TABLE I

RMSE BETWEEN INPUT AND RECONSTRUCTED DATA

Experimental condition	RMSE
Using complete multimodal inputs	1.37
Using deficient inputs (w/o visual)	2.36

(a) From complete inputs

(b) From deficient inputs

Fig. 3. Reconstructed data

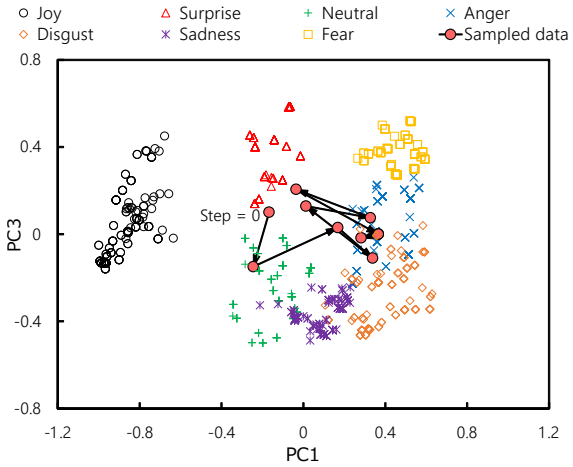

Fig. 4. Transition of recognitions at each 10 step in PC space.

III. Experiment and Results

We evaluated the proposed model using the multimodal interaction dataset as shown in Fig. 2. The dataset includes 6 basic emotions (i.e., joy, surprise, anger, fear, sadness, and disgust) and neutral states stimuli. It is assumed that multimodal inputs does not conflict with each modality in these experiments. After training by the dataset, we use the same model for each experiment.

A. Reconstruction from complete multimodal inputs

The first experiment has been performed to validate a basic ability of the proposed model. We inputted a set of multimodal stimuli for the proposed model to reconstruct sensory data through higher layer. The proposed model sampled activations from lower layers to higher layers sequentially by using input data. After a forward sampling, the model resampled data in the opposite direction. Fig. 3(a) shows the reconstructed data of the visual modality from the angry stimuli (Fig. 2). The root mean squared error (RMSE) between the visual input and the reconstructed data is summarized in Table I. According to Fig. 3(a), the proposed model is able to reconstruct a same emotional expression as well as inputs.

B. Emotion recognition and generation from deficient inputs

The second experiment has been carried out to investigate the advantage of this model that the model can replenish a deficient input through the forward-backward sampling. We inputted multimodal stimuli except a visual input from Fig. 2 and sampled a reconstructed visual data by using a Gibbs sampling method in 10000 steps. Fig. 3(b) depicts the reconstructed data of the visual modality at $step = 10000$. According to Fig. 3(b), the proposed model can imagine the deficient modality data from other modality inputs. The estimation accuracy is lower than the previous experiment. The dataset includes same auditory and tactile signal combination with different visual stimuli. It has an influence on the dispersion of reconstructions.

We illustrates the transition of sampled activations in higher layer at each 10 step from 0 to 100 steps in the principal component (PC) space of outputs in Fig. 4. The each emotional state of the learning dataset distribute like the circumplex model. This figure shows that the model recognized the deficient input as the surprise signal at first recognition process. However, through the sampling method with reconstructed data, the estimated emotional state changed into the anger which is the correct state.

From the above results, we could confirm that the proposed model has advantages which is discussed in section 2. Robots can recognize partners' emotional states and generate emotional expressions by using our model with action modules which converts outputs into robot action.

IV. Conclusion

We discuss the relation between our emotion recognition-generation model and the mirror neuron systems (MNSs) and the mentalizing systems [5]. The result of first experiment shows that our model can mirror the expressions through the abstracted activations like the MNSs. This model automatically mimics others' expressions and emotional states, similar to the emotional contagion. It is clear from the second result that the model can infer the deficient data and update the belief of others' emotional state sequentially through the sampling. This behavior suggests that the reconstruction process relates the MNSs and the estimation process through the sampling relates the mentalizing systems.

For future improvements, we extend the model with action modules by using reconstructed data for real-time human robot interactions. In addition, we verify the detail of relation to the MNSs and the mentalizing systems.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 15J00671, 24000012 ,24119003 and 25700027.

References

[1] J.A. Russell. *Journal of personality and social psychology*, Vol. 39, No. 6, p. 1161, 1980.
[2] T. Horii et al. In *Proc. of IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics*, pp. 1–6, 2013.
[3] G. Hinton. *Momentum*, Vol. 9, p. 1, 2010.
[4] K. Cho et al. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 10–17. 2011.
[5] F. Van Overwalle and K. Baetens. *Neuroimage*, Vol. 48, No. 3, pp. 564–584, 2009.