A Multimodal Infant Behavior Annotation for Developmental Analysis of Demonstrative Expressions

Shigeyoshi Kitazawa [1], Shinya Kiriyama [1], Tomohiko Kasami [2], Shogo Ishikawa [2], Naofumi Otani [3], Hiroaki Horiuchi [1], and Yoichi Takebayashi [4]

[1] Faculty of Informatics, Shizuoka University, Shizuoka, Japan

[2] Graduate School of Informatics, Shizuoka University, Shizuoka, Japan

[3] Graduate School of Science and Engineering, Shizuoka University, Shizuoka, Japan

[4] Graduate School of Science and Technology, Shizuoka University, Shizuoka, Japan

Address3-5-1, Jouhoku, Hamamatsu, 432-8011, Japan

E-mail: { kitazawa@, kiriyama@, kasami@pooh.cs., shogo@pooh.cs., nao@cezanne.cs., horiuchi@, takebay@ }inf.shizuoka.ac.jp

Abstract

We have obtained the valuable findings about the developmental processes of demonstrative expression skills, which is concerned with the fundamental commonsense of human knowledge, such as to get an object and to catch someone's attention. We have already developed a framework to record genuine spontaneous speech of infants. We are constructing a multimodal infant behavior corpus, which enables us to elucidate human commonsense knowledge and its acquisition mechanism. Based on the observation of the corpus, we proposed a multimodal behavior description for observation of demonstrative expressions. We proved that the proposed model has the nearly 90% coverage in an open test of the behavior description task. The analysis using the model produced many valuable findings from multimodal viewpoints; for example, the change of 'line of sight' from 'object to person' to 'person to object' means that the infant has obtained a better way to catch someone's attention. Our intention-based analysis provided us with an infant behavior model that may apply to a likely behavior simulation system.

1. Introduction

Many researches about spoken language acquisition based on our infant behaviors have been conducted for a long time [1, 2]. Most researches, however, study only a single-shot hypothesis, and the test data for the observations is limited in a single modality.

On the other hand, we aim at constructing a "multimodal infant behavior corpus," which annotated comprehensively in the multiple modalities, such as utterance, gesture, and sight. The TalkBank project [3] is accumulating the speech corpus of infants. It includes the multimodal data; however, multimodal observations are only a small part. Deb Roy's group is collecting the infant behavior video data from 0 to 3 years old [4]. They aim to develop a computational framework that simultaneously models referential and functional meaning. Their approach, however, depend on existence of natural language processing models.

In order to create our corpus, we have been holding a regular infant school and recording spontaneous infant behaviors with video and speech. This means that the corpus data increases continuously. Our goal is to represent commonsense knowledge as the computational models, which are applied to the spoken dialogue systems that realize smart and clever man-machine communications by understanding speakers' intentions and emotions appropriately.

In our previous work, the phoneme acquisition process of an infant was investigated [5]. The problem was that the developmental analysis was limited within natural language description. Our corpus data includes a huge number of annotations from multimodal viewpoints, which increases continuously as the practices of the infant school every week. A smarter method to conduct the analysis effectively is indispensable.

In this study, we focused on the development of demonstrative expressions including basic intentions such as to catch someone's attention, to get something, and to mention something. We regard them as an important part of human commonsense knowledge. The fact that the demonstrative expressions are represented with rich observable cues, such as utterance, gesture, and sight is the real advantage of our study.

In the next section, our environments for the observations of infant behaviors and our developed wearable system for speech recording are described. Section 3 explains the method of multimodal behavior annotation for the observations of demonstrative utterances and its evaluation. We show the results of the developmental analysis in Section 4, and conclude the paper in Section 5.

2. A Multimodal Infant Behavior Corpus

2.1 Learning environment

We have an experimental parent-child learning environment for recording [6]. It has two purposes: first, to provide good educational atmosphere to the participants, and second, to provide flexible layouts where we can regularly monitor the infants' behavior and development.

Sixty-minute classes are held three times weekly, each class was involved with three infant-parent pairs, where the infants are of the same age. One teacher is assigned to each class. The first half of a class takes place in a classroom setting where the teacher utilizes various materials, such as clay, crayons, paper, etc. and has the

infants complete various tasks, such as building, drawing or identifying things, usually with their parents' aids (as shown in the left of Figure 1.). For the second half of a class, the teacher and parents discuss on childcare and child learning. During this time, the infants are given various toys and are let to play freely (as shown in the right of the Figure 1). The program also includes reports from their home, i.e. parents' observations on child's development.

