The Developmental Path of Multisensory Perception of Emotion and Phoneme in Japanese Speakers

Hisako W. Yamamoto[1], *Misako Kawahara*[1], *Akihiro Tanaka*[1]

[1] Tokyo Woman's Christian University, Japan

hisako_wy@lab.twcu.ac.jp

Abstract

The purpose of this study is to investigate the developmental path of multisensory perception of emotion and phoneme in Japanese speakers. In Experiment 1, Japanese children from age of 5 to 12 and adults were engaged in Emotion perception task, in which speakers' emotions expressed in their face and voice were congruent or incongruent, and Phoneme perception task, in which auditory and visual phonemic information are congruent or incongruent (i.e., McGurk type movie). Children's judgement weighting on voice information in emotion perception increased over age, whereas the weighting of audiovisual information in phoneme perception remained the same during childhood. Interestingly, adults' emotion perception based on voice information was less than those of 11-12-year-olds. Experiment 2 examined whether adults' multisensory perception was affected by their parenting experience. The results showed that parents who were rearing their children judged speakers' emotion relying on voice information less than non-parents, whereas visual influence in phoneme perception was not different between parents and non-parents; adults' multisensory perception may be affected by daily interaction with their children as for emotion judgement. Taken together, these results show differential integration processes between emotion and phoneme perception.

Index Terms: emotion perception, facial expression, vocal expression, McGurk effect, phoneme perception

1. Introduction

In face-to-face communication, we utilize both visual and auditory information in order to understand others' intention efficiently. Especially, it is well known that emotion and phoneme perception are achieved through such multisensory processes. First, speakers' emotion is conveyed by the face and voice, and integration of these cues enables us to recognize others' emotion appropriately [1]. Recent research has demonstrated cultural differences in integrating audiovisual information in emotion perception. As shown in one study with native Japanese and Dutch people [2], Japanese people's emotion perception tends to be influenced by vocal expression, while Dutch people put weight on facial expression. That is, Japanese people weight auditory cues more than Dutch people in emotion perception. Considering that Japanese young children (5-6-year-olds) tend to judge emotion based on speakers' facial expression than vocal expression [3], this tendency may occur during childhood.

Second, phoneme perception is influenced by lip movements as well as speech sound itself. The most famous phenomenon is known as the McGurk effect, in which the voice /ba/ is played over the face movements for /ga/, then /da/ is perceived [4]. As well as audiovisual emotion perception, a visual influence in phoneme perception differ depending on listeners' cultural and linguistic background. The McGurk effect was induced in Japanese speakers less frequently than English speakers [5], and the difference occurs gradually in accordance with development during childhood [6]. These studies suggested that Japanese speakers tend to be less influenced by lip movement than English speakers in phoneme perception.

Thus, previous studies have demonstrated that Japanese people tend to be influenced by visual information less than Western people both in emotion and phoneme perception. That is, they put weight on auditory cues in integration of audiovisual information. However, it has not been revealed when and how they acquire such a manner of audiovisual integration. The present study investigated Japanese people's developmental path of multisensory emotion and phoneme perception in order to reveal the relation of their culture-specific integration processes.

2. Experiment 1

2.1. Method

2.1.1. Participants

The participants were comprised of eighty-two 5-6-year-olds (*M*=5.6; *SD*=0.5), ninety-two 7-8-year-olds (*M*=7.5; *SD*=0.5), ninety-five 9-10-year-olds (*M*=9.5; *SD*=0.5), eighty 11-12-year-olds (*M*=11.4; *SD*=0.5), and sixty-two adults (30 to 39-year-olds: *M*=36.1; *SD*=2.3). All adult participants were child participants' parents. All participants were native Japanese speakers, recruited in National Museum of Emerging Science and Innovation (Miraikan), Tokyo, Japan.

2.1.2. Apparatus

The experimenters used a 15-inch laptop in order to present and control audiovisual stimuli through Hot Soup Professor 3.4 (Onion software). Participants were seated in front of a laptop at a distance of about 50 cm. Visual stimuli were displayed in the middle of the monitor with a 640×480 pixel size. Sound stimuli were presented through headphones (HDA300, SENNHEISER) at approximately 70 dB SPL at most, adjusted using headphones amplifiers (DAC-HA200, ONKYO) in order to mask the background noise in the experimental laboratory.

