The Development of Eye Gaze Patterns during Audiovisual Perception of Affective and Phonetic Information

Hisako W. Yamamoto[1], *Misako Kawahara*[1, 2], *Akihiro Tanaka*[1]

[1]Tokyo Woman's Christian University
[2]Japan Society for the Promotion of Science
hisako_wy@lab.twcu.ac.jp

Abstract

In face-to-face communication, emotions, as well as phonemes, are perceived through the integration of visual information on the face and auditory information on the voice. Previous studies suggested that children integrate audiovisual information differently from adults. To uncover the mechanism behind this developmental process, we investigated the gaze patterns of Japanese children aged 5 to 12 and adults when looking at speakers' faces after being asked to judge their emotions (emotion perception task) or pronounced syllables (phoneme perception task). The results showed that the participants fixated longer on the speaker's eyes in the emotion perception task than they did in the phoneme perception task. Moreover, the participants' fixation on the speaker's eyes increased with age in emotion perception, while it decreased with age in phoneme perception. This finding suggests that children can shift their attention to a face depending on what they are required to judge, and that this attentional shift becomes more sophisticated with age. We discuss the development of audiovisual integration in terms of the relationship between the participants' gaze and perception.

Index Terms: multisensory perception, emotion, speech perception, gaze, cognitive development, McGurk effect

1. Introduction

The face is a very rich source of social information. We can identify people, detect the direction of their attention, and perceive emotions through observing others' faces. Moreover, even in auditory speech processing, we utilize information from the face, such as lip movement. Thus, looking at others' faces is a very important behavior in social interactions.

In particular, it is necessary to perceive others' emotions accurately for smooth communication. Emotion perception is achieved not only through facial expressions but also through vocal expressions. Moreover, the perception of facial expressions and that of vocal expressions are affected by each other [1]. Thus, emotion perception occurs through an audiovisual integration process. Previous studies have suggested that audiovisual integration in emotion perception differs across cultures and ages. For example, Dutch and Japanese 5–6-year-olds perceive emotion by focusing on facial expressions [2] when they perceive others' emotions, but only Japanese children come to take vocal expression into account more as they age [3]. This developmental path leads to cultural differences in adults in which Japanese adults put more weight on vocal expressions in audiovisual emotion perception than do Dutch adults [4].

What changes along the developmental path of audiovisual emotion perception? The investigation of gaze patterns when people look at faces may give us a clue, considering previous research on audiovisual phoneme perception. Phoneme perception is also achieved through the audiovisual integration of face and voice cues, as demonstrated by the McGurk effect [5], in which the speech sound /ba/ was played with lip movements for /ga/, then perceived as /da/. Audiovisual phoneme perception also develops during childhood. Previous studies reported that the McGurk effect is weaker in children than adults [3][5][6]; that is, children are less influenced by visual information (lip movement) than adults are. To investigate the factors leading to the differences in the McGurk effect, Irwin et al. examined gaze patterns when looking at a face during audiovisual phoneme perception. For example, one study demonstrated that the gaze on a face, especially on the mouth, increases from 5 to 10 years of age during phoneme perception [7]. Additionally, the literature on developmental disorders has debated whether audiovisual perception by children with autism spectrum disorders is related to their gaze on faces [8]. Considering these previous studies, it is possible that children's gaze patterns during audiovisual emotion perception change with age. Therefore, we focus on gaze patterns directed toward the face while participants are being presented with audiovisual stimuli in order to reveal the developmental process of audiovisual integration in the perception of affective and phonetic information.

The present study has three aims. First, we would investigate the development of gaze patterns during audiovisual emotion perception. Second, we would compare the gaze patterns during audiovisual emotion perception with those during audiovisual phoneme perception to examine whether gaze patterns are specific to emotion perception. Third, we would examine the relationship between each participant's gaze duration directed toward the eyes or mouth and their perception of emotions or phonemes.

We asked the participants (young children: 5–8-year-olds; older children: 9–12-year-olds; and adults) to judge speakers' emotions or what they said based on video clips and to respond orally, measuring their fixation on the speaker's face in the clips. Based on previous research analyzing participants' fixation at each timespan that corresponded to speech events [7][8], we divided the time scale of each video clip into three timespans (pre-speech, during-speech, and post-speech) to analyze the time course of the participants' fixation.

