Robots that Learn Language: Developmental Approach to Human-Machine Conversations

Naoto Iwahashi[1,2]

[1] National Institute of Information and Communication Technology,
[2] ATR Spoken Language Communication Research Labs
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
naoto.iwahashi@atr.jp
http://www.slc.atr.jp/~niwaha/

Abstract. This paper describes a machine learning method that enables robots to learn the capability of linguistic communication from scratch through verbal and nonverbal interaction with users. The method focuses on two major problems that should be pursued to realize natural human-machine conversation: a scalable grounded symbol system and belief sharing. The learning is performed in the process of joint perception and joint action with a user. The method enables the robot to learn beliefs for communication by combining speech, visual, and behavioral reinforcement information in a probabilistic framework. The beliefs learned include speech units like phonemes or syllables, a lexicon, grammar, and pragmatic knowledge, and they are integrated in a system represented by a dynamical graphical model. The method also enables the user and the robot to infer the state of each other's beliefs related to communication. To facilitate such inference, the belief system held by the robot possesses a structure that represents the assumption of shared beliefs and allows for fast and robust adaptation of it through communication with the user. This adaptive behavior of the belief systems is modeled by the structural coupling of the belief systems held by the robot and the user, and it is performed through incremental online optimization in the process of interaction. Experimental results reveal that through a practical, small number of learning episodes with a user, the robot was eventually able to understand even fragmental and ambiguous utterances, act upon them, and generate utterances appropriate for the given situation. This work discusses the importance of properly handling the risk of being misunderstood in order to facilitate mutual understanding and to keep the coupling effective.

1 Introduction

The process of human communication is based on certain beliefs shared by those communicating. Language is one such shared belief, and it is used to convey meaning based on its relevance to other shared beliefs [1]. These shared beliefs

are formed through interaction with the environment and other people, and the meaning of utterances is embedded in such shared experiences.

From the perspective of objectivism, if those communicating want to logically convince each other that proposition p is a shared belief, they must prove that the infinitely nested proposition, "They have information that they have information that ... that they have information that p," also holds. However, in reality, all we can do is assume, based on a few clues, that our beliefs are identical to those of the other people we are talking to. In other words, it can never be guaranteed that our beliefs are identical to those of other people. Because shared beliefs defined from the viewpoint of objectivism do not exist, it is more practical to see shared beliefs as a process of interaction between the belief systems held by each person communicating. The processes of generating and understanding utterances rely on the system of beliefs held by each person, and this system changes autonomously and recursively through these two processes. Through utterances, people simultaneously send and receive not only the meanings of their words but also, implicitly, information about each other's system of beliefs. This dynamical process works in a way that makes the belief systems consistent with each other. In this sense, we can say that the belief system of one person couples structurally with the belief systems of those with whom he or she is communicating [2].

Communication by spoken language is one of the most natural methods for human-machine interfaces. The progress made in sensor technologies and in the infrastructure of ubiquitous computing has enabled machines to sense physical environments as well as the behavior of users. In the near future, machines that change their behavior according to the situation in order to support human activities in everyday life will become more and more common, and for this they should feature user-centered intelligent interfaces. One way to obtain such interfaces is through personalization [3], and one of the most essential features of personalized multimodal interfaces is the ability of the machine to share experiences with the user in the physical world. In the future, spoken language interfaces will become increasingly important not only because they enable hands-free interaction but also because of the nature of language, which inherently conveys meaning based on shared experiences as mentioned above. For us to take advantage of such interfaces, language processing methods must make it possible to reflect shared experiences.

However, existing language processing methods, which are characterized by fixed linguistic knowledge, do not satisfy this requirement [4]. In these methods, information is represented and processed by symbols whose meaning has been pre-defined by the machines' developers. In most cases, the meaning of each symbol is defined by its relationship to other symbols, and it is not connected to perception or to the physical world. The precise nature of experiences shared by a user and a machine, however, depends on the situation. Because it is impossible to prepare symbols for all possible situations in advance, machines cannot appropriately express and interpret experiences under dynamically changing sit-

uations. As a result, users and machines fail to interact in a way that accurately reflects shared experiences.

To overcome this problem and realize natural linguistic communication between humans and machines, the methods should satisfy the following requirements.

Scalable Grounded Symbol System: The machines themselves must be able to create a symbol system that reflects their experiences in natural ways. Such a symbol system has to include symbols for perceptual categories, abstract concepts, words, and the map between word sequences (or forms) and meanings (or functions). The information of language, perception, and actions should be processed in an integrative fashion. Perceptual categories should be extracted from this information, and the abstract concepts created based on these categories [5]. The created symbols should then be embedded in an adaptively changing belief system, in which the relations among symbols are represented based on experienced events in the real world. The grounding of the meanings of utterances in conversation in the physical world was explored in [6] and [7], but they did not pursue the learning of grounded symbols.

Belief Sharing: In communications, grounded beliefs held by a user and a machine should ideally be as identical or consistent to each other as possible, with the machine and the user coordinating their utterances and actions to form such beliefs. To achieve such coordination, the machines should include a mechanism that enables the user and machine to infer the state of each other's belief system in a natural way. When a participant interprets an utterance based on their assumptions that certain beliefs are shared and is convinced, based on certain clues, that the interpretation is correct, he or she strengthens the confidence that the beliefs are shared. On the other hand, since the sets of beliefs assumed to be shared by participants actually often contain discrepancies, the more beliefs a listener needs to understand an utterance, the greater the risk that the listener misunderstands it. Therefore, to realize appropriate coupling of belief systems, the computational mechanism should produce utterances so as to control the balance between the transmissions of the meanings of utterances and the information on the state of belief systems. Theoretical research [8] and computational modeling [9] focused on the formation of shared beliefs through the transmission of utterance meanings have attempted to represent the formation of shared beliefs as a procedure- and rule-driven process. In contrast, we should focus on the system of beliefs to be used in the process of generating and understanding utterances in a physical environment; moreover, it is important to represent the formation of this system by a mathematical model to achieve robust communication.

Both of these requirements show that the capability of learning is essential in communications. The cognitive activities related to a scalable grounded symbol system and belief sharing can be observed clearly in the process of language

acquisition by infants as well as in everyday conversation by adults. To focus on learning capabilities in communication, we have been taking on the challenge of developing a method that enables robots to learn linguistic communication capability from scratch through verbal and nonverbal interaction with users [10–12], instead of directly pursuing language processing for everyday conversation.

Language acquisition by machines has been attracting interest in various research fields [13], and several pioneering studies have developed algorithms based on inductive learning by using a set of pairs, where each pair consists of a word sequence and nonlinguistic information about its meaning. In [14–18], visual information, rather than symbolic, was given as nonlinguistic information. Spoken-word acquisition algorithms based on the unsupervised clustering of speech tokens have also been described [19, 15, 17]. In [20, 21], the socially interactive process for the evolution of grounded linguistic knowledge shared by communication agents was examined from the viewpoint of game theory and a complex adaptive system. In [22], a connectionist model for acquiring the semantics of language through the behavioral experiences of a robot was presented, focusing on the compositionality of semantics.

In contrast, the method described in this paper focuses on online learning of a pragmatic capability in the real world through verbal and nonverbal interaction with humans, as well as consideration to the above two requirements. This approach enables a robot to develop the pragmatic capability within a short period of interaction by fast and robust adaptation of its belief system relative to a user. This fast and robust adaptation is a very important feature, since a typical user cannot tolerate extended interaction with a robot that does not possessed communication capability and, moreover, situations in actual everyday conversation continuously change.

