Infant-caregiver interactions affect the early development of vocalization

Minoru Asada[1] and Nobutsuna Endo[1]

***Abstract*— Vocal communication is a unique means to bilaterally exchange messages in real-time. The developmental origin of such communication is the vocal interactions between an infant and a caregiver, and one of the big mysteries is how the infant learns to vocalize the mother tongue of the caregiver. Many theories claim to explain an infant's capability to imitate a caregiver based on acoustic matching. However, the acoustic qualities of the infant and the caregiver are quite different, and, therefore, cannot fully explain the imitation. Instead, the interaction itself may have an important role, but the mechanism is still unclear. In this article, we review studies addressing this problem using constructive approaches based on cognitive developmental robotics.**

I. INTRODUCTION

A unique communication capability of the human species is language because it provides a powerful economy of reference for objects, events, and relationships, which other species cannot achieve by alternate means of communication. Therefore, from an evolutionary perspective, it is still a big mystery as to how human beings acquired language [1]. Further, how infants and children learn to use language needs to be understood from a developmental perspective. In this article, we focus on vocal interactions between an infant and a caregiver, and how the infant learns to vocalize the mother tongue of the caregiver.

Computational modeling has been used to explain the developmental process of speech perception and articulation. Many theories claim to explain an infant's capability to imitate a caregiver based on acoustic matching. However, the acoustic features of the infant and the caregiver are quite different, and, therefore, cannot adequately explain imitation. Instead, the interactions themselves have an important role [2], [3], [4], [5], [6].

Fig. 1. Core ideas of CDR: physical embodiment and social interaction

*This research was supported by Grants-in-Aid for Scientific Research (Research Project Number: 24000012).

[1]Minoru Asada and Nobutsuna Endo are with Dept. of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Yamadaoka 2-1, Suita, Osaka, 565-0871, Japan {asada, endo}@ams.eng.osaka-u.ac.jp

These studies are categorized as constructive approaches based on cognitive developmental robotics (hereafter, CDR) [7], [8]. The central paradigm of the constructivist approaches [9] is to obtain a new understanding through cycles of hypothesis and verification, targeting the issues that are very difficult or almost impossible to solve under existing scientific paradigms. The core idea of CDR is "physical embodiment," and more importantly, "social interaction" that enables information structuring through interactions with other agents. Cognitive development is thought to connect both seamlessly [10], [11] (Fig. 1).

Physical embodiment of the infant-like vocal system is designed with capability of sensorimotor mapping [2], [3], and social interaction with a caregiver has an important role not only in teaching signals to the infant (robot), but also in the entrainment mechanism of the infant's vocalization into that of the caregiver's [4], [5], [6]. These studies reveal the importance of a caregiver's responses to an infant's (robot's) actions, and the mutual feedback between them enables the infant (robot) to vocalize the mother tongue of the caregiver.

The rest of this article is organized as follows. First, the early development of infant speech perception and articulation from observational studies in developmental psychology is briefly reviewed. Next, computational modeling approaches are briefly summarized. Then, the constructive approaches with experiments using real robots and computer simulations are discussed from an issue of how infant-caregiver interactions affect the early development of vocalization.

II. BEHAVIORAL STUDIES ON EARLY DEVELOPMENT OF SPEECH PERCEPTION AND ARTICULATION

In general, an infant's ability to listen to adult voices appears in a language-independent manner from birth and gradually adapts to their mother tongue [12]. Kuhl et al. [13] reported that infants younger than six-month-old can discriminate vowels in any language, but gradually their perception is tuned to their mother tongue, and, therefore, they appear to lose the perceptual capability before they are six-month-old.

In terms of developments that eventually lead to speech production, an infant's utterances are initially quasi-vocalic sounds that resemble vowels and are gradually adapted to his/her caregiver's sounds [14].

In developmental psychology, it was claimed that the infant-caregiver interaction plays an important role in vowel acquisition [15]. From the first month after birth, a mother's speech to her infant differs from that of normal adult

speech, i.e., high in pitch, with many variations that are more pronounced than those of normal speech. It is called "motherese," "infant-directed speech (IDS)," or "baby talks." Liu et al. [16] found that clarity of maternal speech directly affects an infant's early language learning based on the measurement of speech discrimination in infants (6-8 and 10-12-month-olds).

