Impact of Culture on the Development of Multisensory Emotion Perception

Misako Kawahara[1*], *Disa Sauter*[2], *Akihiro Tanaka*[1]

[1]Tokyo Woman's Christian University, Japan
[2]University of Amsterdam, The Netherlands

* misako1753@gmail.com

Abstract

Recent studies have demonstrated that multisensory emotion perception is modulated by culture. Tanaka et al. (2010) showed that Japanese people are more tuned than Dutch people to vocal processing in adults. The present study investigated how such a cultural difference develops in children aged 5-12 years. In the experiment, a face and a voice, expressing either congruent or incongruent emotions, were presented simultaneously on each trial. Participants judged whether the person is happy or angry. The results showed that the rate of vocal responses was higher in Japanese than Dutch in adults, especially when in-group speakers expressed a happy face with an angry voice. The rate of vocal responses was low in both Japanese and Dutch 5-6-year-olds, while it increased over age only in Japanese participants. These results suggest that combinations of facial and vocal emotions have specific meanings and that culture-specific multisensory display rules are acquired with age in childhood.

Index Terms: emotion perception, cultural difference, development

1. Introduction

The purpose of this study is to investigate the development of cultural differences in multisensory emotion perception from face and voice. Recent studies have demonstrated cultural differences in the manner of multisensory emotion perception [1-3]. We examine how people acquire these cultural specificities in multisensory emotion perception.

In speech communication, we use both visual and auditory information. The most famous phenomenon is known as the McGurk effect, in which the voice /ba/ dubbed onto the face movements for /ga/ are perceived as /da/ [4]. This phenomenon shows that auditory (or visual) perception cannot be separated from visual (or audio) perception.

In speech communication, the affective information as well as phonemic information plays an important role. Audiovisual integration also occurs in emotion perception. However, previous studies on emotion perception have mainly focused on either facial or vocal expression as a cue to emotion. We in most social situations do not interact with 'faceless' voices or 'voiceless' faces. When facial expression is congruent with vocal expression, emotion is perceived more easily than when presented alone. Sometimes facial and vocal expressions may be incongruent (as in the case of the McGurk effect in phoneme perception). In that case, facial and vocal information may interfere with each other.

Thus, we receive multimodal information conveyed by face and voice after birth [5]. Previous research has shown that infants by the age of 7 months are able to detect common emotion across modalities [5]. Other study has shown that 9 years children exhibit the crossmodal integration effect, although 5- and 7-year-olds do not when they are presented with incongruent emotional audiovisual information [6]. The results suggest that the age of 9 years is considered a developmental turning point, implying changes in the use of different emotional social cues. These results lead to the question of how the role of the postnatal environment influences the manner of multisensory emotion perception.

In audiovisual speech perception, the developmental onset of inter-language differences have been examined in the McGurk effect. Japanese speakers are less subject to visual influence in the McGurk effect than English speakers [7], and these inter-language differences are displayed with age during childhood [8]. These studies suggest that Japanese speakers tend to be less influenced by lip movement than English speakers in phoneme perception.

Previous studies also have demonstrated that Japanese people tend to be influenced by visual information less than Western people in emotion perception [1-3]. Tanaka and colleagues [1] investigated cultural differences between Japanese and Dutch participants in how multisensory information (i.e., facial and vocal expressions) are integrated. The results demonstrated that Japanese people weighted cues in voices more than Dutch people did, whereas Dutch people weighted cues in faces more than Japanese people did. These results show that multisensory integration of affective information is modulated by perceiver's cultural background in adults. Based on the previous studies in phoneme perception [8], we can speculate that cultural specificities displayed by adults emerge gradually with development in integrating emotional cues from different modalities. No previous research has investigated the age at which cultural differences appear in multisensory emotion perception from face and voice.

