# Modeling Development of Multimodal Emotion Perception Guided by Tactile Dominance and Perceptual Improvement

Takato Horii, Yukie Nagai, *Member, IEEE*, and Minoru Asada, *Fellow, IEEE*

***Abstract*—Humans recognize others' emotional states such as delight, anger, sorrow, and pleasure through their multimodal expressions. However, it is unclear how this capability of emotion perception is acquired during infancy. This paper presents a neural network model that reproduces the developmental process of emotion perception through an infant–caregiver interaction. This network comprises hierarchically structured restricted Boltzmann machines (RBMs) that receive multimodal expressions from a caregiver (visual, audio, and tactile signals in our current experiment) and learn to estimate her/his emotional states. We hypothesize that emotional categories of multimodal stimuli are acquired in a higher layer in the network owing to two important functions: 1) tactile dominance and 2) perceptual improvement. The former refers to that tactile sensors can detect emotional valence of stimuli such as positive, negative, and zero valence more directly than can other sensors due to characteristics of the nerve systems of the skin. This function was implemented as semisupervised learning in the model. The latter refers to developmental changes in the perceptual acuity, which was replicated by refining the variance parameters of the low-layered RBMs. Experimental results demonstrated that tactile dominance and perceptual improvement have the role of facilitating the differentiation of emotional states of multimodal expressions; however, the influences only appear when both functions are included in the model together. Considering our results from the psychological perspective may help to elucidate the neural and social mechanisms of the development of emotion perception.**

***Index Terms*—Computational modeling, development of emotion perception, infant–caregiver interaction, neural networks for development, perceptual development, tactile interaction.**

## I. Introduction

EMOTION perception refers to capabilities of recognizing emotions of others. We, humans, can estimate the other's emotional states such as delight, anger, sorrow, pleasure, and so on from the other's multimodal expressions during interaction [1], [2]. Many researchers studied how humans acquire the capability of emotion perception in infancy (e.g., [3] and [4]) although it is not fully revealed yet. For understanding the developmental changes in emotion perception, several researchers displayed visual, audio, or audiovisual emotional expressions of others to infants and young children (e.g., [5]–[8]). Walker-Andrews [5] argued that human infants have only rudimentary capacities to detect, discriminate, and recognize (or perceive) others' emotional expressions at birth; however, the capabilities rapidly develop during the first year of life. She reviewed many articles about infants' capabilities of emotion perception and suggested that younger infants (around five months old) can detect only primary meaning of emotional expressions of others. On the other hand, older infants can discriminate and recognize the emotional categories of others' expressions (e.g., basic emotions) owing to perceptual development. Grossmann [8] reported that 12-month-old infants showed different reactions to an event-related potential in their brain when they faced angry and happy audiovisual stimuli (i.e., facial and vocal emotional expressions of others). We consider there to be more evidence in support of the developmental process of emotion perception in particular findings that the tactile interaction between infants and their caregivers appear to affect the ability of emotion perception [9]–[11]. However, these studies observed only changes in behaviors when the subjects faced emotional stimuli. Thus, it remains unclear what causes the developmental changes in emotion perception because there are no methods to present the actual emotional states perceived by infants.

Cognitive developmental robotics is a research field aimed at understanding the mechanisms of cognitive development process by synthetic approaches utilizing human-like robots and computer simulations [12]. A number of researchers in robotics have constructed emotional systems and reproduced the abilities and developmental process of emotion (e.g., [6] and [13]–[18]). Blanchard and Cañamero [14] proposed a general perception-action architecture, which takes account of imprinting experiences and reward-based learning methods. Their experimental result of simple human–robot interaction showed that the proposed model acquired affective behavior, driven by parameters of the model (e.g., "distress" caused by a difference between the current and

Manuscript received June 9, 2017; revised November 17, 2017; accepted January 20, 2018. Date of publication February 28, 2018; date of current version September 7, 2018. This work was supported in part by the Grant-in-Aid for JSPS Fellows under Grant 15J00671, in part by the Grant-in-Aid for Specially Promoted Research under Grant 24000012, in part by the Grant-in-Aid for Scientific Research on Innovative Areas under Grant 24119003, and in part by the JST CREST "Cognitive Mirroring" under Grant JPMJCR16E2. *(Corresponding author: Takato Horii.)*

T. Horii and M. Asada are with the Graduate School of Engineering, Osaka University, Osaka 565-0871, Japan (e-mail: takato.horii@ams.eng.osaka-u.ac.jp; asada@ams.eng.osaka-u.ac.jp).

Y. Nagai is with the National Institute of Information and Communications Technology, Osaka 565-0871, Japan.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2809434

imprinted experiences and "comfort" included by reward stimuli). Hiolle *et al.* [16] and Lones *et al.* [18] examined influences of a novel environment and different experiences on the arousal regulation and behavior learning by agents. These studies evaluated how robots' behavior and model parameters (e.g., arousal and comfort) differentiate through the interaction between the robots and environment; however, they did not consider developmental changes in emotion.

To attack this issue with prior studies, we proposed a computational neural network model for reproducing developmental changes in emotion perception in infancy based on infant–caregiver interactions. This model relies on two key ideas based on psychological studies [5], [11]. The first idea is tactile dominance: that is, the fact that the sense of touch can detect the emotional valence of stimuli more directly than other modalities can. The second idea is perceptual development of infants' multimodal sensation (i.e., sense of vision, audio, and touch). Our previous approach [19] only focused on tactile dominance in order to modeling development of multimodal emotion perception. The model was able to differentiate positive and negative emotion; however, it showed unclear differentiation of negative emotions. We expected that the perceptual development might induce detailed differentiation of emotional categories. In Section II, we explain more details of our ideas and hypothesis for representing the development of emotion perception. The proposed model is composed of restricted Boltzmann machines (RBMs), which belong to stochastic neural networks. The first idea, tactile dominance, was modeled as a semisupervised module of the proposed network, while the second idea, perceptual development, was modeled via a learning process of the variance parameters in the distribution of input modules (see Section III for more details). We compared experimental results under four different conditions (i.e., with or without tactile dominance and with or without perceptual development) by using virtual infant–caregiver interactions. Finally, we discuss our model and brain regions related to the module of our model; the validity of tactile dominance as a contributor to the development of emotion perception, which will be discussed in relation to congenital insensitivity to tactile sensation; and the future scope of this line of research in Section V.

## II. Our Hypotheses

### A. Tactile Dominance

Tactile sensation is exceedingly important for infants, and younger infants tend to use touch to interact with their own body and external environment, owing to their undeveloped vision [11], [20]. Interestingly, caregivers also employ the tactile modality more often than others (e.g., vision and audio) when interacting with infants [21]. Touch is also an important modality for emotional communication from a neuroscience perspective. When we suffer pain from tactile stimuli, for instance, two types of nerves in our skin are activated [22], [23]: an A$\delta$-fiber, which is a myelinated fiber, and a C-fiber, which is an unmyelinated fiber. C-fibers are considered more primitive nerves than A$\delta$-fibers. Björnsdotter *et al.* [24] examined the different anatomical mechanisms of C-fibers that transmit positive emotional valence due to the touch. Such C-fibers are called "C tactile (CT) afferents." CT afferents are distributed over hairy skin and respond to gentle contact (e.g., stroking the surface of the skin) at a velocity range of 1–10 cm/s. They concluded that CT afferents help humans experience positive emotions and enhance the social aspects of human–human interaction through skin contact. Importantly, C-fibers, besides detecting the emotional valence of touch, deliver the information not only to the somatosensory area of the brain but also to the limbic system (e.g., the insular cortex and thalamus) that is known as an emotional brain region. However, these communicative functions of touch have been neglected in modeling studies of emotion.

We hypothesize that tactile communication allows infants to perceive the emotional valence, a value of emotional stimuli [e.g., positive, negative, and nonvalued (zero value) information] from others' multimodal expressions. For instance, touching the skin of infants softly might elicit a positive emotional valence, whereas more forceful contact or pinching might elicit a negative emotional valence during the interaction. Consequently, the emotional valence from the sense of touch might aid to perceive emotional categories of other sensory signals during infancy.

### B. Perceptual Development

Humans' sensory organs develop during the fetal period [25], whereas their perceptive faculties develop after birth over the course of one year. For instance, infants' visual acuity increases from birth to six months of age [26]. Furthermore, for auditory perception, infants' ability to discriminate frequencies improves from 3, 6, to 12 months after birth [27].

Walker-Andrews [5] studied the development of emotion perception from the viewpoint of the influence of perceptual development during the first year of life in infancy. By reviewing many studies that considered the development of emotion perception and perceptual development in the sense of vision and audio, she claimed that younger infants (around five months old) can notice primal information of the emotional meaning of others' expressions (e.g., positive or negative). She also claimed that older infants (around one year old) are able to discriminate and then recognize the emotional states (e.g., basic emotions) of others from their expressions because perceptual development differentiates the affective information from the others multimodal expressions.

Relatedly, researchers have explored the influence of perceptual development on category generalization [28], [29]. They compared the tendencies in classifying new objects between children at different ages. The results showed that younger children (around five years old) classified new objects based on holistic similarity, whereas older children (around ten years old) used dimensional similarity (e.g., the size or color of objects) for such classification. They claimed that these differences in the tendency for object classification depended on developmental changes in perceptual resolution.