The learning method applies information from raw speech and visual observations as well as behavioral reinforcement, which is integrated in a probabilistic framework. A system of beliefs belonging to the robot includes speech units like phonemes or syllables, a lexicon consisting of words whose meanings are grounded in vision and motion, simple grammar, non-linguistic beliefs, the representation of the assumption of shared beliefs, and the representation of the consistency between the belief systems of the user and the robot. This belief system is represented by a dynamical graphical model (e.g. [23]), and expands step-by-step through learning. First, the robot learns the basic linguistic beliefs, which comprise speech units, lexicon, and grammar, based on joint perceptual experiences between the user and the robot [10, 12]. Then, the robot learns an entire belief system based on these beliefs online in an interactive way to develop a pragmatic capability [11]. The belief system has a structure that reflects the state of the user's belief system; thus, the learning makes it possible for the user and the robot to infer the state of each other's belief systems. This mechanism works to establish appropriate structural coupling, leading to mutual understanding.

This paper proceeds as follows. Section 2 describes the setting for the robot to learn linguistic communication. The requirements on a scalable grounded symbol system and belief sharing are mainly addressed from Sec. 3 to Sec. 5 and

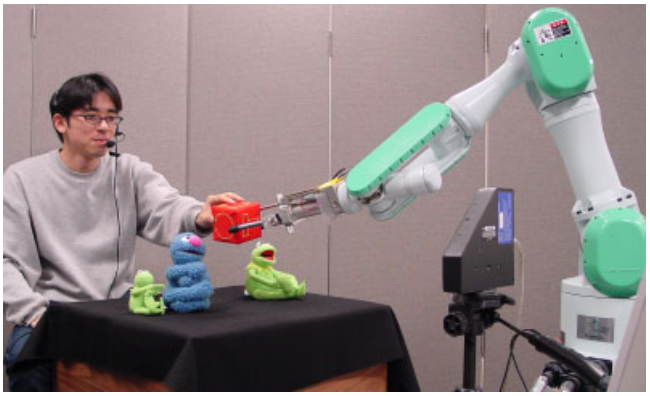


Fig. 1. Interaction between a user and a robot

in Sec. 6, respectively. Section 3 explains the method of learning speech units, followed by Sec. 4, which describes the learning method of words referring to objects, motions, and abstract concepts. Section 5 relates to the learning method of simple grammar. Section 6 addresses the method for learning pragmatic capability, which enables the structural coupling of belief systems held by a robot and a user. Section 7 discusses the findings and mentions future works.

2 Setting for Learning

2.1 Interaction

The spoken-language acquisition task in this work was set up as follows. A camera unit and a robot arm with a hand were set alongside a table, and a participant and the learning robot saw and moved the objects on the table as shown in Fig. 1. The robot arm had seven degrees of freedom and the hand had one. A touch sensor was attached to the robot's hand. The robot initially did not possess any concepts regarding the specific objects or the ways in which they can be moved, nor did it have any linguistic knowledge.

The interactions for step-by-step learning were carried out as follows. First, in learning speech units, a participant spoke for approximately one minute. Second, in learning words that refer to objects, the participant pointed to an object on the table while speaking a word describing it. A sequence of such learning episodes provides a set of pairs, each of which is comprised of the image of an object and the speech describing it. The objects used included boxes, stuffed and wooden toys, and balls (examples are shown in Fig. 2). In addition, in each of the episodes for learning words referring to motions, the participant moved an object while speaking a word describing the motion. Third, in each of the episodes for learning grammar, the participant moved an object while uttering a sentence describing the action. By the end of this learning, the participant and

Fig. 2. Examples of objects used

the robot had shared certain linguistic beliefs consisting of a lexicon and simple grammar, and the robot could understand utterances [3] to some extent.

Finally, in the learning of pragmatic capability, the participant asked the robot to move an object by making an utterance and a gesture, and the robot acted in response. If the robot responded incorrectly, the user slapped the robot's hand. The robot also asked the user to move an object, and the user acted in response. The robot's system of beliefs was formed incrementally, online, through such interaction.

2.2 Speech and image signal processing

A close-talk microphone was used for speech input. The camera unit contained three separate CCDs so that three-dimensional information on each scene could be obtained. The information regarding the position in terms of the depth coordinate was used in the attention-control process.

Speech was detected and segmented based on changes in the short-time power of speech signals, and objects were detected when they were located at a distance of 50-80 cm from the stereo camera unit. All speech and visual sensory output was converted into predetermined features. The speech features used were Mel-frequency cepstral coefficients [24], which are based on short-time spectrum analysis, their delta and acceleration parameters, and the delta of short-time log power. These features (25-dimensional) were calculated in 20-ms intervals with a 30-ms-wide window. The visual features used were position on the table (two-dimensional: horizontal and vertical coordinates), velocity (two-dimensional), L*a*b* components (three dimensions) for the color, complex Fourier coefficients (eight dimensions) of 2D contours for the shape [25], and the area of an object (one dimension) for the size. Trajectory of the object's motion is represented by a time-sequence of its positions.

[3] No function words are included in the lexicon.

3 Learning Speech Units

3.1 Difficulty

Speech is a time-continuous one-dimensional signal. The method learns statistical models of the speech units from such a signal without any transcription on phoneme sequence nor any boundaries between phonemes being given. The difficulty of learning speech units is ascribed to the difficulties of speech segmentation and the clustering of speech segments into speech units.

3.2 Method using Hidden Markov Models

It is possible to cope with the difficulty described above by using Hidden Markov Models (HMMs) and their learning algorithm called the Baum-Welch algorithm [26]. The HMM is a particular form of a graphical model that statistically represents dynamic characteristics of time-series data. It consists of unobservable states, each of which has a probability distribution of observed data, and the probabilities of transitions between them. The Baum-Welch algorithm makes it possible to perform the segmentation, clustering, and learning of HMM parameters simultaneously.

In this method, each speech unit HMM includes three states and allows for left-to-right transitions. Twenty speech unit HMMs were connected to one another to construct a whole speech unit HMM (Fig. 3), in which transitions were allowed from the last states of the speech unit HMMs to their first states. All parameters of this HMM were learned using speech data approximately one minute in length without any phoneme transcriptions. After learning the speech unit HMMs, the individual speech unit HMMs h_1, h_2, h_3, ..., and h_{N_p} were separated from one another by deleting edges between them, and a speech unit HMM set was constructed. The model for each spoken word was represented by connecting these speech unit HMMs.

3.3 Number of speech units

In the above method, the number N_p of speech unit models was determined empirically. However, ideally it should be learned from speech data. Such a method has already been presented [10], which learns the number of speech units and the number of words simultaneously from data comprising pairs of an image of an object and a spoken word describing it. The model performs in a batch-like manner using mutual information between the image and speech observations.

4 Learning Words

4.1 Words referring to objects

Difficulty In general, the difficulty of acquiring spoken words and the visual objects they refer to as their meanings can be ascribed to the difficulties in specifying features and extending them.

