Analysis of the Effect of Infant-Directed Speech on Mutual Learning of Concepts and Language Based on MLDA and Unsupervised Word Segmentation

Miyuki Funada[1], Tomoaki Nakamura[1], Takayuki Nagai[1] and Masahide Kaneko[1]

***Abstract*— Humans in different cultures acquire languages and concepts from perceptual information and utterances conveyed by caregivers. The utterances used by that the caregivers vary from infant-directed speech (IDS) to adult-directed speech (ADS) according to the growth of infants, and humans learn both IDS and ADS during their growth. In our previous work, we proposed a probabilistic model that enables robots to acquire concepts and language by mutually learning the two. In this study, we adopt IDS, in particular words using a repeated pattern of sounds, and ADS as teaching utterances for the mutual learning model. Furthermore, we verify the influence of teaching utterances on concept formation.**

I. INTRODUCTION

Humans in diverse cultures acquire languages and concepts from perceptual information and utterances conveyed by caregivers. When humans acquire languages and concepts, it is believed that the characteristics of the taught language influence the acquisition of the mother tongue, and that different concepts are formed depending on the language and culture. In addition, caregivers speak to infants in a specific manner called infant-directed speech (IDS). This is known to be different from adult-directed speech (ADS), which is used to talk to adults [3]. The characteristics of IDS are that the pitch is higher, the pitch width is wider, and the speech speed is slower compared to those of ADS. It is known that these features of IDS are common phenomena in many countries and regions. Furthermore, in Japanese, caregivers talk to infants using IDS that has morphological features such as onomatopoeia, mimetic words, and words using a repeated pattern of sounds (e.g., "mama" and "dada"), as well as the tendency to use multiple labels to teach the names of single objects [5]. However, the usage frequency of this type of IDS in English-speaking countries is lower than in Japanese-speaking countries. In addition, Japanese children acquire vocabulary more slowly than English children. However, Japanese children have the ability to learn new words accurately from an early stage. It is therefore considered that IDS has an effect on language learning [4].

We define concepts as categories formed by clustering multimodal perceptual information. Moreover, vocabulary and word meanings can be acquired by connecting words with concepts through interaction with others. During this process, words and concepts are learned with a mutual influence on each other. Taniguchi et al. explained the learning process of concepts and language as a "symbol emergence system" [6]. We proposed a model that enables robots to mutually learn concepts and word meanings [8], and showed that concepts can be not only formed from perceptual information in a bottom-up manner but also affected from language in a top-down manner. However, it is not clear how concepts and language affect each other. In this study, we verify the effect of teaching utterances on concept formation through a learning simulation using IDS and ADS in the mutual learning model of concepts and language.

[1]Miyuki Funada, Tomoaki Nakamura, Takayuki Nagai, Masahide Kaneko are with the Department of Mechanical Engineering and Intelligent Systems, The University of Electro-Communications, 1-5-1 Chofugaoka, Chofushi, Tokyo 182-8585, Japan, m_funada@radish.ee.uec.ac.jp

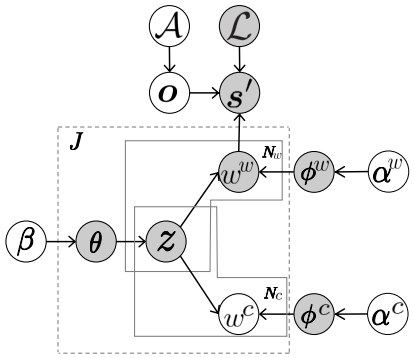

Fig. 1: Graphical model for color concepts and language.