In line with these past findings, we hypothesize that perceptual improvement in multiple modalities also affects the

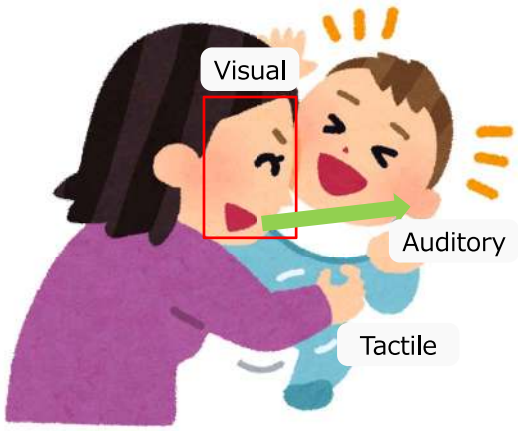


Fig. 1. Example of a face-to-face infant–caregiver interaction. The infant perceives emotional signals from the caregiver using three kinds of modalities: visual, auditory, and tactile. The caregiver's expressions are consistent among the three modalities and induce the same emotional state in the infant as in the caregiver.

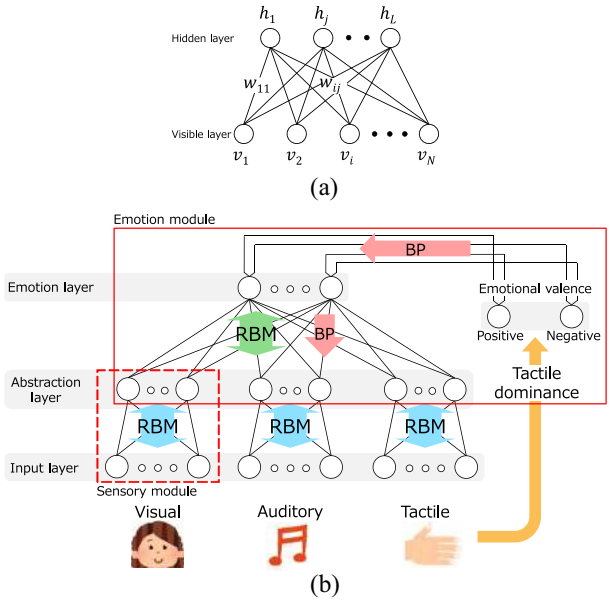


Fig. 2. Computational model for the development of emotion perception based on tactile dominance and perceptual improvement in infancy. $v_i$ and $h_j$ are the activations of the $i$th visible layer and the $j$th hidden layer, respectively, and $w_{ij}$ is connection weight of their relative weight in (a). The three lower RBMs denoted by blue connections (i.e., the region enclosed in broken lines) in (b), constitute sensory modules, which process different modality signals independently. The region enclosed within the red solid line in (b) denotes the emotion module. The red arrows with BP character indicate that the connection weights were modulated by back propagation in the model training (see Section III-C).

development of emotion perception. Specifically, changes in infant's perceptual ability might induce gradual differentiation of emotion perception from ambiguous emotion (e.g., emotional valences) to categorical emotions (e.g., basic emotions [1]).

## III. Computational Model for Development of Emotion Perception

In this section, we first introduce our assumptions for modeling the development of emotion perception in infancy. Then, we describe the basic idea, the proposed model and its computational architecture with the learning process. Finally, the dataset of multimodal emotional expressions simulating infant–caregiver interactions is explained.

### A. Assumptions of the Interaction

We focus on face-to-face multimodal interactions between an infant and a caregiver. Fig. 1 illustrates such an interaction, where the infant perceives stimuli from three kinds of modalities: 1) vision (focusing on the caregiver's face); 2) audition (receiving the caregiver's voice); and 3) touch (detecting the caregiver's touch). For instance, when the caregiver tries to make her/his infant happy, the infant will see the smiling face of the caregiver, hear her happy voice, and feel her gentle touch. Here, we assume that multimodal expressions of the caregiver are consistent among the three modalities. Furthermore, for the sake of simplicity, we suppose that the infant is receiving interaction signals only from the caregiver.

### B. Proposed Model

Fig. 2(a) shows the structure of each RBM, which is a component of the proposed model, and Fig. 2(b) provides an overview of the proposed model. The model comprises two types of modules: sensory and emotion [see Fig. 2(b)]. There are three sensory modules relating to three different sensory modalities (i.e., visual, auditory, and tactile), and these modules process low-level sensory signals observed in infant–caregiver interactions. The emotion module acquires the representation of emotional states by integrating multimodal signals observed in the sensory modules. Our challenge is to propose a biologically and neurologically plausible mechanism for the development of emotion perception. To address this issue, we constructed each module by adopting stochastic neural networks called RBMs [30], [31]. The reason why we employed the RBM is that its learning mechanism corresponds to a well-known theory of the human brain mechanism called free-energy principle [32]. The free-energy principle proposed by Friston hypothesized that the fundamental mechanism of the human brain is the reduction of prediction error by the free-energy minimization. To propose the computational model based on the theory might help us to comprehend the mechanism of the development of emotion perception.

First, we introduce two types of RBM below: the conventional RBM, which deals with binary signals (i.e., Bernoulli–Bernoulli RBM), and another type dealing with continuous signals (i.e., Gaussian–Bernoulli RBM). These two types of RBMs were adopted for the emotion and sensory modules, respectively.

*1) Restricted Boltzmann Machines:* An RBM [30], [31] is a kind of artificial neural network, and consists of two layers [Fig. 2(a)]. One is a visible layer, which receives input signals, and the other is a hidden layer, which represents the latent signals of the input signals. The RBM is able to extract various features from input signals by acquiring latent signals, which can then be used to reconstruct the input signals in an unsupervised manner [33]–[36]. We used RBMs for our model

because this characteristic allows the model to represent emotional states by integrating and abstracting multimodal sensory signals from the infant–caregiver interaction.

Fig. 2(a) illustrates the structure of the RBM. $v_i$ is the activation of the $i$th unit in the visible layer, which receives input signals, while $h_j$ is the activation of the $j$th unit in the hidden layer, which represents the latent signals of the input. Each unit has connections to all other units except for those in the same layer, and all connection weights $w_{ij}$ are symmetrical (i.e., $w_{ij} = w_{ji}$). A Bernoulli–Bernoulli RBM handles only binary signals for both visible (i.e., $v_i \in \{0, 1\}$) and hidden units (i.e., $h_j \in \{0, 1\}$). The activation probabilities for these units are given by

$$p(h_j = 1|\boldsymbol{v}) = g\left(b_j + \sum_i v_i w_{ij}\right) \tag{1}$$

$$p(v_i = 1|\boldsymbol{h}) = g\left(a_i + \sum_j h_j w_{ij}\right) \tag{2}$$

where $g(x)$ is a logistic sigmoid function $1/(1+\exp(-x))$, and $a_i$ and $b_j$ are biases corresponding to the $i$th visible and $j$th hidden units, respectively. Thus, the RBM not only can extract features from input signals as hidden activations [as per (1)] but also can reconstruct the input signals from the extracted features [as per (2)].

The parameters $\theta = \{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{w}\}$ of the RBM are trained through the minimization of reconstruction error between actual input signals and reconstructed input signals calculated in (1) and (2), respectively. This process is replaced by the minimization of cross entropy, denoted by $L$, between two probability distributions: $p(\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \theta)$

$$L = -\sum_x p(\boldsymbol{v}) \log p(\boldsymbol{v}|\boldsymbol{h}, \theta). \tag{3}$$

Cross entropy corresponds to the distance between two probability distributions, where $p(\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \theta)$ are the distributions of actual and reconstructed input signals, respectively. In fact, the minimization of cross entropy in the RBM corresponds to the minimization of the free-energy (i.e., this objective function relates to free-energy principle [32]).

In order to derive update rules for the model parameters, we differentiate cross entropy using the traditional gradient-based method. The update rules for the parameters are given by

$$\Delta w_{ij} = \epsilon_w\left(\langle v_i h_j\rangle_{\text{data}} - \langle v_i h_j\rangle_{\text{recon}}\right) \tag{4}$$

$$\Delta a_i = \epsilon_a(\langle v_i\rangle_{\text{data}} - \langle v_i\rangle_{\text{recon}}) \tag{5}$$

$$\Delta b_j = \epsilon_b\left(\langle h_j\rangle_{\text{data}} - \langle h_j\rangle_{\text{recon}}\right) \tag{6}$$

where the angle brackets $\langle\cdot\rangle_{\text{data}}$ and $\langle\cdot\rangle_{\text{recon}}$ denote the expectations under the distributions of actual and reconstructed input signals, respectively, and $\epsilon_w$, $\epsilon_a$, and $\epsilon_b$ are the learning rates for the corresponding model parameters. We update the parameters by adding their values to subsequent ones in the training (i.e., $w_{ij}^{t+1} = w_{ij}^t + \Delta w_{ij}^t$; here, $t$ is a learning step). For a detailed account of the learning process of RBMs, see [31].

In order to represent the continuous values of sensory signals, we replaced the binary visible units of the RBMs with Gaussian units [37]. The activation probabilities for the visible and the hidden units of this Gaussian–Bernoulli RBM are given as

$$p(h_j = 1|\boldsymbol{v}) = g\left(b_j + \sum_i \frac{1}{\sigma_i^2} v_i w_{ij}\right) \tag{7}$$

$$p(v_i = v|\boldsymbol{h}) = \mathcal{N}\left(v|a_i + \sum_j h_j w_{ij}, \sigma_i^2\right) \tag{8}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes the probability of a normal distribution with a mean $\mu$ and a variance $\sigma^2$, and $\sigma_i$ is the standard deviation associated with the $i$th Gaussian visible unit. The probability function of the hidden units is different from that of (1), because of inclusion of the variance of the visible units. More specifically, visible activation with a small variance is more weighted toward the hidden activations than that with a large variance in (7).