In the current study, we conducted experiments using similar procedure to Tanaka and colleagues [1] for Japanese and Dutch adults and children to examine whether cultural differences appear with age in emotion perception from face and voice. In the experiment, a video clip with voice containing two opposing emotional face and voice (happiness and anger) were presented on each trial. Congruent and incongruent emotional expressions were presented (e.g., a happy face was presented with an angry voice, in the incongruent case). Participants judged the emotion of the speaker as happiness or anger. Since we are interested in the cultural and developmental differences in the degree of focusing on vocal emotion, we calculated *vocal responses* for incongruent expressions. For example, when participants observed happy face with angry voice and responded as anger, the responses were categorized as vocal responses. We compared the percentages of vocal responses in incongruent trials and examined developmental changes in Japanese and Dutch participants.

2. Experiment 1

In Experiment 1, we conducted experiments for university students in Japan and in the Netherlands to examine cultural differences in emotion perception from face and voice. On the basis of findings in Tanaka and colleagues [1], we expected that Japanese participants would weight cues in voices more than Dutch participants.

2.1. Methods

2.1.1. Participants

Thirty-three native speakers of Japanese living in Japan (ages 18–32 years; 16 male, 17 female) and 38 native speakers of Dutch living in the Netherlands (ages 19–30 years; 19 male, 19 female) participated. All Japanese participants had never lived in foreign countries, while we could not control the living history for Dutch participants. All participants had normal hearing and normal or corrected-to-normal vision.

2.1.2. Stimuli

The stimuli were created from simultaneous audio and video recordings of Japanese and Dutch speakers' emotional utterances. Four short fragments with neutral linguistic meaning were uttered by two Japanese and two Dutch female speakers in their native language. Each fragment had an equivalent meaning between Japanese and Dutch. For example, a fragment "Kore nani?" ("What is this?" in English) was uttered by Japanese speakers while an equivalent fragment "Hey, wat is dit?" was uttered by Dutch speakers. Each fragment was uttered with happy or angry emotion. The audio was recorded at a sampling frequency of 48000 Hz. The speaker's visual expressions were recorded from the top of her head to the shoulder. The video frame rate was 29.97 frames per second.

Congruent and incongruent stimuli were created from the original audiovisual fragments. The unchanged fragments served as congruent stimuli. In order to make incongruent stimuli, happy and angry facial expressions were combined with happy and angry vocal expressions for each of the eight utterances in each language (two speakers × four fragments), resulting in a total of 32 bimodal stimuli (16 congruent and 16 incongruent) in each language.

The unimodal stimuli were also created from audio or video recordings of speakers' utterances, resulting in a total of 16 unimodal stimuli (two speaker × four fragments × two language) in each sensory modality (audio only or visual only).

2.1.3. Procedure

Participants were tested in a quiet room. The room was divided into three booths by partitions. Experiments were done for 1 to 3 participants simultaneously. The experiment consisted of four sessions and began with two multisensory sessions, in which speakers (Japanese and Dutch) were different, followed by two unisensory sessions, in which only the faces or voices were presented. A trial consisted of a 1-s fixation point around the speakers' mouth and a simultaneous presentation of dynamic face (happiness or anger) and voice (happiness or anger). The face was displayed on the PC monitor (Dell Latitude3540 in Japan; Dell optiplex 9010 in the Netherlands) and the voice was presented binaurally via headphones (SONY MDR-ZX660 in Japan; IMGStage Line MD-5000DR in the Netherlands) at a comfortable listening level. The order of the session within unisensory and multisensory sessions was counterbalanced between participants. Thus, both Japanese and Dutch participants observed both Japanese and Dutch targets. In all sessions, participants were instructed to categorize the emotion of the speaker in the movies as happiness or anger. No instruction was given on which modality participants were to pay attention to. Therefore, there was no correct answer in incongruent stimuli. Participants responded by pressing either of the two buttons, which were counterbalanced between participants. Each session had 32 trials. In the unisensory sessions, 16 Japanese and 16 Dutch speaker's stimuli were included in a single session.

2.2. Results

Accuracy on the unisensory sessions was high. Participants overall judged emotion from vocal expressions (85.0%) and from facial expressions (96.8%).