In order to capture infant behavior in detail, many pieces of equipment such as cameras and microphones are required. In order to observe infant behavior, we place four cameras that allow us to operate them by remote control shown in Figure 2. Through the Internet, we can follow the movements of each individual infant using the cameras. We have succeeded to develop the ubiquitous environments enabling us to record the behavior of infants wherever they move around.

The whole sixty-minutes sessions are recorded through four cameras placed at different angles and multiple microphones shown in Figure 2, including rucksack microphones worn by each infant shown in Figure 3. The positioning of the cameras and an overview of the classroom and studio is shown in Figure 2. At the time of the writing, we have footage of 51 learning sessions over a year and a half's time.

Figure 1: *Screenshots of our infant school in a classroom (left) and a playroom (right).*

2.2. Infant utterance recording

In order to observe infant utterances, the speech data with less noise and high quality is indispensable. At the beginning, we had recorded the speech data using the microphones embedded in the beams of the yurt equipped in the classroom shown in Figure 2. The signal to noise ratio of recorded speech was not enough for speech analysis because of distance from the source (children). Moreover, we need to identify the speaker of the individual utterance, that is who had spoken what even under multi-speakers' overlapped speeches. Better quality recording equipments need to be developed.

The utterances include important information that explains mental behavior such as intentions or emotions. Therefore, we have developed a wearable speech recorder shown in Figure 3. Two condenser microphones are arranged near both shoulders. Recorded speech is stocked in a voice recorder inside of the rucksack. We have investigated the quality of the speech data recorded with the developed device and found to be sufficiently high enough for speech analysis and also could identify individual utterances as well as murmurous soft voices.

The previous experiments proved that the use of the developed device facilitates the recording of utterances of

Figure 3: *Wearable speech recording device.*

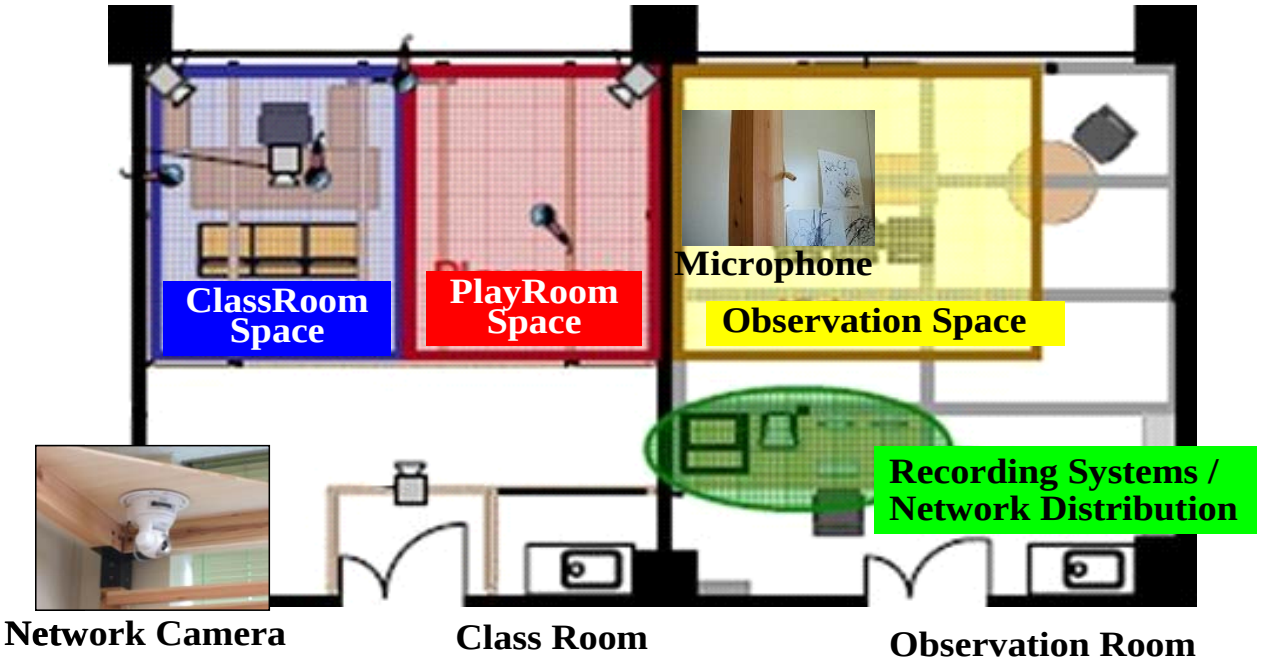


Figure 2: *Infant learning environment layout.*

hyperactive infants with high quality. Utilizing the speech data recorded by the developed device enable us to analyze infant utterances in detail [5].