Furthermore, the update rules must be modified because the data probabilities, $p(\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \theta)$, of the Gaussian–Bernoulli RBM are different from those of the Bernoulli–Bernoulli RBM. The rules for the Gaussian–Bernoulli RBM are given by

$$\Delta w_{ij} = \epsilon_w\left(\left\langle \frac{1}{\sigma_i^2} v_i h_j\right\rangle_{\text{data}} - \left\langle \frac{1}{\sigma_i^2} v_i h_j\right\rangle_{\text{recon}}\right) \tag{9}$$

$$\Delta a_i = \epsilon_a\left(\left\langle \frac{1}{\sigma_i^2} v_i\right\rangle_{\text{data}} - \left\langle \frac{1}{\sigma_i^2} v_i\right\rangle_{\text{recon}}\right) \tag{10}$$

$$\Delta b_j = \epsilon_b\left(\langle h_j\rangle_{\text{data}} - \langle h_j\rangle_{\text{recon}}\right). \tag{11}$$

In addition to model parameters $w_{ij}$, $a_i$, and $b_j$, we need to update another parameter $\sigma_i$ to minimize the cross entropy between $p(\boldsymbol{v})$ and $p(\boldsymbol{v}|\boldsymbol{h}, \theta)$, because these probabilities are modulated by the variance. $\sigma_i$ is updated via a surrogate parameter $z_i$, which is defined as $z_i = \log \sigma_i^2$ because the variance $\sigma_i^2$ takes only positive values. The update rule for $z_i$ is given by

$$\Delta z_i = \epsilon_z e^{-z_i}\left(\left\langle \frac{1}{2}(v_i - a_i)^2 - \sum_j v_i h_j w_{ij}\right\rangle_{\text{data}} - \left\langle \frac{1}{2}(v_i - a_i)^2 - \sum_j v_i h_j w_{ij}\right\rangle_{\text{recon}}\right). \tag{12}$$

Through this modulation, each variance is expected to be closer to the actual variance of the input signals. As the variance becomes closer to 0.0, the more strongly the input signals contribute to the probabilities of the hidden activations. Similarly, the noise in the reconstructed signals decreases with the variance.

*2) Sensory Module:* Each sensory module comprises a Gaussian–Bernoulli RBM because the input signals from the sensors are continuous values. The visible layers receive sensory signals from the corresponding sensors (i.e., visual, auditory, and tactile). Each module processes these signals independently.

Perceptual development, which is one of the factors we proposed to drive the development of emotion perception, was

modeled as modulations in the variance of the visible units. More specifically, the variance $\sigma_i^2$ is refined through the corresponding parameters $z_i$ [see (12)] in order to reduce the error between the actual input signals and the reconstructed input signals from the hidden activations. Early in the model training, the variance of the visible units is large, which causes several Gaussian distributions to cover many input signals [Fig. 3(a)]. Therefore, the hidden layer initially represents rough clusters of input data, which makes the reconstruction signals coarse. Later on in the training, however, the variance is refined, and the region covered by the Gaussian distributions decrease. Such a smaller variance leads to more precise reconstructions compared to when the variance is large [Fig. 3(c)]. In parallel with this, the mean values of the Gaussian distribution should also be updated to become closer to the actual value of the input signals to improve the accuracy. To sum up, the refinement of both variance and mean value of Gaussian distributions reproduces perceptual development, just as in [38]. In Section IV-E, we illustrate the developmental changes in the variance and reconstructed signals in the visual sensory module as an example.

Following the training, the hidden activations of the sensory modules represent abstracted features of the corresponding sensory signals. These activations are then used as input signals for the RBM in the emotion module.

*3) Emotion Module:* The Bernoulli–Bernoulli RBM used in the emotion module is called the multimodal RBM in the proposed model, given that it uses the combined hidden activations of the three sensory modules as input signals. The hidden neurons of the multimodal RBM are connected to the emotion valence layer. The emotion valence layer was introduced into the model to represent tactile dominance. In Section II-A, we mentioned that human skin is equipped with specific nerve fibers (C-fibers) that can detect the emotional valence of touch. C-fibers are activated by specific tactile stimuli (e.g., a gentle stroke with slow velocity or pinch) and transmit the emotional valence of the stimuli (i.e., positive or negative) to the brain regions that process emotion. To emulate this function, we implemented two units in the emotion valence layer that detect and transmit the positive and negative valence to the emotion layer. For example, when stroke stimuli were presented, the activation value of the positive unit was set to one, while that of the negative unit was set to zero. On the other hand, any unpleasant contact set the negative unit to one. When the contact had no emotional valence (e.g., weak pat and touch), both units were set to zero. How emotional valence is detected from various tactile stimuli was defined according to physiological evidence from past studies [22]–[24], [39].

In this module, the multimodal RBM was trained initially by using output signals from whole sensory modules. It updates the parameters of the RBM in order to reduce the reconstruction error in the abstraction layer. The module eventually learns the relationships between the hidden layers of the multimodal RBM and emotional valence in a supervised manner through a back propagation algorithm. In our model, the back propagation mechanism modulates the connection weights not only between the hidden layer and emotion valence nodes but also the hidden and visible layers of the multimodal RBM. We expected that the emotion layer (i.e., the hidden layer of the multimodal RBM) acquires representations of emotional states of multimodal signals.

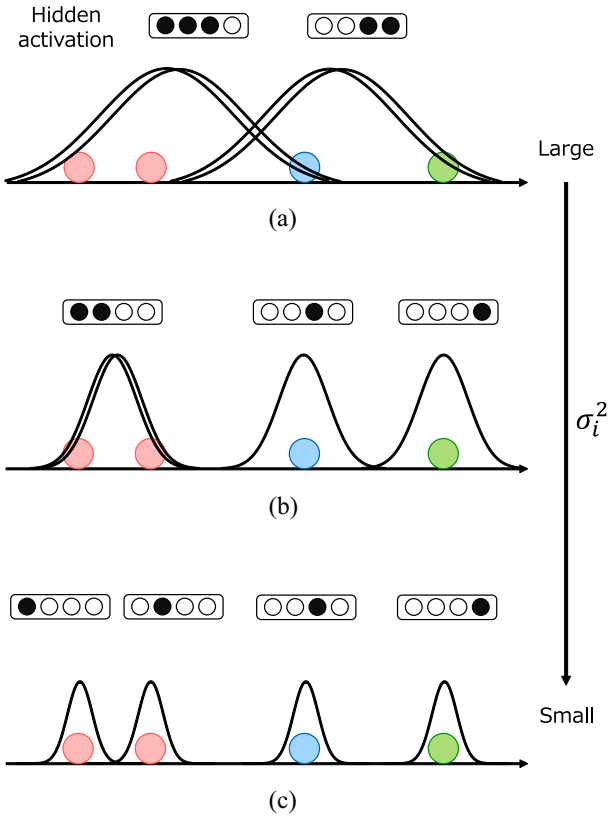


Fig. 3. Examples of the sensory module's behavior through training. The vertical axis and color variation of the circles represent feature value of the sensory signals (e.g., intensity) and the different emotions, respectively. The boxes with black and white circles show the active and inactive hidden units, respectively, and the Gaussian curves represent the Gaussian distributions that correspond to the hidden activations. The variances are refined and the activation patterns of the hidden layer increase as the model training progresses over the (a) early, (b) middle, and (c) later stages.

### C. Learning Process of the Proposed Model

We trained the proposed model by proceeding through ten sequences of the following three phases.

1) The parameters of the RBMs in the sensory modules were trained by using (9)–(12). This phase is illustrated by blue arrows in Fig. 2(b).
2) The multimodal RBM in the emotion module was trained using (4)–(6). This phase is illustrated by the green arrow in Fig. 2(b).
3) The connection weights of the emotional valence units and the multimodal RBM as well as in the multimodal RBM were modulated by back propagation. This phase is illustrated by red arrows in Fig. 2(b).

We continued each training phase for 1000 steps. After the third phase, we began the first phase again. This sequence was repeated ten times.

The model structure and learning method are based on a deep belief net [30] and a multimodal deep belief net [35]. However, these previously used models only executed one

TABLE I
DESCRIPTION OF DATASET SIMULATING INFANT–CAREGIVER INTERACTION

| Emotional state | Visual stimuli | Auditory stimuli | Tactile stimuli (emotional valence) | Number of data |
|---|---|---|---|---|
| Joy | Smiling face | Pitch rise in voice | Stroke (positive) | 150 |
| Surprise | Surprised face | Loud voice | Touch (zero emotion) | 150 |
| Anger | Angry face | Loud voice | Pinch (negative) | 125 |
| Disgust | Worried face | Low tone voice | Pinch (negative) | 125 |
| Sadness | Tearful face | Quite voice | Weak pat (zero emotion) | 125 |
| Fear | Frighten face | High frequency voice | Pat (negative) | 125 |
| Neutral | Neutral face | Neutral voice | Touch (zero emotion) | 125 |

sequence of the training phases; in contrast, we partitioned the training phases because we would like to examine and illustrate the developmental changes in the proposed model.

### D. Multimodal Sensory Signals

We evaluated our model by using an interaction dataset that simulated infant–caregiver interactions. Each interaction datum contains sensory signals from the three modalities (i.e., visual, auditory, and tactile modalities) which may express one of the seven basic emotions (i.e., joy, surprise, anger, disgust, sadness, fear, and neutral). The data were collected by using a robotic system, which consisted of a USB camera, a microphone, and a soft tactile sensor. An experimenter faced the system and expressed emotional expressions, not like generic face-to-face interactions but like infant–caregiver interactions, namely, exaggerated expressions by the caregiver. For instance, auditory signals, especially called infant-directed speech, have salient acoustic features [40], [41] (e.g., wide-range of pitch and fundamental frequency), and tactile signals also have wide-range features [11]. Our dataset includes these characteristics of multimodal signals. We assumed that the caregiver's expressions of emotional states to the infant were consistent across all three modalities in each interaction, and that a given expression would evoke the same emotional state in the infant. That is, one to one correspondence. For example, we assumed that when a caregiver showed the infant a smiling face, the infant would experience joy; furthermore, caregiver's auditory and tactile expressions in that same interaction would also make the infant experience joy. It should be noted that the proposed model is not given with the emotional labels of the input signals (i.e., the seven basic emotional states); instead, the model estimates the emotional states with a help of the emotional valences of the signals (i.e., positive, negative, or zero emotional valence) via tactile dominance.