2. Methods

2.1. Participants

The participants were 62 children who spoke Japanese as their native language. They were divided into a younger group (32 children, mean age: 6 years 10 months; age range: 5–8 years) and an older group (30 children, mean age: 10 years 7 months;

10.21437/AVSP.2019-6

age range: 9–12 years). Additionally, we analyzed 17 adult native Japanese speakers' data as preliminary data[i] (mean age: 20 years 8 months; age range: 18–24 years). All participants were recruited at the National Museum of Emerging Science and Innovation (Miraikan) in Tokyo, Japan.

2.2. Apparatus

The Tobii T60 eye tracker was used along with the Tobii Studio software (version 3.2.1) to present audiovisual stimuli and collect eye tracking data.

Participants were seated in front of the eye tracker, at a distance of about 50 cm. Visual stimuli were displayed at the center of a 17-inch monitor with a resolution of 1024 × 768 pixels. The actual size of the visual stimuli on the display was about 28 × 21 cm. Audio stimuli were presented through headphones (HDA200, SENNHEISER) to mask the background noise in the experimental laboratory, at a comfortable listening level, which was adjusted using a headphone amplifier (DAC-HA200, ONKYO).

2.3. Stimuli and procedure

The experiment was conducted in the experimental laboratory in Miraikan. Prior to each task, the experimenter conducted a calibration of each participant's fixation to the display using the Tobii Studio software.

The experiment consisted of four sessions (Figure 1, emotion perception task - audiovisual session; emotion perception task - visual only session; phoneme perception task - audiovisual session; and phoneme perception task - visual only session). We conducted the visual only session to examine their perceptions and gaze patterns when they were presented with video clips without sound. The order of the sessions was the same among participants, while the order of the test trials of each session was randomized.

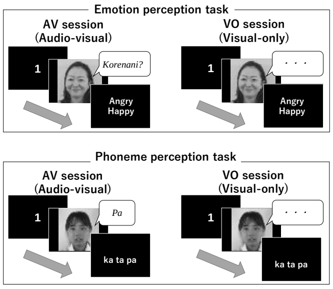


Figure 1. *The flow of a trial in each session*

(In the actual experiment, letters were presented in Hiragana)

2.3.1. Emotion perception task

Audiovisual (AV) session: The audiovisual stimuli were short video clips in which female native Japanese speakers expressed an emotion. In each movie, one speaker expressed happiness or anger through her face and voice. The linguistic information of the voice was emotionally neutral ("*Hai, moshimoshi*" [Hello], "*Sayonara*" [Good-bye], "*Korenani*" [What is this?], or "*Sounandesuka*" [Is that so?]). A total of 32 video clips (2 speakers × 4 emotions [angry face-angry voice/happy face-happy voice: emotions expressed through the face and voice were congruent; angry face-happy voice/happy face-angry voice: emotions expressed through the face and voice were incongruent] × 4 utterances) were used as test stimuli. Two other video clips were used as practice stimuli.

In each trial, a number was presented as a fixation point in order to draw the participants' attention. The fixation point was positioned on the speakers' nose. After 500ms from the onset of the presentation of the number, a video clip and the response alternatives (angry/happy) were presented successively (Figure 1). Subsequently, participants were asked to judge whether the speaker was happy or angry and to respond orally to prevent their gaze from leaving the display between each trial. After they responded, the next test trial began. A total of 32 test trials were conducted after 2 practice trials.

Visual only (VO) session: The visual stimuli were silent versions of the affectively congruent audiovisual stimuli used in the AV session. Thus, the number of visual stimuli and test trials in the VO session was 16 [2 speakers × 2 emotions (angry face, happy face) × 4 utterances]. The procedure was the same as that used in the AV session except that practice trials were not included.

