音声言語情報処理　39－15
(2001. 12. 20)

# ロボットによる発話理解過程に基づく相互信念の形成

宮田　篤人[†,††]　岩橋　直人[†]　榑松　明[††]

† 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
†† 電気通信大学 電気通信学研究科 電子工学専攻
〒182-8585 東京都調布市調布ヶ丘 1-5-1

E-mail: †{miyata,iwahashi}@csl.sony.co.jp, ††kure@apple.ee.uec.ac.jp

**あらまし**　本稿では，人との自然な言語コミュニケーションの基盤となる相互信念を，ロボットが発話理解過程を通して学習する方法について述べる．本方法では，相互信念は，複数の信念と，各信念が人と共有されている確信の強さを示す重み付けによって，表現される．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な信念からなる．実験により，はじめは簡単な言語知識しか持たないアームロボットが人と言語と行動を介したインタラクションを通して相互信念を学習し，断片的であいまいな発話を状況に応じた相互信念を用いて適切に理解して行動できるようになることを示す．本方法は，人とロボットのインタラクションにおいて，言語処理と身体性を反映した認知処理の融合を実現しており，より自然なコミュニケーションの実現のための新しいフレームワークを提供するものと考えている．

**キーワード**　相互信念, 発話理解, ロボット，コミュニケーション, 学習

# Mutual Belief Forming by Robots based on the Process of Utterance Comprehension

Atsuto MIYATA[†,††], Naoto IWAHASHI[†], and Akira KUREMATSU[††]

† Sony Computer Science Labs.
Takanawa Muse. Bldg. 3-14-13, Higashigotanda Shinagawa-ku, Tokyo, 141-0022 Japan
†† Course in Electronic Engineering, University of Electro-Communications
1-5-1, Chofugaoka, Chofu-Shi, Tokyo 182-8585, Japan

E-mail: †{miyata,iwahashi}@csl.sony.co.jp, ††kure@apple.ee.uec.ac.jp

**Abstract**　This paper describes a method of learning a mutual belief, necessary for natural language communication with people, in a process of utterance comprehension by robot. In this method, a system of mutual beliefs is represented by multiple beliefs, and the weights for the confidence that each of the belief is shared with a human. The beliefs dealt with in the method include phonemes, lexicon, grammar, influence of behavioral context, and other nonlinguistic belief. In the experiment, arm-robot with only a basic language knowledge interacts with human by using language and action. Through the interaction, robot learned the mutual beliefs and was eventually able to understand even fragmental and ambiguous utterances according to given situations, and act appropriately. The methods made it possible to integrate the language and cognitive processes, and thus introducing a new framework for natural communication.

**Key words**　mutual belief, utterance comprehension, robot, communication, learning

## 1. Introduction

Language communication in a daily life is based on the mutual beliefs shared by those who are communicating [1]. The mutual beliefs are formed through common experiences with common cognitive ability, and used in the process of utterance production and comprehension. Therefore, utterances are produced with an assumption that a speaker and a listener share similar beliefs concerning meaning. Through these beliefs, a listener can infer the meaning of the utterances. Such mutual beliefs are diverse, since it includes not only linguistic, but also nonlinguistic beliefs, and varies by experiences. So, if we want to enable for humans and robots to communicate with each other the way people do, we need a language processing method that forms mutual beliefs, and uses them appropriately in multi-modal interaction.

This paper presents a method, in which a robot can form mutual beliefs in the multi-modal interaction with a person. Particularly, the method deals with the beliefs which are not directly conveyed by utterances spoken by the person. The robot learns incrementally, in the process of understanding fragmental and ambiguous utterances, according to situations. The learning was carried out by using the information of speech, visual observations and behavioral reinforcement. The beliefs dealt with in the method include those concerning linguistic information - phonemes, lexicon, grammar - and nonlinguistic information - attentional gestures, behavioral context, and task-dependent knowledge.

## 2. Interaction for Forming of Mutual Beliefs

The interactive learning task for forming mutual beliefs was set up as follows. A robot was set next to a table so that the robot and a person sitting at the table could see and move the objects on the table (Fig. 1). The robot initially had certain basic linguistic beliefs, including a lexicon with a small number of items and a simple grammar, and could understand utterances to some extent. By speaking slowly and pausing briefly between words, and using attentional gestures, the person asked the robot to move objects (stuffed-toys). If the robot responded incorrectly, the person slapped the robot's hand. The robot would then respond by acting in a different way. Through a sequence of such reinforcing interaction, the robot incrementally learned an expanded set of mutual beliefs to understand even fragmental and ambiguous utterances according to given situations.