III. MODELING APPROACHES TO VOCAL COMMUNICATION

Computational modeling for vocal communication could be classified into three categories as shown in Fig. 2:

1) No interaction: Motor control ability develops through self-monitoring of vocalizations [17], [18], [19].
2) Caregiver's scaffolding: Statistical estimation of a caregiver's vowel categories from the caregiver's vocalizations [20], [21], [22], [23], [24], [25], [26].
3) Mutual interaction: Self-organization (one of the most popular unsupervised learning method for clustering data without knowing the class memberships of the input data) of shared vowels through imitative interaction [27], [28].

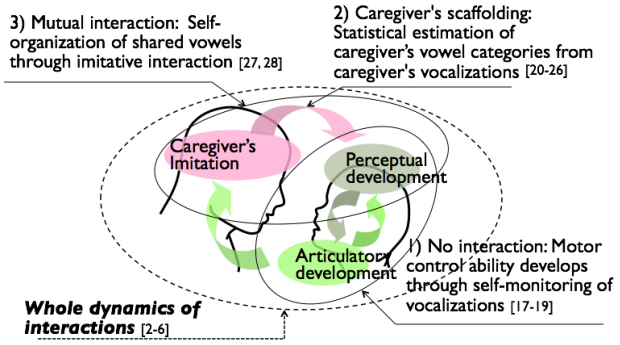

Fig. 2. Modeling approaches for vocal communication (adapted from [29])

There have been many other computational models of speech development, and Rasanen [30] thoroughly reviewed them in addition to the models described above. These models assumed that acoustic matching is an unproblematic mechanism for learning to pronounce speech sounds. Some models also ignored or downplayed the correspondence problem that arises from the different sizes of adult and infant vocal tracts [31] and the inevitable differences in sound qualities that result [6].

IV. MODELING WHOLE DYNAMICS OF VOCAL INTERACTIONS

In order to address the issue of finding correspondence, the whole dynamics of vocal interaction between an infant and a caregiver should be considered, which is indicated as a large broken ellipse in Fig. 2. The key idea is a caregiver's affirmative bias, i.e., the caregiver's anticipations for her infant can bias her perception and imitation as well. Moreover, Rochat [32] claimed that a caregiver's affirmative interpretation and imitation of infant's immature behavior develop infant's social abilities. Here, we introduce several attempts in this category.

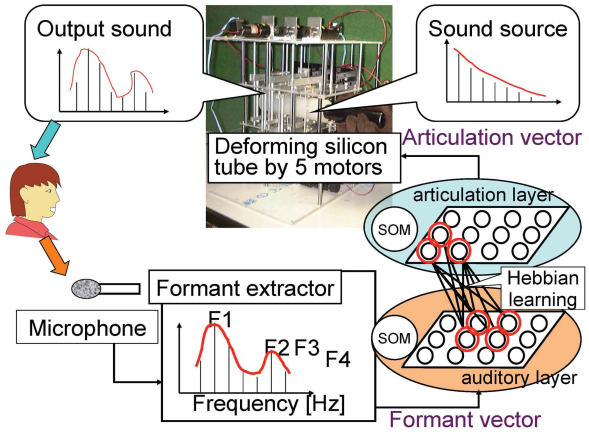

Fig. 3. An overview of the system: "Burpy"

Yoshikawa et al. [2] have addressed this issue in human-robot vocal interaction and demonstrated the importance of being imitated by a human caregiver, whose body is different from that of the robot's, as well as subjective criteria of the robot such as ease of articulation. Fig. 3 shows a vocal robot called "Burpy" which consists of an articulation part and an auditory one with their corresponding layers. Both layers are self-organized and connected by Hebbian learning through parrot-like teaching by a caregiver.

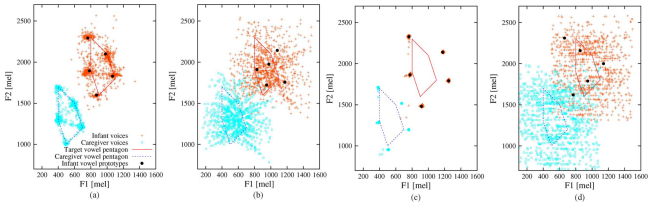

Fig. 4. Different learning results under several conditions. Apexes of red pentagons represent target vowels of infant, in other words, clearest vowels in the infant's vowel region, and black dots represent the infant vowel prototypes after learning. (a) Both biasing elements; (b) only automirroring bias; (c) only sensorimotor magnets; (d) no biasing elements. (adapted from [4])

Ishihara et al. [4] modeled the mechanism of imitation underlying caregiver-infant interaction by focusing on potential roles of the caregiver's imitation in guiding an infant's vowel development. Two kinds of the caregiver's possible biases are used. The first one represents a caregiver's sensorimotor bias such as perceptual magnet effect [33], and the second is based on what we call "automirroring bias," by which the heard vowel is much closer to the expected vowel because of the anticipation of being imitated. The results are shown in Fig. 4 indicating how these two biases worked.