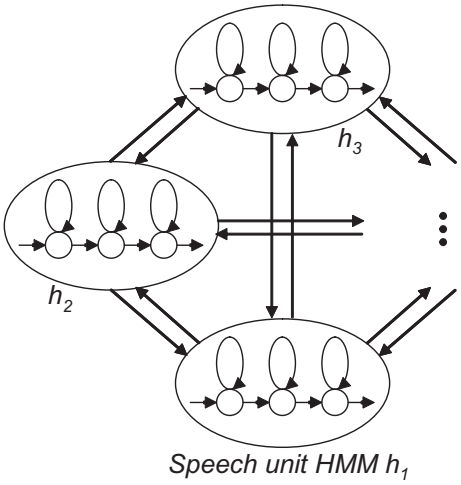


Fig. 3. Structure of HMM for learning speech units

Specification: The acoustic features of a spoken word and the visual features of an object to which it refers should be specified using spatiotemporally continuous audio-visual data. For speech, this means that a continuously spoken utterance is first segmented into intervals, after which acoustic features are extracted from one of the segmented intervals. For objects, this means that an object is first selected for a given situation, and then the spatial part of the object is segmented; after that, visual features are extracted from the segmented part of the object.

Extension: In order to create categories for a given word and its meaning, it is necessary to determine what other features fall into the category to which the specified features belong. This extension of the features of a word's referent to form the word's meaning has been investigated through psychological experiments [27]. When shown an object and given a word for it, human subjects tend to extend the features of the referent immediately to infer a particular meaning of the word, a cognitive ability called *fast mapping* (e.g. [28]), although such inference is not necessarily correct. For machines, however, the difficulty in acquiring spoken words arises not only from the difficulty in extending the features of referents but also from that in understanding spoken words. This is because the accuracy of speech recognition by machines is currently much lower than that by humans, meaning it is not easy for machines to determine whether two different speech segments belong to the same word category.

Learning Method The method described here mainly addresses the problem of extension, in which learning is carried out in an interactive way [12]. The user shows a physical object to the robot and at the same time speaks the name of

Fig. 4. A scene in which utterances were made and understood

the object or its description. The robot then decides whether the input word is a word in its vocabulary (whether it is a *known* word) or not (whether it is an *unknown* word). If the robot judges that the input word is an unknown word, it enters the word into its vocabulary. If the robot judges that it cannot make an accurate decision, it asks the user a question to confirm whether the input word is part of its vocabulary. For the robot to make a correct decision, it uses not only speech but also visual information about the objects to make an accurate decision about an unknown word. For example, when the user shows an orange and says the word /ɔrinʒ/, even if the speech recognizer outputs an unknown word /areːʒ/ as the first candidate, the system can modify it to the correct word /ɔrinʒ/ in the lexicon using visual clues. Such a decision is carried out by using a function that represents the confidence that an input pair of image o and speech s belongs to each existing word category w and is adaptively changed online.

Each word or lexical item to be learned includes statistical models, $p(s|w)$ and $p(o|w)$, for the spoken word and an object image category for its meaning. The model for each image category $p(o|w)$ is represented by a Gaussian function in a twelve-dimensional visual feature space (in terms of shape, color and size), and learned based on a Bayesian method (e.g. [29]) every time an object image is given. The Bayesian method makes it possible to determine the area in the feature space that belongs to an image category in a probabilistic way, even if only a single sample is given. Learned words include those that refer to the whole objects, shapes, colors, sizes, and combinations of them. The model for each spoken word $p(s|w)$ was represented by a concatenation of speech unit HMMs; this extends a speech sample to a spoken word category.

4.2 Words referring to motions

The concept of motion of moving objects represents the time-varying spatial relation between a trajector and a landmark [30]. In Fig. 4, for instance, if the stuffed toy in the middle and the box at the right are considered landmarks, the movements of the trajector are understood as *move over* and *move onto*,

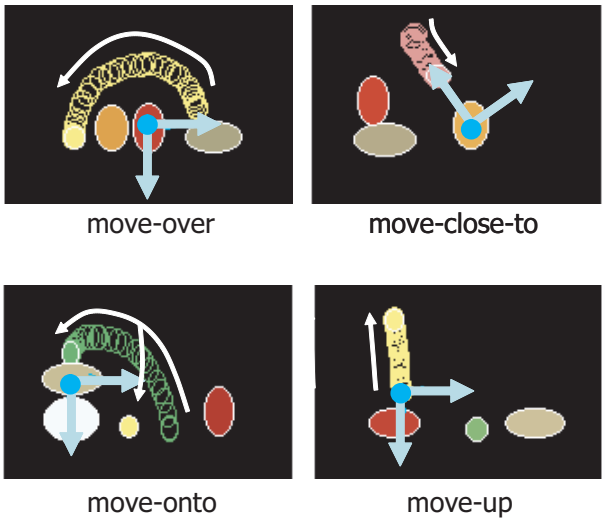


Fig. 5. Examples of trajectories of objects moved in the learning episodes, and selected landmarks and coordinates

respectively. The robot has to infer the landmark selected in each scene, which is not observed in the learning data. In addition, the coordinates in the space should be determined to appropriately represent the graphical model for each concept of a motion.

In the proposed method [31], the concepts regarding motions are represented by probability density functions of the trajectory u of moved objects. The probability density function $p(u|o_{t,p}, o_{l,p}, w)$ for the trajectory of the motion referred by word w is represented by a HMM given the positions $o_{t,p}, o_{l,p}$ of a trajector and a landmark. The HMMs of the motions are learned while the coordinates and the landmarks are being inferred based on the EM algorithm, in which a landmark is taken as a latent variable. Examples of inferred landmarks and coordinates in the learning of some motion concepts are shown in Fig. reffig:example motion.

The trajectory for the motion referred by a word is generated by maximizing the output probability of the learned HMM, given the positions of a trajector and a landmark. This maximization is carried out by the algorithm described in [32].

A graphical model of the lexicon containing words referring to objects and motions is shown in Fig. 6

4.3 Abstract meanings

The categories that are learned by the previously mentioned methods are formed directly from perceptual information. However, we have to consider words that

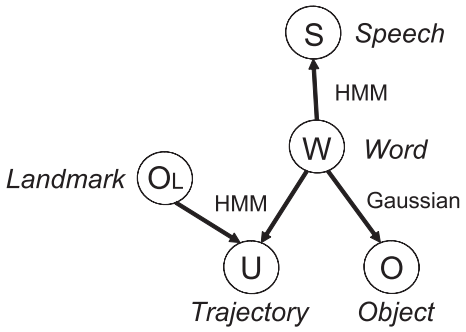


Fig. 6. A graphical model of a lexicon containing words referring to objects and motions

refer to concepts whose levels of abstractness are higher and that are not formed directly from perceptual information, such as "tool," "food," and "pet." In a study on the abstract nature of symbols' meanings [33], it was shown that chimpanzees could learn the lexigrams (graphically represented words) that refer to not only individual object categories (e.g. "banana," "apple," "hammer" and "key") but also the functions ("tool" and "food") of the objects. They could also learn the connection between the lexigrams referring to these two kinds of concepts and generalize it appropriately to connect new lexigrams for individual objects to one of the lexigrams for functions.

A method enabling robots to have this capability of chimpanzees was proposed in [34]. In that method, the motions given to objects are taken as their functions. The main problem is the decision regarding whether the meaning of a new input word is for a concept formed directly from perceptual information or for a function of objects. Because these two kinds of concepts are allocated to the states of different nodes in the graphical model, the problem becomes the selection of the structures of the graphical model. This selection is performed by the Bayesian principle with the calculation of posterior probabilities using the variational Bayes method [35].