In this article, analyses focused on the responses on the incongruent trials in the multisensory sessions (i.e., vocal responses). To examine cultural differences in vocal responses, we performed Group (Japanese and Dutch) × Stimuli (Japanese happy face + angry voice, Japanese angry face + happy voice, Dutch happy face + angry voice, Dutch angry face + happy voice) mixed-factor analysis of variance (ANOVA) on mean vocal responses in the incongruent conditions. The main effect of Group was significant [F (1, 69) = 8.03, p = .006, η^2 = .10]. Results showed that the mean vocal responses were higher in Japanese (19.8%) than in Dutch (10.9%) (Figure 1). These results are consistent with our previous study, which showed that Japanese are more attuned to vocal expressions than Dutch by using cross-modal bias paradigm [1]. The main effect of Stimuli [F (3, 207) = 24.67, p < .001, η^2 = .26] was also significant. More importantly, the interaction was significant [F (3, 207) = 35.39, p < .001, η^2 = .34]. Simple main-effects analyses showed that the vocal responses were not different in Dutch participants [F (3, 207) = 1.34, p = .26], but were different in Japanese participants [F (3, 207) = 54.94, p < .001]. Bonferroni a posteriori tests showed that the vocal responses were especially high in the combination of Japanese happy face and angry voice in Japanese participants (Figure 2). These results suggest that the degree of focusing on voice was varied by the combination of facial and vocal emotions in Japanese participants. Specifically, when they were presented with Japanese happy faces with angry voices they tend to interpret expressions as anger.

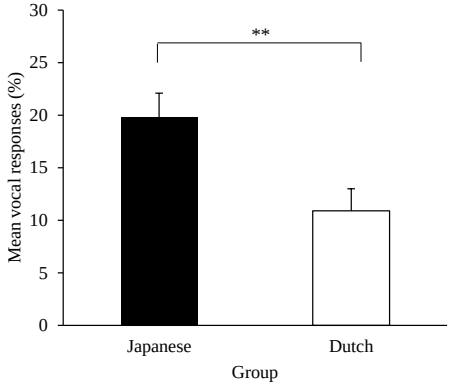

Figure 1: *Vocal responses (mean percentage of vocal responses in the incongruent conditions) in Japanese and Dutch participants. Error bars represent standard errors.*

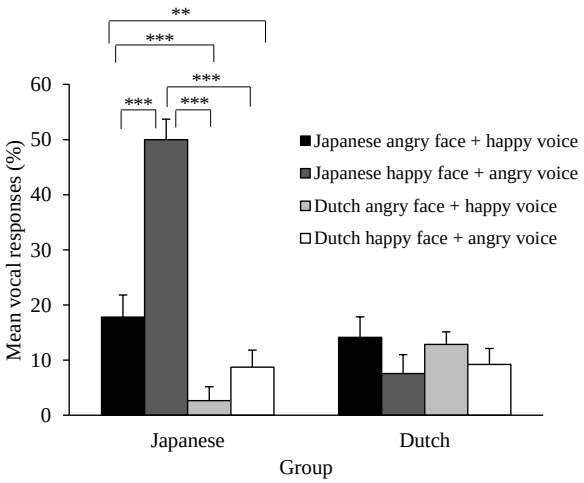

Figure 2: *Vocal responses for each type of stimuli in Japanese and Dutch participants. Error bars represent standard errors.*

3. Experiment 2

To examine whether cultural differences in adults are observed in children, Japanese and Dutch children of 5 and 6 years were tested and the results were compared with those of the adults in Experiment 1. Based on the previous research that inter-language differences were not found at 6 years old yet in the McGurk effect [8], we predicted that cultural differences are not observed at 5 and 6 years old. Based on the findings that accuracy of vocal emotion recognition was lower in younger children than older children [9], we predicted that both cultural groups mainly judge speakers' emotion from facial expressions.

3.1. Methods

3.1.1. Participants

Besides the 33 Japanese and the 38 Dutch adults described in Experiment 1, Japanese children living in Japan (57 children: 28 boys, 29 girls) and Dutch children living in the Netherlands (11 boys, 20 girls, means age 5.53 years) between the ages 5 and 6 years participated. All Japanese children had never lived in foreign countries, while we could not control the living history for Dutch childrens. All participants had normal hearing and normal or corrected-to-normal vision.