2.1.3. Stimuli and Procedure

The experiment was conducted at the experimental laboratory in Miraikan.

[Emotion perception task] Audiovisual stimuli were short movies in which a speaker expressed her emotions. In each movie, a female Japanese speaker expressed happiness or anger

in their face and voice. The linguistic information of voice was emotionally neutral ("*Hai, moshimoshi*" (Hello), "*Sayonara*" (Good-bye), "*Korenani*" (What is this?), or "*Sounandesuka*" (Is that so?)). A total of 32 movies (two speakers × four emotions (angry face and voice, happy face and voice (congruent), angry face and happy voice, happy face and angry voice (incongruent)) × four utterances) were used as test stimuli. In addition to them, other two movies were used as practice stimuli.

On each trial, a fixation point was displayed at the center of the monitor for 500 ms, and a signal sound (440 Hz pure tone lasting 100 ms) was played simultaneously. After 500ms from the onset of the presentation of the fixation point, a movie and a blank display were presented successively. Participants were asked to judge whether the woman was happy or angry and respond by pressing keys (D or K). After 500 ms from the participant's response, the next test trial began. A total of 32 test trials including 16 congruent trials and as many incongruent trials was conducted, following two practice trials. The order of test trials was randomized.

[Phoneme perception task] Audiovisual stimuli were short movies in which a Japanese male or female speaker pronounced one syllable (/ka/, /pa/, or /ta/). Test stimuli contained twelve congruent (lip movement and sound were congruent) and six McGurk type movies (/ka/ lip movement combined with /pa/ sound).

On each trial, a fixation point was displayed at the center of the monitor for 800 ms, and a signal sound (440 Hz pure tone lasting 100 ms) was played simultaneously, then a blank display was presented 500 ms. After 1300ms from the onset of the presentation of the fixation point, a movie and a blank display were presented successively. Participants were asked to judge whether the speaker said /ka/, /pa/, or /ta/ and respond by pressing keys (Z, V or M). After 500 ms from the participant's response, the next test trial began. Each congruent movie was presented once (12 congruent trials) while each McGurk type movie was presented twice (12 McGurk trials), resulting in a total of 24 trials. Congruent trials were included in order to avoid participants' response bias. The order of test trials was randomized.

2.2. Results and Discussion

[Emotion perception task] The voice responses, which indicates the rate of participants' emotion judgement based on speaker's voice, were shown in Figure 1. To examine age difference, we performed an *Age* (5-6y, 7-8y, 9-10y, 11-12y, adults) × *Emotion* (congruent, incongruent) mixed-factor analysis of variance (ANOVA) on the voice responses. The interaction between Age and Emotion ($F(4, 406)=7.53$, $p<.001$) and the main effect of Age ($F(4, 406)=15.22$, $p<.001$) and Emotion ($F(1, 406)=3805.65$, $p<.001$) was significant.

Simple main effect analysis showed that voice responses were different among ages both on incongruent ($F(4, 406)=11.60$, $p<.001$) and congruent ($F(4, 406)=4.40$, $p=.002$) trials. Shaffer's post hoc t-tests revealed that 5-6-year-olds' and 7-8-year-olds' voice responses were less than those of 9-10-year-olds and 11-12-year-olds. Moreover, 9-10-year-olds' voice responses were less than those of 11-12-year-olds in incongruent trials. Moreover, adults' voice responses were less than those of 11-12-year-olds. Overall, voice responses increased as development during childhood, while it declined at adults. Table 1 shows voice responses in each combination (congruent trials: AngryFace / AngryVoice, HappyFace / HappyVoice; incongruent trials: AngryFace / HappyVoice; HappyFace / AngryVoice) of audiovisual information. In congruent trials, 5-6-year-olds' voice responses were less than those of 9-10-year-olds and adults.

