camera and a depth sensor were mounted on the robot's arm (Fig. 4 (a)), and images of objects were captured to use as visual information. A dense, scale-invariant feature transform (DSIFT) [39] was computed from each image. Each feature vector was quantized using 500 representative vectors and converted into a 500-dimensional histogram.

Haptic information : w^h

Haptic information was obtained using a BarrettHand mounted on the arm of the robot and a tactile array sensor mounted on its hand (Fig. 4 (b)). The robot grasped objects and obtained a time series of sensor values. The sensor values were approximated by a sigmoid function the parameters of which were used as feature vectors [38]. Finally, these feature vectors were quantized and converted into a 15-dimensional histogram.

Auditory information : w^a

A microphone was mounted on the robot's hand, and sound was recorded by shaking the given objects (Fig.4 (c)). The sound was divided into frames, and a 13-dimensional Mel-frequency cepstrum coefficient (MFCC) was computed at each frame. The sound was thus converted into a 13-dimensional feature vector. With regard to other information, feature vectors for these were quantized and converted into a 50-dimensional histogram.

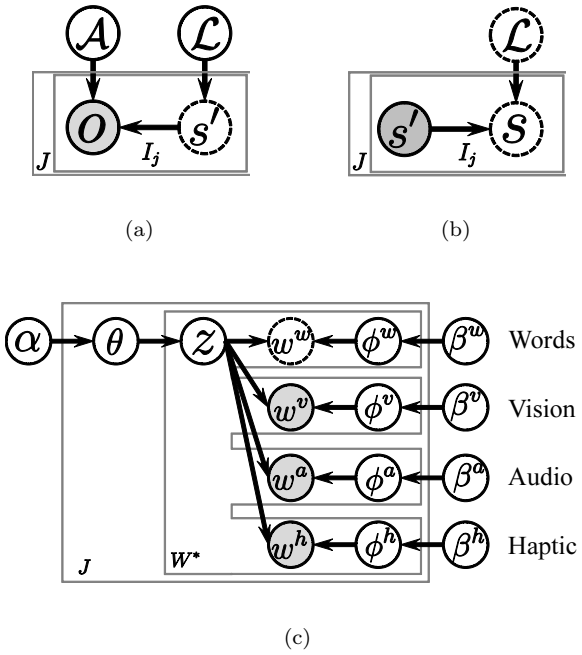
User utterances : o

A user described features of the given object while the robot observed it, and speech signals were recorded to forward to the robot.

IV. Inference of model parameters

As described above, the tasks of concept formation and learning the language model correspond to parameter estimation in Fig.3. However, the integrated model in Fig.3 is too complex to simultaneously estimate all parameters. We thus approximate the model by dividing it into three parts, as shown in Fig.5. The language model

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems



(a) (b) (c)

Fig. 5. Language and object concept acquisition model. (a)Speech recognition. (b)Language model. (c)Word generation.

and the concepts of objects can be learned by estimating the parameters of each part. The gray nodes represent observable information, and the dashed nodes represent latent parameters in Fig.5. It is impractical to conduct batch learning for all objects every time a new object is obtained because it takes long to compute the model parameters. Therefore, we propose a method to learn the model online.

The parameters can be estimated by the following iterative process:

1.Speech Recognition

Fig.5(a) shows the model for speech recognition. Here, acoustic model parameter $\mathcal{A}$ and language model parameter $\mathcal{L}$ are assumed to be known. The N phoneme sequences $\boldsymbol{s}'_{1:N}$ are sampled from the given teacher's utterances $\boldsymbol{o}$.

$$\boldsymbol{s}'_{1:N} \sim p(\boldsymbol{s}'_{1:N}|\boldsymbol{o},\mathcal{A},\mathcal{L}) \propto p(\boldsymbol{o}|\mathcal{A},\boldsymbol{s}'_{1:N})p(\boldsymbol{s}'_{1:N}|\mathcal{L}) \tag{7}$$

$p(\boldsymbol{o}|\mathcal{A},\boldsymbol{s}'_{1:N})$ represents the acoustic likelihood of the given teacher's utterances and $p(\boldsymbol{s}'_{1:N}|\mathcal{L})$ represents the prior probability of $\boldsymbol{s}'_{1:N}$ for the given language model $\mathcal{L}$. However, Julius [40], the speech recognition software we use, cannot perform such sampling. Therefore, the sampling procedure is approximated, and we use the n-best speech recognition results that can be obtained from Julius, rather than sampled phoneme sequences. In our experiments, we used the standard acoustic model in the Julius package.

2. Learning the language model

The parameter of the language model $\mathcal{L}$ can be calculated by maximizing $P(\boldsymbol{o}|\mathcal{A},\mathcal{L})$, which is the probability of generating the teacher's utterances $\boldsymbol{o}$ in Fig.3, as follows:

$$\begin{aligned}\mathcal{L} &= \underset{\mathcal{L}}{\operatorname{argmax}}\, p(\boldsymbol{o}|\mathcal{A},\mathcal{L}) \\ &= \underset{\mathcal{L}}{\operatorname{argmax}} \int p(\boldsymbol{s}|\mathcal{L})p(\boldsymbol{o}|\boldsymbol{s},\mathcal{A})d\boldsymbol{s}\end{aligned} \tag{8}$$

It is impossible, however, to directly compute the probability because the integral over $\boldsymbol{s}$ requires the sum of all phoneme combinations. Hence, the probability $p(\boldsymbol{o}|\boldsymbol{s},\mathcal{A})$ is considered to be sufficiently small for most phoneme sequences, and only the n best phoneme sequences $\boldsymbol{s}'_{1:N}$ are used to calculate the above equation. Therefore, the language model is separated from the other parts, as shown in Fig.5(b). That is, we compute $\mathcal{L}$ by maximizing the probability that $\boldsymbol{s}_{1:N}$ is generated from the n best sequences $\boldsymbol{s}'_{1:N}$ of speech $\boldsymbol{o}$, instead of maximizing the probability that speech $\boldsymbol{o}$ is generated.

$$\mathcal{L},\boldsymbol{s}_{1:N} = \underset{\mathcal{L},\boldsymbol{s}_{1:N}}{\operatorname{argmax}}\, p(\boldsymbol{s}_{1:N}|\boldsymbol{s}'_{1:N},\mathcal{L}) \tag{9}$$

This maximization can be carried out by the pseudo-online NPYLM [7], which is described later.

3. Word Generation

It is possible to generate words related to the concepts of objects in the same manner as the language model is learned, if we can compute the following probability:

$$\begin{aligned}\boldsymbol{w}^w &= \underset{\boldsymbol{w}^w}{\operatorname{argmax}}\, p(\boldsymbol{w}^w|\boldsymbol{o},\mathcal{A},\mathcal{L},\boldsymbol{w}^{v,a,h},\beta^w,\alpha) \\ &= \underset{\boldsymbol{w}^w}{\operatorname{argmax}} \int p(\boldsymbol{o}|\boldsymbol{s},\mathcal{A},\mathcal{L}) \\ &\qquad \times p(\boldsymbol{s}|\boldsymbol{w}^w,\mathcal{L})p(\boldsymbol{w}^w|\boldsymbol{w}^{v,a,h},\beta^w,\alpha)d\boldsymbol{s}\end{aligned} \tag{10}$$