3.1.2. Stimuli and Procedure

Procedure was almost identical to that in Experiment 1. Children were tested in a quiet room. The room was divided into six booths by partitions in Japan. Experiments were conducted for 1 to 6 participants simultaneously (but instruction was given individually) for Japanese children. While the adults took about 20 min to complete the experiment, young children needed more time (about 30 min). The experiments for Dutch children were conducted individually at University of Amsterdam.

3.1.3. Results

Children as young as 5-6 years were proficient in judging facial expressions (92.6%). In contrast, accuracy in the voice-only session (67.3%) was lower than that of face-only session at 5-6 years old.

To examine developmental changes in the pattern of vocal responses, we conducted separate Age (5-6 years or adult) × Group (Japanese or Dutch) ANOVA on mean vocal responses in each type of incongruent stimuli.

In Japanese happy face with angry voice condition (Figure 3-a), the main effects of Age [F (1, 155) = 39.10, $p < .001$, $\eta^2 = .20$] and Group [F (1, 155) = 54.07, $p < .001$, $\eta^2 = .26$] were significant. Importantly, the interaction was significant [F (1, 155) = 41.12, $p < .001$, $\eta^2 = .21$]. The simple main effect of Group was significant in adult [F (1, 155) = 89.04, $p < .001$, $\eta^2 = .37$], but was not significant in 5-6 years [F (1, 155) = 0.47, $p = .49$, $\eta^2 = .003$]. These results showed that cultural differences seen in adults were not found in 5-6 years old. The simple main effect of Age was significant in Japanese [F (1, 155) = 89.16, $p < .001$, $\eta^2 = .40$], but not in Dutch participants [F (1, 155) = 0.01, $p = .91$, $\eta^2 < .001$]. These results demonstrate that only Japanese shift their attention from facial to vocal expressions with age.

In Japanese angry face with happy voice condition, the main effects of Age [F (1, 155) = 0.42, $p = .52$, $\eta^2 = .003$], Group [F (1, 155) = 0.73, $p = .40$, $\eta^2 = .005$]), and the interaction [F (1, 155) = 0.01, $p = .91$, $\eta^2 < .001$] were not significant (Figure 3-b).

In Dutch happy face with angry voice condition, the main effects of Age [F (1, 155) = 1.34, $p = .25$, $\eta^2 = .009$], Group [F (1, 155) = 0.51, $p = .48$, $\eta^2 = .003$]), and the interaction [F (1, 155) = 0.25, $p = .62$, $\eta^2 = .002$] were not significant (Figure 3-c).

In Dutch angry face with happy voice condition (Figure 3-d), the main effect of Group [F (1, 155) = 5.40, $p = .021$, $\eta^2 = .034$] and the interaction [F (1, 155) = 5.67, $p = .018$, $\eta^2 = .04$] were significant. The main effect of Age was not significant [F (1, 155) = 0.32, $p = .57$, $\eta^2 = .002$]). Simple main-effects analyses showed that adults' vocal responses were slightly but significantly higher than those of children in Dutch participants [F (1, 155) = 3.95, $p = .05$, $\eta^2 = .025$], but were not different significantly from those of children in Japanese participants [F (1, 155) = 1.83, $p = .18$, $\eta^2 = .01$].

These findings suggest that young children focus on facial expressions universally. Importantly, our results demonstrate that Japanese participants shift their attention from facial to vocal expressions, though Dutch people do not. One of the interpretations to the results is that Japanese participants are more accurate in perceiving angry voice of own language than Dutch participants, especially from 7-9 years. To examine this possibility, we compared the accuracy for the angry voice of own language between Japanese people and Dutch people. If we explain the cultural differences in vocal responses by the cultural differences in accuracy for the angry voice of own language, the accuracy should have be different between Japanese and Dutch participants in adults, but not in 5-6 years.

