音声言語情報処理　39－13
(2001. 12. 20)

〔招待論文〕

# ロボットによる言語獲得

岩橋直人[†]

† ソニーコンピュータサイエンス研究所
〒141-0022 東京都品川区東五反田 3-14-13 高輪ミューズビル

E-mail: †iwahashi@csl.sony.co.jp

**あらまし**　本稿では，ロボットが人との言語コミュニケーションの基盤となる相互信念を学習する方法について，概略を述べる．本方法では，音声，視覚，行動の情報が，確率のフレームワークで統合的に処理され，相互信念の学習が，共同知覚とインタラクションに基づいて unsupervised に行われる．学習される相互信念は，音韻，語彙，文法，行動コンテキストの影響，およびその他の非言語的信念からなる．実験において，はじめは言語知識を持っていなかったロボットが，学習により，断片的であいまいな発話でさえ状況に応じて適切に理解して行動できるようになった．本方法は，言語における身体性とダイナミクスを相互信念の学習過程に反映させることを可能としており，人とロボットのより自然なコミュニケーションを実現するために拡張可能であると考えられる．

**キーワード**　ロボット，言語，相互信念，コミュニケーション，学習

# Language Acquisition by Robots

Naoto IWAHASHI[†]

† Sony Computer Science Labs.
Takanawa Muse. Bldg. 3-14-13, Higashigotanda Shinagawa-ku, Tokyo, 141-0022 Japan

E-mail: †iwahashi@csl.sony.co.jp

**Abstract**　This paper summarizes the methods by which a robot can learn mutual beliefs necessary for language communication with people. The learning is carried out in unsupervised ways based on joint perception and interaction, combining the information of raw speech and visual observations and behavioral reinforcement in probabilistic framework. The beliefs delt with in the methods include phonemes, lexicon, grammar, the influence of behavioral context, and other nonlinguistic belief. In experiments a robot that initially had no linguistic knowledge was eventually able to understand even fragmental and ambiguous utterances according to given situations, and act appropriately. The methods made it possible to reflect the embodied and dynamic aspects of language in the learning process, and they can be extended to provide more natural communication between people and robots.

**Key words**　robot, language, mutual belief, communication, learning

## 1. Introduction

Language communication in daily life is based on the mutual beliefs shared by those who are communicating [1]. The mutual beliefs are formed through common experiences based on common cognitive ability, and are used in the process of utterance production and comprehension. Such mutual beliefs are diverse including not only linguistic but also nonliguistic beliefs, and changes with the experiences. So if we want to make it possible for human and robot to communicate with the same way people do, we need a language-processing paradigm that reflects the cognitive ability of human and the common experiences shared by a person and a robot.

This paper summarizes the methods by which a robot can learn the mutual beliefs needed for multimodal language communication with a person (for details, see [2]~[6]). The learning is based on joint perception and interaction, and it uses the information of raw speech and visual observations and behavioral reinforcement, and this information is integrated in a probabilistic framework. The learned mutual beliefs include phonemes, lexicon, grammar, the influence of behavioral context, and task-specific knowledge.

## 2. Learning Task

The learning task in the present work is set up as follows. A robot is set alongside a table, and a person and the robot see and move the objects on the table as shown in Fig. 1. The robot initially has no concepts about the specific objects or the ways they can be moved, nor does it have any linguistic knowledge. The person teaches the robot by speaking into a microphone, slowly and pausing briefly between words, while pointing to or moving the objects on the table. After the robot learns basic linguistic mutual beliefs through a sequence of such learning episodes, then the person asks the robot to move objects. If the robot responds wrongly, the person slaps the robot's hand. Then, the robot acts in a different way. Through a sequence of such reinforcing episodes, the robot expands the mutual beliefs until it can understand even fragmental utterances.

## 3. Algorithm Outline

The robot directs its attention to objects that have been put on the table, the ones being pointed to by the person, and the moving ones. This joint perception is one basis of learning. When attention is given to objects and the person speaks, the observations of those objects are associated with that speech. The associated speech and image data constitutes a set of pairs that is used for learning. The interaction is the other basis of learning, and the reinforcement information is given by slapping the robot's hand.