Here, $\boldsymbol{w}^{v,a,t}$ denotes visual, audio, and haptic information obtained from the object. However, it is difficult to compute this probability because of integration over $\boldsymbol{s}$. We assume that $P(\boldsymbol{o}|\boldsymbol{s},\mathcal{A})$ is negligibly small for most phoneme sequences, and hence separate the model used to generate words from the language model by utilizing the n best results $\boldsymbol{s}_{1:N}$, as shown in Fig.5(c). The latent model parameters in Fig.2 become known parameters in Fig.5(c) by being estimated from observed information $\boldsymbol{w}^v, \boldsymbol{w}^a, \boldsymbol{w}^h$, and words can thus be generated. Therefore, we do not maximize the probability that $\boldsymbol{w}^w$ is generated from the teacher's utterances $\boldsymbol{o}$, but rather the probability that $\boldsymbol{w}^w$ is generated from the n best word sequences $\boldsymbol{s}_{1:N}$:

$$\begin{aligned}\boldsymbol{w}^w &\sim p(\boldsymbol{w}^w|\boldsymbol{o},\mathcal{L},\mathcal{A},\boldsymbol{w}^{v,a,h},\beta^w,\alpha) \\ &\approx \frac{p(\boldsymbol{w}^w_n|\boldsymbol{w}^{v,a,h},\beta^w,\alpha)}{\Sigma_{n=1}^N p(\boldsymbol{w}^w_n|\boldsymbol{w}^{v,a,h},\beta^w,\alpha)} \qquad (1 \le n \le N)\end{aligned} \tag{11}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

where $\boldsymbol{w}_n^w$ represents the bag-of-words representation of $\boldsymbol{s}_n$.

4. Object Concept Formation

We can obtain relevant words $\boldsymbol{w}^w$ affected by the concept and speech recognition assigned to each object by the above process. The concepts of objects can be formed by learning the model shown in Fig.2, which is separated from the speech recognition and language model. Learning the concepts of objects is equivalent to estimating the parameters β^* that maximize the probability that multimodal information $\boldsymbol{w}^*$ is generated. These parameters can be inferred by PFoMLDA, which is an online extended version of MLDA.

We use the language model where all phonemes have the same probability as the initial language model parameter $\mathcal{L}_0$. We then estimate the parameters through the above process each time a new object is provided.

In Sec. V, details of language model learning are described, followed by details of object concept formation in Sec. VI. The entire learning algorithm for the proposed integrated model is outlined in Sec. VII.

Algorithm 1
Pseudo-online NPYLM (for a single input)

function oNPYLM($s'|\hat{\boldsymbol{W}}^w$)
  // L and T are predefined parameters of oNPYLM
  A new phoneme sequence $\boldsymbol{s}'$ is entered:
  $\mathbf{w}(\boldsymbol{s}') \sim p(\mathbf{w}|\boldsymbol{s}', \hat{\boldsymbol{W}}^w)$
  Add $\boldsymbol{s}'$ to $\mathbf{S}$
  Add $\mathbf{w}(s')$ to $\hat{\boldsymbol{W}}^w$
  if $|\mathbf{S}| > L$ **then**
    Remove the oldest sentence from $\mathbf{S}$
  end if
  Blocked Gibbs sampler:
  for $t \leftarrow 1$ to T **do**
    for all $\boldsymbol{s}'$ in $\mathbf{S}$ **do**
      Remove $\mathbf{w}(\boldsymbol{s}')$ from $\hat{\boldsymbol{W}}^w$
      $\mathbf{w}(\boldsymbol{s}') \sim p(\mathbf{w}|\boldsymbol{s}', \hat{\boldsymbol{W}}^w)$
      Add $\mathbf{w}(\boldsymbol{s}')$ to $\hat{\boldsymbol{W}}^w$
    end for
  end for
  return $\mathbf{w}(\boldsymbol{s}')$
end function

V. Learning the Language Model

A language model can be obtained by segmenting recognized phoneme sequences into words. $\mathcal{L}$ is a parameter that maximizes the probability that word sequences $\boldsymbol{s}$ are generated by segmenting recognized sentences $\boldsymbol{s}'$:

$$\mathcal{L}, \boldsymbol{s} = \underset{\mathcal{L}, \boldsymbol{s}}{\operatorname{argmax}} P(\boldsymbol{s}|\boldsymbol{s}', \mathcal{L}) \tag{12}$$

In this paper, we utilize pseudo-online NPYLM (oNPYLM) [7], which is an extension of NPYLM [5] to online learning. For the sake of the completeness of this paper, we explain these methods in the following subsections.

A. Hierarchical Pitman-Yor Language Model

The Hierarchical Pitman-Yor Language Model (HPYLM) is an n-gram language model that applies the hierarchical Pitman-Yor process. HPYLM computes the probability that a word w appears following a context h as follows:

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\gamma + \sum_w c(w|h)} + \frac{\gamma + d \cdot \sum_w t_{hw}}{\gamma + \sum_w c(w|h)} p(w|h') \tag{13}$$

where h' denotes the context of $(n-1)$ gram and $p(w|h')$ denotes the probability that w will appear following the context prior to h. The probability can be calculated recursively. Moreover, $c(w|h)$ denotes the number of occurrences w in context h, and d and γ denote hyperparameters of the Pitman-Yor process.

B. Nested Pitman-Yor Language Model

If a lexicon is given, probability $p(w|h')$ can be set to be the inverse of the number of words in case of unigram. However, it is difficult to calculate this probability without a predefined dictionary because all substrings can in principle be words. NPYLM solves this problem by using a character HPYLM as a base measure of the word unigram. Thus, this model is called "nested" because the character HPYLM is embedded as a base measure of the word HPYLM. NPYLM can rapidly segment sentences using a blocked Gibbs sampler and dynamic programming.

C. Pseudo-online NPYLM

NPYLM usually requires a large learning dataset for word segmentation, as well as considerable computational time. In our model, however, the robot has to segment sentences given assigned a target object by a user, and there are at most five to 10 sentences. We extend NPYLM to pseudo-online NPYLM to solve this problem. oNPYLM stores word sequences $\hat{W}^w$ obtained by segmenting phoneme sequences as a parameter and applies the blocked Gibbs sampler to only the final L sequences when a new phoneme sequence is provided. The algorithm for pseudo-online NPYLM is shown in Algorithm.1. $\boldsymbol{s}'$ and $\boldsymbol{w}(\boldsymbol{s}')$ here denote a new phoneme sequence and the segmented results of $\boldsymbol{s}'$, respectively. $\hat{\boldsymbol{W}}^w$, T, and L represent the parameter of the model, the number of sampling iterations, and the number of sentences to which sampling is applied, respectively. In the experiment, we set $L = 200$.

VI. Learning Concepts of Objects Online

We use Particle Filter-based online MLDA (PFoMLDA) [6], which is an extension of MLDA for online learning, for object concepts formation. PFoMLDA is described in the following subsection.

A. Particle Filter-based oMLDA

We calculate each parameter in Fig.2 using Gibbs sampling. The category z_{mi}, which is the category concerning

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

the i-th item of information regarding the modality m of the target object, is sampled from the following posterior distribution:

$$p(z_{mi} = k|\boldsymbol{z}\ , \boldsymbol{w}^m, \alpha, \beta^m) \propto (N_k^{-mi} + \alpha)\frac{N_{mw^mk}^{-mi} + \beta^m}{N_{mk}^{-mi} + W^m\beta^m} \quad (14)$$

where N_k represents the number of times category k is assigned to all information regarding the target object, and N_{mw^mk} represents the number of times category k is assigned to multimodal information item w^m of the object. W^m denotes the number of dimensions of the object's modality m. The superscript with the "-" sign indicates the exception to the feature. N_k and N_{mk} are computed as follows:

$$N_k = \Sigma_{nw^m} N_{mw^mk} \quad (15)$$

$$N_{mk} = \Sigma_{w^m} N_{mw^mk} \quad (16)$$

Finally, parameters $\hat{\theta}_{kj}$ and $\hat{\phi}^m_{w^mk}$ are estimated as follows:

$$\hat{\theta}_{kj} = \frac{N_{kj} + \alpha}{N_j + K\alpha} \quad (17)$$

$$\hat{\phi}^m_{w^mk} = \frac{N_{mw^mk} + \beta^m}{N_{mk} + W^m\beta^m} \quad (18)$$

