OBSERVATIONS OF THE SPOKEN LANGUAGE ACQUISITION PROCESS BASED ON A MULTIMODAL INFANT BEHAVIOR CORPUS

Ryo Tsuji[1] *Tomohiko Kasami*[1] *Shogo Ishikawa*[1] *Shinya Kiriyama*[2]

Yoichi Takebayashi[2] *Shigeyoshi Kitazawa*[2]

[1]Graduate School of Informatics, Shizuoka University
[2]Faculty of Informatics, Shizuoka University
3-5-1, Johoku, Hamamatsu-shi, Shizuoka, 432-8011, Japan

Email: {cs1057, cs2027, cs1007}@s.inf.shizuoka.ac.jp, {kiriyama, takebay, kitazawa}@inf.shizuoka.ac.jp

ABSTRACT

We have developed a framework to record spontaneous speech of infants. Using the framework, we have accumulated infant speech data, and proved that the data is quite efficient for the explication of the spoken language acquisition process. We aim at constructing the "multimodal infant behavior corpus," which contributes to the elucidation of human commonsense knowledge and its acquisition mechanism. We previously established the environments to record infant behaviors as multimodal data. We have newly developed a wearable speech recording system and succeeded to record infant utterances with high quality. We have held an infant school once a week for 10 months, and accumulated infant speech data more than 100 hours long. We observed infant utterances in the aspects of acoustic and prosodic features. Through an acoustic observation, we have succeeded to analyze alteration of the pronunciation skills focused on demonstratives that appear quite often in infant utterances. As the result of a prosodic observation, we have also obtained knowledge of how infants enrich and diversify the ways to explain their intentions or emotions corresponding to their growth.

Index Terms: phoneme acquisition, emotional expressions, spoken language acquisition, infant behavior observation.

1. INTRODUCTION

Many researches about spoken language acquisition based on the observations of infant behaviors have been conducted [1][2]. As the recent development of brain science technologies, much scientific knowledge based on the observations of brain functions has been accumulated. For instance, observations by means of the NIRS (Near Infrared Spectrometer) that measures changes of blood current are energetically conducted [3]. Most researches of the above approaches, however, study only a single-shot hypothesis, and test data for observations is limited in a single modality.

On the contrary, we aim at observing infant behaviors in multimodal viewpoints and elucidating human commonsense knowledge and its acquisition mechanism from various perspectives. We have been developing the "multimodal infant behavior corpus," which annotated comprehensively in multiple modalities, such as utterance, gesture, and sight. In order to create the corpus, we have been holding a regular infant school and recording spontaneous infant behaviors with video and speech. Our goal is to represent human commonsense knowledge as the computational models, which are applied to the spoken dialogue systems that realize smart and clever man-machine interactions by understanding speakers' intentions and emotions appropriately.

This paper describes observations of the spoken language acquisition process through the analysis of infant utterances utilizing the developing infant behavior corpus.

In the next section, we describe our environments for infant behavior observations and our developed wearable system for speech recording. Results of infant utterance observation are explained in detail in Section 3. We conclude the paper in Section 4.

2. ENVIRONMENTS FOR INFANT UTTERANCE OBSERVATIONS

2.1 Practice of an infant school and record of infant behaviors

Aiming at constructing the "multimodal infant behavior corpus," we already have the system to hold an infant school regularly [4]. In order to manage the school successively, the agreement of parents on the intention of the school is essential. We have invited an expert of infant education as a collaborator. The expert supervises the school, who advises us on how to enrich activities effective in the sound growing of infants. He also gives parents solutions for their problems about parenting. We have succeeded to establish win-win relationships among infants, parents, and researchers.

For the purpose of recording spontaneous infant behaviors naturally, the school was constituted of a cedar yurt called "Cedar PAO(yurt)" as shown in Figure 1. The yurt enables us to arrange many cameras and microphones freely. Results of subjective evaluation verified that the environments provide infants and parents with a calm and comfortable atmosphere produced by cedar lumber and enable infants to behave naturally with little awareness of recording devices.