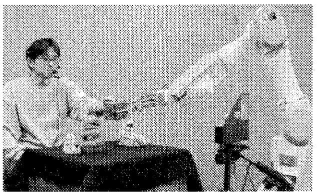


Fig. 1 System configuration

The robot learns the mutual beliefs in unsupervised way by using the information provided by the raw speech and visual observations and behavioral reinforcement, and these mutual beliefs are represented by a graphical model including hidden Markov models (HMMs). It first learns the speech units like phonemes and the lexicon, which consits of the lexical items for the concepts on the objects, simultaneously by using a set comprising the images of static objects and the word utterances describing those objects. It also learns the lexical items for the concepts of motions by using a set comprising the images, in each of which the person is moving an object, and the word utterances describing those motions. Then it learns the grammar by using a set comprising images, in each of which the person is moving an object, and the sentence utterances describing those scenes. In this process, the concepts of motions work to represent the trajector-landmark relationships between the individual concepts correspoding to the words in the sentence utterances. Finally, it learns the mutual beliefs in the process of understanding fragmental utterances. It does this by using the reinforcement information in an iterative way.These mutual beliefs can consist not only of linguistic information but also of non-linguistic information related to behavioral context and task-specific knowledge.

## 4. Experimental Setup

The robot had a arm (seven degrees freedom) with a hand (one degree of freedom), and a camera unit. The camera unit contained three separate CCDs so that three-dimensional information about the scenes could be obtained. A close-talk microphone was used for speech input, and the speech was represented by using Mel-scale cepstrum coefficients and their delta parameters (twenty-five dimensional). The visual observations were represented by using the such features as position on the table (two-dimensional: horizontal and vertical coordinates), velocity (two-dimensional), color (three-dimensional: L*a*b* parameters), size (one-dimensional), and shape (two-dimensional). The system's attention was restricted to objects within 90 cm of the camera unit, and a person using eleven stuffed toys and four boxes as objects taught language to the system under acoustic conditions typical of an office environment.

## 5. Learning of Phonemes and the Lexicon

### 5.1 Algorithm

Let $C = \{c_1, c_2, ..., c_M\}$ be the set of $M$ lexical items. Suppose that $M$ is unknown. Suppose that a spoken word $S$ and corresponding object image $V$ occur at the same time, and that each spoken word and object image pair correspond to a lexical item in $C$. Also suppose that the set of the pairs of a word utterance sample and an object image sample, $D_l = \{(s_1, v_1), ..., (s_{N_l}, v_{N_l})\}$, is given as learning data. Then we want to estimate $M$ as well as probability density functions $p(S|c_i)$ and $p(V|c_i)$ $(i = 1, ..., M)$, for the probabilistic model $L$ of the lexicon.

Because the difference between the features of the spoken word samples in a lexical item ordinary does not reflect the difference between the features of the object image samples in the lexical item, we can assume that in each lexical item the speech features $S$ and image features $V$ are independent. That is,

$$p(S,V|c_i) = p(S|c_i)p(V|c_i), \quad (i = 1, ..., M). \qquad (1)$$

Therefore, the joint probability density function $p(S, V)$ can be written as

$$p(S,V) = \sum_{i=1}^{M} p(S|c_i)p(V|c_i)P(c_i). \qquad (2)$$

If the number of lexical items, $M = m$, is given, the estimate of the lexicon including $m$ items can be obtained by maximizing the likelihood of joint probability density function $p(S, V)$ as

$$\tilde{L_m} = \underset{L_m}{\operatorname{argmax}} \prod_{i=1}^{N_l} p(s_i, v_i|L_m). \qquad (3)$$

Because $M$ is not actually given, $M$ has to be estimated. The lexicon represents a mapping between words and object image categories, and if this mapping is to be efficient in language communication, it can be assumed to maximizes the mutual information $I(S, V)$ between spoken word $S$ and object image $V$ with the smallest number of lexical items. Therefore, the estimate $\tilde{L}$ of the lexicon is obtained by choosing from among the estimates $\tilde{L_m}$ maximizing the estimate of mutual information the one with the smallest $m$. The estimate of mutual information is obtained by leave-one-out cross-validation.