It is likely that the learned concepts will widely fluctuate because of the order of objects in online learning. Hence, we adopt a forgetting rate λ, and solve the problem by allowing the robot to forget part of the model every time it learns a new object.

$$N_{mw^mk(j+1)} = (1 - \lambda)N_{mw^mkj} \quad (19)$$

Eq.(19) denotes the forgetting rate of the parameters of the model. Furthermore, we introduce a particle filter to oMLDA and extend it to PFoMLDA in order to make it more robust against initial parameters and the influence of the learning order. PFoMLDA deals with these problems by constructing a variety of models with different forgetting rates λ and initial parameters. We then select a model based on the likelihood of the word provided by the user because word information is considered correct information. The probability can be computed by the following equation:

$$p(w^w|\boldsymbol{w}^{v,a,h}) = \int \Sigma_z p(w^w|z)p(z|\theta)p(\theta|\boldsymbol{w}^{v,a,h})d\theta \quad (20)$$

Of all the models, we select ones with higher likelihood based on Eq.(20). The selected models continue to be learned without changing the forgetting rate and the initial value. Models with lower likelihoods are rejected and replaced with ones with the highest likelihoods. It thus becomes possible for the robot to learn objects online and reduce the influence of the order of objects. The algorithm used to learn each object is summarized in Algorithm.2. Fig.6 shows a schematic figure of the online learning system. In the experiment detailed later, we used 50 particles.

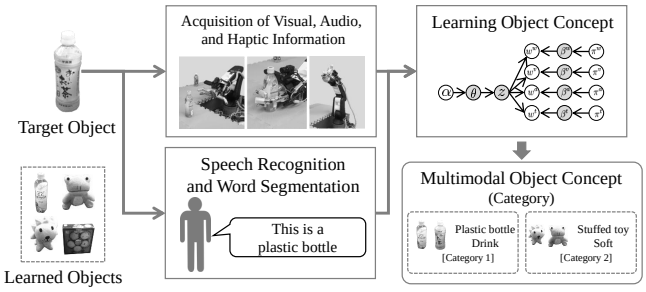


Fig. 6. Online learning system.

Algorithm 2 PFoMLDA (for single object)

function PFoMLDA ($\boldsymbol{w}^{v,a,h}, \boldsymbol{w}^w|\Theta$)
  // $\Theta = \{N_{mw^mk}|m \in (v,a,h,w), 1 \leq k \leq K\}$
  // λ, α and β^m are predefined model parameters
  for all m, w^m, k **do**
    $N_{mw^mk} \leftarrow (1-\lambda)N_{mw^mk}$
  end for
  The following process is repeated until convergence
  for all m, i (of new input data) **do**
    for $k \leftarrow 1$ to K **do**
      $P[k] \leftarrow P[k-1] + (N_k^{-mi} + \alpha)\frac{N_{mw^mk}^{-mi} + \beta^m}{N_{mk}^{-mi} + W^m\beta^m}$
    end for
    $u \leftarrow$ random value $[0, 1]$
    for $k \leftarrow 1$ to K **do**
      if $u < P[k]/P[K]$ **then**
        $z_{mi} = k$, break
      end if
    end for
  end for
  return Θ
end function

VII. The learning algorithm

When we are told "This is a plastic bottle" when shown one, we can correctly recognize it even if we hear "This is a p**d**ast**i**c bottle," for instance. We can predict the natural sentence that follows "This is" using linguistic knowledge. Moreover, we utilize the knowledge that the object in front of us is denoted by the term "plastic bottle." Thus, language and concepts are closely related, and it is important to learn both at the same time. The language model and the concepts of objects can be learned together by applying PFoMLDA and oNPYLM to the model described in Sec. IV. The proposed algorithm is as follows:

The robot can obtain the N best phoneme sequences $\boldsymbol{s}'^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}$ from utterances $\boldsymbol{o}_{j,1:I_j}$ for the j-th object with language model parameter $\mathcal{L}_{j-1}$. The subscripts $1:I_j$ and $1:N$ denote a set of I_j utterances and a set of the N best recognition results, respectively. However, recognition using $\mathcal{L}_{j-1}$ is not always correct. Thus, the robot also obtains $\boldsymbol{s}'^{\mathcal{A}}_{j,1:I_j}$ recognized by the use of the language model $\mathcal{L}_0$, where all phonemes have the same probability of occurring. Word sequences $\boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}$ and $\boldsymbol{s}^{\mathcal{A}}_{j,1:I_j}$ are then computed by segmenting $\boldsymbol{s}'^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}$ and $\boldsymbol{s}'^{\mathcal{L}}_{j,1:I_j,1:N}$ by using

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

oNPYLM.

$$\boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N} \sim \text{oNPYLM}(\boldsymbol{s}'^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}|\hat{W}^w) \quad (21)$$

$$\boldsymbol{s}^{\mathcal{A}}_{j,1:I_j} \sim \text{oNPYLM}(\boldsymbol{s}'^{\mathcal{A}}_{j,1:I_j}|\hat{W}^w) \quad (22)$$

where $\text{oNPYLM}(\cdot|*)$ denotes a function that segments a sentence into words using oNPYLM with parameter $*$, and $\hat{W}^w$ represents word sequences selected based on the concepts of objects. A new language model parameter $\mathcal{L}_j$ is then computed from $\boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}$ and $\boldsymbol{s}^{\mathcal{A}}_{j,1:I_j}$ as follows:

$$\mathcal{L}_j \sim \text{LM}(\boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}, \boldsymbol{s}^{\mathcal{A}}_{j,1:I_j}, \hat{W}^w) \quad (23)$$

where $\text{LM}(\cdot)$ denotes a function to compute a parameter of the language model. The utterance is recognized again with the new language model parameter $\mathcal{L}_j$, and the n best word sequences $\boldsymbol{s}^{\mathcal{L}_j}_{j,1:I_j,1:N}$ are computed as well as Eq.(21) and Eq.(22). The appropriate word sequence is sampled from the n best word sequences based on the probability Eq.(20) that words are generated from multimodal information.

For all i

$$\hat{n} \sim \frac{P(\boldsymbol{w}^{\mathcal{L}_j}_{jin}|\boldsymbol{w}^{v,a,h}_j, \Theta_{j-1})}{\Sigma^N_{n=1} P(\boldsymbol{w}^{\mathcal{L}_j}_{jin}|\boldsymbol{w}^{v,a,h}_j, \Theta_{j-1})} \quad (1 \leq n \leq N) \quad (24)$$

$$\boldsymbol{s}_{ji} = \boldsymbol{s}^{\mathcal{L}_j}_{ji\hat{n}} \quad (25)$$

$$\boldsymbol{w}^w_{ji} = \boldsymbol{w}^{\mathcal{L}_j}_{ji\hat{n}} \quad (26)$$

where $\boldsymbol{w}^{\mathcal{L}_j}_{jin}$ denotes the n-th word sequence of the i-th utterance for the j-th object and the bag-of-words representation of $\boldsymbol{s}^{\mathcal{L}_j}_{jin}$, which is an element of $\boldsymbol{s}^{\mathcal{L}_j}_{j,1:I_j,1:N}$. Θ_{j-1} denotes the parameter of the concept of the relevant object.

Following this, the parameter of the concept of the relevant object is updated from the selected word sequence $\boldsymbol{w}^w_{j,1:I_j}$.

$$\Theta_j \sim \text{PFoMLDA}(\boldsymbol{w}^{v,a,h}_j, \boldsymbol{w}^w_{j,1:I_j}|\Theta_{j-1}) \quad (27)$$

Finally, $\boldsymbol{s}_{j,1:I_j}$ is added to the word sequences $\hat{W}^w$.