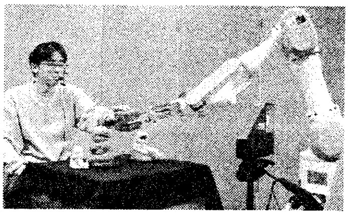


Fig. 1 System set-up

The mutual beliefs were learned in the process of understanding fragmental and ambiguous utterances. The confidence that each belief was shared between the robot and the person was strengthened when the robot showed misunderstanding of a utterance in its first response, but understood the utterance correctly by using the belief in the second response, invoked by being slapped.

An example, using mutual beliefs in utterance production by person, and comprehension by robot in the task is as follows. In the scene shown in Fig. 2, the object on the left, *Kermit*, has just been put onto the table. When a person would like to move Kermit onto the box, the person may specifically say *"Kermit box move-onto"*[1] In this situation, if the person assumed that the robot had a belief that a box is something for something to be moved onto, he may say *"Kermit move-onto"*, using a fragmental utterance. Moreover, if the person assumed that the robot believed that an object moved in the previous action was likely to be the next target for movement, he might just say *"move-onto"*. To understand these fragmental utterances, the robot had to have similar beliefs, and knew that robot shared these beliefs with the person.

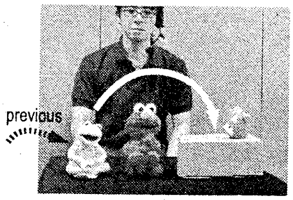


Fig. 2 Example of forming mutual belief

## 3. Representation of Mutual Beliefs

In our algorithm, the system of mutual beliefs consists of beliefs with a confidence that each belief is shared between the robot and the person. This system is represented by probabilistic models including gaussian distribution and hidden Markov models(HMMs). The confidence in each belief is represented by a weighting factor for the output of probabilistic model for each belief. The beliefs dealt with are those concerning lexicon and grammar, behavioral context, and motion-object relationships.

### 3.1 Lexicon and Grammar

Let $L$ denote the parametric model for the lexicon including lexical items $c_i, i = 1, \ldots M$. Each item consists of a pair of concepts and a word. The speech $s$ for the word and the image $v$ for the concept in lexical item $c_i$ are respectively represented by distributions $p(s|c_i)$ and $p(v|c_i)$. The lexicon $L$ includes the concepts of static images of the stuffed toys and the concepts of motions. The distributions for the concepts of static image of the stuffed toys are represented by gaussian distributions, and the distributions for the concepts of motions and the distributions for words are both represented by HMMs.

Let $G$ denote the grammar. We assume that each phrase in a sentence utterance describes a landmark, trajector, or motion, and that the conceptual structure $z$ in each sentence is expressed with semantic attributes - [motion], [trajector],

[1] The robot had the lexicon without any functional words, and fairly simple grammar that represented the occurrence probabilities of the order of the constituents characterized by semantic attributes in a sentence. The lexicon and grammar was learned in supervised way by the robot (for details, see [2]~[6]).

$$\Psi(s,t,u,q,L,G,R,B,\Gamma) \equiv \max_{l,z} \left( \underbrace{\gamma_1 \log p(s|z,L,G)}_{\text{Speech}} + \underbrace{\gamma_2\{logp(t|W_t,L) + \log p(l|W_l,L)\}}_{\text{Object}} + \underbrace{\gamma_2 \log p(u|W_m,L)}_{\text{Motion}} \right. \tag{1}$$

$$\left. + \underbrace{\gamma_3 \log p(t,l|W_m,R)}_{\text{Motion-Object Relationship}} \quad + \underbrace{\gamma_4\{f(t,q,B) + f(l,q,B)\}}_{\text{Behavioral Context}} \right)$$

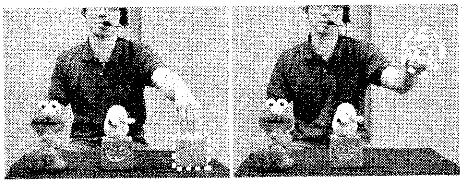
and [landmark]. For the scene in Fig. 2, the corresponding spoken sentence would be a sequence of spoken words, "big Kermit brown box move-onto", and the conceptual structure might be

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

where the right column contains the spoken words, and the left column corresponds to trajector, landmark, and motion. Let $y$ denote the order of the semantic attributes so that it represents the order of constituents with the semantic attributes in a sentence. For instance, in the given example of a spoken sentence, the order is [trajector]-[landmark]-[motion]. The grammar $G$ is represented by a set of occurrence probabilities for possible combinations of order as is represented by a set of occurrence probabilities for possible combinations of order as $G = \{P(y_1), P(y_2), ..., P(y_k)\}$.