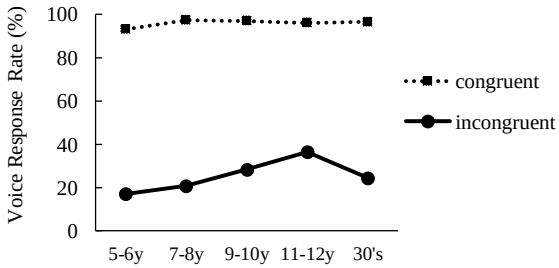

Figure 1 The developmental change in voice response rates in Emotion perception task

Table 1 Voice response rates in each combination of audiovisual information in Emotion perception task (%)

	Congruent		Incongruent	
	AngryFace/ AngryVoice	HappyFace/ HappyVoice	AngryFace/ HappyVoice	HappyFace/ AngryVoice
5-6y	92.5	93.6	15.2	19.1
7-8y	98.0	96.6	18.1	23.5
9-10y	98.9	95.0	18.8	37.9
11-12y	98.9	93.1	24.2	48.8
30's	99.6	93.3	8.7	40.1

[Phoneme perception task]

*McGurk trial*s As for responses of McGurk trials, we regarded participants' /ka/, /ta/, /pa/ responses as visual, fusion, and auditory responses, respectively. We showed participants' fusion responses in Figure 2. The result of one-way ANOVA on fusion responses showed that the main effect of Age was significant ($F(4, 406)=27.92$, $p<.001$). Shaffer's post hoc t-test revealed that adults' responses were influenced by lip movement more than those of children ($ps<.001$). There was no significant difference among age groups of children.

Congruent trials As for congruent trials, one-way ANOVA revealed that the main effect of Age on correct responses was significant ($F(4, 406)=4.05$, $p=.003$). Shaffer's post hoc t-tests revealed that 5-6-year-olds' correct responses were less than those of 11-12-year-olds and adults.

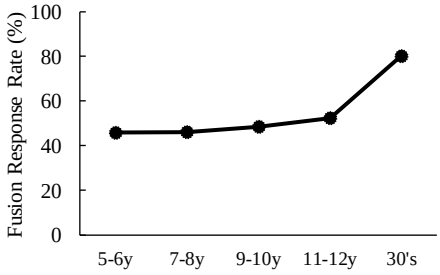

Figure 2 The developmental change in the occurrence of the McGurk effect (fusion response rates) in Phoneme perception task

Thus, the results showed that the developmental path of the acquisition of integration of audiovisual information is different between emotion perception and phoneme perception. The emotion judgement weighting on voice information increased during the childhood, whereas the developmental change in audiovisual phoneme perception was observed only between children and adults.

One interesting result was that adults' voice response rate in emotion perception declined compared with 11-12-year-olds. There are two possibilities of this change. One possibility is that their decline was due to aging. Another possibility is that their multisensory emotion perception come to be similar to their children through communication with them considering that all adult participants in Experiment 1 were parents of child participants. In order to explore this possibility, we examined whether parenting experience affected in multisensory perception in Experiment 2.

3. Experiment 2

3.1. Method

Participants were Japanese adults ages 30 to 39 years old including 30 parents (*M*=36.5; *SD*=1.7) and 23 people who had no their own children (non-parents: *M*=34.7; *SD*=2.8). Stimuli and procedure were the same with Experiment 1.

3.2. Results and Discussion

[Emotion perception task] We performed a *Group* (parents, non-parents) ×*Emotion* (congruent, incongruent) mixed-factor ANOVA on the voice responses. Results are shown in Figure 3 and Table 2. The interaction ($F(1, 51)= 5.25$, $p=.026$) and the main effect of Emotion ($F(1, 51)= 419.44$, $p<.001$) was significant. Simple main effect analysis showed that parents' voice responses were marginally significantly lower than non-parents ($F(1, 51)= 3.12$, $p=.083$) in incongruent trials. As for congruent trials, parents judged speakers' emotion more correctly than those of non-parents ($F(1, 51)= 5.32$, $p=.025$).