2.3.2. Phoneme perception task

AV session: The audiovisual stimuli were short video clips in which male or female Japanese speakers pronounced one syllable (/ka/, /pa/, or /ta/). Eighteen congruent stimuli in which the lip movement and the sound were congruent (six speakers × three syllables) and six McGurk-type stimuli in which the lip movement /ka/ was combined with a /pa/ sound were created. Each McGurk-type stimulus was presented twice in all of the trials. To equalize the number of congruent trials with that of the incongruent trials, 12 video clips of 18 congruent stimuli were presented once. That is, the whole set of test trials consisted of 12 congruent trials and 12 incongruent (McGurk-type) trials. Another three congruent clips were used as practice stimuli, and the other three congruent clips were not presented in the AV session.

In each trial, a number was presented as a fixation point. The fixation point was positioned on the speaker's nose. After 500ms from the onset of the presentation of the number, a video clip and the response alternatives (ka/ta/pa) were presented successively (Figure 1). Subsequently, the participants were asked to judge whether the speaker pronounced /ka/, /ta/, or /pa/, and to respond orally. After they responded, the next test trial began. A total of 24 test trials were conducted, following 3 practice trials.

VO session: The visual stimuli were silent versions of the congruent audiovisual stimuli used in the AV session. Thus, a total of 18 trials were conducted in the VO session. The procedure was the same as that used for the AV session, except that practice trials were not included.

3. Results

3.1. Gaze data

Two specific areas of interest, the eye area and the mouth area, were established for analysis of the participants' eye gaze (Figure 2). The division line of the two areas was drawn over the top of each speaker's nose. Additionally, to examine the time course of the gaze, we divided each video clip into three time periods: the pre-speech, during-speech, and post-speech periods. We calculated the participants' rates of fixation on the eye area by dividing the eye area fixation duration with the sum of the eye area and mouth area fixation durations in each time window. That is, the rate of fixation on the mouth area can be calculated by subtracting that on the eye area from 100%. If participants fixated on the mouth area for a long time, their rates of fixation on the eye area would decrease.

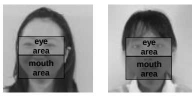


Figure 2. *A sample of a bitmap indicating areas of interest*

3.1.1. Emotion perception task

The gaze data of each task are presented in Figure 3.

[AV session] We performed a 3 (Age) × 3 (Time) × 4 (Emotion) mixed-factor analysis of variance (ANOVA) on the rates of fixation on the eye area during the AV session. The results showed a significant main effect of time (F(2, 142) = 54.76, p < .001) and emotion (F(3, 213) = 7.79, p < .001), and an interaction between age and time (F(4, 142) = 4.27, p = .003). The simple main effect of age was significant for the during-speech period (F(2, 71) = 3.95, p = .024). The post hoc analysis (Shaffer's modified sequentially rejective Bonferroni procedure) showed that adults fixated on the eye area longer than younger (p = .020) and older children (p = .027) in the during-speech period. Thus, all participants shifted their gaze from the eyes to the mouth when a speaker began to speak, but the amount of children's gaze shift was larger than that of adults.

The simple main effect of time was significant across all age groups (ps < .05). The post hoc analysis revealed that older children and adults' fixation on the eye area returned to the pre-speech level in the post-speech period (ps > .3), while young children's fixation in the post-speech period did not reach the level of the pre-speech period (p = .003). These results indicated that participants' gaze on the eye area recovered after the speech, but younger children's recovery was smaller than that of other groups.

Additionally, the interaction between time and emotion was marginally significant (F(6, 426) = 2.08, p = .055). The simple main effect of emotion was significant for the pre-speech period (F(3, 213) = 11.87, p < .001). The post hoc analysis showed that the participants gazed at the eye area of angry face-angry voice (Figure 3(a)) and angry face-happy voice people (Figure 3(c)) more than the happy face-happy voice (Figure 3(b)) and happy face-angry voice ones (Figure 3(d)) (ps < .05) in the pre-speech period; the participants fixated on the eye area longer while watching angry faces than happy faces. These results suggest that the gaze pattern is dependent on the facial expression rather than the emotional congruency.