### 3.2 Behavioral context

Here, the behavioral context is any behavior that can be used to predict the contents that utterances are likely to describe. The behavioral context particularly includes the previous action and the current attentional gesture. 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 $f(o,q)$. In the system, attentional gestures are categorized into two types, pointing and holding (Fig. 3). We distinguish between the two types of gesture because a holding gesture may only indicate an attempt to hold an object in its position, whereas a pointing gesture is used to indicate direct attention.



(a) Pointing (b) Holding

Fig. 3 Attentional gestures

$f(o,q)$ takes $b_p$ as its value, if $o$ is being pointed, $b_h$ if $o$ is being held, $b_c$ if $o$ is involved as trajector or landmark in a previous action, and otherwise, 0.

### 3.3 Motion-Object relationship

Let $R$ denote a parameter set representing a belief concerning a motion-object relationship. The belief concerning the relationship between motion $W_m$, and the features of the trajector and landmark objects, $t$ and $l$, involved in an action, is represented by $p(t,l|W_m,R)$ as a gaussian distribution of vector $o_{t,l} = [o_t, o_t - o_l, o_l]^T$. Here, $R$, $o_t$ and $o_l$ denote, respectively, the parameter set representing this belief, the features of the trajectory, and the landmark object.

## 4. Utterance Comprehension

In this paper, we define utterance comprehension as inferring the action which the utterance describes. Utterances are understood by using beliefs relating to the situation. Situation includes the allocation of objects on the table, attentional gestures used during utterance, and objects used in the previous action. An action is represented by the trajector $t$ and trajectory of the motion $u$. Given behavioral context $q$, the beliefs (lexicon $L$, grammar $G$, and the effect of behavioral context $B$, motion-object relationship $R$), and the confidence of the beliefs $\Gamma = [\gamma_1, \ldots, \gamma_4]$ , the corresponding action understood to reflect the meaning of speech $s$ is determined by maximizing the decision function (1).

## 5. Learning of Parameters for Mutual Beliefs

Let $s_i$ denote the $i$th utterance during the course of learning. And let $\{t_i, u_i\}$ denote the correct response expected by the human.

The parameters for the beliefs, $L$, $G$, $R$ and $B$, are learned by using Bayesian learning method. This learning takes place after each episode in which the robot showed correct understanding of utterance $s_i$ in its first response or in the second response.

The parameters for the confidence, $\Gamma$, are optimized incrementally through the sequence of episodes so as to minimize the number of decision errors. If the robot showed correct understanding of utterance $s_i$ in its first response or in the second response, loss $l_i$ is given as

$$l_i = \Psi(s_i, t_i', u_i', q_i, L, G, R, B, \Gamma) - \Psi(s_i, t_i, u_i, q_i, L, G, R, B, \Gamma) \tag{2}$$

where

$$(t_i', u_i') = \underset{(t,u)}{\operatorname{argmax}}\, \Psi(s_i, t, u, q_i, L, G, R, B, \Gamma) \tag{3}$$

If the robot showed correct understanding of $s_i$ neither in the first nor second responses, the robot cannot obtain the information of the correct action $\{t_i, u_i\}$ and loss $l_i$ is set to 0 for convenience. The loss $l_i$ is used to calculate global loss at the $i$th episode, $L_i = \sum_{j=1}^{i} l_j$. The parameters for the mutual beliefs are learned after each episode in which the first response is incorrect and the second response is correct. The global loss $L_i$ is minimized by a gradient descent algorithm, and parameters are updated until the parameter converges.

## 6. Experiments

### 6.1 Condition

The experiment was done using a set of data including speech, scenery with objects, and behavioral context(pointing gestures, holding gestures, and previous actions). Along with each set of data, a correct response labeled by human, was given. In the experiment, response by the robot for a set of data was automatically checked against pre-labeled correct response, therefore, we were able to execute a simulated experiment.

