

The Developmental Analysis of Demonstrative Expression Skills Utilizing a Multimodal Infant Behavior Corpus

Shinya Kiriyama¹, Ryo Tsuji², Tomohiko Kasami², Shogo Ishikawa²
Naofumi Otani³, Hiroaki Horiuchi¹, Yoichi Takebayashi⁴, and Shigeyoshi Kitazawa¹

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

{kiriyama@, ryo_t@pooh.cs., kasami@pooh.cs., shogo@pooh.cs.,
nao@cezanne.cs., horiuchi@, takebay@, kitazawa@}inf.shizuoka.ac.jp

Abstract

We have succeeded to obtain the valuable findings about 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 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 utilizing the corpus, we proposed a multimodal behavior description model for the effective 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 results 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. Furthermore, the results of intention-based analysis provided us with an infant behavior model which is possible to apply to construct a behavior simulation system.

Index Terms: spoken language acquisition, multimodal behavior corpus, demonstrative expressions, goal-oriented observation.

1. Introduction

Many researches about spoken language acquisition based on the observations of 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, the results of multimodal observations are few. 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 is, however, to use the existing models of natural language processing. In order to create the 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 natural language description was the only method for the developmental analysis. Our corpus data includes a huge number of annotations from multimodal viewpoints and increases continuously as the practice of the infant school. A smart method to conduct the analysis more effectively is indispensable.

In this study, we focused on the development of demonstrative expression skills, because they convey the 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 cues, such as utterance, gesture, and sight, which are easy to observe, encourages us to focus on them.

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 infants [6]. Figure 1 shows screenshots. It has two purposes: first, to provide good education to the participants, and second, to provide us with a setting where we can regularly monitor the infants' behavior and development.



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

Three sixty-minute classes are held weekly, each class consisting of three infant-parent pairs, where the infants are of the same age. There is one teacher per 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 the parents' aid. For the second half of a class, the teacher and parents discuss child care and child learning. During this time, the infants are given various toys and are let to play freely. The program also includes reports on what is happening at homes, including parents' observations on child's development.

The whole sixty-minute sessions are recorded by four cameras placed at different angles and multiple microphones, including rucksack microphones worn by each infant. 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.

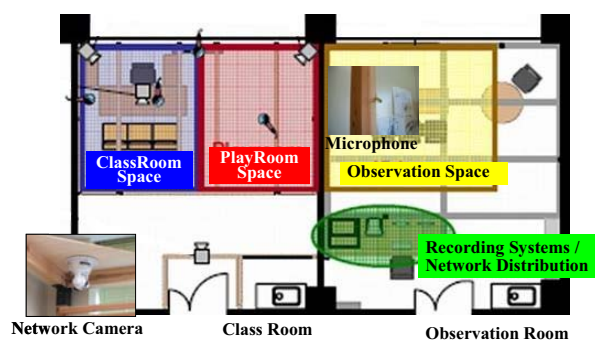


Figure 2: Infant learning environment layout

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.

We have developed the wearable speech recording device shown in Figure 3. Two condenser microphones are arranged near both shoulders and the recorded speech is stocked in the voice recorder stored inside of the rucksack.

The previous experiments proved that the use of the developed device facilitates the recording of utterances of hyperactive infants with high quality. Utilizing the speech data recorded by the developed device enable us to analyze infant utterances in great detail [5].



Figure 3: The wearable speech recording device.

3. Multimodal infant behavior annotation

3.1. Procedure

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

(1) **Extract utterances which 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 minute class room for each month was used to analyze the developmental process 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 utterance, 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 by natural language.

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

(6) **Screen utterances which can be obviously categorized into the decided 'intention' item** out of the 240 utterances. 65 utterances survived the 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 F_0 and intensity), awareness (rising F_0), emphasis (high average and rising intensity), assertion (falling F_0), question (rising F_0 at ending), calling out (gentle falling F_0).
- **Line of sight:** object, person, object to object, object to person, person to object.
- **Gesture:** point by hand, point by finger, point by finger in detail, point by finger and tap, pass, get, show.

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 that the feature of each item can not be described within the set of words, for the screened 65 utterances as the closed test. For the open test, 32 utterances of the same infant which have the obvious target intentions extracted newly and randomly from the corpus, and 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 description items were successfully described by using the proposed model. The utterances of meaningless words increased the number of OOV for the ‘utterance’ item. The indication of direction by ‘line of sight’ and the ‘gesture’ of leaning forward are examples of OOV. These behaviors appeared in nearly the last month. The increase of the corpus data will provide us with the revision of the model.