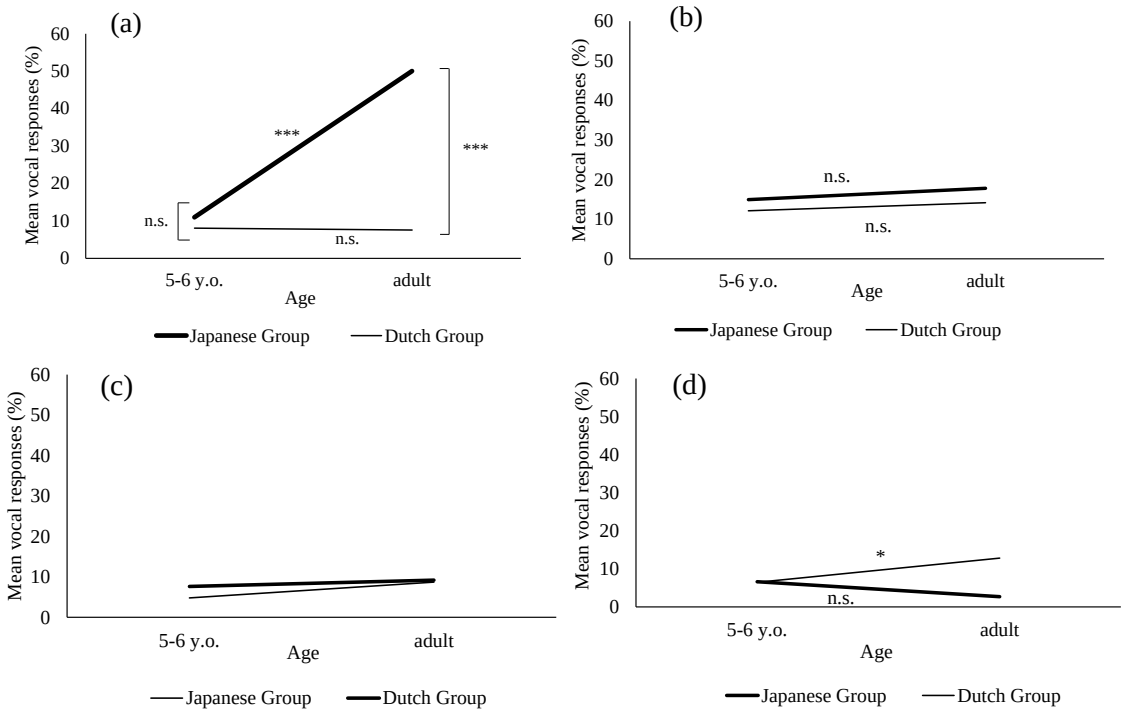

Figure 3: *(a) Vocal responses for Japanese happy face with angry voice in Japanese and Dutch participants. (b) Vocal responses for Japanese angry face with happy voice in Japanese and Dutch participants. (c) Vocal responses for Dutch happy face with angry voice in Japanese and Dutch participants. (d) Vocal responses for Dutch angry face with happy voice in Japanese and Dutch participants.*

However, our results were contrary to the expectations. Results showed that accuracy of Japanese participants (74.8%) was higher than those of Dutch participants (59.3%) in 5-6 years [F (1, 155) = 14.85, $p < .001$, $\eta 2 = .087$], but was not different significantly in adults (Japanese = 92.4%; Dutch = 89.3% [F (1, 155) = 0.37, $p = .54$, $\eta 2 = .002$]). These findings suggest that we cannot explain the development of cultural difference by accurate in perceiving angry voice of own language.

4. Experiment 3

In Experiment 3, we examined developmental changes of the emotion perception in Japanese participants. We additionally conducted experiments for Japanese children between 7 and 12 years and the results were compared with those of the Japanese adults in Experiment 1 and children between 5 and 6 years in Experiment 2. We expected that Japanese participants gradually shift their attention from facial to vocal expressions.