Howard and Messum [6] addressed the issue using Elija, a computational model of an infant. First, through unsupervised active learning, Elija began by discovering motor patterns, which produced sounds. Next, native speakers of English, French and German played the role of Elija's caregiver, and Elija memorized the caregiver's responses

and reacted to the memorized patterns. This interaction was expanded to word teaching.

V. DISCUSSION

According to the developmental process of speech perception and articulation, we discuss the following issues pertaining to the above studies of whole dynamics [2], [3], [4], [5], [6].

A. Self-learning at early stage

Infant-caregiver interactions start from the beginning, observed as motherese or infant-directed speech. However, infants tend to be affected by their caregivers/ mother tongues after 6 or 8 months. This period could be a mixture between self-learning and interaction with the caregiver because it is not plausible that infants are born with speech perception and articulation skills. For the constructive approaches, two styles are observed.

1) Separation of self-learning from interaction: Elija [6] took this type of learning for the convenience of computation and to make clear the roles of different learning schemes. Several properties such as salience/diversity (selected by the caregiver in [5]) and effort ("toil" in [2]) were used.

2) No self-learning process: Miura et al. [5] focused on the process of selection and correspondence of the vowels with initially fixed motor patterns without the self-learning process. Their method is same as Elija's process after the self-learning because the learned (Elija) or initially fixed motor patterns (Miura's) do not change during the interaction process with the caregiver. In contrast, Burpy [2] took an interaction process from the beginning without the self-learning process.

B. Affirmative bias of caregivers

Howard and Messum [6] analyzed the interactions through phonemic transcriptions of the caregivers' utterances and found that the caregivers interpreted Elija's output within the framework of their native languages. Ishihara et al. [4] formalized such caregiver's affirmative biases as "sensorimotor magnets" and "automirroring bias" by which the heard vowel is much closer to the expected vowel because of the anticipation of being imitated (Fig. 4).

Computer simulated results of the caregiver-infant interaction showed the sensorimotor magnets help form small clusters and the automirroring bias shapes these clusters to become clearer vowels in association with the sensorimotor magnets.

C. Strategies of learners

Burpy (Fig. 3) [2] has slightly modified Hebbian learning so that the size of the final cluster could be small by introducing the criterion of "toil" parameter (less deformation and less energy consumption) mentioned above. After the learning process, this association plays the role of a mirror neuron system (MNS) for Burpy to remind its articulation vector when it hears one of the caregiver's vowels.

Ishihara et al. [4] represented the learner's vowel primitives as a Gaussian mixture network (GMN), and its parameters changed during the interactions with a caregiver, which indicates the developmental process of finding the correspondence of vowels between the learner and the caregiver.

In these two studies, the caregivers are assumed to be ideal imitators who always respond to the learner's utterance with their own corresponding vowel. However, in reality, caregiver's imitation is less than 20% in the interaction between mothers and their 7- to 10-month-old infants [34].

Howard and Messum [6] reported that the caregivers' (four English, two German, and two French) responses were almost reformulation (more than 90%) and contained little mimicry (less than 10%). Therefore, Elija has a strategy to memorize the responded patterns from the caregiver, and responds with the most similar pattern. Through many cycles of such feedback, Elija is expected to statistically converge its responses and to consolidate its memory patterns to respond appropriately.

Based on the data by [34], Miura et al. [5] adopted a learning method based on the automirroring bias on the learner's side with a self-evaluation mechanism to find the correspondence with less frequent imitative caregivers. The automirroring bias is the robot's anticipation of being imitated by its caregiver, and has a role of narrowing the candidates for the correspondence.

D. Research platforms

Physical embodiment is one of the core ideas of cognitive developmental robotics [7], and in the case of vocal interaction, it corresponds to two types: physical articulation systems and virtual ones. Both are important parts of all systems mentioned above.