Speech was represented by using mel-scale cepstrum coefficients and their delta parameters (32 dimensional). Static object features captured from the camera device were represented by their size (one dimensional), color (three dimensional: $L^*$,$a^*$,$b^*$), and shape (two dimensional: width/height, squareness). Motion was represented by a sequence of coordinates (two dimensional: vertical and horizontal) and velocity (two dimensional). For an attentional gesture in a behavioral context, because pointing induces direct attention, we defined $b_p$ as having an enough large value 100.

Each of motion-object relationship model $R$ has been initialized with 100 randomly selected objects.

Initially, confidence measure was set to a given parameter ($\gamma_1 = \gamma_2 = 0.5, \gamma_3 = \gamma_4 = 0.0$). Note that the confidence measure in this experiment was defined such that ($\gamma_1 + \gamma_2 + \gamma_3 = 1$). Also, $\gamma_4$ was separated into two individual confidence measures ($\gamma_{4,1}, \gamma_{4,2}$), where $\gamma_{4,1}$ indicate confidence for gestural attention ($\gamma_{4,1} = \gamma_4 \cdot b_h$), and $\gamma_{4,2}$ indicate confidence for behavioral context($\gamma_{4,2} = \gamma_4 \cdot b_c$).

Interactions for the experiment were classified into three categories depending on the level of difficulty. where the level of difficulty is different.

In the experiment, we used two sequences of utterances, *sequence A* and *sequence B* (128 utterances for each set).

i) *Sequence A*

In *sequence A*, no information was omitted from the utterances used in the first 32 episodes. Figure 4(a) shows an example of this type of interaction. The utterance is “Kermit Elmo put-beside”, in response to which the human expected the robot to put *Kermit* beside *Elmo.*

The utterances from episodes 33 through 64 required an understanding of the behavioral context. Figure 4(b) shows an example of this type of interaction. The utterance is “Green toy-box blue toy-box jump-over”, where there are two green toy-boxes in a scene, one held by the human and the other on the table. Because the human expects the holding action to attract attention to the held object, the correct action is for the robot to grab the green toy-box held by the human and make it jump over the blue toy-box.

For episodes 65 through 128, the human made a fragmental utterance. Figure 4(c) shows an example of this type of interaction. In this example, the utterance is “move-onto”, while *Kermit* is held by the human. The response expected of the robot is to take *Kermit* from from the human's hand and move it onto the toy-box.

ii) *Sequence B*

In *sequence B*, fragmental utterances were used in all 128 episodes. For comparison, the last 64 episodes in *sequence B* were the same as in *sequence A*.

## 6.2 Results

The error rate for *sequence A* in Fig. 6(a) clearly show that the robot was able to communicate better with learning. The

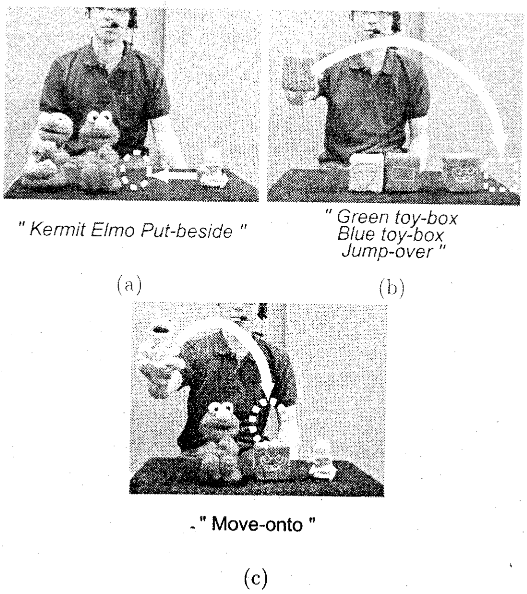


Fig. 4 Utterance types

effect of learning can be seen in Fig. 7(a), where a sharp distribution in the motion-object relationship belief represent a belief for squareness of landmark object in motion “move-onto”. Also, the increased confidence in the motion-object relationship belief (Fig. 8(c)) after episode 64 show that robot is learning to use the motion-object relationship belief. However, the error rate for *sequence B* after episode 96 (Fig. 6(b)) did not match that for *sequence A*, despite the same episode sequences being used. The reason for this difference is illustrated in Figs. 7(a),(b). With many successful episodes in *sequence A*, the robot had already developed a belief by episode 64. Although, with *sequence B*, a lack of successful episodes early on meant more episodes were needed for the robot to develop a belief.

Figures 5(a)-(c) show generated actions as the results of the utterance comprehension after learning. The differences in the calculated log probabilities between the first and the second candidates for the decision are also shown.