This principle for the learning of the probabilistic model for the lexicon is also applied to the learning of the probabilistic models for speech-units. From the speech-unit sets that maximize the estimate of mutual information is chosen the one with the smallest number of speech units. In a concrete optimization algorithm, Hidden Markov models (HMMs), each of which has left-to-right state structure, are used to represent the speech-units, which compose probabilistic models for spoken words. Multivariate normal *p.d.f.*s are used to represent the static concepts of object images. The details of this learning method have been described in [3].

In addition, the lexical items for the motion of moving objects are learned as the concepts of relation between a trajector and a landmark. In Fig. 2, for instance, if the stuffed toy in the middle and the box at the right side are considered landmarks, the movement of the trajector is understood as *jump over* and *move onto*.

The concepts of motions are represented by HMMs for the trajectories of moved objects in appropriate coordinates based on the positions of a trajector and a landmark. Because the information on the appropriate coordinates and the landmark selected in each scene is not observed in the learning data, the HHMs of the motions are learned while the coordinates and the landmarks are being inferred [4].

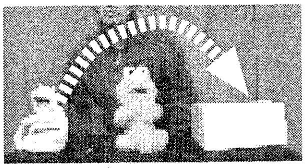


Fig. 2 Example of a dynamic image

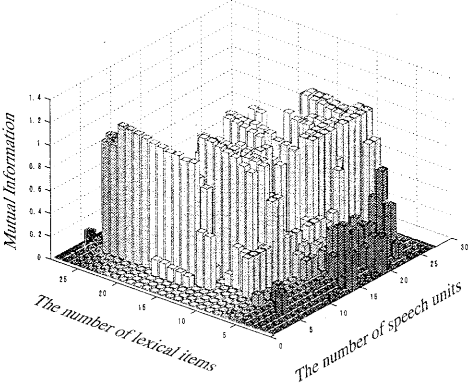


Fig. 3 The estimated value of mutual information vs. the numbers of speech-units and lexical items

### 5.2 Experiment

The learning data included one hundred ninety pairs of spoken-word and object image samples, each of which corresponded to one of ten lexical items. The words for lexical items were randomly selected from a Japanese dictionary. We can see in Fig. 3, which shows the estimates of the mutual information obtained when the number of speech units and the number of lexical items were changed, that the estimates of mutual information was maximized when eight speech-units and fifteen lexical items were used. The lexicon acquired is listed in Table 1, where the acquired words are represented by the sequence of the acquired speech units, which are denoted by indices. We can see that seven lexical items were suitably learned in such a way that there was a one-to-one mapping between the acquired lexical items and the lexical items used in learning data. And we can see that some speech units seem to correspond to particular phonemes: speech units *2* and *8* are respectively mapped to vowels */o/* and */e/*.

## 6. Learning of Grammar

### 6.1 Algorithm

The set of the pairs of dynamic action image and sentence utterance samples, $D_g = \{(s_1, v_1), (s_2, v_2), \ldots, (s_{N_g}, v_{N_g})\}$, is given as learning data. It is supposed that each utterance is based on stochastic grammar $G$ and describes the corresponding image. The grammar $G$ is learned by maximizing the likelihood of the joint probability density function $p(s, v)$ of utterance $s$ and dynamic image $v$. The joint probability density function $p(s, v)$ is represented with an internal structure that