5 Learning Grammar

5.1 Difficulty

In learning grammar using moving images of actions and speech describing them, the robot should detect the correspondence between a semantic structure in the moving image and a syntactic structure in speech. However, such semantic and syntactic structures are not observable. While we can extract an enormous number of structures from a moving image and speech, we ideally select the ones for which the correspondence between them is the most appropriate. The grammar should be statistically learned using such correspondences, and inversely used to extract the correspondence.

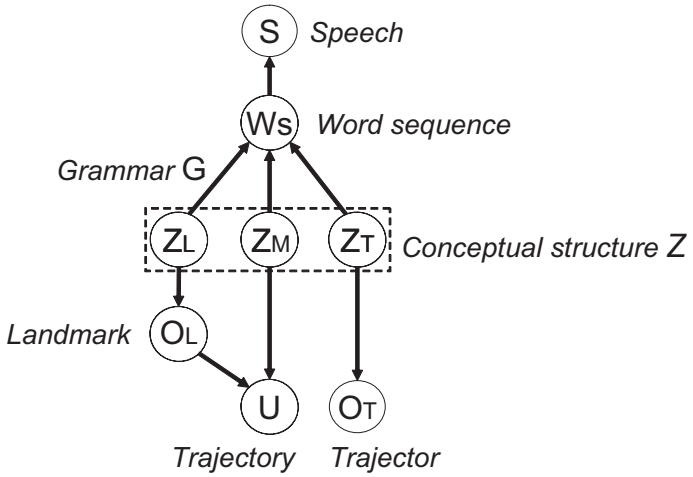


Fig. 7. Graphical model of lexicon and grammar

5.2 Learning method

The set comprising triplets of a scene O before an action, the action a, and a sentence utterance s describing the action, $\mathcal{D}_g = \{(s_1, a_1, O_1), (s_2, a_2, O_2), \ldots, (s_{N_g}, a_{N_g}, O_{N_g})\}$, is given in this order as learning data. Scene O_i includes the set of positions $o_{j,p}$ and features $o_{j,f}$ concerning color, size, and shape, $j = 1, ..., J_i$, of all objects in the scene. The action a_i is represented by a pair, (t_i, u_i), of trajector object t_i and the trajectory u_i of its movement.

It is assumed that each utterance is generated based on the stochastic grammar G. The grammar G is learned by maximizing the likelihood of the joint probability density function $p(s, a, O; L, G)$, where L denotes a parameter set of the lexicon. This function is represented by a graphical model with an internal structure that includes the parameters of the grammar G and the conceptual structure z that the utterance represents (Fig. 7).

The conceptual structure used here is expressed with three attributes as the elements in an image schema - [motion], [trajector], and [landmark] - that are initially given to the system, and they are fixed. For instance, when the image is the one shown in Fig. 4 and the corresponding utterance is the sequence of spoken words *"large Kermit brown box move-onto"*, the conceptual structure might be

$$\begin{bmatrix} Z_T \text{ [trajector]} & : \textit{large Kermit} \\ Z_L \text{ [landmark]} & : \textit{brown box} \\ Z_M \text{ [motion]} & : \textit{move-onto} \end{bmatrix},$$

where the right-hand column contains the spoken word sub-sequences referring to trajector, landmark, and motion, in a moving image. Let y denote the order of conceptual attributes, which also represents the order of the constituents

with the conceptual attributes in an utterance. For instance, in the above utterance example, the order is [trajector]-[landmark]-[motion]. The grammar is represented by the set comprising occurrence probabilities of the possible orders as $G = \{P(y_1), P(y_2), ..., P(y_k)\}$. By assuming $p(z, O\ ; L, G)$ is constant, the joint log-probability density function is written as

$$\begin{aligned}
&\log p(s, a, O\ ; L, G) \\
&= \log \sum_z p(s|z\ ; L, G) p(a|z, O\ ; L, G) p(z, O\ ; L, G) \\
&\approx \alpha \max_{z,l} \Bigg(\log p(s|z\ ; L, G) && \text{[Speech]} \\
&\quad + \log p(u|o_{t,p}, o_{l,p}, W_M\ ; L) && \text{[Motion]} \\
&\quad + \log p(o_{t,f}|W_T\ ; L) + \log p(o_{l,f}|W_L\ ; L) \Bigg) \quad . && \text{[Static Image of Object]}
\end{aligned} \tag{1}$$

where α is a constant value of $p(z, O; L, G)$. Furthermore, t and l are discrete variables across all objects in each moving image, and represent, respectively, a trajector object and a landmark object. In addition, W_M, W_T, and W_L are, respectively, word sequences corresponding to the motion, trajector, and landmark in the conceptual structure z.

The estimate $\tilde{G}_i$ of grammar G given ith learning data is obtained as the maximum values of the posterior probability distribution as

$$\tilde{G}_i = \underset{G}{\operatorname{argmax}} \, p(G\ |\mathcal{D}_g^i\ ; L) \quad . \tag{2}$$

where $\mathcal{D}_g^i$ denotes learning sample set $\{(s_1, a_1, O_1), (s_2, a_2, O_2), \ldots, (s_i, a_i, O_i)\}$. An utterance asking the robot to move an object is understood using the lexicon L and the grammar G, and one of the objects in the current scene O is accordingly grasped and moved by the robot arm. The algorithm that understands speech s infers the conceptual structure $z = (W_T, W_L, W_M)$ and generates action $\tilde{a} = (\tilde{t}, \tilde{u})$ as

$$\tilde{a} = \underset{a}{\operatorname{argmax}} \log p(s, a, O\ ; L, \tilde{G}) \quad . \tag{3}$$

The robot arm is controlled according to the generated trajectory $\tilde{u}$.

6 Learning Pragmatic Capability Based on Coupling of Belief Systems

6.1 Difficulty

As mentioned in Sec. 1, a pragmatic capability relies on the capability to infer the state of another participant's belief system. The computational mechanism

should enable the robot to adapt its assumption of shared beliefs rapidly and robustly through verbal and nonverbal interaction. It also should control the balance between transmissions of the meaning of utterances and the information on the state of belief systems. The following is an example of generating and understanding utterances based on the assumption of shared beliefs. Suppose that in the scene shown in Fig. 4 the object on the left, Kermit, has just been put on the table. If the user in the figure wants to ask the robot to move Kermit onto the box, he may say, *"Kermit box move-onto"*. In this situation, if the user assumes that the robot shares the belief that the object moved in the previous action is likely to be the next target for movement and the belief that the box is likely to be something for the object to be moved onto, he might just say *"move-onto"*. To understand this fragmental utterance, the robot has to possess similar beliefs. If the user knows that the robot has acted as he has asked in response, he would strengthen the confidence that the beliefs he has assumed to be shared are really shared. Inversely, when the robot wants to ask the user to do something, the beliefs that it assumes to be shared are used in the same way. We can see that the former utterance is more effective than the latter in transmitting the meaning of the utterance, while the latter is more effective in transmitting the information on the state of belief systems.

6.2 Representation of a system of beliefs

To cope with the above difficulty, a system of beliefs needs to consist of the following two parts:

Shared belief function, which represents the assumption of shared beliefs and is composed of a set of belief modules with values (local confidence) representing the degree of confidence that each belief is shared by the robot and the user.

Global confidence function, which represents the degree of confidence for the shared belief function.