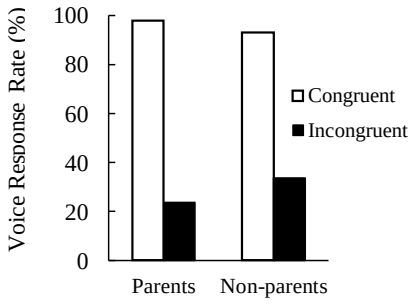

Figure 3 Parents and non-parents' voice response rates in Emotion perception task

Table 2 Parents and non-parents' voice response rates in each combination of audiovisual information in Emotion perception task (%)

	Congruent		Incongruent	
	AngryFace/ AngryVoice	HappyFace/ HappyVoice	AngryFace/ HappyVoice	HappyFace/ AngryVoice
Parents	98.8	97.1	7.9	39.6
Non-parents	98.9	87.0	13.0	54.3

[Phoneme perception task]

McGurk trials As for responses in McGurk trials, the result of one-way ANOVA on fusion responses showed that the main effect of Group was not significant ($F(1, 51)<.001$, $p=.987$). The frequency of the occurrence of the McGurk effect was not different between parents and non-parents (Figure 4).

Congruent trials One-way ANOVA revealed that the main effect of Group on correct responses in congruent trials was marginally significant ($F(1, 51)=3.88$, $p=.054$).

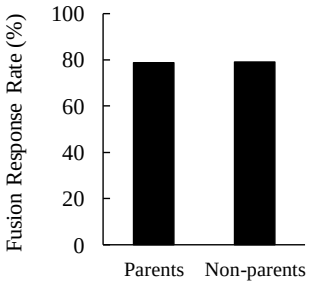

Figure 4 Parents and non-parents' occurrence of the McGurk effect (fusion response rates) in Phoneme perception task

4. General Discussion

The two experiments showed that Japanese people's multisensory emotion and phoneme perception develops along different paths.

In emotion perception, Japanese children shifted the cue from speakers' facial expression to vocal expression during their childhood. This developmental path is consistent with the previous study with Japanese children [3]. Although facial expression is more salient than vocal expression [7], Japanese people tend to express their emotions (especially the negative ones) in their face less clearly than Western people [8]. For this reason, voice can be useful information for Japanese people to recognize others' real intention precisely. Japanese children may learn this cultural manner in communication and come to interpret others' emotion taking emotional voice into account in addition to the facial expression.

Considering these children's results and the results of previous research [2], in which Japanese adult participants (21-29 years-old) were influenced by vocal expression more than Dutch participants, adult participants' judgement in Experiment 1 of the present study seems to be odd. Their judgement was influenced by facial expression more than 11-12 year-olds; adult participants' emotion perception was rather similar to younger children. The results of Experiment 2 showed that parents tend to judge speakers' emotion weighting facial expression more than non-parents. Taken together, the way of adults' judgement may be modified by their daily interaction with children. In order to reveal this mechanism, further experiments with various age groups are needed.

In contrast to the developmental path of emotion perception, the developmental change in induction of the McGurk effect was not observed during childhood. This is consistent with the Sekiyama and Burnham's study [6] demonstrating that English speakers' visual influence increased over age during their childhood, while such developmental change was not observed in Japanese-speaking children. Nevertheless, the results of present study differ from Sekiyama and Burnham's study in that increase of visual influence was

clearly observed between children and adults. This difference may be due to methodological differences employed in these studies. In the present study, a fixation point and a signal sound was followed by audiovisual stimuli. Moreover, we used a headphone in presentation of voice, while Sekiyama and Burnham's study used a loudspeaker. Since these ways of presenting can induce participants to fixate visual information stronger than in the previous study, the developmental change between children and adults might be clearly observed. Other possibility is that the difference in ages of adult participants may have influenced on the results. The age of adult participants of Sekiyama and Burnham's study was from 18 to 29. This is younger than those of the present study. Other study reported that Japanese people aged from 60 to 65 years were more strongly influenced by visual information in phoneme perception than Japanese people aged from 19 to 21 years [9]. Based on these studies, it is possible that participants in the present study, ranging from 30 to 39 years old, was affected by visual information more strongly than adults in the twentieth. In any case, Japanese people's audiovisual phoneme perception appears to develop later than emotion perception. Additionally, the difference between parents and non-parents was not observed in the induction of the McGurk effect, as shown in Experiment 2. These results suggest that experience of interaction with children have less impact on multisensory phoneme perception than emotion perception.