3. Multimodal infant behavior annotation

3.1. Procedure

We took the following steps to make multimodal annotations about demonstrative expressions;

(1) **Extract those utterances that have at least one feature among the following from the corpus**: demonstrative utterance (e.g. *this* or *that*), pointing by hand, or pointing by finger. The speech and video data of 30 minutes classroom for each month was used to analyze the utterances of an infant for 10 months (14 to 23 months old). As the result, 240 utterances were picked up.

(2) **Make natural language descriptions** of background situations, contents of utterances, prosodic features, and actions for each utterance.

(3) **Consider what kinds of features are necessary** to explain the change of demonstrative expressions by the natural language descriptions, and decide items for the description of features. Six items have chosen; intention, age, utterance, prosody, line of sight, and gesture.

(4) **Arrange the natural language descriptions** based on the selected items. Each description consists of a pair of an index of the six items and an explanation in natural language.

(5) **Decide the sets of words for each item** by analyzing the arranged descriptions.

(6) **Screen utterances that can be obviously categorized into the decided 'intention' item** out of the 240 utterances. then 65 utterances survived after our screening.

(7) **Convert the natural language descriptions** for the 65 utterances into the new format based on the proposed model.

3.2. A multimodal behavior description model

As the results of Step (5) in Section 3.1, we propose a model to describe multimodal infant behavior of demonstrative expressions;

- **Intention:** want, request, opinion, discovery.
- **Age:** 14 month, 15 month , 16 month …
- **Utterance:** single vowel, demonstrative, single noun (except for demonstratives), more than a word.
- **Prosody:** normal (flat in F_0 and intensity), awareness (rising F_0), emphasis (higher average and rising intensity), assertion (falling F_0), question (rising F_0 at the end), calling out (gentle falling F_0).
- **Line of sight:** object, person, object to object, object to person, person to object.
- **Gesture:** pointing by hand, pointing by finger, pointing details by finger, pointing by finger and tapping, passing, getting, showing.

For each item, one of key words was selected from each entry for description.

3.3. Evaluation of the proposed model

In order to verify the value of the proposed description model, we have conducted two experimental evaluations; a closed and an open test. We have checked the number of 'out of vocabulary (OOV),' which means the feature of each item was indescribable within the set of words. The screened 65 utterances used for the closed test. For the open test, 32 utterances of the same infant having the obvious target intentions extracted newly and randomly from the corpus were annotated by the proposed model. The same evaluation was conducted for the 32 utterances. As shown in Table 1, the results proved that almost 90% of described items were describable using the proposed model. The meaningless words raised the number of OOV in the 'utterance' item. The directions by 'lines of sight' and 'gestures' such as leaning forward are examples of OOV. These behaviors began to appear mostly in the last month. The increase of the corpus data will push us to revise the model.

Table 1. *Evaluation results of the proposed description model.*

Test	Coverage rate (Total number)	Numbers of OOV			
		Utterance	Prosody	Sight	Gesture
Closed	91.2% (65*4)	16	0	4	3
Open	89.8% (32*4)	10	0	1	2

4. The developmental analysis of demonstrative expression

We have analyzed the development of demonstrative expression skills by the following two steps; (1) Observations of each feature. (2) Investigation of individual 'intention.' The following two subsections describe the result of each step, respectively.

4.1 Feature-based analysis

The natural language description with the index information of feature descriptions (produced by Step (4) in Section 3.1) was observed. The results for each of the four features (utterance, prosody, line of sight, and gesture except for 'intention' and 'age') are shown in Figure 4.

Utterances: The example shows that an utterance /*ue*/ in 14 month develops through /*koe*/ or /*koko*/ up to an two words phrase /*kore irete*/ in 22 month of his age.

Prosody: The same example as above that was spoken in emphatic prosody in 14 month of age became a controlled calling out.