Table 1. *The 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 skills

We have analyzed the development of demonstrative expression skills by the following two steps; (1) observe developmental processes for each feature. (2) Investigate developmental changes in each individual ‘intention.’ The following two subsections describe the result of each step, respectively.

4.1. Feature-based analysis

The natural language description data with the index information of feature description items (produced by Step (4) in Section 3.1) was used to the observation. The results for each item of the four features (except for ‘intention’ and ‘age’) are shown in Figure 4.

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. The appearance of ‘point by finger and tap’ shows that the infant has grown enough to express his intention clearly.

4.2. Intention-based analysis

The annotation data based on the proposed model, of total 97 utterances used in the evaluation in Section 3.3, was investigated for each ‘intention.’ We have succeeded to find the various developmental changes as follows:

- **Want:** after 20 month, ‘assertion’ and ‘calling out’ of the ‘prosody’ feature increase.
- **Request:** This first appears in 17 month. The ‘gesture’ changes from ‘point by hand’ to ‘point by finger,’ and finally to ‘pass.’
- **Opinion:** This first appears in 16 month. In 18 month, the ‘point by finger’ an object followed by the ‘person’ sight appears. The ‘more than a word’ utterance and the ‘question’ prosody appear in 22month.
- **Discovery:** The ‘prosody’ changes from ‘awareness’ and ‘emphasis’ to ‘assertion’ and ‘calling out’ which need the consciousness of others. The ‘gesture’ of ‘point by finger in detail’ appears in 16 month.

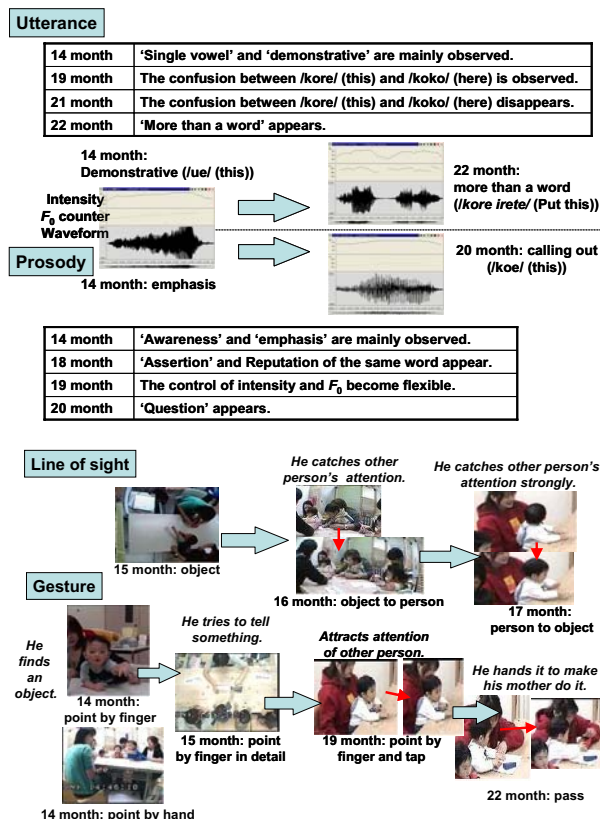


Figure 4: *The results of the developmental analysis of demonstrative expression skills for the features of utterance, prosody, line of sight, and gesture.*

4.3. Discussions

The investigation in Section 4.2 revealed the relationships between the observation results in Section 4.1 and the ‘intentions.’ The cost of the annotation by the proposed method was remarkably reduced in comparison with the annotation by natural language. These facts support 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’ decide a simulated behavior represented by the 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 the ‘intention’ feature. We continue the studies of goal-oriented methodologies for behavior analysis.

5. Conclusions

We 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 for the demonstrative expression observation. We showed that the proposed model has the nearly 90% coverage in an open test of the behavior description task. The analysis results using the model produced many valuable findings from multimodal viewpoints. Especially, the results of intention-based analysis provided us with an infant behavior model which is possible to apply to construct a behavior simulation system. In future, we enhance the behavior description model by continuing the goal-oriented observations.

6. 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 Kiriya, 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.