[VO session] We performed a 3 (age) × 3 (time) × 2 (emotion) mixed-factor ANOVA on the rates of fixation on the eye area during the VO session. The results showed significant main effects of age (F(2, 71) = 4.50, p = .015) and time (F(2, 142) = 64.06, p < .001), and an interaction between age and emotion (F(2, 71) = 3.95, p = .024). The simple main effect of age was significant when participants were presented with an angry face (F(2, 71) = 3.30, p = .043; Figure 3(e)), and the post hoc analysis showed that the differences between adults and children were marginally significant (younger children-adults: p = .037; older children-adults: p = .076). The simple main effect of age was also significant when the participants were presented with a happy face (F(2, 71) = 5.60, p = .006; Figure 3(f)), and the post hoc test showed that younger children gazed at the eye area for less time than older children (p = .035) and adults (p = .007). These results indicated that the participants' fixation on the eye area increased with age, especially when they were presented with a happy face.

3.1.2. Phoneme perception task

[AV session] We performed a 3 (age) × 3 (time) × 2 (stimuli: congruent and McGurk-type) mixed-factor ANOVA on the rates of fixation on the eye area calculated as with the results of the emotion perception task. The results revealed a significant interaction between age and time (F(4, 142) = 3.65, p = .007), an interaction between time and stimuli (F(2, 142) = 3.22, p = .043), and a main effect of time (F(2, 142) = 23.23, p < .001). The main effect of age was marginally significant (F(2, 71) = 2.54, p = .086). The simple main effect of age was significant for the during-speech period (F(2, 71) = 4.10, p = .021), and the post hoc test revealed that younger children's fixation on the eye area was longer than that of older children (p = .040) and adults (p = .040); older children and adults gazed at the mouth area longer than younger children while watching a speaking face. Moreover, as for the pre-speech period, the simple main effect of age was marginally significant (F(2, 71) = 3.07, p = .053). Before the speech, older children (p = .095) and adults (p = .095) tended to fixate on the mouth area longer than younger children.

The simple main effect of stimuli was significant during the post-speech period (F(1, 71) = 7.91, p = .006). This finding

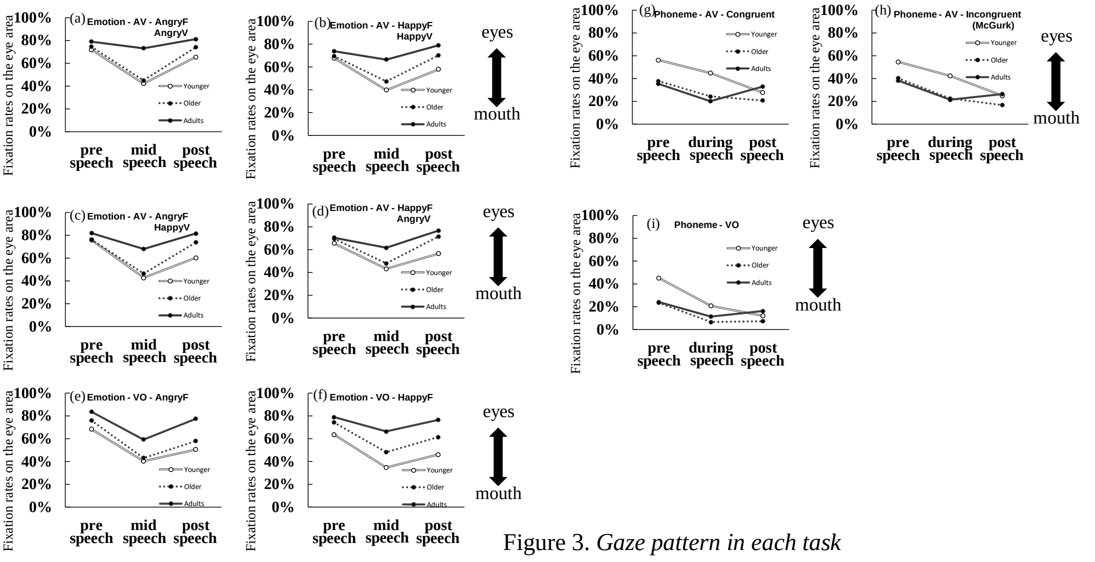


Figure 3. *Gaze pattern in each task*

indicates that the participants fixated on the mouth area for a longer duration when they were presented with McGurk-type stimuli than when they were presented with congruent stimuli (Figure 3 (g)(h)).