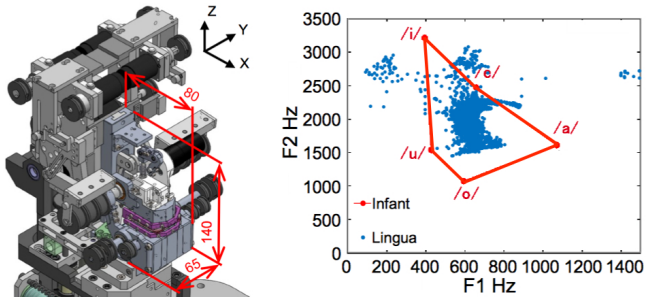

Fig. 5. Lingua [left] and its formant frequencies [right](adapted from [36])

Vocalization is generally well known as an outcome from a modulation of a source of sound energy by a filter function determined by the shape of the vocal tract; this is often referred to as the "source-filter theory of speech production" [35]. Yoshikawa et al. [2] implemented this theory into Burpy by using a vibrator as a sound source and silicon rubber tube as a vocal tract whose shape is deformed by five electric motors. Miura et al. [3] improved Burpy by replacing the sound source with an air compressor and an artificial vocal band, and added a lip at the front end of the vocal tract. The length of the robot's vocal tract changed from 170 [mm] (average vocal tract length of the human male) to 116 [mm].

Endo et al. [36] developed an infant-like vocal robot, "Lingua," as a controllable vocal platform that affords a model of real infant vocalization. The results of the preliminary experiments showed that the robot could vocalize almost the same ranges of fundamental frequencies and vowel-like utterances as an infant. Lingua needs additional improvements and will be used for experiments of interaction with human caregivers (Fig. 5).

Recent progress in articulation simulator in terms of anatomical structure, function, and motor control is striking (e.g., [37]). Elija's motor control system [6] incorporates a Maeda articulatory speech synthesizer [38], [39]. It was supposed that the articulation simulators were not good at real-time response to human subjects. Elija has partially solved this issue by separating the self-learning process off-line, and selecting one of the fixed motor patterns during the real-time interactions.

REFERENCES

[1] Terrence W. Deacon. *The Symbolic Species: The co-evolution of language and the brain*. W. W. Norton & Company, New York, London, 1998.

[2] Yuichiro Yoshikawa, Junpei Koga, Minoru Asada, and Koh Hosoda. A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science*, 15(4):245–258, 2003.

[3] Katsushi Miura, Yuichiro Yoshikawa, and Minoru Asada. Unconscious anchoring in maternal imitation that helps finding the correspondence of caregiver's vowel categories. *Advanced Robotics*, 21:1583–1600, 2007.

[4] H. Ishihara, Y. Yoshikawa, K. Miura, and M. Asada. How caregiver's anticipation shapes infant's vowel through mutual imitation. *IEEE Transactions on Autonomous Mental Development*, 1(4):217–225, 2009.

[5] Katsushi Miura, Yuichiro Yoshikawa, and Minoru Asada. Vowel acquisition based on an auto-mirroring bias with a less imitative caregiver. *Advanced Robotics*, 26:23–44, 2012.

[6] Ian S. Howard and Piers Messum. Learning to pronounce first words in three languages: An investigation of caregiver and infant behavior using a computational model of an infant. *PloS One*, 9(10):e110334: 1–21, 2014.

[7] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: a survey. *IEEE Transactions on Autonomous Mental Development*, 1(1):12–34, 2009.

[8] A. Cangelosi and M. Schlesinger. *Developmental Robotics - From Babies to Robots* -. MIT Press, 2015.

[9] Takashi Hashimoto, Takashi Sato, Masaya Nakatsuka, and Masanori Fujimoto. Evolutionary constructive approach for studying dynamic complex systems. In Giuseppe Petrone and Giuliano Cammarata, editors, *Recent Advances in Modelling and Simulation*, chapter 7. I-Tech Books, 2008.

[10] Minoru Asada. Can cognitive developmental robotics cause a paradigm shift? In Jeffrey L. Krichmar and Hiroaki Wagatsuma, editors, *Neuromorphic and Brain-Based Robots*, pages 251–273. Cambridge University Press, 2011.

[11] Minoru Asada. Towards artificial empathy. *International Journal of Social Robotics*, 7:19–33, 2015.