*Table. 1 The lexical items acquired by the proposed method*

| acquired word | spoken word | concept |
|---|---|---|
| *1 3* | *aru* | *Barba* |
| *5 1 4 7 8* | *kanke:* | *Kermit* |
| *6 5 2 3 5 7* | *kyo:iku* | *red* |
| *5 1 6 3* | *kuwashi:* | *blue* |
| *2 5 6 3* | *ko:shiki* | *green* |
| *2 3 5* | *koNya* | *big* |
| *6 5 1 3 5 2* | *zairyo:* | *small* |
| *4 8 6 7* | *ueru* | *Dumbo* |
| *4 8 7* | *ueru* | *Dumbo* |
| *2 5 1 3 8* | *okage* | *Elmo* |
| *2 5 1 7 3 8* | *okage* | *Elmo* |
| *5 1 4 1 3 5 4 7 3* | *kagayaku* | *Grover* |
| *5 4 1 8 5 4 7 3* | *kagayaku* | *Grover* |
| *5 4 1 3 5 4 7 3* | *kagayaku* | *Grover* |
| *5 1 4 1 3 5 1 4 7 3* | *kagayaku* | *Grover* |
| *6 5 1 4 1 3 5 4 7 3* | *kagayaku* | *Grover* |

includes the parameters of the grammar $G$ and the conceptual structure $z$ that the utterance means.

The conceptual structure used here is expressed with semantic attributes - [motion], [trajector], and [landmark] - which are initially given to the system and are fixed. For instance, when the image is the one shown in Fig. 2 and the corresponding utterance is the sequence of spoken words, *'big Kermit brown box move-onto'*, the conceptual structure might be

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

where in the right-hand column are the spoken word subsequences corresponding to trajector, landmark and motion.

Let $y$ denote the order of semantic attributes, which represents the order of constituents with the semantic attributes in an utterance. For instance, in the above utterace example, the order is [trajector]-[landmark]-[motion]. Suppose that grammar is represented by the set of the occurrence probabilities of the possible orders as $G = \{P(y_1), P(y_2), ..., P(y_k)\}$. The joint probability density function conditioned by the estimated lexicon parameters $\tilde{L}$ and the grammar $G$ is written as

$$\begin{aligned} & p(s, v|\tilde{L}, G) \\ & = \max_z p(S|z, \tilde{L}, G) p(V|z, \tilde{L}) \\ & = \max_{z,l} \Big\{ p(S|z, \tilde{L}, G) \\ & \qquad \times p(u|t, l, W_m, \tilde{L}) p(t|W_t, \tilde{L}) p(l|W_l, \tilde{L}) \Big\} \quad (4) \end{aligned}$$

where $t$, $l$ and $u$ are respectively a trajector object, a landmark object, and the trajectory of the movement of $t$ in the image $v$. $W_m$, $W_t$, and $W_l$ are respectively word sequences corresponding to the motion, trajector, and landmark in the conceptual structure $z$.

The estimate of grammar $G$ is obtained by maximizing the likelihood of this function with regard to the learning data as

$$\tilde{G} = \operatorname*{argmax}_G \prod_{i=1}^{N_g} p(s_i, v_i|\tilde{L}, G). \quad (5)$$

Here the concept of motion represents the relationship between a trajector object and a landmark object in the form of $p(u|t, l, W_m, \tilde{L})$.

An utterance asking the robot to move a object is understood using the lexicon $\tilde{L}$ and the grammar $\tilde{G}$ which have been learned so far, and one of the objects in the current scene is accordingly grasped and moved by the robot arm. The algorithm understanding speech $s$ infers the conceptual structure $Z$ and generates the dynamic image $\tilde{v}$ of the action, which consists of the trajectory $u$ of trajector $t$, as

$$\tilde{v} = \operatorname*{argmax}_v p(s, v|\tilde{L}, \tilde{G}). \quad (6)$$

The robot arm is controlled according to the generated trajectory $u$. The utterance-understanding algorithm has been described in detail elsewhere [5].

### 6.2 Experiments

Seventy-two utterance and dynamic image pairs were given as learning data. The average number of words in an utterance was 3.5 and the average number of objects in an image was 4.7. The lexicon consisted of twenty-one lexical items: 14 for static concepts of objects and 7 for motions. The utterances were rather simple. For example, one dynamic image showed a person moving a small Elmo onto a green box, and the corresponding sentence was '*move-onto small Elmo green box*.' The estimated values $\tilde{P}(y)$ of the occurrence probabilities $P(y)$ of attribute order $y$ are listed in Table 2.