[VO session] A 3 (age) × 3 (time) mixed-factor ANOVA on the rates of fixation on the eye area revealed a significant interaction between age and time (*F*(4, 142) = 4.56, *p* = .002) and main effects of age (*F*(2, 71) = 4.23, *p* = .018) and time (*F*(2, 142) = 34.19, *p* < .001). The simple main effect of age was significant in the pre-speech (*F*(2, 71) = 6.58, *p* =. 002) and during-speech (*F*(2, 71) = 3.91, *p* = .024) periods. The post hoc analysis showed that older children (*p* = .004) and adults (*p* = .016) fixated on the mouth area for longer than younger children in the pre-speech period, and that older children fixated on the mouth area longer than younger children in the during-speech period (*p* = .022). These results indicated that the participants' fixation on the mouth before and during speech increased with age (Figure 3 (i)).

3.1.3. Comparison of the gaze between tasks

To compare the fixation patterns between tasks, we conducted a 3 (age) × 2 (task: emotion and phoneme) × 2 (session: AV and VO) mixed-factor ANOVA on the fixation rate on the eye area. The results showed that the interaction between age and task (*F*(2, 71) = 13.72, *p* < .001) and the main effects of task *F*(1, 71) = 204.44, *p* < .001) and session (*F*(1, 71) = 41.82, *p* < .001) were significant. The simple main effect indicated that the participants fixated for longer on the eye area in the emotion perception task than in the phoneme perception task, both in the AV session and the VO session and across all age groups (younger children: *F*(1, 31) = 35.53, *p* < .001; older children: *F*(1, 29) = 210.14, *p* < .001; adults: *F*(1, 11) = 29.91, *p* < .001). Thus, even younger children changed their gaze pattern depending on the task, although the difference between tasks seemed to become more salient with age.

Finally, the correlation coefficients for the relationship between participants' rates of fixation on the eye area in the emotion perception task and that in the phoneme perception task were significant (AV: *r* = .55, *p* < .001; VO: *r* = .34, *p* = .003) when they were controlled with each participant's age. These data suggested that participants who fixated for longer on the eye area during emotion perception tended to fixate on the same area during phoneme perception (Figure 4).

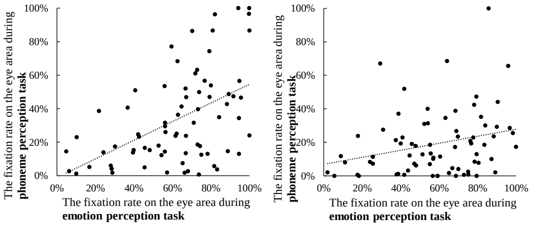


Figure 4. *Scatter plot showing correlation between the fixation of tasks (Left: AV session, Right: VO session)*

3.2. Relationship between perception and gaze data

3.2.1. Emotion perception task

3.2.1.1 Emotion perception

The results on the participants' emotion perception in the emotion perception task are presented in Figure 5.

AV session: We calculated the participants' face response rates, which indicated the rate of responses based on the speaker's facial expression, as for the AV session - incongruent trials (e.g., a response of "happy" to a happy face-angry voice stimulus in an incongruent trial).

To examine the differences among age groups, we performed a 3 (age: younger children, older, and adults) × 2 (emotion: angry face-happy voice and happy face-angry voice) mixed-factor ANOVA on the face response rates. The results showed that the interaction between age and emotion (*F*(2, 76) = 5.88, *p* = .004) and the main effect of emotion (*F*(1, 76) = 15.14, *p* < .001) were significant. The simple main effect of age was significant when the participants were presented with a happy face-angry voice stimuli (*F*(2, 76) = 3.33, *p* = .041); the face response rate to a happy face with an angry voice differed among groups. The post hoc analysis showed that the differences between younger children's face response rates and those of older children (*p* = .074), and the difference between those of younger children and those of adults (*p* = .054), were marginally significant. This finding indicates that, when presented with a happy face and an angry voice, older children and adults tended to judge "angry" more frequently than younger children did (Figure 5(b)). This tendency is consistent with previous findings [4][5]. Given that our previous data demonstrated that even 5-6-year-olds distinguished angry from happy voices [4], younger children's judgement weighting on faces cannot be explained only by their disability to perceive vocal emotion. As for the congruent trials, a 3 × 2 (emotion: angry face-angry voice and happy face-happy voice) ANOVA revealed a main effect of emotion (*F*(1, 76) = 7.47, *p* = .008), suggesting that participants' response accuracy was higher when they were presented with an angry face-angry voice than a happy face-happy voice (Figure 5(a)).