Thus, the present study showed that the developmental path of multisensory perception is different between emotion and phoneme perception both in ages and the effect of people's parenting experience. This may reflect possible differences in the integration processes between these tasks.

First, communication manners may have more impact on the process of emotion than phoneme perception. Since the way of expressing emotion is culturally various [8], emotion judgment is dependent on display rules in perceivers' environment. Then, the developmental change in multisensory emotion perception may be strongly related with the acquisition of communication manners. Moreover, this can explain the effect of parenting experience on emotion perception. Children who are acquiring communication manners of their culture may express their emotion differently from adults. For this reason, adults interacting with children may attune their criteria of emotion judgment to children's rules. In contrast to emotion perception, phoneme perception seems to be irrelevant with such communication manners. Phoneme perception itself does not allow of interpretation of ambiguous stimuli such as emotion. For this reason, multisensory phoneme perception may not be affected by participants' parenting experience.

Second, in Phoneme perception task, it is necessary to focus on auditory information, whereas in Emotion perception task, the experimenter did not instruct to pay attention to specific modality. Then, the integration in emotion perception seems to be based on either reliable modality, while it is the process of adding visual information to auditory information in phoneme perception. This task demand may lead to the difference in the developmental change.

Third, the development of unimodal perception may differ between emotion and phoneme perception. As for visual perception, the lipreading skill needed in phoneme perception and the ability to distinguish facial expression may be totally different. Similar difference can be pointed in auditory perception. Phoneme is segmental vocal information, whereas emotion is mostly expressed in a suprasegmental aspect of voice. Thus, it is worth investigating the relation between unimodal perception skill and multisensory emotion and phoneme perception in Japanese speakers in future research.

The present study showed different developmental paths in audiovisual emotion and phoneme perception, suggesting that their integration processes are different. However, even though the process is different, it is possible that the perceivers' habits in communication such as fixating other's face affect multisensory perception both in emotion and phoneme. This possibility and the concrete processes of two multisensory perception need to be examined in the further study.

5. Acknowledgements

We thank participants in the experiments. We also thank Miraikan staffs and volunteer staffs from Tokyo Women's Christian University for assistance with data collection. This work was supported by Strategic Information and Communications R & D Promotion Program (SCOPE) (No. 10210311) from the Ministry of international affairs and communications of Japan, and by JSPS KAKENHI (No. 15H02714).

6. References

[1] B. de Gelder and J. Vroomen, "The perception of emotions by ear and by eye," Cognition & Emotion, vol.14, no.3, pp.289-311, 2000.

[2] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder, "I feel your voice. Cultural differences in the multisensory perception of emotion," Psychological Science, vol. 21, no. 9, pp. 1259-1262, 2010.

[3] M. Kawahara, D. A. Sauter, and A. Tanaka, "Development of cultural differences in emotion perception from face and voice," Poster presented at 31st International Congress of Psychology, 2016.

[4] H. McGurk, and J. McDonald, "Hearing lips and seeing," Nature, vol. 264, pp. 746–748, 1976.

[5] K. Sekiyama and Y. Tohkura, "McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility," Journal of Acoustic Society of America, vol. 90, no.4, pp. 1797–1805, 1991.

[6] K. Sekiyama and D. Burnham, "Impact of language on development of auditory-visual speech," Developmental Science, vol. 11, no. 2, pp. 306-320, 2008.

[7] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder, "Cross-cultural differences in the multisensory perception of emotion", Poster presented at International Conference on Auditory-Visual Speech Processing, 2010.

[8] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," Science, vol. 164, no. 3875, pp.86-88, 1969.

[9] K. Sekiyama, T. Soshi, and S. Sakamoto, "Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults", Frontiers in Psychology, vol.5, no.323, 2014.