In Fig. 5(a), the human said “Lift”, indicating that the robot should lift the Kermit in his hand. The Kermit held by the human was chosen as the first candidate. The details of the log probability show that, for this example, the behavioral context belief based on the human's hand was effective.

In Fig. 5(b), the object on the left, *Barba*, had been put onto the table in the previous action. The person said “move-onto”, meaning to put *Barba* on the toy-box. The use of the beliefs on behavioral context and motion-object relationship were effective to obtain the correct comprehension..

In Fig. 5(c), the object on the right, the big *Kermit*, had been put onto the table in the previous action. The human said “Grover small Kermit jump-over”, meaning that *Grover* should jump-over the small *Kermit* on the toy-box. The result log probabilities show a large difference in the belief concerning object concepts because the object belief in the first candidate fits the description of “small Kermit” better. In the first two examples, the behavioral context belief was a

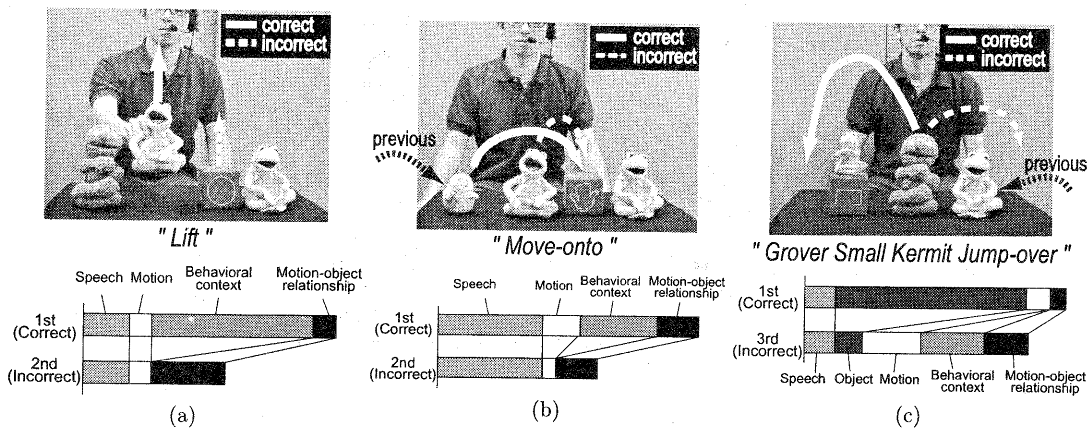


Fig. 5 Results of interaction

factor in understanding an utterance, but in this example the object belief which was much stronger than the behavioral context belief, was effective for correct comprehension.

## 7. Discussion

Although the experiment results showed that the robot could learn the mutual beliefs, which is useful to understand ambiguous utterances. It is interesting to investigate whether a human is able to assume similar mutual beliefs through the interaction. Future work includes such investigation, and the expansion of the methods for more natural language communication between people and robots.

## 8. Related works

Recently communication between a human and a robot has been attracting interest, and there have been the several studies applying a theory of mind. In [8] [9], an infant-like humanoid robot was developed to interact with humans, but the interaction did not include speech. In [7], the importance of gestures in utterance comprehension was shown in the communication of human and humanoid robots. Although, [10] included no mind-reading mechanism in the robot, the robot could use gestures to resolve visual ambiguities in dialogue.

In all of these studies, however, the mutual beliefs were pre-defined before interaction with a human, and could not be changed and expanded for natural communication.

## 9. Conclusion

The method making it possible for a robot to learn the mutual beliefs in the process of utterance comprehension in multi-modal interaction with a human was presented. Once mutual beliefs are established between a human and a robot, the robot can understand even a difficult utterance, such as a fragmental utterance. Just as humans achieve communication by using a set of mutual beliefs, the robot system is capable of achieving communication by building mutual beliefs and using theses beliefs to understand utterances according to situations.

### Bibliography

[1] D. Sperber and D. Wilson、"Relevance: Communication and Cognition.", Oxford, Basil Blackwell, 1986

[2] N.Iwahashi, Language acquisition by robot, Tech. Rep. of IEICE, SP, 2001.12.