VO session: The participants' response accuracy in the VO session was also examined. A 3 (age) × 2 (emotion: angry face and happy face) ANOVA on the response accuracy showed a significant main effect of emotion (*F*(1, 76) = 18.24, *p* < .001), while the interaction between age and emotion (*F*(2, 76) = 1.39, *p* = .256) and the main effect of age (*F*(2, 76) = 0.98, *p* = .379) were not significant. Thus, the participants' accuracy in perceiving angry faces was higher than that of happy faces, regardless of age group (Figure 5(a)).

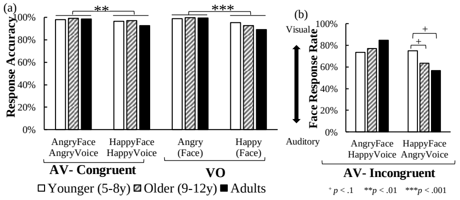


Figure 5. *Results on emotion perception*

3.2.1.2 Relationship with gaze

We calculated Pearson's correlation coefficients between perception and the rate of fixation on the eye area in each period in the AV and VO sessions. However, no significant correlations were found when we controlled for each participant's age (*rs* < .24).

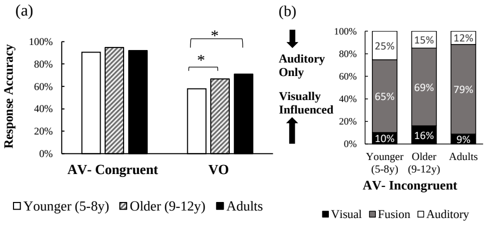
3.2.2. Phoneme perception task

3.2.2.1 Phoneme perception

The results on the participants' phoneme perception in the phoneme perception task are presented in Figure 6.

AV session: We calculated the visual response (/ka/), auditory response (/pa/), and fusion response (/ta/) rates for the McGurk-type trials and performed a one-way ANOVA on each response rate (Figure 6(b)). The auditory response rates differed among groups ($F(2, 76) = 4.14$, $p = .020$), and the post hoc analysis revealed that younger children's auditory response rates were higher than older children's ($p = .040$) and adults' ($p = .040$). As for the fusion response rates, the difference among age groups was marginally significant ($F(2, 76) = 2.83$, $p = .065$). The post hoc analysis revealed that adults' fusion response rates were marginally significantly higher than younger ($p = .060$) and older children's ($p < .1$). The visual response rates did not differ among age groups (F $(2, 76) = 1.79$, $p = .174$). In the congruent trials, the response accuracy did not differ among groups ($F(2, 76) = 1.09$, $p = .341$, Figure 6(a)). Taken together, the influence of visual information on the responses increased with age.

VO session: We performed a one-way ANOVA on the accuracy in the VO session. The main effect of age was significant ($F(2, 76) = 7.29$, $p = .001$). The post hoc analysis revealed that younger children's accuracy was lower than that of older children ($p = .007$) and adults ($p = .002$). These results suggest that children's lipreading skills are acquired during childhood (Figure 6(a)).



*$p < .05$

Figure 6. *Results on phoneme perception*

3.2.2.2 *Relationship with gaze*

The correlation coefficient between the auditory response rate and the rates of fixation on the eye area in the pre-speech period was significant ($r = .26$, $p = .025$) when we controlled for each participant's age, suggesting that participants who fixated on the mouth area were influenced by visual information. As for the VO session, the correlation coefficient between response accuracy and the rates of fixation on the eye area in the during-speech period was significant ($r = -.24$, $p = .039$) when we controlled for each participant's age, suggesting that participants who fixated on the mouth area for a longer duration tended to perceive emotions more accurately in the VO session. These results indicated that a gaze directed toward the mouth area has an impact on audiovisual phoneme perception [7].