Line of sights: Changes in 'line of sight' from 'object to person' to 'person to object' mean that the infant has got a better way to catch someone's attention.

Gestures: An appearance of 'pointing by finger and tapping' shows that the infant has grown enough to express his intention clearly. At 20 month of age, a more sophisticated strategy was observed that a child passes something to his mother to make her do a task for him.

4.2 Intention-based analysis

The annotation data based on the proposed model consisting of total 97 utterances used in the evaluation in

Section 3.3 was investigated according to individual 'intention.' We have found the various developmental changes as follows:

- **Want:** After 20 month, 'assertion' and 'calling out' have increased in the 'prosody' feature.
- **Request:** This first appeared in 17 month. The 'gesture' changed from 'pointing by hand' to 'pointing by finger,' and finally to 'passing it to the hand of his mother.'
- **Opinion:** This first appeared in 16 month. In 18 month, 'pointing (an object) by finger' appeared when he judged 'a person's' opinion by tracking the person's eyes. 'More than a word' utterances and 'question' prosodies appeared in 22 month.
- **Discovery:** 'Prosody' changed from 'awareness' and 'emphasis' to 'assertion' and 'calling out' which demands consciousness of others. 'Gestures' of 'pointing details by finger' appeared in 16 month.

4.3. Discussions

Investigations in Section 4.2 revealed the relationships between the observation results in Section 4.1 and the 'intentions' in 4.2. The costs of annotation by the proposed method were remarkably reduced in comparison with that of annotation by natural language. These facts support advantages of the proposed description model.

We plan to apply the results of developmental analyses into the construction of an infant behavior simulation system, which provides parents and teachers with infants' possible reactions in various situations. The inputs of 'intention' and 'age' will reduce possible behavior accompanied with likely features of 'utterance,' 'prosody,' 'line of sight,' and 'gesture,' as the output of the system.

For the improvement of the model, further considerations of intentions or goals of each behavior are indispensable. In the screening process (at Step (6) in Section 3.1), 175 utterances out of 240 remained unlabelled in terms of 'intention' features. We shall continue this kind of goal-oriented methodologies for behavior analysis.

5. Conclusion

We have proved that our multimodal infant behavior corpus is useful to analyze the developmental processes of demonstrative expression skills, which concern the fundamental human commonsense knowledge, such as to get an object and to catch someone's attention. We proposed a multimodal behavior description model to observe the demonstrative expressions. We showed that the proposed model has nearly 90% coverage in an open test of behavior description tasks. The analysis results using the model produced many valuable findings from multimodal viewpoints. Especially, the results of intention-based analysis provided us with a model of infant that is possible to apply to construct a behavior simulation system. In the future, we shall enhance the behavior description model by continuing our goal-oriented observations.

6. Acknowledgements

The authors would like to thank members of Kitazawa Laboratory and members of Takebayashi Laboratory for their assistance in the preparations of this paper.

7. References

[1] Oller, D. K., "Metaphonology and infant vocalizations," Precursors of Early Speech, pp.21-35, 1986.