Table I describes the infant–caregiver interaction dataset, while Fig. 4 shows a sample of the actual multimodal signals in the dataset. The first row of the figure [i.e., Fig. 4(a), (d), and (g)] illustrates the visual, auditory, and tactile sensory signals for joyful emotion. The second and third rows show the signals for angry and neutral emotions, respectively. In the following sections, we provide further the details on the multimodal sensory signals and their features for the sensory modules.

*1) Visual Stimuli:* The visual stimuli used were facial expressions produced by an experimenter playing a parent. We cropped the face region from images captured using a USB camera, and each face image was converted to a gray scale image of size $30 \times 30$ pixels. The experimenter expressed facial expressions of the seven basic emotions, and each emotional face had ten variations. Fig. 4(a)–(c) shows examples of the converted facial images for joyful, angry, and neutral emotions, respectively. The shapes of the eyebrow, eye, and mouth represented emotional characteristics of facial images. For instance, in many of the facial images of joy and surprise in our dataset, the mouth was opened. By contrast, for the other emotional states, the mouth was closed. The first 20 principal components (PCs) with contribution rates above 98% were extracted from all of the converted images by the PC analysis (PCA) method in order to reduce the dimensions of the facial data. We utilized these 20 PCs as input signals for the visual sensory module.

*2) Auditory Stimuli:* Auditory stimuli were single mora voices expressed as "Maa" corresponding to the seven basic emotions recorded from the same experimenter as for the visual stimuli, and each emotional voice had ten variations. The reason that we used a single mora voice was that the acoustic characteristics were enhanced by simple utterances as well as infant-directed speech [41], and we wanted to simplify the stimuli as much as possible for the experimental setting. The graphs in the center column of Fig. 4 show the raw signals (i.e., sound waves) of voices corresponding to joyful, angry, and neutral emotions. To extract features from these signals, we divided each signal into ten even sections and calculated acoustic features namely, the change in the fundamental frequency (F0) and the power of the F0 for those signals for each section (i.e., there are 20 features). All features were normalized in each section to fit the Gaussian distribution at a zero mean and a unit variance. We used these 20 features of audio signals as input values for the auditory sensory module. Fig. 5 shows the example of extracted acoustic features from audio signals in Fig. 4. As evident in Fig. 5, joyful and angry voices were shorter than were neutral voices. The power of F0 indicated by blue bars was detected only during two consecutive sections. Joyful and angry voices, on the other hand, had similar characteristics such as a short duration and high-intensity sounds as seen in Figs. 4 and 5. This implies that only some emotional voices can be discriminated in low-level

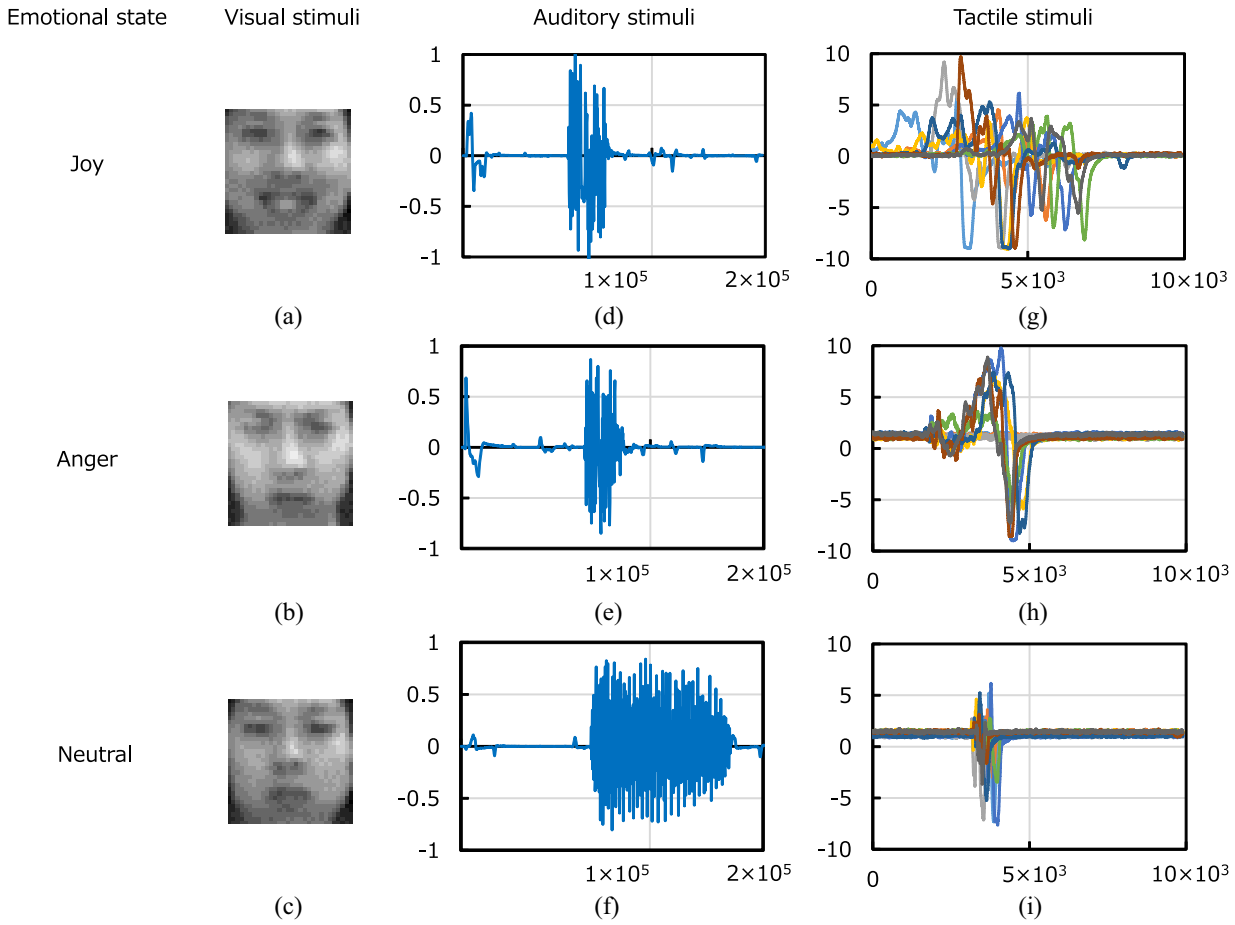


Fig. 4. Samples of multimodal signals in our interaction dataset. (a)–(c) Visual stimuli. (d)–(f) Auditory stimuli. (g)–(i) Tactile stimuli.

acoustic characteristics, whereas others are difficult to be discriminated.

*3) Tactile Stimuli and Emotional Valence:* Tactile stimuli and emotional valence were some of the most important signals in this experiment, as they were used to verify one of our hypotheses. We collected tactile stimuli simulating those used in infant–caregiver interaction via a human-skin-like tactile sensor. The tactile sensor, which was developed based on [42], is composed of two materials: 1) polyvinylidene difluoride (PVDF) films and 2) elastomer. More specifically, nine PVDF films (i.e., nine channels of the tactile sensor), which output voltage depending on the change rate of deformation (i.e., velocity of contact force), were sandwiched between two layers of human-skin-like elastomers (EXSEAL Company Ltd). Fig. 4(g)–(i) provides examples of nine sensor output signals corresponding to joyful (stroke), angry (pinch), and neutral (touch) tactile stimuli, respectively. Each sensory signal was smoothed with a moving average filter based on the previous 100 samples to reduce the noise. For instance, the stroke stimulus activated sensors for a longer duration than did other forms of contact, and the sensor output signals did not synchronize with each other [Fig. 4(g)] because the contact point moved over large areas very slowly. On the other hand, the pinch and touch stimuli activated sensors synchronously because their contact points did not move. Furthermore, the dynamic deformation of the sensor during the pinch stimulus was evident by the large values for the sensors' signals.

We calculated nine features from the sensory signals to extract the characteristics of tactile stimuli. Fig. 6 shows the relationships between a single sensory signal and the calculated features from a stroke stimulus. First, we extracted five features: (i) the maximum absolute velocity of contacts; (ii) the number of code (i.e., a direction of the signal) changes in the signals; and the intensity of (iii) low, (iv) middle, and (v) high frequency bands (low: 1–60 Hz; middle: 61–100 Hz; and high: 101–200 Hz) from the raw signal. Next, we calculated the integral values of the signal in terms of time to estimate the contact force and extracted the remaining four features: (vi) the duration of contact; (vii) the duration of contact with a strong force; (viii) the maximum force of the contact; and (ix) the number of sensors that detected contact. The maximum features [i.e., (i) and (viii)] were calculated from the whole channel (i.e., nine channels) values. The number of sensors that detected contact [i.e., (ix)] was determined by counting the sensors where the integral value exceeded a threshold. The duration features [i.e., (vi) and (vii)] were logical disjunction operated between all channels. The thresholds for contact detection and strong contact detection were 0.8 and 1.5, respectively. The other features [i.e., (ii)–(v)] were averaged across the whole channel, individually. These nine features were determined based on our knowledge of tactile receptors and the most important