4. Discussion

4.1. The development of gaze patterns

The first aim of the present study was to investigate the development of gaze patterns during audiovisual emotion perception. In the emotion perception task, participants of all age groups shifted their fixation from the eye to the mouth area when the speaker started to speak, and returned it to the eye area after the speech. We found two age differences in this gaze pattern. First, the amount of children's shift to a speaking mouth was larger than that of adults. Both younger and older children shifted their gaze to the mouth area clearly when speakers started to move their mouth, while adults held their shift to the mouth to a minimum, keeping their fixation on the eye area. Second, older children and adults returned their gaze to the eye area after speech more quickly than younger children. Thus, the gaze data from the emotion perception task suggested that children's gazes tend to be drawn to speaking mouths, but that they become able to keep this to a minimum with age. This results in the fixation on the eyes increasing with age during audiovisual emotion perception. The tendency for children to fixate more on a speaker's eye area as they age was observed in one previous study [9], although it demonstrated an age difference among infants and did not focus on emotion perception.

4.2. Comparison of gaze patterns between tasks

The second aim was to compare the gaze patterns during audiovisual emotion perception with those during audiovisual phoneme perception. Participants of all age groups fixated on the mouth area for a longer time during the phoneme perception task than during the emotion perception task. This difference between tasks is consistent with previous research [10] which demonstrated that adults tend to focus on the eyes when they are asked to judge emotion compared to when they are asked to report spoken words. This is due to the importance of eye area in emotion judgement. Moreover, it is suggested that this tendency comes to be salient with age in the present study.

In contrast, the results of phoneme perception suggested that participants come to fixate on the mouth area longer with age. The results are apparently inconsistent with previous studies. Lewkowicz and Hansen-Tift demonstrated that 12-month-olds and adults tended to fixate on the eye area of speaking face compared to 10-month-olds [9]. Irwin et al. showed that 7-8- and 9-10-years-olds tend to fixate on the mouth longer than adults during speech [7]. Thus, previous studies suggested that participants' fixation on the mouth area of speaking face decreases with age. The inconsistency between the present children's study and infant study [9] may be explained by the difference in developmental stage. But how can we explain the inconsistency in the similar age? One possibility is that adults posited their first fixation on the mouth area more frequently in the present study than in Irwin's study. That is, adults seem to prepare their fixation on the mouth area before the onset of speech in the present study. The differential results between these stimuli may be due to the length of video clips considering that the duration of our stimuli was shorter than Irwin's stimuli. We cannot compare the first gaze position at the beginning of each video clips between the present study and Irwin's study directly. However, it is important to investigate the effect of the duration of video clips in the future study. In contrast, the fixation pattern of our data is similar to Irwin's study in that the amount of adults' gaze shift to the mouth is relatively small during the playing of each video clip. Taken together, our study showed that people come to take an appropriate strategy to focus on the eyes in emotion perception and on the mouth in phoneme perception with age. It seems that people come to develop sophisticated gaze patterns that suit the task at hand.

Although the participants shifted their attention depending on the task, there was a significant correlation between participants' rates of fixation during tasks. Participants who fixated on the eye area during emotion perception tended to also fixate on it during phoneme perception. This is an interesting finding because it may account for the cultural differences in audiovisual integration demonstrated in previous studies. In a previous cross-cultural study, it was revealed that Japanese people give more weight to auditory information than do Western people during the perception of both emotion [4] and phonetic information [11]. This culture-specific integration may be caused by cultural differences in gaze patterns directed toward the face. For example, if Japanese people do not fixate on others' faces for a long time in daily communication compared to Western people, this habit may affect the weighting of visual and auditory information in the perception of both emotion and phonemes. This could not be analyzed further using the present data because the faces took up the whole display, but a gaze analysis of stimuli including bodies or one of actual face-to-face interactions may allow this hypothesis to be tested in a future study.