4.1. Methods

4.1.1. Participants

Besides the 33 Japanese adult described in Experiment 1 and the 57 Japanese children between the ages of 5 and 6 years described in Experiment 2, 122 Japanese children between 7 and 12 years living in Japan participated. All Japanese children had never lived in foreign countries. All participants had normal hearing and normal or corrected-to-normal vision. The children were divided into three groups, 5-6 years, 7-9 years (56 children: 37 boys, 19 girls) and 10-12 years (66 children: 39 boys, 27 girls).

4.1.2. Stimuli and Procedure

Procedure was identical to those in Experiment 1 and 2 for Japanese participants.

4.2. Results

To examine developmental changes of vocal responses, we performed Speaker (Japanese or Dutch) × Combination (happy face with angry voice, or angry face with happy voice) × Age (5-6 years, 7-9 years, 10-12 years, or adults) ANOVA on mean vocal responses in the incongruent conditions. Importantly, the three-way interaction [F (3, 208) = 17.09, p < .001, $\eta 2 = .20$] was significant. The main effects of Speaker [F (1, 208) = 171.39, $p < .001$, $\eta^2 = .45$], Combination [F (1, 208) = 66.18, $p < .001$, $\eta^2 = .24$], Age [F (3, 208) = 6.14, $p < .001$, $\eta^2 = .08$.], the two-way interactions between Speaker × Age [F (1, 208) = 13.02, $p < .001$, $\eta^2 = .16$], Combination × Age [F (3, 208) 18.41, $p < .001$, $\eta^2 = .21$], and Speaker × Combination [F (1, 208) = 74.09, $p < .001$, $\eta^2 = .26$] were also significant.

To examine developmental changes in the pattern of vocal responses, we conducted subordination analysis for the three-way interaction. ANOVA showed that vocal responses increased during childhood only in Japanese happy face with angry voice condition (F (3, 208) = 26.23, $p < .001$, $\eta^2 = .27$). Bonferroni a posteriori tests showed that vocal responses were different among all age groups except between 10-12 years and adults. These results show that Japanese children are gradually attuned to voices with age, only when they observe Japanese happy face with angry voice (Figure 4). We speculate that this

finding may be linked to the acquisition of emotional display rules in Japanese.

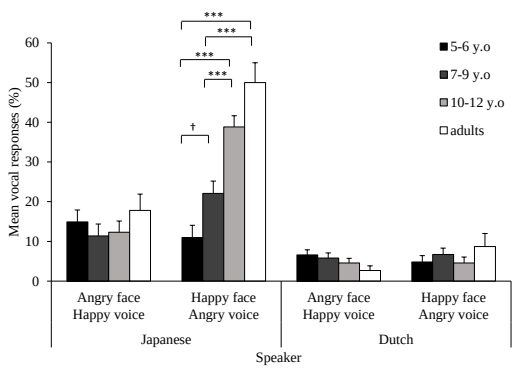

Figure 4: *Vocal responses for each type of stimuli among Japanese age groups. Error bars represent standard errors.*

5. Discussion

Our findings provide the first evidence that cultural differences appear developmentally in multisensory emotion perception from face and voice. Our results demonstrate that young children focus on facial expressions universally. Importantly, we found that Japanese participants gradually shift their attention from facial to vocal expressions during childhood, whereas Dutch participants do not shift their attention to facial expressions.

The vocal responses were not different between 5-6 years children and adults in Dutch participant, but increased with ages in Japanese participants. These results demonstrate the development of cultural specificities in multisensory integration of affective information. Interestingly, Japanese adults focused on vocal emotion when they observed Japanese happy face with angry voice, but not in other types of combination of face and voice. Since the vocal responses were different between 5-6 years and 7-9 years in Japanese participants, we speculate that Japanese children get more attuned to vocal emotion than Dutch children around this age. Our results are in line with the findings that Japanese speakers use visual information less than English speakers do in auditory-visual speech perception [7] and that this inter-language differences emerge between 6 and 8 years [8].