II. MUTUAL LEARNING OF CONCEPTS AND LANGUAGE

The purpose of the model used in [8] was object concept formation, and the structure of the model was considerably complicated. In this study, we deal with the simpler model of color concept formation to analyze the influence of language on concepts. Fig. 1 illustrates a graphical model that integrates language models and color concepts. In Fig. 1, gray nodes denote unobserved nodes, o is the speech uttered by a human, s denotes the recognition result of this speech by the acoustic model with a parameter $\mathcal{A}$ and the language model with a parameter $\mathcal{L}$, and w^w is represents word information, and the bag of words (BoW) representation of s, which can be obtained by dividing s into words using the language model $\mathcal{L}$. In addition, w^c denotes perceptual information, which is color information in this paper. The information w^* is generated from a multinomial distribution with parameter ϕ^*. Furthermore, z denotes the category of color, and is generated from the multinomial distribution with parameter θ, and α^* and β are parameters of the Dirichlet distribution, which is the prior distribution of ϕ^* and θ. The

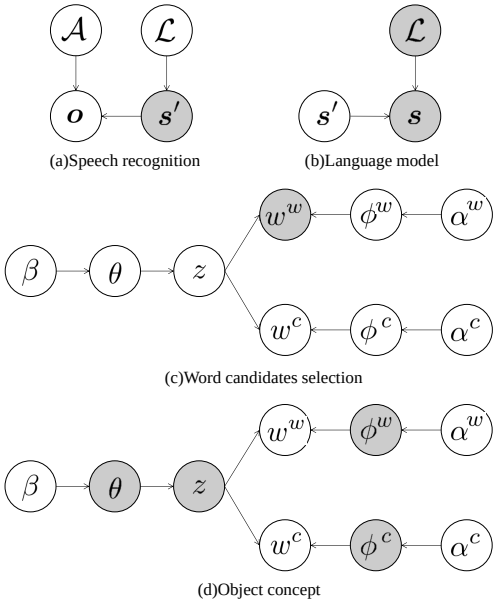

Fig. 2: Approximate graphical model for color concepts and language.

size of the training dataset is denoted by J, and N_c and N_w are the number of occurrences of information of the modalities c and w, respectively. The robot can acquire the concept z by categorizing the perceptual information w^* in an unsupervised manner by estimating the parameters of this model. In addition, s and z are connected by w^w. Therefore, this model integrates speech recognition, word grounding and concept acquisition, and the robot can mutually learn concepts and language models. However, this model is highly complicated and it is difficult to learn all at the same time. In this study, we approximate the model. Fig. 2 illustrates an approximate graphical model obtained by disassembling Fig. 1 into four parts, where the parameters of each are learned in an unsupervised manner.

A. Speech recognition

Fig. 2(a) illustrates a speech recognition model. In Fig. 2(a), $\mathcal{A}$ and $\mathcal{L}$ are known and $\boldsymbol{s}'_{1:N}$ denotes the n-best strings obtained by recognizing all teaching utterances $\boldsymbol{o}$ given for all objects.

$$\boldsymbol{s}'_{1:N} \sim P(\boldsymbol{s}'_{1:N}|\boldsymbol{o}, \mathcal{A}, \mathcal{L}). \tag{1}$$

Julius was used for speech recognition, and the Julius standard acoustic model was adopted.

B. Word segmentation

The language model parameter $\mathcal{L}$ is obtained by maximizing the probability $P(\boldsymbol{o}|\mathcal{A}, \mathcal{L})$ that generates utterances $\boldsymbol{o}$ given for all objects.

$$\begin{aligned} \mathcal{L} &= \underset{\mathcal{L}}{\operatorname{argmax}} P(\boldsymbol{o}|\mathcal{A}, \mathcal{L}) && (2) \\ &= \underset{\mathcal{L}}{\operatorname{argmax}} \int P(\boldsymbol{s}|\mathcal{L}) P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A}) d\boldsymbol{s}. && (3) \end{aligned}$$

The integration over $\boldsymbol{s}$ means to take the sum of all possible combinations of characters, and this is difficult to calculate directly. To deal with this problem, we assume that probabilities $P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A})$, with the exception of some strings are negligibly small. Therefore, we utilize the n-best recognized strings $\boldsymbol{s}'_{1:N}$ of $\boldsymbol{o}$. Using this approximation, the language model is separated from the speech recognition model, as shown in Fig. 2(b). $\mathcal{L}$ is approximately obtained by maximizing the probability that generates word sequences $\boldsymbol{s}_{1:N}$ from the n-best recognized strings $\boldsymbol{s}'_{1:N}$.