4.3. The relationship between gaze patterns and perception

The third aim was to examine the relationship between gaze patterns and perception of emotions or phonemes. As for the emotion perception task, when a speaker expressed happiness on her face and anger in her voice, adults tended to judge her emotion as anger more than children did. This suggests that in perceiving emotion, people give more weight to vocal expression as they age. This direction of development was consistent with previous studies [4][5], although a significant difference was not observed in the present study[ii]. However, we could not find a significant correlation between the participants' gaze patterns and their perception. Therefore, it has not yet been revealed whether their emotion perception from faces and voices are related to the development of their gaze patterns directed toward the face. The weighting of vocal expression may develop independently from gazes directed toward the face.

To the contrary, a significant correlation between participants' gaze patterns and their perception in audiovisual phoneme perception was observed. The results on phoneme perception demonstrated that the auditory responses in the AV session decreased and response accuracy in the VO session increased with age; thus, people come to be influenced by lip movement with age. Even when we controlled for each participant's age, there was a significant positive correlation between participants' auditory response rate and the duration of their fixation on the eyes. That is, participants' responses influenced by the visual information were negatively correlated with the duration of their fixation on the eyes. Moreover, there was a negative correlation between their accuracy and the duration of their fixation on the eyes in the VO session. These results suggest that as they age, children come to utilize lip movement in phoneme perception by focusing on the visual information derived from a speaker's mouth. Thus, the relationship between gaze patterns and perception is clearly observed in audiovisual phoneme perception, while it is still obscure in audiovisual emotion perception.

There are some limitations in the present study. First, it is impossible to rule out the possibility that the order of sessions and the property of stimuli might have impact on their fixations. In the present study, we used different stimuli between tasks and the order of sessions was constant among participants. To confirm that differential instructions (emotion or phoneme) contributed to the differential gaze patterns, we should use the same stimuli in all sessions (e.g., the speaker is saying "pa" with angry voice and face) and counterbalance the order of tasks. Second, the results obtained here are limited to Japanese people. Given cultural differences in audiovisual integration between East Asian and Westerners, it is possible that the children of Westerners may develop differently. For example, they may come to fixate on both a speaker's eyes and mouth during audiovisual emotion perception. Further cross-cultural studies are needed to elucidate the development of audiovisual integration.

5. Acknowledgements

We thank the participants in the experiments. We also thank the staff at Miraikan and the volunteer staff from Tokyo Woman's Christian University for their assistance with the data collection. This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas No. 17H06345 "Construction of the Face-Body Studies in Transcultural Conditions," JSPS KAKENHI (No. 18K13124), and the Scholarship Fund for Young/Women Researchers.

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] M. Kawahara, D. A. Sauter, and A. Tanaka, "Impact of Culture on the Development of Multisensory Emotion Perception," Proceedings of the 14th International Conference on Auditory-Visual Speech Processing, D2. S5. 2, 2017.

[3] H. W. Yamamoto, M. Kawahara, and A. Tanaka, "The developmental path of multisensory perception of emotion and phoneme in Japanese speakers," Proceedings of the 14th International Conference on Auditory-Visual Speech Processing, D2. S5. 1, 2017.

[4] 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.

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

[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] J. Irwin, L. Brancazio, and N. Volpe, "The development of gaze to a speaking face," The Journal of the Acoustical Society of America, vol.141, no.5, pp. 3145-3150, 2017.

[8] J. R. Irwin and L. Brancazio, "Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders," Frontiers in Psychology, vol.5, no.397, pp.1-10, 2014.

[9] D. J. Lewkowicz and A. M. Hansen-Tift, "Infants deploy selective attention to the mouth of a talking face when learning speech," Proceedings of the National Academy of Sciences of the United States of America, vol.109, no. 5, pp.1431-1436, 2012.

[10] J. N. Buchan, M. Paré, and K. G. Munhall, "Spatial statistics of gaze fixations during dynamic face processing," Social Neuroscience, 2, 1-13, 2007.

[11] 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.

[i] We conducted the experiment with 17 adults, but 5 participants' data were excluded from the analysis due to the low recording level of the gaze data.

[ii] In the present study, participants were presented with video clips of an enlarged size and asked to respond orally. This approach may have led to results in which the participants tended to put more weight on visual information than in previous studies, and any related age difference would not have been detected. To investigate this, we need to examine the effect of the size of audiovisual stimuli in a future study.