The school consists of two classes; one is for the one-year-old and the other is for the two-year-old. Each class has three pairs of infants and their mothers. A 50 minutes class is held once a week for each. The school started in June, 2005. More than 60 classes have been held, and the length of recorded video and speech data far exceeds 100 hours.

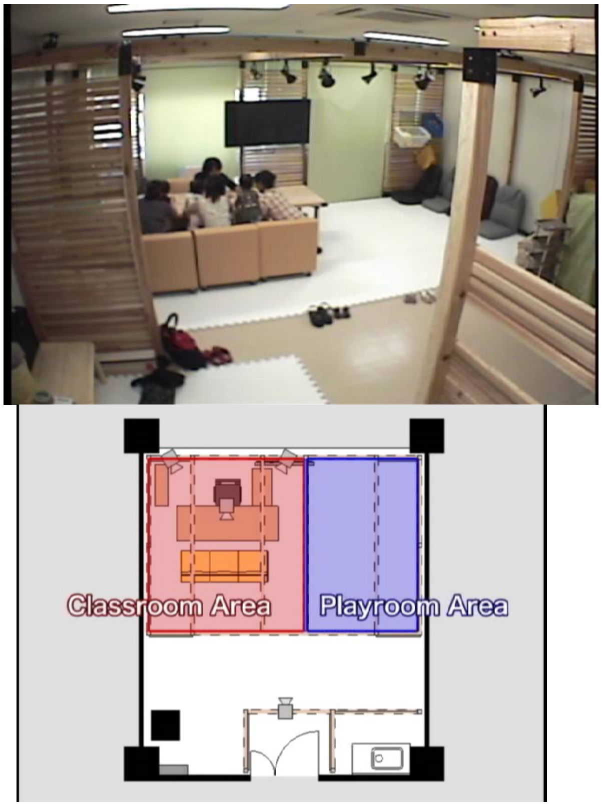

Figure 1. An infant school constituted of a cedar yurt called "cedar PAO(yurt)" (upper), and a sketch of the school (bottom).

2.2 A framework of infant utterances recording

In order to observe infant utterances, speech data with less noise and high quality is indispensable. At the beginning, we had recorded speech data using microphones embedded in the beams of the yurt. Speech data recorded in our old environment had the following problems: (1) large ambient noise was included, (2) sound volume was not stable, and (3) speaker identification was difficult.

To cope with these problems, we have developed a wearable speech recording device shown in Figure 2. Two condenser microphones are arranged near both shoulders. Recorded speech is stocked in a voice recorder stored inside of the rucksack.

We have investigated the quality of speech data recorded by the developed device. Figure 3 shows the difference in quality between speech data recorded by the proposed wearable microphone and by the original environment embedded microphone. As for ambient noise (1), the noise level was improved 17 dB in comparison with the original microphone. About the stability of sound volume (2), the graphs of speech intensity plot indicated that the range of change in the intensity was reduced by using the proposed microphone. Speaker identification ability (3) was also evaluated and the result revealed that the multi channel speech data recorded by the proposed microphones enabled us to distinguish speakers easily for each utterance.

As described above, we have proved that the use of the developed device facilitates the recording of hyperactive infants' utterances with high quality. Utilizing the speech data recorded by the developed device enable us to analyze infant utterances in great detail.

Furthermore, because the developed device is wearable, we can apply it to recording speech data not only at the school, but also at other various situations of everyday life such as homes and parks.

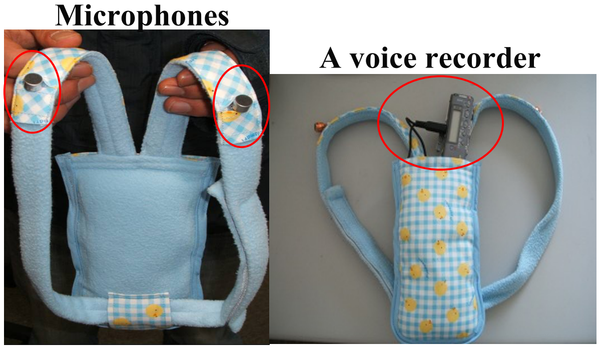

Figure 2. A wearable speech recording device.