The robot incrementally acquires the language model and the concept of the object by updating language model parameter $\mathcal{L}$ and object concept parameter Θ each time it is given a new object. However, the robot does not have a parameter of the object concept Θ, and hence the probability Eq.(24) cannot be computed when it first learns the object. Hence, we select words with the highest score from the n best word candidates in case of the first object.

The above process is summarized in Algorithm.3, and the baseline online learning system in Fig.6 can be extended to that in Fig.7.

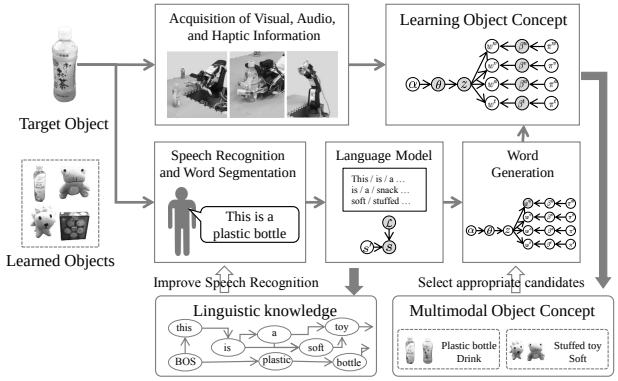


Fig. 7. Extended online learning system.

Algorithm 3 Learning language model and concepts of objects together

function LEARN($\boldsymbol{w}^{v,a,h}_j, \boldsymbol{o}_{j,1:I_j}|\mathcal{L}_{j-1}, \Theta_{j-1}, \hat{\boldsymbol{W}}^w_{j-1}$)
  Update language model:
    $\boldsymbol{s}'^{\mathcal{A}}_{j,1:I_j} \leftarrow \text{Recognize}(\boldsymbol{o}_{j,1:I_j}|\mathcal{A}, \mathcal{L}_0)$
    $\boldsymbol{s}'^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N} \leftarrow \text{Recognize}(\boldsymbol{o}_{j,1:I_j}|\mathcal{A}, \mathcal{L}_{j-1})$
    $\boldsymbol{s}^{\mathcal{A}}_{j,1:I_j} \sim \text{oNPYLM}(\boldsymbol{s}'^{\mathcal{A}}_{j,1:I_j}|\hat{\boldsymbol{W}}^w_{j-1})$
    $\boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N} \sim \text{oNPYLM}(\boldsymbol{s}'^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}|\hat{\boldsymbol{W}}^w_{j-1})$
    $\mathcal{L}_j \sim \text{LM}(\boldsymbol{s}^{\mathcal{A}}_{j,1:I_j}, \boldsymbol{s}^{\mathcal{L}_{j-1}}_{j,1:I_j,1:N}, \hat{\boldsymbol{W}}^w_{j-1})$
  Obtain word information:
    $\boldsymbol{s}'^{\mathcal{L}_j}_{j,1:I_j,1:N} \leftarrow \text{Recognize}(\boldsymbol{o}_{j,1:I_j}|\mathcal{A}, \mathcal{L}_j)$
    $\boldsymbol{s}^{\mathcal{L}_j}_{j,1:I_j,1:N} \sim \text{oNPYLM}(\boldsymbol{s}'^{\mathcal{L}_j}_{j,1:I_j,1:N}|\hat{\boldsymbol{W}}^w_{j-1})$
    $\boldsymbol{w}^w_{j,1:I_j,1:N} = \text{BoW}(\boldsymbol{s}^{\mathcal{L}_j}_{j,1:I_j,1:N})$
  for $i \leftarrow 1$ to I_j **do**
    Select words:
      $\hat{n} \sim P(\boldsymbol{w}^w_{jin}|\boldsymbol{w}^w_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j, \Theta_{j-1})$
      $\boldsymbol{w}^w_{ji} \leftarrow \boldsymbol{w}^w_{ji\hat{n}}$
      $\boldsymbol{s}_{ji} \leftarrow \boldsymbol{s}_{ji\hat{n}}$
  end for
  Learn concept of object:
    $\Theta_j \sim \text{PFoMLDA}\ (\boldsymbol{w}^v_j, \boldsymbol{w}^a_j, \boldsymbol{w}^h_j, \boldsymbol{w}^w_{j,1:I_j}|\Theta_{j-1})$
  $\hat{\boldsymbol{W}}^w_j = \boldsymbol{s}_{j,1:I_j} + \hat{\boldsymbol{W}}^w_{j-1}$
  return $\mathcal{L}_j, \Theta_j, \hat{\boldsymbol{W}}^w_j$
end function

VIII. Experiments

We conducted experiments to test our proposed method. We used 50 objects belonging to 10 categories, as shown in Fig.8. In the experiments, multimodal information, explained in Sec. III-B, and teaching utterances were provided by the user while the robot observed the objects. In the proposed method, teaching utterance $\boldsymbol{o}$ is recognized and converted into a phoneme sequence based on language model parameter $\mathcal{L}$ according to Algorithm 3 every time a new item of object information $\boldsymbol{w}^{v,ah}, \boldsymbol{o}$ is obtained. Then, the phoneme sequence is segmented into words by NPYLM with parameter $\hat{\boldsymbol{W}}^w$. Finally, the object concept is learned from the multimodal information, and words as parameters $\mathcal{L}$, $\hat{\boldsymbol{W}}^w$ and Θ affect each other. In order to test the proposed model, speech recognition accuracy was used to assess $\mathcal{L}$, segmentation accuracy to test the value of $\hat{\boldsymbol{W}}^w$, and object classification accuracy was used

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

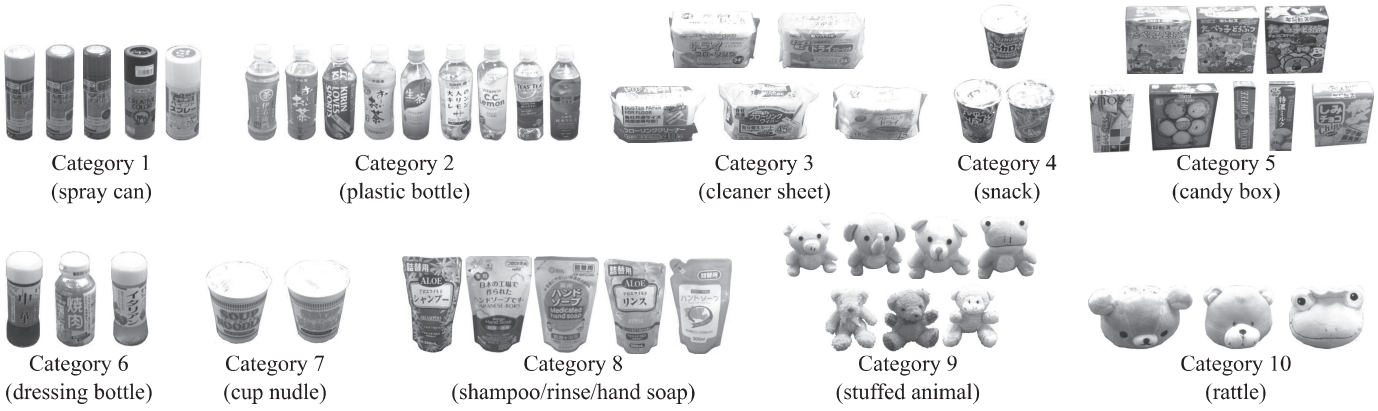


Fig. 8. Objects used in the experiments.

to assess Θ. To note the variation in learning, the robot was made to learn the same objects twice. Therefore, it learned 100 times (50 objects $\times$ 2) in total. We conducted experiments using the following four methods to evaluate the proposed method:

(a) A method that used only word sequences obtained by NPYLM by segmenting the phoneme recognition results (Acoustic model only).
(b) A method that used word sequences obtained by NPYLM by segmenting the 1 best result of speech recognition, and updated the language model using this (Language model only).
(c) A method that used word sequences selected from the n best recognition results and updated the language model. (Proposed method).
(d) A method that used word sequences obtained by NPYLM by segmenting correct phoneme sequences $\boldsymbol{s}_{\text{correct}}$ provided manually by a human teacher and free from phoneme errors (Correct sequence).