Table. 2 The estimated probabilities in the grammar

| attribute order $y$ | $\tilde{P}(y)$ | $P(y)$ |
|---|---|---|
| [motion] [trajector] [landmark] | 0.38 | 0.46 |
| [motion] [landmark] | 0.25 | 0.25 |
| [motion] | 0.17 | 0.17 |
| [motion] [trajector] | 0.17 | 0.13 |
| [motion] [landmark][trajector] | 0.04 | 0.00 |

Example of the action generated by correct inference is shown with the differences of calculated log probabilities between the first and fifth candidates of the decision in Fig. 4. We can see that the difference of the probabilities with respect to the trajectory of the motion was effective in correcting the error in speech recognition.

## 7. Forming of Mutual Beliefs

### 7.1 Algorithm

The system of mutual beliefs is represented by the sum of weighted beliefs, each weighting value representing the confidence that each belief is shared between the robot and the person. The mutual beliefs are learned in the process of the understanding of fragmental utterances. Each belief is learned incrementally during the learning

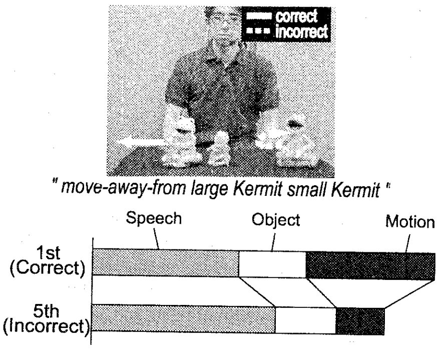


Fig. 4 Example of utterance comprehension

course. The confidence on a belief is strengthened, when the robot shows misunderstanding a utterance in its fisrt response, and understand it correctly by using the belief in the second response invoked by being slapped. In addition to dealing with linguistic information, the algorithm deals with nonlinguistic information, two examples of which are the following:

- The possibility that object $o$ is involved as a trajector or landmark in the action described by the current utterance, given behavioral context $q$, is represented by function $f(o,q)$. It takes as its value $b_h$ if $o$ is being held, $b_c$ if $o$ is involved in a previous action, and 0 otherwise.
- The relationship between motion $W_m$ and the features of the involved objects, $t$ and $l$, is represented by gaussian $p(t,l|W_m)$.

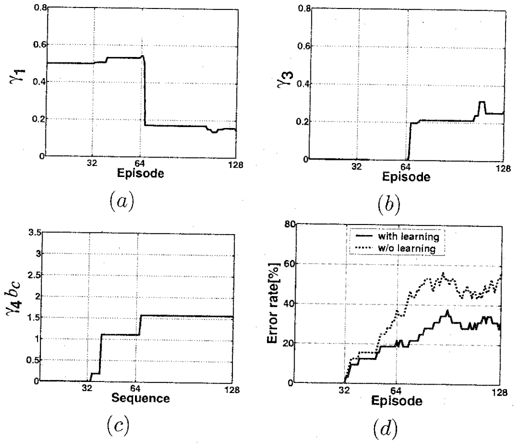
The corresponding behavior $v = \{u,t\}$ understood to be the meaning of speech $s$ is determined by maximizing the following decision function:

$$\begin{aligned}
&\Psi(s,t,u,q,\tilde{L},\tilde{G},R,B,\Gamma)\\
&\equiv \max_{l,z}\Big[\gamma_1 \log p(s|z,\tilde{L},\tilde{G})\\
&\quad +\gamma_2\left\{\log p(u|W_m,\tilde{L}) + \log p(t|W_t,\tilde{L}) + \log p(l|W_l,\tilde{L})\right\}\\
&\quad +\gamma_3 \log p(t,l|W_m,R)\\
&\quad +\gamma_4\left\{f(t,q,B) + f(l,q,B)\right\}\Big] \qquad (7)
\end{aligned}$$

where $\Gamma = \{\gamma_1,\ldots,\gamma_4\}$ are the set of weights, $R$ is the set of prameters for $p(t,l|W_m)$, and $B$ is the set of the paremeters $\{b_h, b_c\}$. $\Gamma$ are learned incrementally based on minimum decision error criterion.