Such a belief system is depicted in Fig. 8. The beliefs we used are those concerning speech, motions, static images of objects, behavioral context, and motion-object relationship. The behavioral context and motion-object relationship are represented as follows.

Motion-object relationship $B_R(o_{t,f}, o_{l,f}, W_M; R)$ **:** The motion-object relationship represents the belief that in the motion corresponding to motion word W_M, feature $o_{t,f}$ of object t and feature $o_{l,f}$ of object l are typical for a trajector and a landmark, respectively. This belief is represented by a conditional multivariate Gaussian probability density function, $p(o_{t,f}, o_{l,f}|W_M; R)$, where R is its parameter set.

Effect of behavioral context $B_H(i, q; H)$ **:** The effect of behavioral context represents the belief that the current utterance refers to object i, given behavioral context q. Here, q includes information on whether object i was a trajector or a landmark in the previous action and whether the user's current

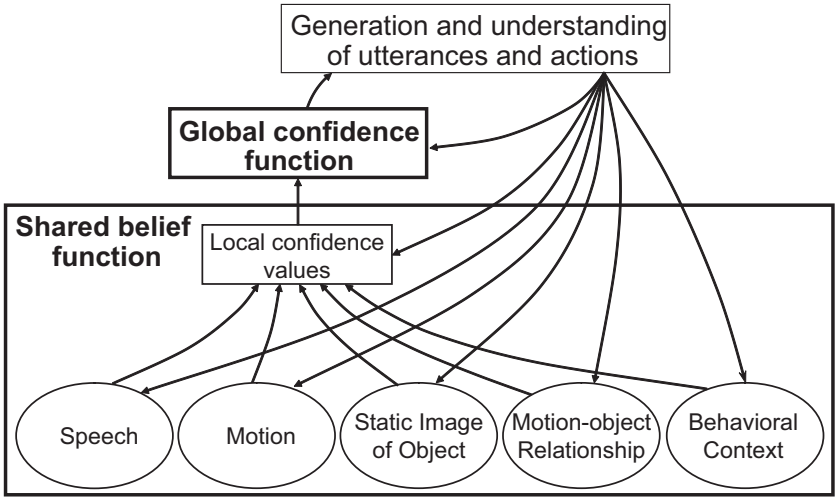


Fig. 8. Belief system of the robot that consists of shared belief and global confidence functions.

gesture is referring to object i. This belief is represented by a parameter set H.

6.3 Shared belief function

The beliefs described above are organized and assigned local confidence values to obtain the shared belief function used in the processes of generating and understanding utterances. This shared belief function Ψ is the extension of $\log p(s, a, O; L, G)$ in Eq. 1. The function outputs the degree of correspondence between utterance s and action a, and it is written as

$$
\begin{aligned}
&\Psi(s, a, O, q, L, G, R, H, \Gamma) \\
&= \max_{l,z} \Bigg(\gamma_1 \log p(s|z\ ;\ L, G) && \text{[Speech]} \\
&\quad + \gamma_2 \log p(u|o_{t,p}, o_{l,p}, W_M\ ;\ L) && \text{[Motion]} \\
&\quad + \gamma_2 \Big(\log p(o_{t,f}|W_T\ ; L) + \log p(o_{l,f}|W_L\ ; L) \Big) && \text{[Static Image of Object]} \\
&\quad + \gamma_3 \log p(o_{t,f}, o_{l,f}|W_M\ ; R) && \text{[Motion-Object Relationship]} \\
&\quad + \gamma_4 \Big(B_H(t, q\ ; H) + B_H(l, q\ ; H) \Big) \Bigg) \quad . && \text{[Behavioral Context]}
\end{aligned}
\tag{4}
$$

where $\Gamma = \{\gamma_1, \ldots, \gamma_4\}$ is a set of local confidence parameters for beliefs corresponding to the speech, motion, static images of objects, motion-object relationship, and behavioral context. Given O, q, L, G, R, H, and Γ, the corresponding action, $\tilde{a} = (\tilde{t}, \tilde{u})$, understood to be the meaning of utterance s, is determined by maximizing the shared belief function as

$$\tilde{a} = \arg\max_{a} \Psi(s, a, O, q, L, G, R, H, \Gamma) \quad . \tag{5}$$

6.4 Global confidence function

The global confidence function f outputs an estimate of the probability that the robot's utterance s will be correctly understood by the user, and it is written as

$$f(d) = \frac{1}{\pi} \arctan\left(\frac{d - \lambda_1}{\lambda_2}\right) + 0.5 \quad , \tag{6}$$

where λ_1 and λ_2 are the parameters of this function. Input d of this function is a margin in the value of the output of the shared belief function between an action that the robot asks a user to do and other actions in the process of generating an utterance. Margin d in generating utterance s to refer to action a in scene O under behavioral context q is defined as

$$\begin{aligned} &d(s, a, O, q, L, G, R, H, \Gamma) \\ &= \Psi(s, a, O, q, L, G, R, H, \Gamma) - \max_{A \neq a} \Psi(s, A, O, q, L, G, R, H, \Gamma) \quad . \end{aligned} \tag{7}$$

The examples of the shapes of global confidence functions are shown in Fig. 9. Clearly, a large margin increases the probability of the robot being understood correctly by the user. If there is a high probability of the robot's utterances being understood correctly even when the margin is small, we can say that the robot's beliefs are consistent with those of the user. The example of a shape of such a global confidence function is indicated by "strong". In contrast, the example of a shape in the case when a large margin is necessary to get a high probability is indicated by "weak". When the robot asks for action a in scene O under behavioral context q, the robot generates utterance $\tilde{s}$ so as to bring the value of the output of f as close as possible to the value of parameter ξ, which represents the target probability of the robot's utterance being understood correctly. This utterance can be represented as

$$\tilde{s} = \arg\min_{s} \left| f(d(s, a, O, q, L, G, R, H, \Gamma)) - \xi \right| \quad . \tag{8}$$

The robot can increase its chance of being understood correctly by using more words. On the other hand, if the robot can predict correct understanding with a sufficiently high probability, it can manage with a fragmental utterance using a small number of words.

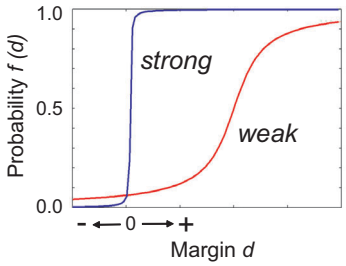


Fig. 9. Examples of the shapes of global confidence functions

6.5 Learning methods

The shared belief function and the global confidence function are learned separately in the processes of utterance understanding and utterance generation.

The decision function is learned incrementally, online, through a sequence of episodes, each of which comprises the following steps.

1. Through an utterance and a gesture, the user asks the robot to move an object.
2. The robot acts on its understanding of the utterance.
3. If the robot acts correctly, the process is terminated. Otherwise, the user slaps its hand.
4. The robot acts in a different way.
5. If the robot acts incorrectly, the user slaps its hand. The process is terminated.