Method (a) was the baseline method used in a past study [7], (b) only updated the language model without considering the concept of the object, (c) was our proposed method, and (d) represented ideal conditions and the performance limit. Considering variations in the learning results, each method was learned three times. The following shows the average of the three learning results. The utterances provided to the robot were in Japanese.

A. Accuracy of Speech Recognition

To assess the language models learned using methods (b) and (c), the robot recognized all speech signals for all objects by using the language model learned at each step. We determined the accuracy of speech recognition by calculating the edit distance from $\boldsymbol{s}_{\text{correct}}$ as correct sentences:

$$Accuracy = 1 - \frac{\text{LS}(s_1, s_2)}{\max(\text{len}(s_1), \text{len}(s_2))} \tag{28}$$

where $\text{LS}(s_1, s_2)$ is a function that returns the edit distance between s_1 and s_2, and $\max(a, b)$ is a function that returns the longer of the two sequences.

TABLE I
Example of improved phoneme recognition.

The Number of Learning Steps	Recognition Result
Step 1	*e* i *go ro ko*
Step 2	*e* i *go ro* mi
Step 12	nu *e* i gu ru mi
Step 36	nu i gu ru mi

Fig.9 shows speech recognition accuracy values obtained by each method at each step of learning. The score for method (a) was constant at approximately 65% because it did not update its language model. On the contrary, the scores for methods (b) and (c) were lower than that for method (a) when learning commenced. This is because language model $\mathcal{L}$, calculated using a few samples, worked against language model $\mathcal{L}_0$, where all phonemes had the same probability of occurrence. However, as learning progressed, these scores improved; finally, the scores for each method converged to 73%. The score for the proposed method (c) was higher by about 2% than that for method (b), although the difference in scores between methods (b) and (c) was small. It was natural for method (b) to improve its accuracy of phoneme recognition because it updated language models by utilizing the 1 best result. However, the score for method (c) was higher because the n best results of speech recognition included candidates with fewer errors, and method (c) could select these candidates. Table I shows an instance of improved phoneme recognition of "nu i gu ru mi," which means "stuffed animal" in Japanese, by our proposed method. These results suggest that the robot was able to accurately recognize speech by learning the language model through online learning.

It might appear that the approximately 2% improvement in phoneme recognition is insignificant. However, the consistency of the speech recognition, which is discussed in the next section, improves with mutual learning. One can confirm that the classification accuracy remarkably improves in the proposed method and, in fact, the performance of the categorization is comparable to the correct phoneme sequence case, as we will presently see.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

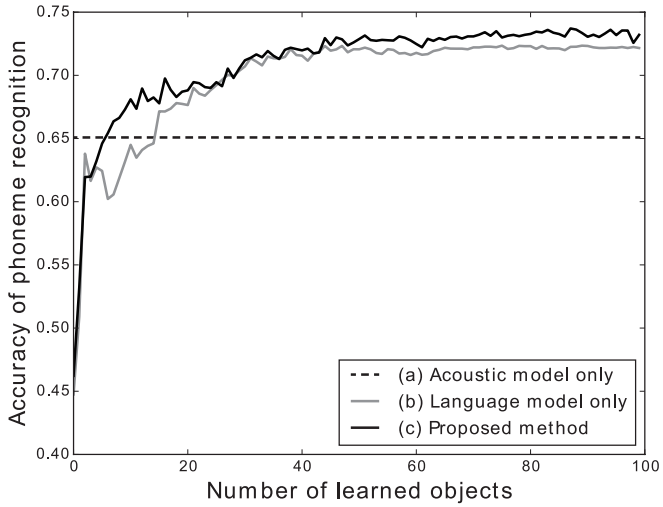


Fig. 9. Accuracy of speech recognition for each method.

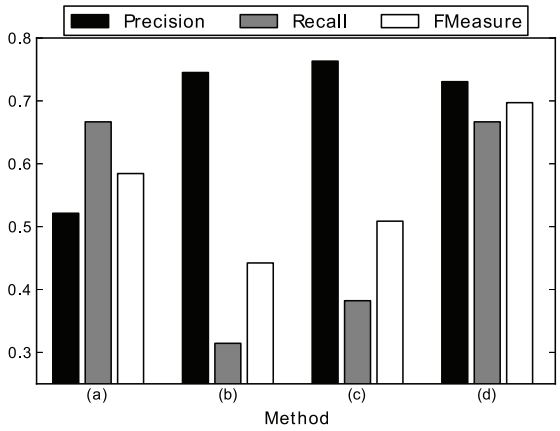


Fig. 10. Accuracy of word segmentation for each method.

B. Word Segmentation

Table II shows examples of speech recognition and word segmentation by method (a), method (b), and the proposed method (c): "Ko re wa ha n do so u pu no e ki ta i" means "This is liquid for hand soap" and "Ko re wa do u bu tsu no nu i gu ru mi" means "This is a stuffed animal." The italic letters in Table II denote phoneme errors.

To calculate the accuracy of word segmentation, we utilized dynamic programming to match the strings of word sequences and the correct word sequences segmented by a human teacher. A correctly estimated segmentation point was considered a true positive (TP), whereas a point incorrectly assigned as a segment was considered a false positive (FP). Similarly, a point that was correctly determined as not being a segment point was considered a true negative (TN), whereas a segment point incorrectly assigned as not being a cut point was considered a false negative (FN). We calculated the precision, recall, and F-measure of the segmentation as follows:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, \tag{29}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}, \tag{30}$$

$$F = \frac{2PR}{P + R}, \tag{31}$$

where N_{TP}, N_{FP}, and N_{TN} represent the number of points assessed as TP, FP, and FN. Table III shows an example where a recognized phoneme sequence "ABCD" and a segmented sequence "A/BC/D" are evaluated in the case where the answer is "AB/C/D." Fig.10 shows these results. The recall rate for method (a) was higher than those for methods (b) and (c). However, the precision of method (a) was the lowest of all methods. This is because the oNPYLM in method (a) tended to segment speech recognition results into shorter phoneme sequences, since it was difficult to correctly estimate segments because of phoneme errors. On the contrary, the precision of methods (b) and (c) was higher than that of method (a), whereas the recall was lower. This is because the robot identified frequently appearing patterns, such as pronoun-be-verb and preposition-noun, as one word, as shown in Table II. Method (d) recorded the highest score of all methods tested.