### 7.2 Experiments

$R$ was learned when the robot acts correctly according to each utterance by using the Bayesian learning method. And $B$ and $\Gamma$ were learned when the robot acts incorrectly in the first response and acts correctly next. At the beginning of the course of learning were given utterances that were complete sentences ( e.g., "move-onto green kermit red box"). Then sentences were gradually getting fragmental ( e.g. "move-onto" ). The changes of the values of $\gamma_1$, $\gamma_3$, and $\gamma_4 b_c$ are shown in Fig. 5 (a)-(c). We can see that each value was adapted according to the ambiguity of the given sentences. Figure 5 (d) shows the comprehension error rates during the course of the episodes, along with the error rates obtained by using the initial parameter values without learning.



(a) (b) (c) (d)

Fig. 5 The change of the values of weights (a)-(c), and error rate (d), during the learning course.

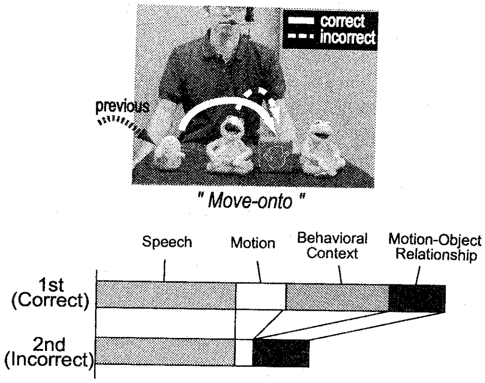


Fig. 6 Example of the comprehension of fragmental utterance

In order to understand fragment utterances correctly the robot had to use the beliefs which have been learned enough and its weight has been made large. Such beliefs can be assumed to be mutual beliefs of the person and the robot. An example of the action generated by correct inference is shown with calculated log probabilities in Fig. 6, where we can see that each non-linguistic belief was appropriately used in understanding the utterance. Details of this learning are described in [6]

## 8. Discussion

In the presented methods, the initial setting for the learning was decided by taking account of the generality and efficiency of the learning. The robot could give its attention to the objects in particular states. The conceptual attributes – [motion], [trajector], and [landmark] – were given beforehand because they would be general and essential in linguistic and other cognitive processes. We may use different conceptual attributes to make the process of language acquisition more general and efficient. The initial setting would depend on the task and the situation that the robot has to manage.

Although the experimental results showed that the robot could learn the mutual beliefs through the interactive with a person, the learned mutual beliefs were actually the mutual beliefs assumed by the robot. It is interesting to investigate whether the person is able to assume similar mutual beliefs through the interaction with the robot.

The method described here can be improved by extending them to learn the lexicon, grammar, and nonlinguisitc beliefs simultaneous, to learn autonomously, to deal with continuous speech.

## 9. Related work

Language acquisition by machines has been attracting interest in various research areas [7], and there have been several pioneering studies. Siskind's algorithm [8] learned a word-to-meaning mapping by using a set of pairs each consisting of a sentence and a collection of its possible meanings represented symbolically with Jackendoff-style expression, and it successfully addressed the problems due to homonyms and to noisy learning data. There have also been some studies on the use of semantic information in the learning of syntactic rules [9], [10]. Visual rather than symbolic information has also been used in word-to-meaning learning tasks [11]~[13], and the judgment of whether or not the system's response is appropriate has also been used in [14]. A spoken-word acquisition algorithm based on the unsupervised clustering of speech tokens has already been described [15], [16]. Furthermore, an algorithm for the learning of stochastic regular grammar in a visually grounded way was presented in [12]. In all these algorithms, however, some categories of phonemes and meanings or some values for threshold for the clustering must be specified beforehand. Thus they are neither expandable nor adaptive. Nor do they deal with spatiotemporal information about the meanings of utterances, even though it should be processed in natural communication.

There have also been many interesting reports on human-robot language communication, such as [17]~[21].