Our results can be interpreted within the framework of culture-specific display rules and decoding rules [10, 11]. Display rules are the culturally shared norms of how, when, and where to show emotions. Decoding rules are culturally prescribed rules that manage the perception and interpretation of other's emotional expressions. Our findings may reflect the differences of display and decoding rules in multisensory emotion perception between Japanese and Western cultures. In the present study, vocal responses were higher in Japanese than Dutch in adults, especially when Japanese speakers expressed a happy face with an angry voice. This expression can be seen as the state of masking or concealing their anger with smile. These findings may be linked to the fact that Japanese participants smile to mask their negative emotion when another person is present [10]. We propose to extend the existing theories of display and decoding rules in facial expressions [10, 11] to multisensory affective expressions from face and voice. We suppose that there are display rules hiding negative emotion by smile and decoding rules reading the intentions from voice in Japanese culture. In this study, Japanese participants paid attention to voices, especially when they were presented with Japanese expressions (i.e., in-group speakers' emotional expressions). These results imply that Japanese participants do not pay attention to voices when they judge emotional expressions of out-group members, and that they judge those of in-group members using culture-specific rules shared within Japanese culture.

One of the alternative interpretations to the results is that Japanese participants are more sensitive to anger of other person than Dutch participants. East Asian people values the social harmony more than Western people [12]. Japanese participants may maintain the harmony by being more sensitive to the anger of other person. This negativity bias can modulate the emotion perception in Japanese participants. In this study, Japanese participants tend to interpret expressions as anger when they were presented with Japanese happy faces with angry voices. They also tend to interpret expressions as anger when they were presented with Japanese angry faces with happy voices. These results can be explained by the negativity bias. However, we cannot explain the results that Japanese participants tend to interpret expressions as happiness when they observed Dutch happy faces with angry voices. If we explain the cultural differences by the negativity bias, Japanese participants should have focused on angry cues regardless of culture of speakers, but these were not shown in our results. These findings suggest that Japanese participants are not always sensitive to anger. Rather, our findings imply that Japanese participants use the culture-specific rules only when they judge emotional expressions of in-group members.

6. Conclusions

In conclusion, our results revealed the development of cultural differences in multisensory emotion perception. We speculate that these findings may be linked to the cultural specificities of display and decoding rules. At the moment, it is not clear whether Japanese people express their feeling as happy face with angry voice in natural settings. Further study is necessary to examine the link between the pattern of expression and perception in multisensory emotional communication.

7. Acknowledgements

We thank all 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).

8. References

[1] 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*, **21**, 1259-1262, 2010.

[2] P. Liu, S. Rigoulot, M. D. Pell, "Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence," *Neuropsychologia*, **67**, 1-13, 2015.

[3] P. Liu, S. Rigoulot, and M. D. Pell, “Cultural differences in on-line sensitivity to emotional voices: comparing East and West,” *Frontiers in Human Neuroscience*, **9**, 311, 2015.
[4] H. McGurk, and J. McDonald, “Hearing lips and seeing,” Nature, vol. 264, pp. 746–748, 1976.
[5] T. Grossmann, “The development of emotion perception in face and voice during infancy,” *Restorative Neurology and Neuroscience*, **28**, 219–236, 2010.
[6] S. Gil, J. Hattouti, and V. Laval, “How children use emotional prosody: Crossmodal emotional integration?” *Developmental psychology*, **52**, 1064, 2016.
[7] 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 the Acoustical Society of America*, **90**, 1797-1805, 1991.
[8] K. Sekiyama, and D. Burnham, “Impact of language on development of auditory-visual speech perception,” *Developmental Science*, **11**, 306-320, 2008.
[9] D. A. Sauter, C. Panattoni, and F. Happé, “Children's recognition of emotions from vocal cues,” *British Journal of Developmental Psychology*, **31**, 97-113, 2013.
[10] P. Ekman, “Universals and cultural differences in facial expressions of emotion” In J. Cole (Ed.), *Nebraska symposium on motivation*. Lincoln: University of Nebraska Press, pp. 207-282, 1972.
[11] P. Ekman, and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*, Englewood Cliffs, NJ: Prentice-Hall, 1975.
[12] H. R. Markus, and S. Kitayama, “Culture and the self: Implications for cognition, emotion, and motivation,” *Psychological Review*, **98**, 224-253, 1991.