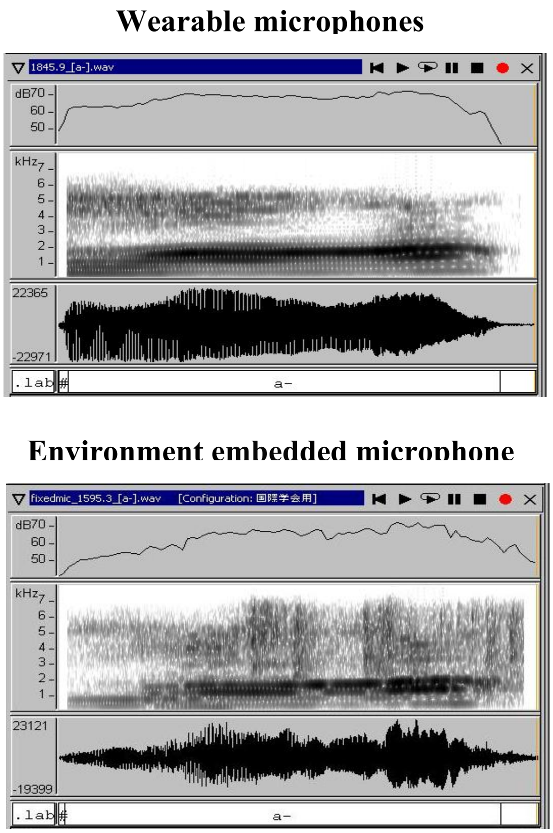

Figure 3. Comparison between the wearable microphone (upper) and the environment embedded microphone (bottom) for an utterance /a:/.

3. ANALYSIS OF INFANT UTTERANCES

The spoken language acquisition has the following two aspects; one is how to learn the phoneme system of the native language, and the other is how to get the ways to express intentions or emotions. We inspected how much the recorded data of infant utterances was effective in observing in those aspects.

We have two classes that differ in the age of infants. Observation of the phoneme system acquisition was conducted using speech data from the one-year-old class. As for observing the acquisition of expressions of intentions or emotions, speech data from the two-year-old class was utilized. This is because a large change in the learning of the phoneme system should be observed in relatively early stage of the growth, and because the variety of expressions of intentions or emotions should be enriched after the acquisition of words.

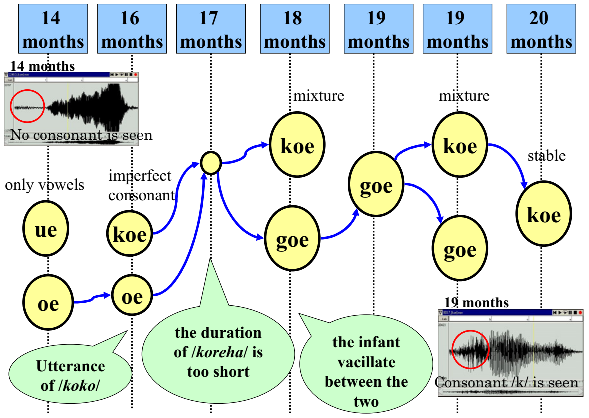

Figure 4. Changes of phoneme arrangement in the utterances /*kore*/(this) of a one-year-old infant.

3.1 Observation of phoneme acquisition

Changes of the phoneme system acquisition have been observed focused on a one-year-old infant. A demonstrative /*kore*/(this) in Japanese was selected as the target for the observation, because the word appeared most frequently in the utterances of infants. Changes of the phoneme arrangement of utterances were traced.

Observation results are shown in Figure 4. 14-month-old utterances consist of vowels only. After 2 months, utterances include a consonant, but the pronunciation is imperfect. From 17-month-old, the infant utters /*koreha*/ means 'this is,' but the duration of the utterances is too short and each phoneme is not acquired perfectly. After a month, the pronunciation of consonants becomes stable. At the same time, some mutters which utter /*koe*/ and /*goe*/ continuously are observed, which shows that the infant vacillate between the two. After that, some utterances include /*goe*/, however, the utterances at 20 months old become /*koe*/ only.