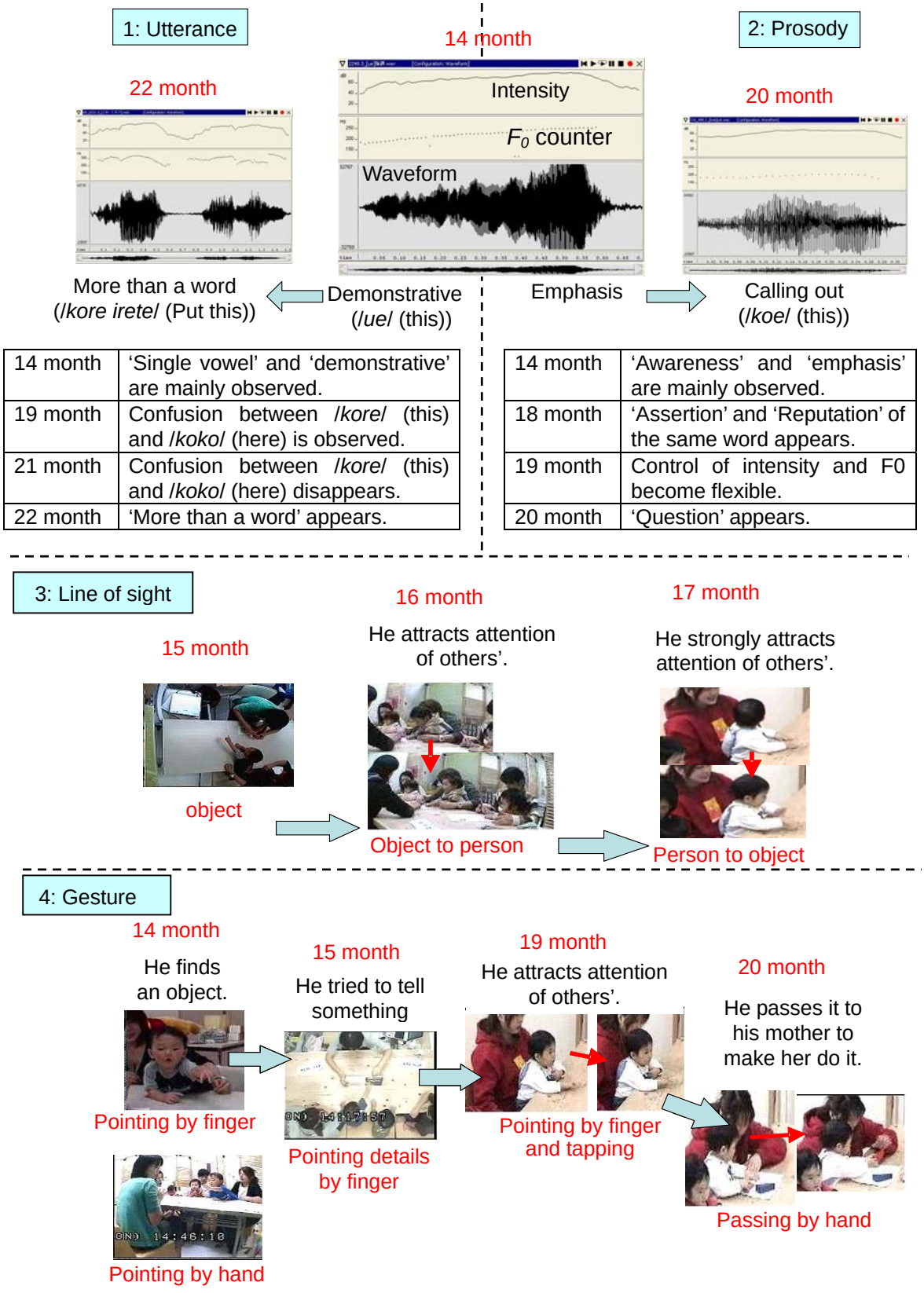
[2] K. Ejiri （1998） Relationship between rhythmic behavior and canonical babbling in infant development, Phonetica 54, 226-237.

[3] MacWhinney, B., Bird, S., Cieri, C., & Martell, C. (2004). TalkBank: Building an open unified multimodal database of communicative interaction. In LREC 2004 (pp. 525-528). Lisbon: LREC.

[4] Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak. (2006), 'The Human Speechome Project,' the Proceedings of the 28th Annual Cognitive Science Conference.

[5] Ryo Tsuji, Tomohiko Kasami, Shogo Ishikawa, Shinya Kiriyama, Yoichi Takebayashi, Shigeyoshi Kitazawa, "Observations of the Spoken Language Acquisition Process Based on a Multimodal Infant Behavior Corpus," Interspeech2006, 2006-9.

[6] Yoichi Takebayashi: Multimodal Knowledge Contents Design from the Viewpoint of Commonsense Reasoning. Proceedings of GSIS International Symposium on Information Sciences of New Era: Brain, Mind and Society. Sendai, Japan 2005.



14 month	'Single vowel' and 'demonstrative' are mainly observed.
19 month	Confusion between /*kore*/ (this) and /*koko*/ (here) is observed.
21 month	Confusion between /*kore*/ (this) and /*koko*/ (here) disappears.
22 month	'More than a word' appears.

14 month	'Awareness' and 'emphasis' are mainly observed.
18 month	'Assertion' and 'Reputation' of the same word appears.
19 month	Control of intensity and F0 become flexible.
20 month	'Question' appears.

Figure 4: *Results of the developmental analysis of demonstrative expressions for the features of 1: utterance, 2: prosody, 3: line of sight, and 4: gesture.*