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

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

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

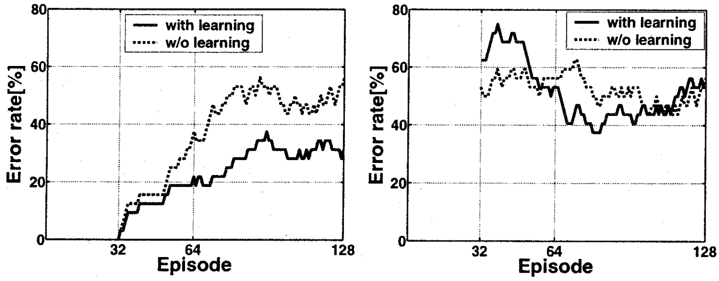
[6] T.Haoka, N.Iwahashi, Speech understanding based on cognitibe language knowledge, Proc. Acoustic Society of Japan Spring Meeting Vol.I, 159-160, 2001.

[7] T. Ono and M. Imai, "Reading a robot's mind: A model of utterance understanding based on the theory of mind mechanism.", In Proceedings of AAAI-2000, pp.142-148, 2000.

[8] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot", 1999 International Joint Conference on Artificial Intelligence, 1999.

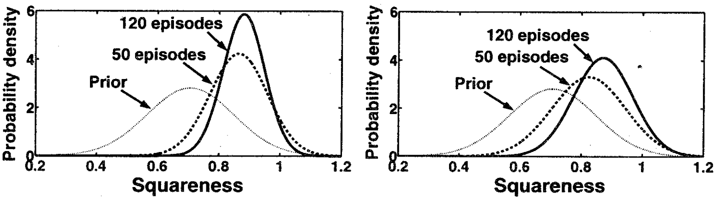
[9] H. Kozima, "Attention-sharing and behavior-sharing in human-robot communication", IEEE International Workshop on Robot and Human Communication (ROMAN-98, Takamatsu), pp.9-14, 1998.

[10] T. Takahashi, S. Nakanishi, Y. Kuno, Y. Shirai,"Human-Robot Interface by Verbal and Nonverbal Communication",IEICE 98-HI-76, pp.37-42.



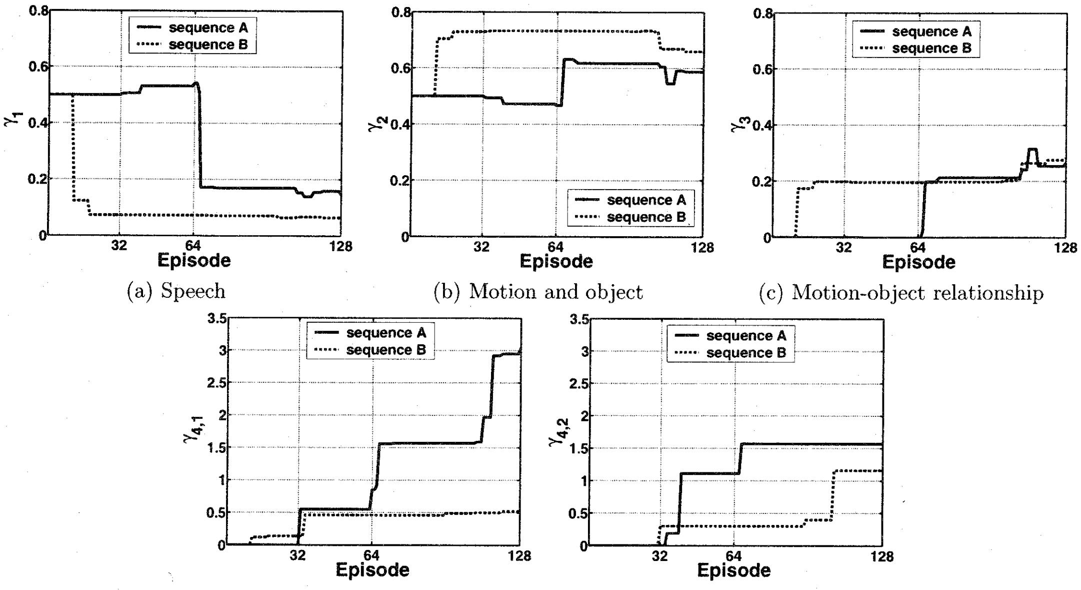
(a) Sequence A (b) Sequence B

Fig. 6 Error rate with different episode sequence



(a) Sequence A (b) Sequence B

Fig. 7 Motion-object relationship with different episode sequence



(a) Speech (b) Motion and object (c) Motion-object relationship

(d) Gestural attention (e) Behavioral context

Fig. 8 Confidence measure for speech, motion and object, motion-object relationship, gestural attention and behavioral context