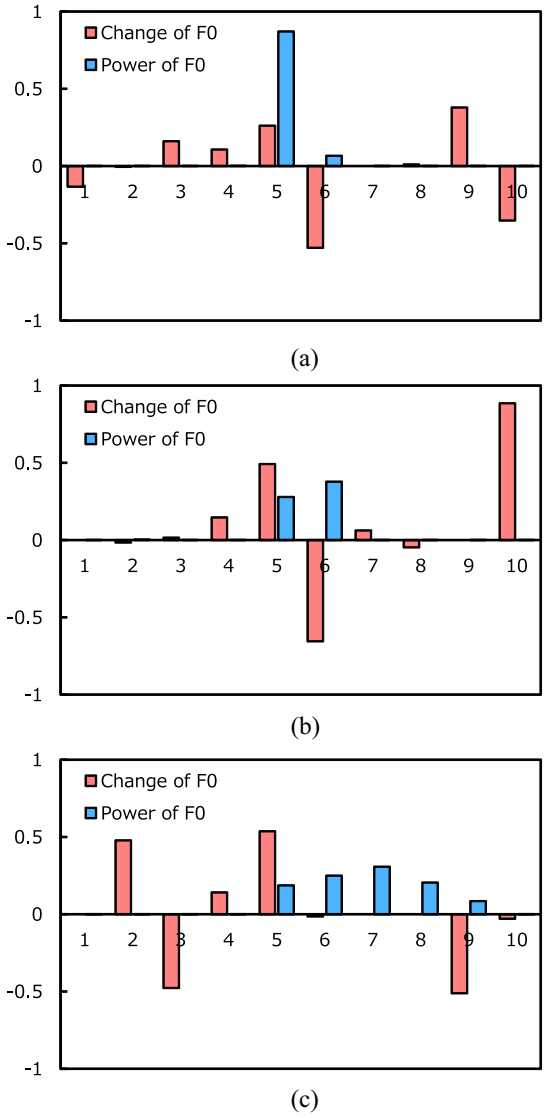


Fig. 5. Extracted auditory features from Fig. 4(d)–(f) (i.e., joyful, angry, and neutral voices). The horizontal axis represents the divided sections, and the vertical axis shows the normalized value of each feature. The red and blue bars indicate the change of the F0 between the current section and the previous section and the power of the F0 in each section, respectively. (a) Joy. (b) Anger. (c) Neutral.

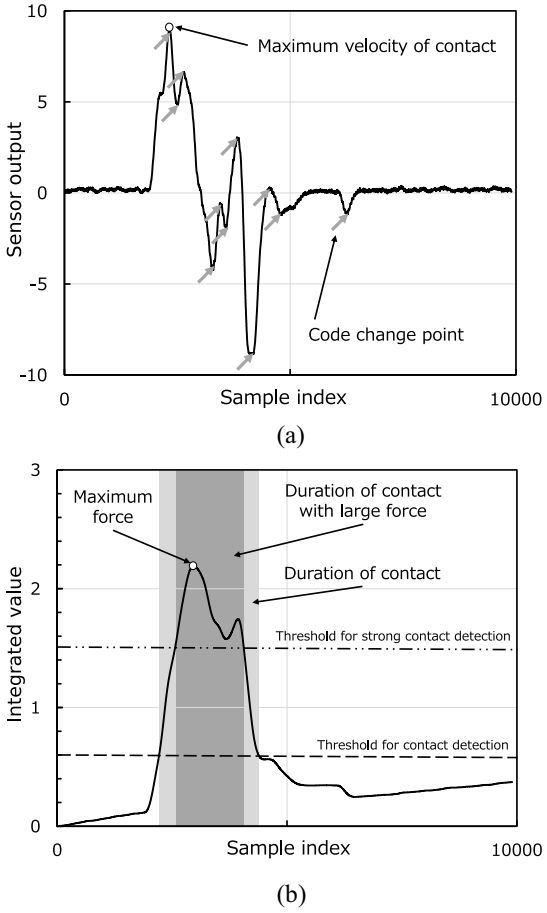


Fig. 6. Relationships between sensor signals and extracted features. (a) Smoothed sensor signal (using the moving average filter) with features (i) and (ii). (b) Integrated values of sensor signal (a) with features (vi), (vii), and (viii).

properties of touch [11], [43], and they were used as input signals for the tactile sensory module (i.e., the tactile sensory module has nine input nodes).

In this experiment, we used three kinds of emotional valence: 1) positive; 2) negative; and 3) zero. Emotional valences were predetermined for each tactile stimulus by the designer based on the inherent properties of C-fibers in the human skin [22], [24], [39]. Specifically, the stroke stimuli induced a positive emotional valence, while the touch and weak pat stimuli corresponded to zero emotional valence. The pinch and pat stimuli were considered to generate a negative emotional valence because the high pressure is known to activate C-fibers and evoke pain in humans. Table I indicates all relations between tactile stimuli and emotional valences. These emotional valence signals were represented as neuron activations of the emotion valence layer in the emotion module. As described in Section III-B3, positive emotional valence set the positive unit of the emotion valence layer to one while set the negative unit of that to zero. On the other hand, negative valence set the negative unit to one. When the emotional valence signal was zero valence, both units were set to zero.

## IV. Experiment and Result

Section IV-A outlines the experimental conditions used to verify our hypotheses regarding tactile dominance and perceptual development. Four conditions were designed to investigate the roles of these two functions. Then, in Sections IV-B–IV-E, we demonstrate the influences of these functions on the development of emotion perception by comparing emotion differentiation under the four conditions.

### *A. Experimental Conditions and Settings*

We conducted experiments under the following four conditions to investigate how emotion develops differently with or without each or both of these two functions of interest.

1) A model with tactile dominance and perceptual development (i.e., wTD-wPD condition).
2) A model with only perceptual development (i.e., w/oTD-wPD condition).

3) A model with only tactile dominance (i.e., wTD-w/oPD condition).
4) A model without either function (i.e., w/oTD-w/oPD condition).

The condition 1 (wTD–wPD) included both functions. Tactile dominance was modeled using emotional valence units (which represent positive, negative, and zero emotional valence signals based on tactile stimuli), whereas perceptual development was achieved by refinement of the variance of the input nodes in the sensory modules. The initial values of the variance were set to 1.0 and modulated by (12). This modulation represented the development of perceptual abilities from the immature sensation to the mature one (see Section III-B2). This was the main condition for verifying our hypotheses.

The condition 2 (w/oTD–wPD) omitted tactile dominance from wTD–wPD condition. In this condition, we assumed that the infant was not able to perceive positive or negative valences from tactile stimuli; this disorder has been observed in infants that are born without tactile nerves [44], [45]. To replicate this situation, we removed the emotional valence units and their connections from the emotion module, and therefore skipped the third phase of the training process (see Section III-C). We used this condition to assess the role of tactile dominance in the development of emotion perception by comparing the results with the first condition.

The condition 3 (wTD–w/oPD) excluded perceptual development instead of tactile dominance. In this condition, we assumed that the infant's perception had matured at the start of the developmental processes. To represent this we fixed the variance of sensory modules' input nodes to 0.01 and excluded the modulation of the variance in (12). We used this condition to verify the influence of perceptual development on the development of emotion perception.

In condition 4 (w/oTD–w/oPD), we excluded both functions from wTD–wPD condition. In other words, the emotional valence units and refinement process of the variance of the input nodes were removed from the proposed model as in the conditions 2 and 3, respectively.

We utilized the dataset of the simulated infant–caregiver interaction described in Section III-D. All data were used for the model learning and the visualization of results. The parameters for the proposed model are listed in Table II. We carried out the model leaning for ten times with the different initial values of network weights under four conditions.

TABLE II
PARAMETERS OF THE SENSORY MODULES AND EMOTION MODULE

| Parameter | Explanation | Value |
|---|---|---|
| $v_n^v$ | Number of visible nodes of visual sensory module | 20 |
| $h_n^v$ | Number of hidden nodes of visual sensory module | 10 |
| $v_n^a$ | Number of visible nodes of auditory sensory module | 20 |
| $h_n^a$ | Number of hidden nodes of auditory sensory module | 10 |
| $v_n^t$ | Number of visible nodes of tactile sensory module | 9 |
| $h_n^t$ | Number of hidden nodes of tactile sensory module | 10 |
| $\epsilon_w^s$ | Learning rate for weights of sensory module | 0.001 |
| $\epsilon_a^s$ | Learning rate for biases of visible units of sensory module | 0.001 |
| $\epsilon_b^s$ | Learning rate for biases of hidden units of sensory module | 0.001 |
| $\epsilon_z$ | Learning rate for log-variance | 0.001 |
| $v_n^e$ | Number of visible nodes of emotion module | 30 |
| $h_n^e$ | Number of hidden nodes of emotion module | 20 |
| $l_n^e$ | Number of nodes of emotion valence layer | 2 |
| $\epsilon_w^e$ | Learning rate for weights of emotion module | 0.01 |
| $\epsilon_a^e$ | Learning rate for biases of visible units of emotion module | 0.01 |
| $\epsilon_b^e$ | Learning rate for biases of hidden units of emotion module | 0.01 |
| $\eta$ | Learning rate for weights on back propagation | 0.0001 |

## B. Results

We first present the experimental results under all four conditions. Then, in Section IV-C, we compare the results with a focus on tactile dominance to test our first hypothesis, while in Section IV-D, we compare the results with a focus on the perceptual improvement to test the second hypothesis.

To visualize and evaluate the acquired representations of emotion, we carried out the PCA on the activations of the emotion layer. Fig. 7 shows one example of the PCA result under the four conditions. We selected first three PCs and illustrated the first and second PCs in Fig. 7(a), (c), (e), and (g) and the first and third PCs in Fig. 7(b), (d), (f), and (h). All plotted data are labeled using the emotional states of input signals. Note that these labels were not used for the model learning.