In order to assess the performance by simultaneously considering speech recognition and word segmentation, a "word entropy" was introduced. Entropy represents the dispersion of words generated from a given utterance. Therefore, the smaller the entropy, the more consistent the word generation. Word entropy H_l was calculated using the multinomial distribution obtained from frequency counts of words, which were recognized using the language model $\mathcal{L}_l$ and segmented using parameter $\hat{W}_l^w$ at the l-th learning step:

$$H_l = -\sum_{j,i,k} P(w_{jikl}) \log P(w_{jikl}), \tag{32}$$

where H_l represents the word entropy at the l-th learning step, and $P(w_{jikl})$ is the probability of the occurrence of the k-th word of the i-th utterance given the j-th object. Fig.11 shows examples of multinomial distributions. It should be noted that the components of multinomial distributions were sorted in descending order for method (a) and the proposed method (c) following the final step of learning. It can be seen that method (a) generated a multinomial distribution with a longer tail as it segmented many words with minor differences. This led to higher word entropy. Fig.12 represents a transition of word entropy for each method. The entropy increased for all methods, since the number of words in the robot's lexicon increased as learning progressed. In early stages of learning, the methods (b) and (c) yielded very small entropy values. This is because the acquired language model overfitted small amounts of data in the early stage of learning, and speech recognition results were thus biased. One can see that the proposed method yielded the lowest entropy values of the three methods. This indicates that the proposed method generated words highly consistently because it learned the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

TABLE II
EXAMPLES OF SPEECH RECOGNITION AND WORD SEGMENTATION

Teacher Utterance	kore (this) / wa (is) / ha n do so u pu (hand soap) / no (for) / e ki ta i (liquid) (This is liquid for hand soap.)
(a)	*o* re wa / ha n do / so u pu / no e *e i* / *chi* / ta *e*
(b)	*ho* re wa ha n do so / pu no e *i chi* ta *e*
(c)	ko re wa/ ha n do *zo bu*/ *n ro* e ki ta i
Teacher Utterance	ko re (this) / wa (is) / do u bu tsu (animal) / no (of) / nu i gu ru mi (stuffed) (This is a stuffed animal.)
(a)	ko re wa / do u *a* / bu tsu no / nu *gi* / *zu* ru mi
(b)	ko re wa / do u *a* / bu tsu no nu *gi zu* ru mi
(c)	ko re wa do *o* bu tsu no / nu i gu ru mi

TABLE III
EXAMPLE OF WORD SEGMENTATION EVALUATION

Result	A	/	B	C	/	D
Correct	A	B	/	C	/	D
Evaluation	TN	FP	FN	TN	TP	TN

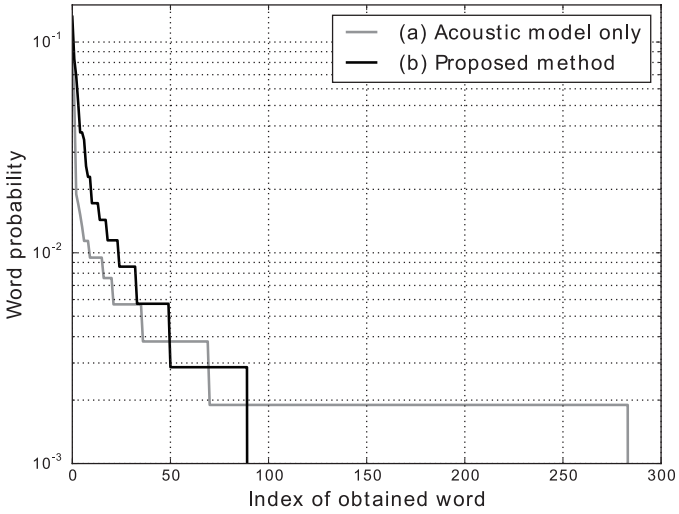


Fig. 11. Example of word distribution used to calculate word entropy.

language model and selected candidate words based on the acquired concepts.

These results indicate that the robot was able to obtain more significant words by improving the accuracy of speech recognition by updating the language model. oNPYLM in methods (b) and (c) often segmented words incorrectly, e.g., a preposition and a noun were connected. This is because the number of utterances given to the robot was small, and this problem can be solved by increasing the variety of expressions, such as by including "That is," "This was," "This has," and so on. Moreover, the precision and recall of method (c) were higher than those of method (b), and the proposed method improved the accuracy of word segmentation. This is because the proposed method selected more appropriate word sequences by using the learned concepts of objects, whereas method (b) used the 1-best word sequences without considering concepts. This result means that the robot could consistently recognize the teacher's utterances by referring to its learned model. That is, the robot was able to obtain proper word sequences by learning the language model and the concepts of objects.

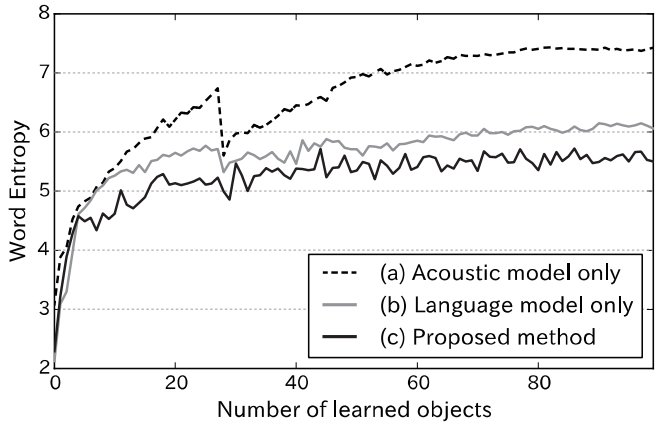


Fig. 12. Entropy for each method.

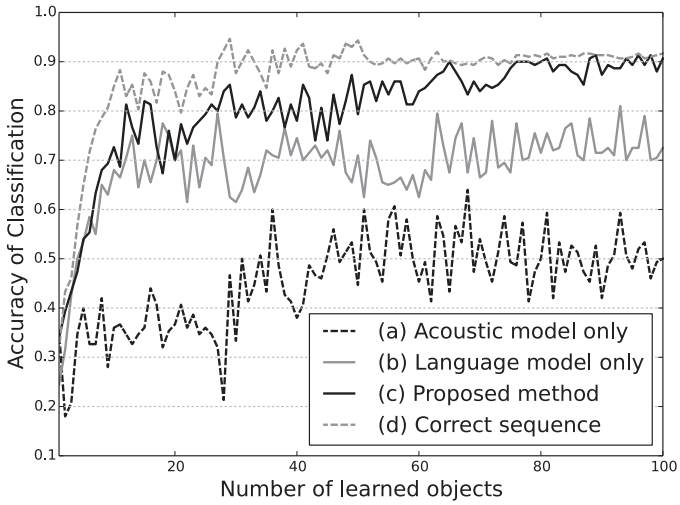


Fig. 13. Accuracy of object classification.

C. Learning the Concepts of Objects

The robot formed concepts of objects for each method. Fig.13 shows the accuracy of object classification for each method at each learning step against the correct classification shown in Fig.8. The accuracy of method (b) was

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

TABLE IV
COMPARISON BETWEEN ONLINE AND BATCH SCORES

		Online	Batch
Accuracy of Speech Recognition		0.732	0.791
Accuracy of Word Segmentation	P	0.763	0.915
	R	0.382	0.560
	F	0.509	0.695
Accuracy of Classification		0.907	0.940

higher than that of the previously proposed method (a), that of the proposed method (c) was higher still than that of method (b), and was approximately 40% higher than that of method (a). Moreover, the accuracy of the proposed method was as high as that of method (d), which used correct phoneme sequences. Fig.14 shows the classification results of the final learning step for the proposed method (c), where objects in squares denote the robot's classification mistakes. These included mistaking "candy box" for "cup noodle" (Category 7) and "hand soap" and "shampoo" (Category 8) because the features of these objects are similar—hardness, sound, and packaging. In a similar manner, mistakes in understanding the expression "cleaner sheet" were caused by similar issues. However, the robot accurately classified objects in general, which means that it can learn the concepts of objects as we do. This is because the robot was able to recognize the teacher's utterances more accurately by learning the language model and the concepts of objects. This result shows that it is important to learn language and concepts together.

D. Comparison between Online Method and Batch Method

We compared the performance of the online method with the batch method. Table IV shows the accuracy of speech recognition, word segmentation, and object classification for each.

