Selective integration based on subjective consistency facilitates simultaneous development of vocal imitation and lexicon acquisition

Yuki Sasamoto **Yuichiro Yoshikawa** **Minoru Asada**

Asada Synergistic Intelligence Project, ERATO, JST
Graduate School of Eng., Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
yuki.sasamoto@ams.eng.osaka-u.ac.jp, yoshikawa@jeap.org, asada@jeap.org

1. Introduction

In human-language communication, vocalization is one of the most efficient channels because humans can share a large lexicon within a short duration of time. Human infants start to understand caregiver's words from eight months of age and produce their first word by the end of their first year. Meanwhile, they exhibit mimicry of adults' single vowels by eight months of age as well as that of adults' strings of consecutive vowels by 14 months of age(Jones, 2007). Therefore, the development processes of lexicon acquisition and vocal imitation seem to overlap each other. Furthermore, we conjecture that these processes might facilitate each other. For example, the ability of vocal imitation could help infants to vocalize unheard words, and the knowledge of a lexicon and its correspondence to objects could help them to imitate partially inaudible words. What kinds of mechanisms underlie the developmental processes of such complementary abilities?

Synthetic studies have attracted wide attention as one of the most promising approaches to resolving such questions of developmental mechanisms (Asada et al., 2009). In previous work, the development of lexical acquisition (Roy and Pentland, 2002, Yoshikawa et al., 2008) and that of vocal imitation (Kanda et al., 2007, Miura et al., 2007) have been modeled as learning processes. However, such studies on theses abilities have generally been conducted separately, and thus their interaction has remained unexplored.

In this paper, we propose a method for simultaneous development of vocal imitation and lexicon acquisition through mutually associative cross-modal mapping using "subjective consistency." Subjective consistency of a signal from each pathway in the mapping is calculated by its proximity to those from others and used as a contribution rate in integrating signals. The integrated vector is used as a learning signal that is expected to ignore errors of the caregiver along with the learning progress of other pathways.

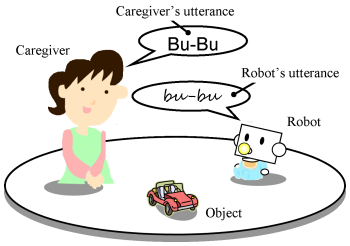

Figure 1: Assumed environment of caregiver-robot interaction

2. Assumptions

A robot and a caregiver take turns in an environment that includes objects (Fig.1). In each step, it looks at either the caregiver or any of the objects and decides whether to utter. Then, the caregiver selects either of three types of behaviors: replying, showing, and describing. The behavior of the caregiver is successful based on the pre-determined probability of each type (p_R, p_S, p_D).

Through such interactions, it learns connection-weight matrixes between nodes in two different layers, namely those between one's own phonemes and the caregiver's phonemes $\boldsymbol{W}^{uv}$ (imitation mapping), those between the caregiver's phonemes and objects $\boldsymbol{W}^{vo}$ (word-listening mapping), and those between objects and one's own phonemes, $\boldsymbol{W}^{ou}$ (word-producing mapping) (Fig.2).

3. Selective combination based on subjective consistency

We propose a method of selective combination to create a reliable learning signal based on subjective consistency. Let $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ be external input vectors to the i-th and j-th layers and $\boldsymbol{x}^{ij}$ be a direct prediction vector of $\boldsymbol{x}^j$ from $\boldsymbol{x}^i$ by the mapping with $\boldsymbol{W}^{ij}$. Furthermore, suppose that there

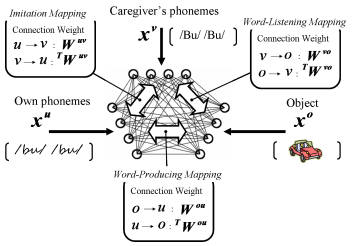

Figure 2: Mutually associative cross-modal mapping

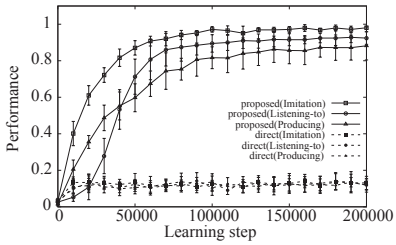

Figure 3: Transition of performance of each mapping

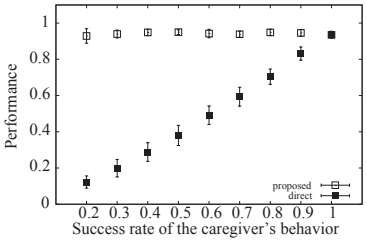

Figure 4: Final performance with respect to the success rate of a caregiver's behaviors