The robot adapts the values of parameter set R for the belief about the motion-object relationship, parameter set H for the belief about the effect of the behavioral context, and local confidence parameter set Γ. Lexicon L and grammar G were learned beforehand as described in the previous sections, and they were fixed. When the robot acts correctly in the first or second trials, it learns R by applying the Bayesian learning method using the information of features of trajector and landmark objects $o_{t,f}, o_{l,f}$ and motion word W_M in the utterances. In addition, when the robot acts correctly in the second trial, the robot associates utterance s, correct action a, incorrect action A done in the first trial, scene O, and behavioral context q with one another and makes these associations a learning sample. When the ith sample $(s_i, a_i, A_i, O_i, q_i)$ is obtained based on this process of association, H_i and Γ_i are adapted to approximately minimize the probability of misunderstanding as

$$(\tilde{H}_i, \tilde{\Gamma}_i) = \arg\min_{H,\Gamma} \sum_{j=i-K}^{i} w_{i-j}\, g\big(\Psi\left(s_j, a_j, O_j, q_j, L, G, R_i, H, \Gamma\right) - \Psi\left(s_j, A_j, O_j, q_j, L, G, R_i, H, \Gamma\right)\big), \quad (9)$$

where $g(x)$ is $-x$ if $x < 0$ and 0 otherwise, and K and w_{i-j} represent the number of latest samples used in the learning process and the weights for each sample, respectively.

The global confidence function f is learned incrementally, online, through a sequence of episodes that consist of the following steps.

1. The robot generates an utterance to ask the user to move an object.
2. The user acts according to his or her understanding of the robot's utterance.
3. The robot determines whether the user's action is correct.

In each episode, the robot generates an utterance that brings the value of the output of global confidence function f as close to ξ as possible. After each episode, the value of margin d in the utterance generation process is associated with information about whether the utterance was understood correctly, and this sample of associations is used for learning. The learning is done online incrementally so as to approximate the probability that an utterance will be understood correctly by minimizing the weighted sum of squared errors in the most recent episodes. After the ith episode, parameters λ_1 and λ_2 are adapted as

$$[\lambda_{1,i}, \lambda_{2,i}] \leftarrow (1-\delta)[\lambda_{1,i-1}, \lambda_{2,i-1}] + \delta[\tilde{\lambda}_{1,i}, \tilde{\lambda}_{2,i}], \tag{10}$$

where

$$(\tilde{\lambda}_{1,i}, \tilde{\lambda}_{2,i}) = \arg\min_{\lambda_1,\lambda_2} \sum_{j=i-K}^{i} w_{i-j}(f(d_j\ ; \lambda_1, \lambda_2) - e_j)^2, \tag{11}$$

where e_i is 1 if the user's understanding is correct and 0 if it is not, and δ is the value that determines the learning speed.

6.6 Experimental results

Utterance understanding by the robot Sequence $\mathcal{D}_d$ of quadruplets (s_i, a_i, O_i, q_i), $i = 1, \ldots, N_d$, comprising the user's utterance s_i, scene O_i, behavioral context q_i, and action a_i that the user wants to ask the robot to perform, was used for the interaction. At the beginning of the sequence, the sentences were relatively complete (e.g., "*green kermit red box move-onto*"). Then the lengths of the sentences were gradually reduced (e.g., "*move-onto*") to become fragmental so that the meanings of the sentences were ambiguous. At the beginning of the learning course, the local confidence values γ_1 and γ_2 for speech, static images of objects, and motions were set to 0.5, while γ_3 and γ_4 were set to 0.

R could be estimated with high accuracy during the episodes in which relatively complete utterances were given and understood correctly. In addition, H and Γ could be effectively estimated based on the estimation of R during the episodes in which fragmental utterances were given. Figure 10 shows changes in the values of γ_1, γ_2, γ_3, and γ_4. The values did not change during the first thirty-two episodes because the sentences were relatively complete and the actions in the first trials were all correct. Then, we can see that the value γ_1 for speech decreased adaptively according to the ambiguity of a given sentence, whereas

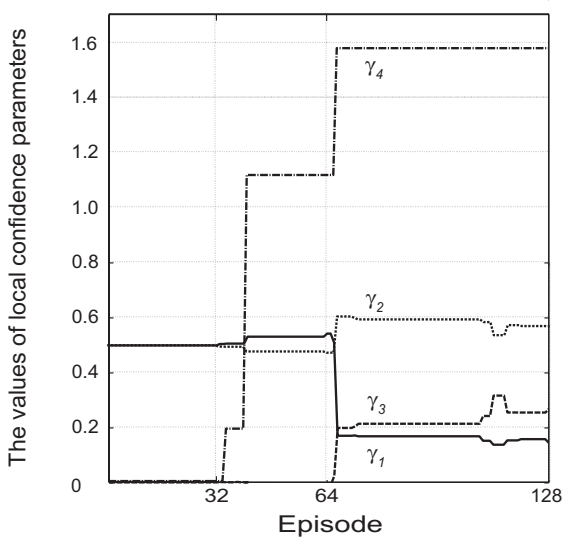


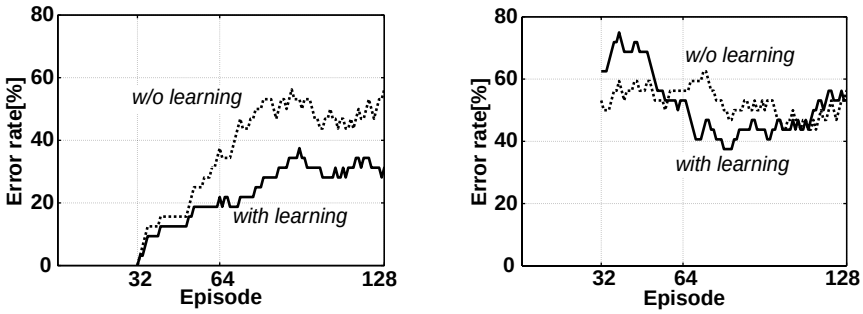
Fig. 10. Changes in the values of local confidence parameters

the values γ_2, γ_3 and γ_4 for static images of objects, motions, the motion-object relationship, and behavioral context increased. This means that nonlinguistic information was gradually being used more than linguistic information.

Figure 11 (a) shows the decision error (misunderstanding) rates obtained during the course of the interaction, along with the error rates obtained for the same learning data by keeping the values of the parameters of the shared belief function fixed to their initial values. In contrast, when fragmental utterances were provided all over the sequence of interaction, the learning was not effective (Fig. 11 (b)) because the robot misunderstood the utterances too often.

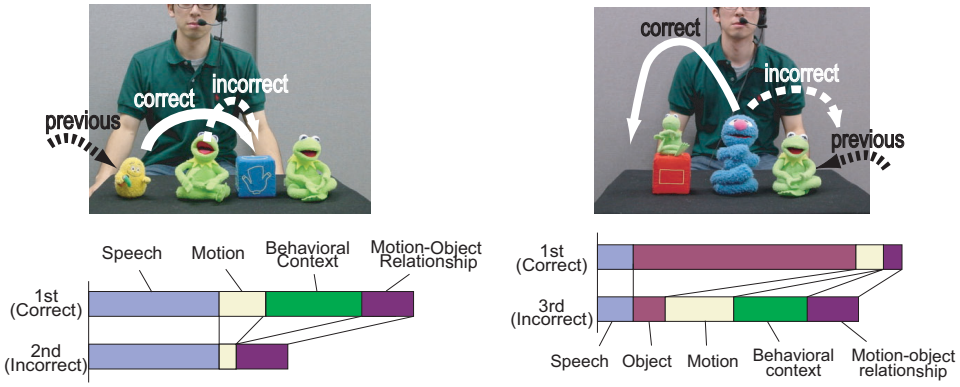
Examples of actions generated as a result of correct understanding are shown together with the output log-probabilities from the weighted beliefs in Figs. 12 (a) and (b), along with the second and third action candidates, which led to incorrect actions. It is clear that each nonlinguistic belief was used appropriately in understanding the utterances according to their relevance to the situations. Beliefs about the effect of behavioral context were more effective in Fig. 12 (a), while in Fig. 12 (b), beliefs about the concepts for the static images of objects were more effective than other nonlinguistic beliefs in leading to the correct understanding.