[12] Janet F. Werker and Richard C. Tees. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 25:121–133, 2002.

[13] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255:606–608, 1992.

[14] P. K. Kuhl and A. N. Meltzoff. Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of Acoustic Society of America*, 100:2415–2438, 1996.

[15] Jennifer A. Schwade Michael H. Goldstein. Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, 19:515–523, 2008.

[16] H.-M. Liu, P. K. Kuhl, and F.-M. Tsao. An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental Science*, 6:F1–F10, 2003.

[17] H. Kanda, T. Ogata, T. Takahashi, K. Komatani, and H. G. Okuno. Continuous vocal imitation with self-organized vowel spaces in recurrent neural network. *Proceedings of IEEE International Conference on Robotics and Automation*, pages 4438–4443, May 2009.

[18] Frank H. Guenther. A neural network model of speech acquisition and motor equivalent speech production running title: Speech acquisition and motor equivalence. *Biological Cybernetics*, 72:43–53, 1994.

[19] Westermann G. and Reck Miranda E. A new model of sensorimotor coupling in the development of speech. *Brain and Language*, 89:393–400, 2004.

[20] B. McMurray, R. N. Aslin, and J. C. Toscano. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12:369–378, 2009.

[21] Gautam K. Vallabha, James L. McClelland, Ferran Pons, Janet F. Werker, and Shigeaki Amano. Unsupervised learning of vowel categories from infant-directed speech. *Proc. of National Academy of Sciences USA*, 104:13273–13278, 2007.

[22] C. Yu, D. Ballard, and R. Aslin. The role of embodied intention in early lexical acquisition. *Cognitive Science*, 2005.

[23] D. Roy and A. Pentland. Learning words from sights and sounds: a computational model. *Cognitive Science*, 26:113–146, 2002.

[24] Yuki Sasamoto, Yuichiro Yoshikawa, and Minoru Asada. Mutually constrained multimodal mapping for simultaneous development: modeling vocal imitation and lexicon acquisition. In *The 9th International Conference on Development and Learning (ICDL'10)*, pages CD–ROM, 2010.

[25] Y. Yoshikawa, T. Nakano, M. Asada, and H. Ishiguro. Multimodal joint attention through cross facilitative learning based on μx principle. In *Proceedings of the 7th IEEE International Conference on Development and Learning*, 2008.

[26] Masaki Ogino, Masaaki Kikuchi, and Minoru Asada. Active lexicon acquisition based on curiosity. In *The 5th International Conference on Development and Learning (ICDL'06)*, 2006.

[27] Pierre-Yves Oudeyer. The self-organization of speech sounds. *Journal of Theoretical Biology*, 233(3):435–449, 2005.

[28] B. de Boer and W. Zuidema. Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, 18:141–154, 2010.

[29] H. Ishihara. http://www.gcoe-cnr.osaka-u.ac.jp/media/handouts/bu04_ishihara.pdf, 2013.

[30] O. Rasanen. Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, 54(9):975–997, 2012.

[31] H. Vorperian and R. Kent. Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, 50:1510–1545, 2007.

[32] Philippe Rochat. *THE INFANT'S WORLD*, chapter 4. Harverd University Press, 2004.

[33] P. K. Kuhl. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50:93–107, 1991.

[34] Julie Gros-Louis, Meredith J. West, Michael H. Goldstein, and Andrew P. King. Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, 30(6):509–516, 2006.

[35] P. Rubin and E. Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter Measuring and modeling speech production. Springer-Verlag, New York, 1998.

[36] Nobutsuna Endo, Tomohiro Kojima, Hisashi Ishihara, Takato Horii, and Minoru Asada. Design and preliminary evaluation of the vocal cords and articulator of an infant-like vocal robot "lingua". In *the IEEE-RAS International Conference on Humanoid Robots*, pages Vol.USB, ThuI2–3.8, 2014.

[37] H. Rasilo, O. Rasanen, and U. K. Laine. Feedback and imitation by a caregiver guides a virtual infant to learn native phonemes and the skill of speech inversion. *Speech Communication*, 55(9):903–931, 2013.

[38] S. Maeda. An articulatory model of the tongue based on a statistical analysis. *Journal of the Acoustical Society of America*, 65:S22, 1979.

[39] S. Maeda. *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulator model*, pages 131–149. Kluwer Academic Publishers, Boston, 1990.