To quantitatively evaluate the representations of emotion in the PC space, we calculated the separation metric, given by

$$J_\sigma = \frac{s_b^2}{s_w^2} \tag{13}$$

$$s_b^2 = \frac{1}{n}\sum_{c=1}^{C} n_c(\boldsymbol{m_c} - \boldsymbol{m})^t(\boldsymbol{m_c} - \boldsymbol{m}) \tag{14}$$

$$s_w^2 = \frac{1}{n}\sum_{c=1}^{C}\sum_{\boldsymbol{x}\in\boldsymbol{\mathcal{X}}_c} n_c(\boldsymbol{x} - \boldsymbol{m}_c)^t(\boldsymbol{x} - \boldsymbol{m}_c) \tag{15}$$

where $s_b^2$ and $s_w^2$ are the between-class and within-class variance, respectively; $C$ and $c$ are the number and index of classes; $n$ and $n_c$ are the number of all data and the number of data belonging to class $c$; and $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}_c$, $\boldsymbol{m}_c$, and $\boldsymbol{m}$ are data belonging to class $c$, the mean of values of $\boldsymbol{\mathcal{X}}_c$, and the mean of all data, respectively. The larger the separation metric is, the greater the separation of the cluster in the PC space is. Fig. 8(a) and (b) summarizes the separation metrics for the categories of emotional valences (i.e., positive, negative, and zero emotional valences) and the seven basic emotions under the four experimental conditions. We calculated the averages and standard deviations of the separation metrics over ten times of experiments with different initial parameters of the network.

## C. Influence of Tactile Dominance on Differentiation of Emotion

We compared the results of the four conditions to demonstrate the influence of tactile dominance on the development of emotion perception. In the first and second PC space under wTD–wPD condition [Fig. 7(a)], the positive (i.e., joy) and the negative (i.e., anger, disgust, and fear) emotional valence

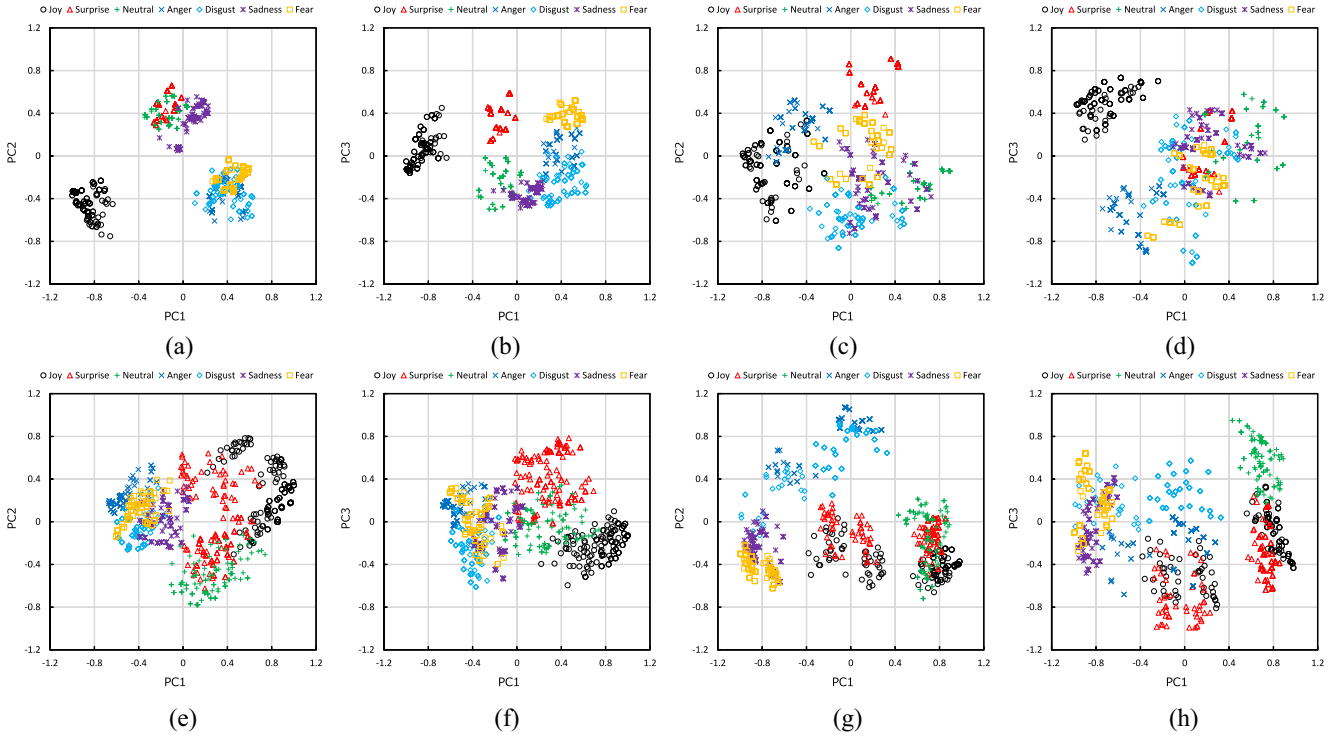


Fig. 7. Acquired low-dimensional representations of emotional stimuli by PCA for the emotion layer activations in the proposed model under the four conditions. (a), (c), (e), and (g) PC1-2 spaces, and (b), (d), (f), and (h) PC1-3 spaces for each condition.

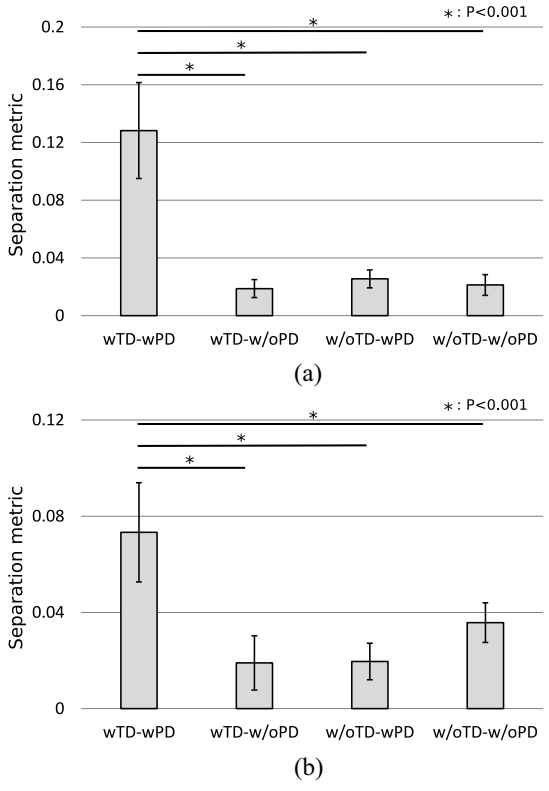


Fig. 8. Separation metrics for the categories of (a) emotional valences (i.e., positive, negative, and zero) and the categories of the (b) seven basic emotions under the four conditions.

clusters are separated by the first PC axis, while the second PC represents the differentiation between the zero emotional valence cluster and the others. Hence, the interaction data are differentiated into fundamental emotional clusters (i.e., positive, negative, and zero emotional valence) in this space. In the first and third PC space [Fig. 7(b)], the vertical axis (i.e., PC3) subdivides the clusters of negative and zero emotional valence into seven emotional states except joy. More specifically, the cluster of negative emotional valence is differentiated into subcategories of fear, anger, and disgust from the top of the graph. The space composed of the first and third PCs shows that the clusters of emotional valence were subdivided into the seven basic emotions. The separation metrics for both emotional categories under wTD–wPD condition showed significant differences between wTD-wPD and other three conditions [in Fig. 8(a) and (b)]. In contrast to the above, the results for the other conditions show unclear differentiation of emotional categories in the PC spaces. Only the PCA results of wTD–w/oPD condition demonstrated that the interaction data weakly differentiated into clusters of emotional valences in both spaces; however, the separation metric for categories of emotional valences showed nonsignificant differences between other conditions.

The comparison of these conditions overall demonstrates that tactile dominance leads to better separation of the emotional categories; however, it is necessary that perceptual development also works in the proposed model as seen in Fig. 7(a) and (b). When tactile dominance was excluded from the model (i.e., w/oTD–wPD and w/oTD–w/oPD conditions), the representation of emotional valence clusters could not be obtained even using the same interaction data. A potential reason is that visual and auditory signals contained ambiguous features in terms of positive and negative valence. For instance, joy and anger had similar auditory characteristics due

to the intensity of those stimuli (Section III-D2). By contrast, when tactile dominance was included, the emotional valence inherent in the tactile stimuli disambiguated such situations.

### *D. Influence of Perceptual Development on Differentiation of Emotion*

We assessed the influence of perceptual improvement on the development of emotion perception. From the comparison of results between wTD–wPD and wTD–w/oPD conditions, we found that the distribution of interaction data, which weakly clustered based on the emotional valences in Fig. 7(f), showed clearly differentiation of the seven basic emotions in Fig. 7(b) owing to the perceptual development. It was clear that perceptual development also facilitated the differentiation of the clusters relevant to the emotional valences as shown in Figs. 7(a) and 8(a), although this effect was not clearly observed in the comparison between w/oTD–wPD and w/oTD–w/oPD conditions.

Taken together, these results indicate that perceptual development does enhance clearer differentiation of emotional categories in terms of both emotional valences and the seven basic emotions; however, the effect appears only when tactile dominance is included together in the model. The result suggests that the two functions facilitate the developmental differentiation of emotion perception but their effects become significant only when they exist in the model together.

### *E. Perceptual Development Produced by Modulation of RBM Parameter $\sigma_i^2$*

We then closely analyzed how perceptual development was reproduced by the modulation of the variance parameters, $\sigma_i^2$, in the sensory RBMs. Fig. 9 shows the transition of $\sigma_i^2$ $(i = 0, \ldots, 20)$ of the vision module over the learning process as an example. All the variances of the visible nodes with perceptual development were initialized at 1.0 and updated using (12). The results showed that the variance parameter $\sigma_i^2$ was properly adjusted through training.