Each score recorded by the online method was lower than that of the batch-type method. However, these scores were considered sufficiently high, except in the case of word segmentation. The accuracy of classification was especially high at approximately 90%, and was comparable to that of the batch method. With regard to the accuracy of word segmentation, that of the online method was significantly worse than that of the batch method. This is because the parameter was estimated from only the latest L phoneme sequences in the case of the online method. In particular, when the robot did not have a sufficient number of phoneme sequences at the early learning phase, it was difficult to estimate the correct cut points, and incorrect words were thus added to the lexicon. Following this, even if the robot obtained a large number of phoneme sequences, incorrect words remained in the lexicon and affected the segmentation of subsequent phoneme sequences. However, this problem can be solved by providing more utterances to the robot and increasing the number of stored sentences L.

Furthermore, an advantage to our proposed online method is the computational time. The computational time of the batch method increases in proportion to the number of learned objects. On the other hand, the computational time of the online method is constant. In fact, it required approximately 46 times longer to learn the 50-th object with the batch method than it did with the online method. With the batch method, more time is needed as more objects are learned. We consider this to be impractical for the robots learning objects interactively in a real environment. By contrast, our online method is efficient for learning more objects in such a situation.

IX. Discussion

The goal of this paper was to develop a framework for robots to learn concepts and a language model in a bottom-up manner. In general, concepts and language are closely connected to each other. There is a chicken and egg problem concerning the two: concept formation requires consistent speech recognition, and speech recognition in turn requires formed concepts. The proposed model captures this interdependence so that robots learn both the language model and concepts better than in the conventional model, which considers only concept formation. Therefore, it is interesting to see how the language model and concepts were jointly learned through the learning process. We will discuss such dynamics of learning involving the language model and concepts.

We examine the likelihood of the model to determine the progression of the learning process. Recent studies on language acquisition have revealed that children refine their learning strategies with later lexical development. The principle of conventionality contributes to gradual convergence towards adult naming patterns [41], [42]. In [43], the authors experimentally examined how children learn the meanings of basic color words, and how they are immersed into the language-specific system of the color lexicon. It is very interesting if we can compare the learning process in robots and humans in terms of interdependence between concepts and language. To see this, the variations in log likelihood over time for different models were observed.

The upper part of Fig.15 shows plots of the log likelihood of multimodal information excluding word information, such as visual, auditory, and tactile information. The lower part of Fig.15 indicates the log likelihood of word information. The yellow curve in Fig.15 (Upper) represents the result of learning using multimodal information excluding word information. In this case, the learning algorithm tried to maximize the log likelihood for given multimodal information. Therefore, the log likelihood rapidly increased and saturated around step 20 of learning. On the contrary, the log likelihoods of methods (a) and (c) gradually increased, since these methods used multimodal information, including words. Now, note that step 60 of the learning process is interesting. In the upper part of Fig.15, the log likelihood values of (a) and (e) monotonically increased, whereas those for (c) began to decline around step 60. Correspondingly, the log likelihood of word information (lower part of Fig.15) shows that the values monotonically decreased in (a). They also decreased

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

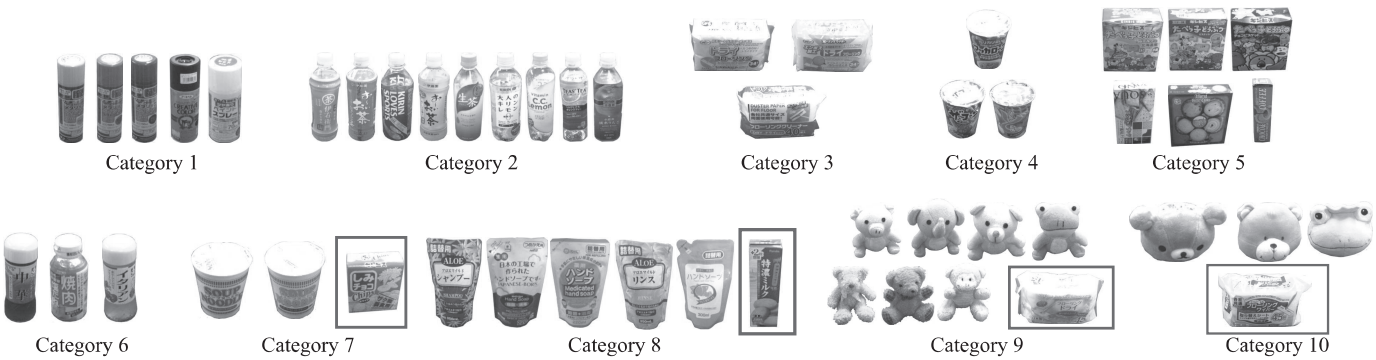


Fig. 14. Classification results.

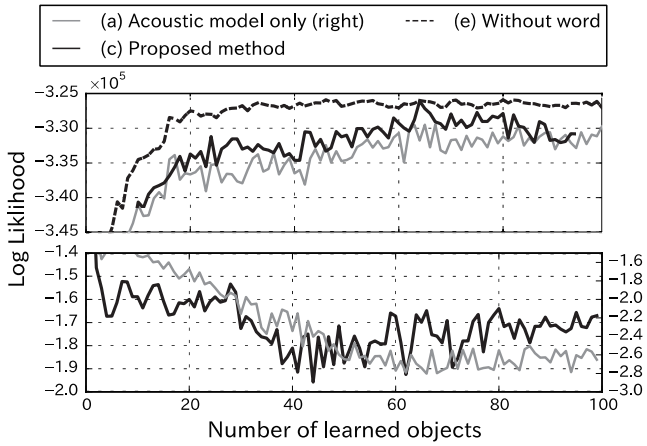


Fig. 15. Log likelihood of visual, audio, and haptic information (Upper), log likelihood of word information (Lower).

in (c), but began increasing after step 60. These results imply that robot and humans share a similar dynamics underlying their language learning mechanisms.

As mentioned above, concepts and language are interdependent in the proposed model. As the language model improves, concepts are refined, and vice versa. This bootstrapping process makes it possible for the robot to learn better concepts and the language model, which results in better performance in speech recognition and object categorization in comparison with methods (a) and (b).

More importantly, these results imply that the concept and language learning processes can be viewed as emergent systems [44]. At early stages of learning, concepts and the language model are built in a bottom-up manner and, after a while, work as constraints to correctly recognize teaching utterances, which refines concepts and the language model. This kind of loop can be seen as an emergent phenomenon.

On the contrary, the proposed algorithm may confuse different words with similar sounds if the utterances used by the user become complex. This difficulty can be resolved by multimodal information that is a key to correctly forming concepts. It also provides an important insight into how users provide teaching utterances to the robot. Roy et al. recently launched the "Human Speechome Project," and revealed that caregivers changed the form of their utterances in order to accommodate the linguistic knowledge of children [8]. We also obtained preliminary results regarding the dynamics of communication between human teachers and concepts/words learned by the robot in [7]. Although this issue is beyond the scope of this paper, we are planning to carry out a long-term experiment to observe the extent to which the robot learns concepts and the language model in real, complex situations.

X. Conclusion

In this paper, we proposed an integrated model for a robot to learn a language model and the concepts of objects online, and estimated its parameters using PFoMLDA and oNPYLM. The robot was able to reduce phoneme recognition errors, which had been a problem in our previously proposed method, and greatly improved the learning of the concepts of objects. Experimental results showed that appropriate knowledge of language and the concepts of objects can be obtained by learning them together. We considered this issue not only from the viewpoint of accuracy of learning, but also from the interrelation between the language model and concepts learned by using word entropy and classification accuracies. Furthermore, the transition of the log likelihood of models over time reveals the bootstrapping learning process of concepts and the language model.

In future work, we plan to conduct long-term experiments where robots learn the concepts of objects and language by interacting with users for several months.

Acknowledgment

This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research on Innovative Areas (Grant Number: 26118003).

References

[1] E. Rosch, "Principles of categorization," *Concepts: core readings*, pp. 189–206, 1999.