Utterance generation by the robot A sequence of triplets (a, O, q) consisting of scene O, behavioral context q, and action a that the robot needed to ask the user to perform was given beforehand for the interaction. In each episode, the



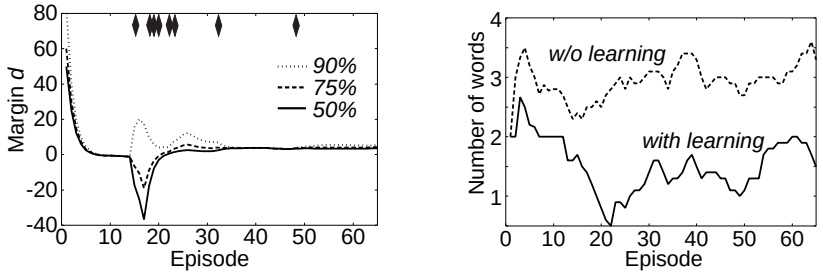
(a) complete → fragmental (b) fragmental → fragmental

Fig. 11. Change in decision error rate



(*a*) "move-onto" (*b*) "grover small kermit move-over"

Fig. 12. Examples of actions generated as a result of correct understanding and the weighted ouput log-probabilities from the beliefs, along with the second and third action candidates, that led to incorrect actions.



(a) (b)

Fig. 13. Changes in the global confidence function (a) and the number of words needed to describe the objects in each utterance (b), $\xi = 0.75$

robot generated an utterance so as to make the global confidence function as close to 0.75 as possible. Even when the target value was fixed at 0.75, we found that the obtained values were distributed widely around it. The initial shape of the global confidence function was set so as to make $f^{-1}(0.9) = 161$, $f^{-1}(0.75) = 120$, and $f^{-1}(0.5) = 100$, meaning that a large margin was necessary for an utterance to be understood correctly. In other words, the shape of f in this case represents weak confidence. Note that when all of the values are close to 0, the slope in the middle of f is steep, and the robot makes the decision that a small margin is sufficient for its utterances to be understood correctly. The shape of f in this case represents strong confidence.

The changes in $f(d)$ are shown in Fig. 13 (a), where three lines have been drawn for $f^{-1}(0.9)$, $f^{-1}(0.75)$, and $f^{-1}(0.5)$ to make the shape of f easily recognizable. The episodes in which the utterances were misunderstood are depicted in the upper part of the graph by the black lozenges. Figure 13 (b) displays the changes in the moving average of the number of words used to describe the objects in each utterance, along with the changes obtained in the case when f was not learned, which are shown for comparison. After the learning began, the slope in the middle of f rapidly became steep, and the number of words decreased. The function became temporarily unstable with $f^{-1}(0.5) < 0$ at around the 15th episode. The number of words then became too small, which sometimes led to misunderstanding. We might say that the robot was overconfident in this period. Finally, the slope became steep again at around the 35th episode.

We conducted another experiment in which the value of parameter ξ was set at 0.95. Figure 14 shows the result of this experiment. It is clear that after approximately the 40th episode the change in f became very unstable, and the number of words became large. We found that f became highly unstable when the utterances with a large margin, d, were not understood correctly.

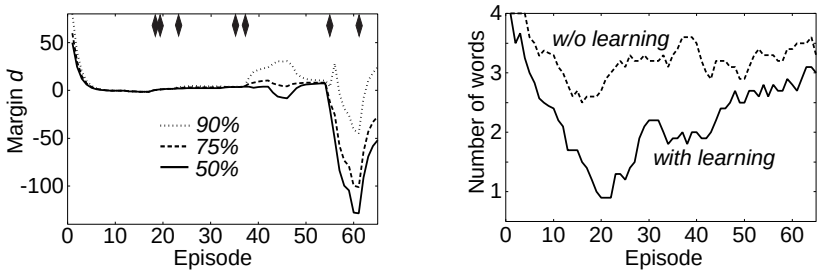


Fig. 14. Changes in the global confidence function (a) and the number of words needed to describe the objects in each utterance (b), $\xi = 0.95$

7 Discussion

Sharing the risk of being misunderstood The experiments in learning a pragmatic capability illustrate the importance of sharing the risk of not being understood correctly between the user and the robot. In the learning period for utterance understanding by the robot, the values of the local confidence parameters changed significantly when the robot acted incorrectly in the first trial and correctly in the second trial. To facilitate the learning, the user had to gradually increase the ambiguity of utterances according to the robot's developing ability to understand them and had to take the risk of not being understood correctly. In the robot's learning period for utterance generation, it adjusted its utterances to the user while learning the global confidence function. When the target understanding rate ξ was set to 0.95, the global confidence function became very unstable in cases where the robot's expectations of being understood correctly at a high probability were not met. This instability could be prevented by using a lower value of ξ, which means that the robot would have to take a greater risk to be understood correctly.

Accordingly, in human-machine interaction, both users and the robots must face the risk of not being understood correctly and thus adjust their actions to accommodate such risk in order to effectively couple their belief systems. Although the importance of controlling the risk of error in learning has generally been seen as an exploration-exploitation trade-off in the field of reinforcement learning by machines (e.g. [36]), we argue here that the mutual accommodation of the risk of error by those communicating is an important basis for the formation of mutual understanding.

Partiality of information and fast adaptation of function An utterance includes only partial information that is relevant to what a speaker wants to convey to a listener. The method interpreted such an utterance by using the belief system under a given situation, and this enabled the robot and the user to adapt to each other rapidly.

In the field of autonomous robotics, the validity of the architecture in which sub-systems are allocated in parallel has been shown [37]. This architecture can flexibly cope with two problems faced by systems interacting with the physical world: the partiality of information and real-time processing [38]. On the other hand, statistical inference using partial information has been studied intensively, particularly in the research on Bayesian networks [39], in which parallel connection of sub-systems is not necessarily important.

The shared belief function Ψ is a kind of Bayesian network in which a small number of weighting values Γ are added to some nodes, and it has an architecture with belief modules allocated in parallel as shown in Eq.4. Due to this structure of the belief system, the method could successfully cope with the partiality of information and enable rapid and robust adaptation of the function by changing weighting values.

Initial setting *No free lunch theory* [40] shows that when no prior knowledge on a problem exists, it is not possible to assume that one learning algorithm is superior to another. That is, there is no learning method that is efficient for all possible tasks. This suggests that we should pay attention to domain specificity as well as versatility.

In the methods described here, the initial setting for the learning was decided by taking into account the generality and efficiency of language learning. The semantic attributes – [motion], [trajector], and [landmark] – were given beforehand because they would be general and essential in linguistic and other cognitive processes. With this setting, however, the constructions the method could learn were limited to those like transitive and ditransitive ones. Overcoming this limitation is a future work.

Integrated learning In the method, speech units, lexicon, grammar, and pragmatic capability were learned step-by-step separately. These learning processes, however, should be carried out simultaneously. In developmental psychology, it has been shown that a pragmatic capability facilitates the process of learning other linguistic knowledge, such as the specification of referents in word learning [41]. The computational mechanism for such cognitive bootstrapping should be pursued.