As described above, we succeeded to observe changes of the phoneme arrangement in detail. Further observations will show how /*koe*/ changes into /*kore*/, and what the triggers of each change are.

3.2 Observation of acquisition of intentional or emotional expressions

The two-year-old utterances had been observed for 6 months. As the results of observations based on prosodic features such as fundamental frequency (F_0) and speech intensity, the following 8 items have been extracted as the features of the utterances.

A. flat F_0.

B. rising F_0.

C. falling F_0.

D. strong average intensity.

E. weak average intensity.

F. high average F_0.

G. with rhythm change.

H. with F_0 change in a phoneme.

The frequency of utterances for each item was counted up for the utterances of 24 to 27 months old (the anterior), and for the utterances of 28 to 30 months old (the posterior). Table 1 showed the results of counting and indicated that the total number of utterances of the posterior was increased corresponding to the growth, and that the features such as 'E: weak average F0,' 'G: with rhythm change,' and 'H: with F0 change' in a phoneme were newly appeared in the posterior.

The analysis result of infant utterances proved that the following intentional or emotional expressions appeared most frequently; question, irritation, emphasis, worry, curiosity, mutter and notice. Table 2 showed the combinations of the prosodic features in Table 1 which represented the above 7 kinds of expressions. It indicated that the variety of expressions increased as the infant grew. The analysis also revealed the following facts that the intention of question was possible to be conveyed by not only 'B: rising F0' but also 'E: weak average intensity,' or 'H: with F0 change in a phoneme' which represented worry or diffidence, and that the features of 'D: weak average intensity' and 'G: with rhythm change' performed bouncy and joyful utterances which conveyed the intention of curiosity.

We continue the more detailed analysis to reveal how infants acquire new intentional or emotional expressions.

Table 1. Variety of prosodic expression patterns of a two-year-old infant and its change corresponding to the growth.

Observation periods		24 to 27 months	28 to 30 months
A	flat F_0.	9	30
B	rising F_0.	2	23
C	falling F_0.	3	20
D	strong average intensity.	6	13
E	weak average intensity.	0	9
F	high average F_0.	10	5
G	with rhythm change.	0	9
H	with F_0 change in a phoneme.	0	3

Table 2. Changes of prosodic expression ways to explain intentions or emotions of a two-year-old infant. Bold & italic letters mean that the way is newly acquired in the posterior period.

Intention or emotion	Combinations of the features in Table 1 for the expressions
question	B ***A+E*** ***A+H***
irritation	***A+D+F+H***
emphasis	A+D+F ***C*** ***B+D***
worry	C+F ***A+E*** ***B+E***
curiosity	A+D+F ***C+D+G***
mutter	***C+E*** ***C+G***
notice	A+D ***A+G*** ***C+D***

4. CONCLUSIONS

A wearable speech recording device has been developed, which enables us to record infant utterances with high quality. In comparison of the original environment embedded microphones, the following three points have been improved; (1) ambient noise, (2) stability of sound volume, and (3) ability of speaker identification.

Using the constructed infant learning environments, infant utterances have been recorded by the developed devices, and the data has been observed in detail. The results of observations verified that the collected data was quite valuable for analyses of the spoken language acquisition process. The results of the one-year-old utterance observation showed a process of the phoneme system acquisition minutely. Through the observation of two-year-old utterances, the results proved that the number and the variety of intentional or emotional expressions increased as infants grew.

In future, the more systematic considerations are required to urge the further detailed analyses, which encourages us to undertake the design of sets of acoustic and prosodic labels for the utterances.

5. 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] Wyatt J S, Cope M, Delpy D T, Richardson C E, Edwards A D, Wray S and Reynolds E O R 1990 Quantization of cerebral blood volume in human infants by near-infrared spectroscopy, J. Appl. Physiol. 68, pp.1086-1091.

[4] S. Ishikawa, S. Kiriyama, and S. Kitazawa, "Construction of an infant's learning corpus based on the observation of utterance in place of parents and children coeducation," JSAI2005 (in Japanese)