$$\mathcal{L}, \boldsymbol{s}_{1:N} = \underset{\mathcal{L}, \boldsymbol{s}_{1:N}}{\operatorname{argmax}} P(\boldsymbol{s}_{1:N}|\boldsymbol{s}'_{1:N}, \mathcal{L}). \tag{4}$$

Here, $\mathcal{L}$ and $\boldsymbol{s}_{1:N}$ can be obtained by utilizing the Nested Pitman-Yor Language Model [12], which is a method of unsupervised word segmentation.

C. Word selection

Word information $\bar{w}^w$ can be generated by considering both the language model and the object concept.

$$\begin{aligned} \bar{w}^w &= \underset{w^w}{\operatorname{argmax}} P(w^w|\boldsymbol{o}, \mathcal{A}, \mathcal{L}, w^c, \alpha^w, \theta, \pi, \beta) && (5) \\ &= \underset{w^w}{\operatorname{argmax}} \int P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A}, \mathcal{L}) \\ &\quad \times P(\boldsymbol{s}|w^w, \mathcal{L}) P(w^w|w^c, \alpha^w, \theta, \pi, \beta) d\boldsymbol{s}, && (6) \end{aligned}$$

where w^c denotes the perceptual information obtained from an object. However, this integration over $\boldsymbol{s}$ is also difficult to calculate directly. To overcome this problem, we assume that with the exception of some strings, probabilities $P(\boldsymbol{o}|\boldsymbol{s}, \mathcal{A}, \mathcal{L})$ are negligibly small, as well as employing word segmentation, and using only the n-best word sequences $\boldsymbol{s}_{1:N}$. Therefore, the word selection model is separated from the word segmentation model, as shown in Fig. 2(c). Furthermore, instead of using Eq. (6), words $\bar{w}^w$ are selected from the n-best word sequences $\boldsymbol{s}_{1:N}$ based on maximizing the probability that the n-best word sequences $\boldsymbol{s}_{1:N}$ are generated from the category z.

III. Simulation Experiment on Mutual Learning of Concept and Language Models

In this study, we verify the influence of IDS and ADS on the learning of the model. In particular, for IDS, we used words employing a repeated pattern of sounds, such as "guruguru" and "tonton", which are often used in Japanese IDS.

A. Experimental Setting

First, 160 unicolor images were generated randomly, and teaching utterances to represent the features of each image were recorded. Therefore, the dataset consisted of 160 pairs of unicolor images and corresponding teaching utterances. Next, 100 pairs of data were used for training, and 60 pairs were used for recognition. Learning was performed by varying the amount of training data used from one to 100. At each step, all recognition data was recognized using the learned model. In the case of ADS, color words such as "Aka" (red) and "Ao" (blue) were used, and speech utterances such as

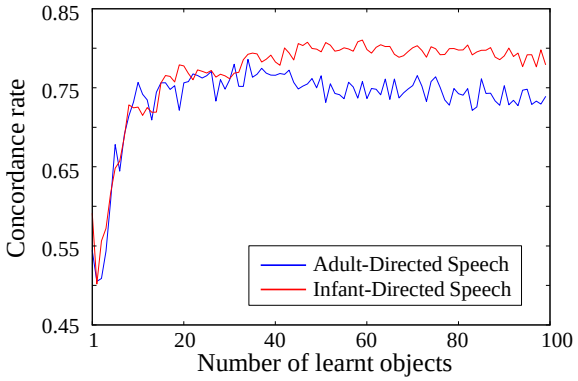

Fig. 3: Transition of speech recognition accuracy of teaching utterances including IDS and ADS.