We also visualized the changes in reconstructed images across the learning steps. Fig. 10(d) shows four randomly selected input images depicting facial images of joy (left top), neutral (right top), anger (left down), and sadness (right down). Fig. 10(a)–(c) shows the reconstructed images from those in Fig. 10(d). For some images, in the early stages of learning [Fig. 10(a) and (b)], the reconstructed images were unclear, making it hard to determine their emotional states. Furthermore, some reconstructed images seemed to represent different emotional states from the input images. For example, the top left in Fig. 10(a) looks similar to an angry face though the input was a happy face. We described the reason for this result in Section III-B2. The Gaussian distribution with a large variance covered many input signals and thus generated highly ambiguous reconstructions. However, in the later stage of learning [Fig. 10(c)], the reconstructed images became more similar to the input images. These results indicate that the sensory modules were able to simulate perceptual development by updating the variance of their visible nodes.

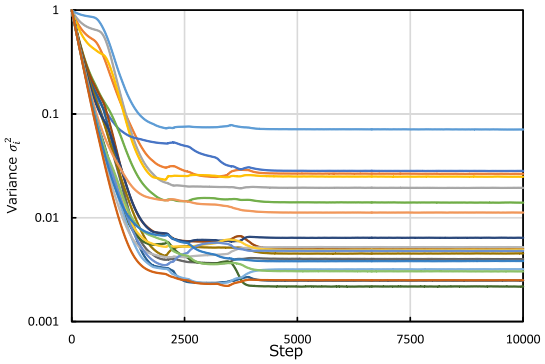


Fig. 9. Transition of visual nodes' variance for the vision sensory module.

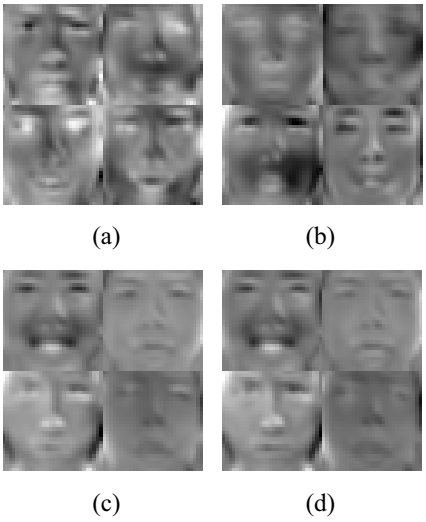


Fig. 10. Examples of input stimuli and reconstructed images during learning process in the visual sensory module. (a) Step 0. (b) Step 1000. (c) Step 10 000. (d) Input data.

## V. Discussion

We proposed a computational neural network model comprising two modules (i.e., the sensory module and the emotion module) to verify our hypotheses regarding the development of emotion perception in infancy. The sensory module processed multimodal sensory signals individually, as in the sensory area of the cerebral cortex (i.e., the visual, auditory, and somatosensory cortices). The emotion module, which was on a higher-level than the sensory modules, was used to integrate the abstracted signals obtained from the sensory modules, and the emotion layer in the module further integrated the information of emotional valences based on tactile dominance. It is known that the superior temporal sulcus (STS) integrates visual, audio, and tactile signals [46] and engages in multimodal information processing for emotion perception [47], [48]. The superior temporal gyrus (STG), which is near the STS, also responds to various nonverbal emotional stimuli [49], and the temporal area of infants' brain perceives and reacts to audiovisual emotional stimuli [50]. There are also known neural connections between the STS and the amygdala, Björnsdotter *et al.* [24] reported that tactile C-fibers deliver the signals of positive and negative touches to a part of the limbic

system, such as the amygdala and insular cortex. All of this prior knowledge suggests that the structure of the proposed model reproduced rough neural connections between human brain regions, at least in relation to emotional processing. In other words, the sensory module corresponds to the sensory areas of the brain, while the emotional valence layer emulates the neural connection between the amygdala and tactile C-fibers. The emotion layer in the emotion module emulates the function, such as perception of categorical emotion in the temporal region (i.e., the STS and the STG). Furthermore, we consider that the proposed model would replicate not only the development of emotion perception but also multimodal sensory processing in general (e.g., object recognition). In fact, our challenge was to design a biologically and neurologically plausible mechanism for the development of emotion perception based on the latest knowledge about a theory of the human brain mechanism (e.g., free-energy principle [32]).

In Section IV-C, we compared the experimental results between the four conditions to investigate how emotion develops differently with and without tactile dominance, in order to test our first hypothesis. The results demonstrated that tactile dominance facilitated the differentiation of emotional categories when perceptual development was also included in the model (i.e., wTD–wPD condition). There are two types of C-fibers in the human skin that perceive emotional valence produced by tactile stimuli. The first type of C-fibers, CT afferents, specializes in the detection of the pleasurable touch. This type of fiber tends to be distributed over hairy skin and is activated by light strokes with a velocity of 1–10 cm/s [24]. The other type of C-fibers responds to chemical substances, thermal stimuli, and otherwise negative stimuli such as tactile stimuli pain. In addition to this second type of C-fibers, there are A$\delta$-fibers, which transmit pain signals from the skin to the brain, especially the somatosensory area. There is a condition called congenital insensitivity to pain with anhidrosis (CIPA), whereby individuals are born without the second type of C-fibers and A$\delta$-fibers; in lacking these fibers, individuals with CIPA are unable to feel pain. Past studies have also shown that patients with CIPA have impairments in the perception, recognition, and modulation of emotion [45]. Danziger *et al.* [44], in an experiment designed to estimate others' emotional states, demonstrated that patients with a similar condition called congenital insensitivity to pain (CIP) inhibited differences in their ratings of others' painful situations or propensity to infer pain from others' facial expressions from control subjects. Note that CIP patients lose only the A$\delta$-fibers and do not lack the C-fibers in their skin; therefore, they are able to roughly detect pain (i.e., negative sensations) through the C-fibers, unlike CIPA patients. These past findings suggested that tactile C-fibers play an important role in emotion perception and understanding others' emotion. In this paper, the experimental conditions where tactile dominance was excluded seemed to simulate the characteristics of patients with CIPA, which suggests that our results both demonstrate the effects of CIPA and the importance of C-fibers for the development of emotion perception.

In Section IV-D, we assessed the influence of sensory improvement on the development of emotion perception. The comparison results of the four conditions showed that the perceptual development facilitated the differentiation of emotional categories when tactile dominance was also included in the proposed model (i.e., wTD–wPD condition). This result supports the claim of Walker-Andrew's [5] study that perceptual development increases the differentiation of information for affect. The idea of perceptual development was modeled by refining the variance of the input nodes in the sensory modules across the training process. The experimental results in Section IV-E further demonstrated that the sensory modules simulated perceptual development by modulating the variance parameters.

In summary, the proposed model represented the development of emotion perception through learning of caregivers' visual, audio, and tactile expressions during interactions. The experimental results demonstrated that both tactile dominance and perceptual development have the role for facilitating the development of emotional perception; however, its influence appears only when both functions are integrated into the model together. Connecting our results to behavioral studies in physiology may help to elucidate the neural and social mechanisms of the development of emotion perception. On the other hand, humans' emotion is affected not only by their external senses but also by their behavior and internal, physiological systems (e.g., the endocrine system). Note that we take such systems into account as in other studies in cognitive developmental robotics [6], [13], [14], [16]–[18] when modeling the development of emotion perception. Additionally, it seems essential to examine the development of emotion perception by integrating physiological indices using nonparametric Bayesian models [51]. To address these future issues, we need to combine the results of our prior analyses for more accurate modeling of emotion development.

## VI. Conclusion

This paper presented a modeling study of the development of emotion perception in infancy. We hypothesized that tactile dominance and perceptual development contribute to the development of emotion perception. The proposed model learned the virtual infant–caregiver interaction, and the experimental results were compared between the four conditions (with and without tactile dominance and perceptual development). Our results suggested that tactile dominance and perceptual development facilitated the differentiation of emotional states when both mechanisms were integrated into the proposed model.

## References

[1] C. E. Izard, *The Psychology of Emotions*. New York, NY, USA: Springer, 1991.
[2] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychol. Rev.*, vol. 110, no. 1, pp. 145–172, 2003.
[3] L. A. Sroufe, *Emotional Development: The Organization of Emotional Life in the Early Years*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[4] M. Lewis, "The self in self-conscious emotions," *Ann. New York Acad. Sci.*, vol. 818, no. 1, pp. 119–142, 1997.
[5] A. S. Walker-Andrews, "Infants' perception of expressive behaviors: Differentiation of multimodal information," *Psychol. Bull.*, vol. 121, no. 3, pp. 437–456, 1997.