[2] A. L. Silvia Coradeschi and B. Wrede, "A Short Review of Symbol Grounding in Robotic and Intelligent Systems," *German Journal on Artificial Intelligence*, vol. 27, no. 2, pp. 129–136, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

[3] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of Word Meanings in LDA-Based Multimodal Concepts," *Advanced Robotics*, vol. 25, pp. 2189–2206, 2012.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] 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*, vol. 1, 2009, pp. 100–108.

[6] 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*, 2012, pp. 1623–1630.

[7] T. Araki, T. Nakamura, and T. Nagai, "Long-term learning of concept and word by robots: Interactive learning framework and preliminary results," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. pp.2280–2287, 2013.

[8] B. C. Roy, M. C. Frank, and D. Roy, "Exploring word learning in a high-density longitudinal corpus," in *Annual Meeting of the Cognitive Science Society*, 2009, pp. 2106–2111.

[9] 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*, 2014, pp. 600–607.

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

[11] R. Taguchi, Y. Yamada, K. Hattori, T. Umezaki, M. Hoguro, N. Iwahashi, K. Funakoshi, and M. Nakano, "Learning placenames from spoken utterances and localization results by mobile robot," in *INTERSPEECH'11*, 2011, pp. 1325–1328.

[12] S. Wermter, C. Weber, M. Elshaw, C. Panchev, H. Erwin, and F. Pulvermüller, "Towards Multimodal Neural Robot Learning," *Robotics and Autonomous Systems*, vol. 47, no. 2, pp. 171–175, 2004.

[13] B. Ridge, D. Skocaj, and A. Leonardis, "Self-supervised Cross-Modal Online Learning of Basic Object Affordances for Developmental Robotic Systems," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 5047–5054.

[14] T. Ogata, S. Nishide, H. Kozima, K. Komatani, and H. Okuno, "Inter-Modality Mapping in Robot with Recurrent Neural Network," *Pattern Recognition Letters*, vol. 31, no. 12, pp. 1560–1569, 2010.

[15] S. Lallee and P. F. Dominey, "Multi-modal Convergence Maps: From Body Schema and Self-Representation to Mental Imagery," *Adaptive Behavior*, vol. 21, no. 4, pp. 274–285, 2013.

[16] O. Mangin and P.-Y. Oudeyer, "Learning Semantic Components from Subsymbolic Multimodal Perception," in *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics*, 2013, pp. 1–7.

[17] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *IEEE International Conference on Computer Vision*, 2005, pp. 17–20.

[18] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 264–271.

[19] L. Fei-Fei, "A bayesian hierarchical model for learning natural scene categories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 524–531.

[20] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, 2009, pp. 1903–1910.

[21] F. Endres, C. Plagemann, C. Stachniss, and W. Burgard, "Unsupervised discovery of object classes from range data using latent dirichlet allocation," in *Robotics: Science and Systems*, 2009.

[22] J. Sinapov and A. Stoytchev, "Object category recognition by a humanoid robot using behavior-grounded relational learning," in *IEEE International Conference on Robotics and Automation*, 2011, pp. 184–190.

[23] L. Natale, G. Metta, and G. Sandini, "Learning haptic representation of objects," in *IEEE International Conference on Intelligent Manipulation and Grasping*, 2004.

[24] R. Russell, "Object recognition by a 'smart' tactile sensor," in *Australian Conference on Robotics and Automation*, 2000.

[25] A. Schneider, J. Sturm, C. Stachniss, M. Reisert, H. Burkhardt, and W. Burgard, "Object identification with tactile sensors using bag-of-features," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 243–248.

[26] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[27] Y. Yang, A. Guha, C. Fermuller, and Y. Aloimonos, "A cognitive system for understanding human manipulation actions," *Advances in Cognitive Sysytems*, vol. 3, pp. 67–86, 2014.

[28] X. Yu, C. Fermuller, C. L. Teo, Y. Yang, and Y. Aloimonos, "Active scene recognition with vision and language," in *IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 810–817.

[29] C. L. Teo, Y. Yang, H. Daumé III, C. Fermüller, and Y. Aloimonos, "Towards a watson that sees: Language-guided action recognition for robots," in *IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 374–381.

[30] M.-M. Cheng, S. Zheng, W.-Y. Lin, V. Vineet, P. Sturgess, N. Crook, N. J. Mitra, and P. Torr, "Imagespirit: Verbal guided image parsing," *ACM Transactions on Graphics*, vol. 34, no. 1, p. 3, 2014.

[31] M. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent Dirichlet allocation," in *Advances in Neural Information Processing Systems*, vol. 23, 2010, pp. 856–864.

[32] I. Sato, K. Kurihara, and H. Nakagawa, "Deterministic single-pass algorithm for lda," in *Advances in neural information processing systems*, 2010, pp. 2074–2082.

[33] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[34] K. Canini, L. Shi, and T. Griffiths, "Online inference of topics with latent Dirichlet allocation," in *International Conference on Artificial Intelligence and Statistics*, vol. 5, 2009, pp. 65–72.

[35] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. Suppl. 1, pp. 5228–5235, 2004.

[36] T. Nakamura, T. Nagai, and N. Iwahashi, "Multimodal Object Categorization by a Robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 2415–2420.

[37] ——, "Grounding of word meanings in multimodal concepts using LDA," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 3943–3948.

[38] T. Araki, T. Nakamura, T. Nagai, K. Funakoshi, M. Nakano, and N. Iwahashi, "Autonomous acquisition of multimodal information for online object concept formation by a robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1540–1547.

[39] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *ACM International Conference on Multimedia*, 2010, pp. 1469–1472.

[40] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2009, pp. 131–137.

[41] E. Ameel, B. Malt, and G. Storms, "Object naming and later lexical development: From baby bottle to beer bottle," *Journal of Memory and Language*, vol. 58, pp. 262–285, 2008.

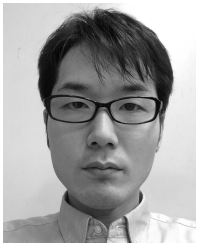
[42] N. Saji, M. Imai, H. Saalbach, Y. Zhang, H. Shuc, and H. Okada, "Word learning does not end at fast-mapping: Evolution of verb meanings through reorganization of an entire semantic domain," *Cognition*, vol. 118, pp. 45–61, 2011.

[43] N. Saji, M. Asano, M. Oishi, and M. Imai, "How do children construct the color lexicon? :restructuring the domain as a connected system," in *Annual Meeting of the Cognitive Science Society*, 2015, pp. 2080–2085.

[44] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: A survey," *arXiv:1509.08973*, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2552579, IEEE Transactions on Cognitive and Developmental Systems

Joe Nishihara received his bachelor's degree from the University of Electro-Communications in 2014. He received his master's degree from the Graduate School of Electro-Communications, University of Electro-Communications in 2016. He is a member of RSJ.



Tomoaki Nakamura received his BE, ME, and Dr. of Eng. degrees from the University of Electro-Communications in 2007, 2009, and 2011. From 2011 to 2012, He was a research fellow of the Japan Society for the Promotion of Science. In 2013, he worked for Honda Research Institute Japan Co., Ltd. He is currently an assistant professor at the University of Electro-Communications. He is a member of RSJ and JSAI.

Takayuki Nagai received his BE, ME, and DE degrees from the Department of Electrical Engineering, Keio University, in 1993, 1995, and 1997, respectively. Since 1998, he has been with the University of Electro-Communications where he is currently a professor of the Graduate School of Informatics and Engineering. From 2002 to 2003, he was a visiting scholar at the Department of Electrical Computer Engineering, University of California, San Diego. Since 2011, he has also been a visiting researcher at Tamagawa University Brain Science Institute. He has received the 2013 Advanced Robotics Best Paper Award. He is a member of the IEEE, RSJ, JSAI, IEICE and IPSJ.