"Kore wa Aka dayo" (this is red.) were conveyed. On the other hand, in the case of IDS, each color word was replaced by words using a repeated pattern of sounds to simulate IDS. For example, "Aka" (red) was replaced by "guruguru," and teaching utterances such as "Kore wa guruguru dayo" (this is guruguru.) were conveyed. Therefore, the difference in teaching utterances between ADS and IDS were only in the word that represents the color feature, and all other conditions were kept the same. This experiment allowed the exploration of the influence IDS has on learning.

B. Speech Recognition Accuracy

Fig. 3 illustrates the speech recognition accuracy of the speeches uttered using IDS and ADS, as the amount of training data was increased. In Fig. 3, the result of learning using ADS is drawn with a blue line, and the result of learning that used IDS is drawn with a red line. One can see that the speech recognition of ADS is more accurate than that of IDS in the early stage of learning. However, in the later stages of learning, the speech recognition of IDS is more accurate than that of ADS. This fact indicates that IDS has a positive influence on the mutual learning of concepts and language. It has been suggested that IDS has a positive influence on infants' language learning, and these experiment confirms that. We believe that the effects of IDS can be explained computationally using the proposed model.

IV. CONCLUSION

V. CONCLUSION

In this study, we conducted a concept learning simulation using IDS and ADS to verify the effect of teaching utterances on the mutual learning of concepts and language. The experimental results showed that the speech recognition of ADS was more accurate than that of IDS in the early learning stage. However, as learning progressed, the speech recognition of IDS became more accurate than that of ADS. In the future, we intend to investigate the influence of each teaching utterance on concept learning in more further detail.

ACKNOWLEDGMENTS

This research was supported by JST CREST and a Grant-in-Aid for Scientific Research JP16H02835 funded by Japan Society for the Promotion of Science.

REFERENCES

[1] Ferguson. C., and C. Snow., Talking to children: Language input and acquisition, Cambridge University Press, London, 1977.
[2] Martin. A., Igarashi. Y., Jincho. N., and Mazuka. R., Utterances in infant-directed speech are shorter, not slower, Cognition 156, pp. 52-59, 2016.
[3] Mazuka. R., Igarashi. Y., Martin. A., and Utsugi. A., Infant-directed speech as a window into the dynamic nature of phonology, Laboratory Phonology; 6(3-4), pp. 281-303, 2015.
[4] Okumura. Y., Kobayashi. T., and Oshima-Takane. Y., Child Language Development: The Differences between Japanese and English, NTT Technical Review, vol.14, no.11, 1-7, Nov. 2016.
[5] Murase. T., and Kobayashi. T. The use of multiple labels in Japanese mothers' speech to toddlers: Baby talk and adult speech, Proceedings of the 15th European Conference on Developmental Psychology, pp. 513-516, 2012.
[6] Taniguchi. T., Nagai. T., Nakamura. T., Iwahashi. N., Ogata. T., and Asoh. H., Symbol Emergence in Robotics: A Survey, Advanced Robotics, Vol. 30, pp. 706-728, 2016.
[7] Taniguchi. T., Symbol Emergence in Robotics for Long-Term Human-Robot Collaboration, IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, 2016.
[8] Nakamura. T., Nagai. T., Funakoshi. K., Nagasaka. S., Taniguchi. T. and Iwahashi. N., Mutual Learning of an Object Concept and Language Model Based on MLDA and NPYLM, in IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 600-607, 2014.
[9] Blei. D. M., Ng. A. Y., and Jordan. M. I. Latent dirichlet allocation, Journal of machine Learning research, Vol. 3, pp. 993-1022, 2013.
[10] Nakamura. T., Araki. T., Nagai. T. and Iwahashi. N., Grounding of Word Meanings in LDA-Based Multimodal Concepts, Advanced Robotics, Vol. 25, pp. 2189-2206, 2012.
[11] Araki. T., Nakamura. T., Nagai. T., Nagasaka. S., Taniguchi. T., and Iwahashi. N., Online Learning of Concepts and Words Using Multimodal LDA and Hierarchical Pitman-Yor Language Model, in IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.1623-1630, 2012.
[12] Mochihashi. D., Yamada, T., and Ueda. N., Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Vol. 1, pp. 100-108, 2009.