[6] J. Nadel and D. Muir, *Emotional Development: Recent Research Advances*. Oxford, U.K.: Oxford Univ. Press, 2005.
[7] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, *Handbook of Emotions*. New York, NY, USA: Guilford Press, 2008.
[8] T. Grossmann, "The development of emotion perception in face and voice during infancy," *Restorative Neurol. Neurosci.*, vol. 28, no. 2, pp. 219–236, 2010.
[9] M. Peláez-Nogueras *et al.*, "Infants' preference for touch stimulation in face-to-face interactions," *J. Appl. Develop. Psychol.*, vol. 17, no. 2, pp. 199–213, 1996.
[10] M. Peláez-Nogueras, T. M. Field, Z. Hossain, and J. Pickens, "Depressed mothers' touching increases infants' positive affect and attention in still-face interactions," *Child Develop.*, vol. 67, no. 4, pp. 1780–1792, 1996.
[11] M. J. Hertenstein, "Touch: Its communicative functions in infancy," *Human Develop.*, vol. 45, no. 2, pp. 79–94, 2002.
[12] M. Asada *et al.*, "Cognitive developmental robotics: A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, May 2009.
[13] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Auton. Robots*, vol. 12, no. 1, pp. 83–104, 2002.
[14] A. Blanchard and L. Cañamero, "From imprinting to adaptation: Building a history of affective interaction," in *Proc. 5th Int. Workshop Epigenetic Robot.*, 2005, pp. 23–30.
[15] C. Hasson, P. Gaussier, and S. Boucenna, "Emotions as a dynamical system: The interplay between the meta-control and communication function of emotions," *Paladyn*, vol. 2, no. 3, pp. 111–125, 2011.
[16] A. Hiolle, M. Lewis, and L. Cañamero, "Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment," *Front. Neurorobot.*, vol. 8, p. 17, May 2014.
[17] A. Lim and H. G. Okuno, "The MEI robot: Towards using motherese to develop multimodal emotional intelligence," *IEEE Trans. Auton. Mental Develop.*, vol. 6, no. 2, pp. 126–138, Jun. 2014.
[18] J. Lones, M. Lewis, and L. Cañamero, "From sensorimotor experiences to cognitive development: Investigating the influence of experiential diversity on the development of an epigenetic robot," *Front. Robot. AI*, vol. 3, p. 44, Aug. 2016.
[19] T. Horii, Y. Nagai, and M. Asada, "Touch and emotion: Modeling of developmental differentiation of emotion lead by tactile dominance," in *Proc. IEEE 3rd Joint Int. Conf. Develop. Learn. Epigenetic Robot.*, Osaka, Japan, 2013, pp. 1–6.
[20] M. J. Hertenstein, R. Holmes, M. McCullough, and D. Keltner, "The communication of emotion via touch," *Emotion*, vol. 9, no. 4, pp. 566–573, 2009.
[21] A. D. L. Jean, D. M. Stack, and A. Fogel, "A longitudinal investigation of maternal touching across the first 6 months of life: Age and context effects," *Infant Behav. Develop.*, vol. 32, no. 3, pp. 344–349, 2009.
[22] R. J. Traub and L. M. Mendell, "The spinal projection of individual identified A-delta-and C-fibers," *J. Neurophysiol.*, vol. 59, no. 1, pp. 41–55, 1988.
[23] L. Fabrizi *et al.*, "A shift in sensory processing that enables the developing human brain to discriminate touch from pain," *Current Biol.*, vol. 21, no. 18, pp. 1552–1558, 2011.
[24] M. Björnsdotter, I. Morrison, and H. Olausson, "Feeling good: On the role of C fiber mediated touch in interoception," *Exp. Brain Res.*, vol. 207, nos. 3–4, pp. 149–155, 2010.
[25] R. M. Bradley and C. M. Mistretta, "Fetal sensory receptors," *Physiol. Rev.*, vol. 55, no. 3, pp. 352–382, Jul. 1975.
[26] V. Dobson and D. Y. Teller, "Visual acuity in human infants: A review and comparison of behavioral and electrophysiological studies," *Vis. Res.*, vol. 18, no. 11, pp. 1469–1483, 1978.
[27] L. W. Olsho, E. G. Koch, and C. F. Halpin, "Level and age effects in infant frequency discrimination," *J. Acoust. Soc. America*, vol. 82, no. 2, pp. 454–464, 1987.
[28] L. B. Smith and D. G. Kemler, "Developmental trends in free classification: Evidence for a new conceptualization of perceptual development," *J. Exp. Child Psychol.*, vol. 24, no. 2, pp. 279–298, 1977.
[29] L. B. Smith, "Perceptual development and category generalization," *Child Develop.*, vol. 50, no. 3, pp. 705–715, 1979.
[30] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[31] G. E. Hinton, "A practical guide to training restricted Boltzmann machines," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Rep. UTML TR 2010-003, 2010.
[32] K. Friston, "The free-energy principle: A unified brain theory?" *Nat. Rev. Neurosci.*, vol. 11, no. 2, pp. 127–138, 2010.
[33] S. Sukhbaatar, T. Makino, K. Aihara, and T. Chikayama, "Robust generation of dynamical patterns in human motion by a deep belief nets," in *Proc. Asian Conf. Mach. Learn.*, 2011, pp. 231–246.
[34] J. Ngiam *et al.*, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 689–696.
[35] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, pp. 1–8.
[36] N. Srivastava and R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2949–2980, 2014.
[37] K. Cho, A. Ilin, and T. Raiko, "Improved learning of Gaussian–Bernoulli restricted Boltzmann machines," in *Artificial Neural Networks and Machine Learning–ICANN 2011*. Heidelberg, Germany: Springer, 2011, pp. 10–17.
[38] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Adv. Robot.*, vol. 20, no. 10, pp. 1165–1181, 2006.
[39] R. Schmidt, M. Schmelz, H. Torebjörk, and H. O. Handwerker, "Mechano-insensitive nociceptors encode pain evoked by tonic pressure to human skin," *Neuroscience*, vol. 98, no. 4, pp. 793–800, 2000.
[40] A. Fernald, "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?" *Child Develop.*, vol. 60, no. 6, pp. 1497–1510, 1989.
[41] M. Spinelli, M. Fasolo, and J. Mesman, "Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes," *Develop. Rev.*, vol. 44, pp. 1–18, Jun. 2017.
[42] Y. Tada and K. Hosoda, "Acquisition of multi-modal expression of slip through pick-up experiences," *Adv. Robot.*, vol. 21, nos. 5–6, pp. 601–617, 2007.
[43] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—From humans to humanoids," *IEEE Trans. Robot.*, vol. 26, no. 1, pp. 1–20, Feb. 2010.
[44] N. Danziger, K. Prkachin, and J. Willer, "Is pain the price of empathy? The perception of others' pain in patients with congenital insensitivity to pain," *Brain*, vol. 129, no. 9, pp. 2494–2507, 2006.
[45] Y. Indo, "Nerve growth factor and the physiology of pain: Lessons from congenital insensitivity to pain with anhidrosis," *Clin. Genet.*, vol. 82, no. 4, pp. 341–350, 2012.
[46] M. S. Beauchamp, N. E. Yasar, R. E. Frye, and T. Ro, "Touch, sound and vision in human superior temporal sulcus," *Neuroimage*, vol. 41, no. 3, pp. 1011–1020, 2008.
[47] S. Campanella and P. Belin, "Integrating face and voice in person perception," *Trends Cogn. Sci.*, vol. 11, no. 12, pp. 535–543, 2007.
[48] R. Watson *et al.*, "Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration," *J. Neurosci.*, vol. 34, no. 20, pp. 6813–6821, 2014.
[49] B. Kreifelts, T. Ethofer, W. Grodd, M. Erb, and D. Wildgruber, "Audiovisual integration of emotional signals in voice and face: An event-related fMRI study," *Neuroimage*, vol. 37, no. 4, pp. 1445–1456, 2007.
[50] T. Grossmann, T. Striano, and A. D. Friederici, "Crossmodal integration of emotional information from face and voice in the infant brain," *Develop. Sci.*, vol. 9, no. 3, pp. 309–315, 2006.
[51] T. Horii, Y. Nagai, and M. Asada, "Toward analysis of emotional development using physiological and behavioral data," in *Proc. HRI Workshop HRI Bridge Between Robot. Neurosci.*, 2014, pp. 47–48.

**Takato Horii** is currently pursuing the Ph.D. degree with the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Osaka, Japan.

His current research interests include computational modeling of emotion development, emotional human–robot interaction, and tactile sensing.

Mr. Horii was a Research Fellow of the Japan Society for the Promotion of Science.

**Yukie Nagai** (M'09) received the master's degree from Aoyama Gakuin University, Tokyo, Japan, in 1999, and the Ph.D. degree from Osaka University, Osaka, Japan, in 2004, both in engineering.

She was a Post-Doctoral Researcher with the National Institute of Information and Communications Technology (NICT), Kyoto, Japan, from 2004 to 2006, and Bielefeld University, Bielefeld, Germany, from 2006 to 2009. She was then a Specially Appointed Associate Professor with Osaka University, from 2009 to 2017, and became a Senior Researcher with NICT in 2017. She was a Project Leader of MEXT Grant-in-Aid for Scientific Research on Innovative Areas Computational Modeling of Social Cognitive Development and Design of Assistance Systems for Developmental Disorders from 2012 to 2017, and has been the Project Leader of JST CREST Cognitive Mirroring: Assisting People With Developmental Disorders by Means of Self Understanding and Social Sharing of Cognitive Processes since 2016. Her current research interests include computational modeling of human cognitive functions such as self-other cognition, imitation, and joint attention, and design of assistant systems for developmental disorders.

**Minoru Asada** (M'88–F'05) received the B.E., M.E., and Ph.D. degrees in control engineering from Osaka University, Osaka, Japan, in 1977, 1979, and 1982, respectively.

He is a Professor with Osaka University, Suita, Japan, in 1995, where he has been a Professor with the Department of Adaptive Machine Systems, Graduate School of Engineering, since 1997. He is one of the Founders of RoboCup, and the Former President of the International RoboCup Federation from 2002 to 2008. Since 2005, he has been the Research Director of ASADA Synergistic Intelligence Project at Exploratory Research for Advanced Technology by Japan Science and Technology Agency.

Dr. Asada was a recipient of several awards such as the Best Paper Award at the 1992 IEEE/RSJ International Conference on Intelligent Robots and Systems and the Commendation by the Minister of Education, Culture, Sports, Science and Technology, Japanese Government as Person of Distinguished Services to Enlightening People on Science and Technology. He is currently a Principal Investigator of Grants-in-Aid for Scientific Research entitled Constructive Developmental Science Based on Understanding the Process from Neurodynamics to Social Interaction.