## 10. Conclusion

A framework making it possible to reflect the embodied and dynamic aspects of language in the learning process was presented. It combines linguistic process and other cognitive process and it deals with speech information, visual information, and behavioral information in a unified way. It also combines phonetic, syntactic, semantic, and pragmatic functions in language processing. It can adaptively learn the mutual belief needed for utterance comprehension in multimodal communication and can understand ambiguous and fragmental utterances. It could be extended to provide more natural communication between people and robots.

### Acknowledgements

I thank Dr. H. Iida for discussion and support. I thank Mr. T. Haoka, Mr. A. Miyata, Mr. K. Kim, Mr. M. Tamura, and Mr. A. Toyoizumi for their cooperative work.

### References

[1] D.Sperber, D.Wilson, Relevance, 2nd Edition, Blackwell, 1995.

[2] N.Iwahashi, Language acquisition through a human-robot interface, Proc. of Int. Conf. Spoken Language Processing, 2000.

[3] K.Kim, N.Iwahashi, The acquisition of linguistic speech units with hierarchical structure based on the integration of perceptual infomation, Proc. Japan Acoustical Society Spring Mtg. Vol.I, 99-100, 2001.

[4] T.Haoka, N.Iwahashi, Learning of the reference-point-dependent concepts on movement for language acquisition, Tech. Rep. of IEICE PRMU2000-105, 2000.

[5] T.Haoka, N.Iwahashi, Speech understanding based on cognitibe language knowledge, Proc. Acoustic Society of Japan Spring Meeting Vol.I, 159-160, 2001.

[6] A.Miyata, N.Iwahashi, et al., Mutual belief forming by robots based on the process of utterance comprehension, Tech. Rep. of IEICE, SP, 2001.12.

[7] M.R.Brent, Advances in the computational study of language acquisition, Cognition, 61, 1-61, 1996.

[8] J.M.Siskind, A computational study of cross-situational techniques for learning word-to-meaning mappings, Cognition, 61, 39-91, 1996.

[9] P.Langley, Language acquisition through error recovery, Cognition and Brain Theory 5, 221–225, 1982.

[10] R.C.Berwick, The acquisition of syntactic knowledge, MIT Press, 1985.

[11] M.G.Dyer, V.I.Nenov, Learning Language via Perceptual/Motor Experiences, Proc. Annual Conf. of the Congnitive Science Society, 400-405, 1993.

[12] S.Nakagawa, M.Masukata, An acquisition system of concept and grammar based on combining with visual and auditory information, Trans. Information Society of Japan, 10-4, 129–137, 1995.

[13] T.Regier, The Human Semantic Potential, MIT Press, 1997.

[14] A.L.Gorin, S.E.Levinson, et al., Adaptive acquisition of language, Computer Speech and Language 5, 101-132, 1991.

[15] A.L.Gorin, S.E.Levinson, et al., An experiment in spoken language acquisition, IEEE Trans. Speech and Audio Processing 2-1, 224-240, 1994.

[16] D.Roy, Integration of speech and vision using mutual information, Proc. Int. Conf. Acoustics, Speech and Signal Processing, 2369-2372, 2000.

[17] T.Winograd, Understanding Natural Language, Academic Press New York, 1972.

[18] Shapiro, C.S., Ismail, H.O., J.F.Santore, Our Dinner with Cassie, Working Notes for AAAI 2000 Spring Symp. on Natural Dialogues with Practical Robotic Devices, 57-61, 2000.

[19] T.Ono, M.Imai, et al., A Model of Embodied Communications with Gestures between Humans and Robots, Annual Mtg. Cognitive Science Society, 2001.

[20] T.Matsui, H.Asoh, et al., A speech dialogue system of the office mobile robot Jijo2, Trans. Japan Robotics Society, 18-2, 300-307, 2000.

[21] T.Inamura, N.Inaba, et al., Integration Model of Learning Mechanism and Dialogue Strategy based on Stochastic Experience Representation using Bayesian Network, Proc. Int. Workshop Robot and Human Interactive Communication, 27-29, 2000.