Prerequisites for conversation Language learning can be regarded as a kind of role reversal imitation [42]. To coordinate roles in a joint action among participants, they should read the intentions of the others. It is known that in the very early stage of development infants become able to understand the intentional actions of others [43] and even to understand that others might have beliefs different from the ones held by themselves [44].

The method described here enabled the robot to understand the user's utterances, act, and make utterances to ask the user to act. The roles in this speak-and-act task, however, were given to the robot and the user beforehand,

and they knew it. For the robot to learn the conversational (speak-and-speak) capability, the robot should find its role in a joint action by itself and coordinate it with the user.

Psychological investigation The experimental results showed that the robot could learn the system of beliefs that the robot had assumed the user had. Because the user and the robot came to understand fragmental and ambiguous utterances, they must have shared similar beliefs and must have been aware of that. It would be interesting to investigate through psychological experiments the dynamics of belief sharing between users and robots.

8 Conclusion

A developmental approach to language processing for grounded conversations was presented. It can cope with two major requirements that existing language processing methods cannot satisfy: a scalable grounded symbol system and belief sharing. The proposed method enabled a robot to learn a pragmatic capability online in a short period of verbal and nonverbal interaction with a user by rapid and robust adaptation of its grounded belief system.

Ackonwledgements I would like to thank Komei Sugiura and an anonymous reviewer for comments on earlier drafts of this manuscript. This work was supported by a research grant from the National Institute of Informatics.

References

1. Sperber, D., Wilson, D.: Relevance (2nd Edition). Blackwell (1995)
2. Maturana, H.R.: Biology of language – the epistemology of reality. In Miller, G.A., Lenneberg, E., eds.: Psychology and Biology of Language and Thought – Essay in Honor of Eric Lenneberg. (1978) 27–64
3. Negroponte, N.: Being Digital. Alfred A. Knopf Inc. (1995)
4. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A.: Toward conversational human-computer interaction. AI Magazine (2001)
5. Johnson, M.: The Body in the Mind - The Bodily Basis of Meaning, Imagination, and Reason. University of Chicago Press (1987)
6. Winograd, T.: Understanding Natural Language. Academic Press New York (1972)
7. Shapiro, C.S., Ismail, O., Santore, J.F.: Our dinner with Cassie. In: AAAI 2000 Spring Symposium on Natural Dialogues with Practical Robotic Devices. (2000) 57–61
8. Clark, H.: Using Language. Cambridge University Press (1996)
9. Traum, D.R.: A computational theory of grounding in natural language conversation. Doctoral dissertation, University of Rochester (1994)
10. Iwahashi, N.: Language acquisition through a human-robot interface by combining speech, visual, and behavioral information. Information Sciences **156** (2003)

11. Iwahashi, N.: A method of coupling of belief systems through human-robot language interaction. In: IEEE Workshop on Robot and Human Interactive Communication. (2003)
12. Iwahashi, N.: Active and unsupervised learning of spoken words through a multimodal interface. In: IEEE Workshop on Robot and Human Interactive Communication. (2004)
13. Brent, M.R.: Advances in the computational study of language acquisition. Cognition (61) (1996) 1–61
14. Dyer, M.G., Nenov, V.I.: Learning language via perceptual/motor experiences. In: Proc. of Annual Conf. of the Congnitive Science Society. (1993) 400–405
15. Nakagawa, S., Masukata, M.: An acquisition system of concept and grammar based on combining with visual and auditory information. Trans. Information Society of Japan **10**(4) (1995) 129–137
16. Regier, T.: The Human Semantic Potential. MIT Press (1997)
17. Roy, D.: Integration of speech and vision using mutual information. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing. (2000) 2369–2372
18. Steels, L., Kaplan, K.: Aibo's first words. the social learning of language and meaning. Evolution of Communication **4**(1) (2001) 3–32
19. Gorin, A., Levinson, S., Sanker, A.: An experiment in spoken language acquisition. IEEE Trans. on Speech and Audio Processing **2**(1) (1994) 224–240
20. Steels, L., Vogt, P.: Grounding adaptive language games in robotic agents. In: Proc. of the Fourth European Conf. on Artificial Life. (1997)
21. Steels, L.: Evolving grounded communication for robots. Trends in Cognitive Science **7**(7) (2003) 308–312
22. Sugita, Y., Tani, J.: Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. Adaptive Behavior (**13**(1)) 33–52
23. Jordan, M.I., Sejnowski, T.J., eds.: Graphical Models - Foundations of Neural Computation. The MIT Press (2001)
24. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing **28**(4) (1980) 357–366
25. Persoon, E., Fu, K.S.: Shape discrimination using Fourier descriptors. IEEE Trans Systems, Man, and Cybernetics **7**(3) (1977) 170–179
26. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics **41**(1) (1970) 164–171
27. Bloom, P.: How children learn the meanings of words. MIT Press (2000)
28. Imai, M., Gentner, D.: A crosslinguistic study of early word meaning – universal ontology and linguistic influence. Cognition **62** (1997) 169–200
29. DeGroot, M.H.: Optimal Statistical Decisions. McGraw-Hill (1970)
30. Langacker, R.: Foundation of cognitive grammar. Stanford University Press, CA (1991)
31. Haoka, T., Iwahashi, N.: Learning of the reference-point-dependent concepts on movement for language acquisition. Tech. Rep. of the Institute of Electronics, Information and Communication Engineers PRMU2000-105 (2000)
32. Tokuda, K., Kobayashi, T., Imai, S.: Speech parameter generation from HMM using dynamic features. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing. (1995) 660–663
33. Savage-Rumgaugh, E.: Ape Language – From Conditional Response to Symbol. Columbia Univ. Press (1986)

34. Iwahashi, N., Satoh, K., Asoh, H.: Learning abstract concepts and words from perception based on Bayesian model selection. Tech. Rep. of the Institute of Electronics, Information and Communication Engineers PRMU-2005-234 (2006)
35. Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes. In: Int. Conf. on Uncertainty in Artificial Intelligence. (1999) 21–30
36. Dayan, P., Sejnowski, T.J.: Exploration bonuses and dual conrol. Machine Learning **25** (1996) 5–22
37. Brooks, R.: A robust layered control system for a mobile robot. IEEE Journal of Robotics and Automation (1) (1986) 14–23
38. Matsubara, H., Hashida, K.: Partiality of information and unsolvability of the frame problem. Japanese Society for Artificial Intelligence **4**(6) (1989) 695–703
39. Pearl, J.: Probabilistic reasoning in intellignet systems: Networks of Plausible Inference. Morgan Kaufmann (1988)
40. Wolpert, D.H.: The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In Wolpert, D.H., ed.: The mathematics of Generalization, Addison-Wesley, Reading, MA (1995)
41. Tomasello, M.: The pragmatics of word learning. Cognitive Studies **4**(1) (1997) 59–74
42. Carpenter, M., Tomasello, M., Striano, T.: Role reversal imitation and language in typically developing infants and children with autism. INFANCY **8**(3) (2005) 253–278
43. Behne, T., Carpenter, M., Call, J., Tomasello, M.: Unwilling versus unable – infants' understanding of intentional action. Developmental Psychology **41**(2) (2005) 328–337
44. Onishi, K.H., Baillargeon, R.: Do 15-month-old infants understand false beliefs? Science **308** (2005) 225–258