is another layer labeled by k that receives a vector from the i-th layer and outputs an indirect prediction vector $\boldsymbol{x}^{kj}$ to the j-th layer. Three vectors, $\boldsymbol{x}^j$, $\boldsymbol{x}^{ij}$, and $\boldsymbol{x}^{kj}$, are regarded as potentially having information for learning $\boldsymbol{W}^{ij}$. A integrated vector $\hat{\boldsymbol{x}}^j$ is calculated as $\hat{\boldsymbol{x}}^j = f(\boldsymbol{x}^j, \boldsymbol{x}^{ij}, \boldsymbol{x}^{kj}) = \lambda_j \boldsymbol{x}^j + \lambda_{ij} \boldsymbol{x}^{ij} + \lambda_{kj} \boldsymbol{x}^{kj}$, where λ_n $(n = j, ij, kj)$ represents a subjective consistency of each vector. Here, each vector's subjective consistency indicates how close it is to other vectors, and it is calculated by $\lambda_n = \exp\left(-\prod_{l,l\notin n} \left\|\boldsymbol{x}^n - \boldsymbol{x}^l\right\|/\sigma^2\right) / \sum_{m,m\in\{i,ij,kj\}} \exp\left(-\prod_{o,o\notin m} \left\|\boldsymbol{x}^m - \boldsymbol{x}^o\right\|/\sigma^2\right)$, where σ is the parameter of sensitivity for consistencies. The integrated vector is used as a learning signal. It is expected not only to basically bias the learning of $\boldsymbol{W}^{ij}$ to predict the current signal $\boldsymbol{x}^j$ from $\boldsymbol{x}^i$ but also to ignore $\boldsymbol{x}^j$ when it seems to involve errors of the caregiver along with the learning progress of the other pathways ($\boldsymbol{W}^{ij}$ and $\boldsymbol{W}^{kj}$).

4. Simulation

We conducted a series of computer simulations to show the validity of the proposed method for mutually associative cross-modal mappings. We assumed that the number of objects was 39 and the number of phonemes was 37. The parameter *sigma* was empirically set to 1.0 for good performance. We compared the learning performances of the proposed method (hereinafter *proposed*) to those of another method without integration based on subjective consistency for updating the connection matrix (hereinafter *direct*).

We ran 10 sets of simulation with 200,000-step interaction for different sets of parameters p_R, p_S and p_D. These parameters were set to be equal with each other and varied from 0.2 to 1.0. Figure 3 shows the transitions of the average performance of each mapping over different sets of simulation, where $p_R = p_S = p_D = 0.2$. Figure 4 shows the final performances of the entire learning process with respect to the success rate of the caregiver's behaviors. This is calculated from the average performances among all three mappings. We can see that the performance of both methods is high in the case of a high success rate of the caregiver's behaviors. However, the performances of *direct* (filled symbol) becomes worse along with the decrease in the success rate, while that of *proposed* (blank symbol) remains high against the decrease of the success rate.

5. Conclusion

In this paper, we proposed a method to combine several sources of a learning signal for mutually associative cross-modal mappings, which is formed by an external input and internal outputs from possible streams of mapping within it. The subjective consistency of each signal, which evaluates how close it is to other signals, is used to weight it to calculate the combined signal. The proposed method makes it possible to successfully ignore the external input in the case where the caregiver fails to give examples of correct mapping, which is presumed to be typical in real caregiver-infant interaction.

References

Asada, M., Hosoda, K., Kuniyoshi, Y., Ishiguro, H., Inui, T., Yoshikawa, Y., Ogino, M., and Yoshida, C. (2009). Cognitive developmental robotics: a survey. *IEEE Trans. on Autonomous Mental Dev.*, 1(1):12–34.

Bates, E., Dale, P. S., and Thal, D. (1995). *The Handbook of Child Language*, chapter 4: Individual Differences and their Implications for Theories of Language Development, pages 96–151. Blackwell Publishing.

Jones, S. S. (2007). Imitation in infancy the development of mimicry. *Psycho. Sci.*, 18(7):593–599.

Kanda, H., Ogata, T., Komatani, K., and Okuno, H. G. (2007). Vocal imitation using physical vocal tract model. In *Proc. of the 2007 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 1846–1851.

Miura, K., Yoshikawa, Y., and Asada, M. (2007). Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories. *Advanced Robotics*, 21:1583–1600.

Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113–146.

Yoshikawa, Y., Nakano, T., Ishiguro, H., and Asada, M. (2008). Multimodal joint attention through cross facilitative learning based on μx principle. In *Proceedings of the 7th International Conference on Development and Learning*, pages 226–231.