顔と声を用いる感情知覚と音韻知覚のプロセスは共通か独立か
Does perception of phonetic and affective information from facial and vocal cues share the same integration process?

山本 寿子[†], 河原 美彩子[†‡], 田中 章浩[†]
Hisako W. Yamamoto[†], Misako Kawahara[†‡], Akihiro Tanaka[†]
[†]東京女子大学, [‡]日本学術振興会特別研究員(DC)
[†]Tokyo Woman's Christian University, [‡]Japan Society for the Promotion of Science
hisako_wy@lab.twcu.ac.jp

Abstract

Previous studies have suggested that Japanese people place weight on auditory information in multisensory perception of affective and phonetic information. Is this because each type of perception shares the same integration process? To examine this hypothesis, we conducted three studies on the commonality between audiovisual perception of affective and phonetic information. Study 1 investigated the correlation between weighting on auditory information in audiovisual emotion perception and that on audiovisual phoneme perception using adult Japanese and Dutch speakers. The results showed that the correlation coefficient was low, suggesting that these two types of audiovisual perception have no relationship in the integration process. Study 2 explored the developmental changes in audiovisual perception of affective and phonetic information with Japanese-learning 5- to 12-year-olds and demonstrated that the developmental patterns were different between ages and in direction. In Study 3, we examined whether the incongruency of audiovisual phoneme information can affect emotion perception or vice versa with adult Japanese speakers. The results showed that each type of perception was not affected by the other types of perception. Overall, the results suggest that audiovisual perception of affective and phonetic information are processed independently and that the tendency of Japanese individuals to place greater weight on audiovisual information may be caused by another factors.

Keywords — Multisensory, Audiovisual Integration, Emotion, Phoneme

1. Introduction

It is important to properly recognize other's emotion for smooth communication. Such communicative cues come from both facial and vocal cues. The emotional value of vocal cues can affect the categorization or perception of facial expression [1] and vice versa [2][3]. Thus, emotion is perceived by the integration of information from facial (visual information) and vocal (auditory information) cues. Moreover, such audiovisual integration is also observed in the processing of linguistic information: phoneme perception. A famous example is the McGurk effect [4]: when sound /ba/ (/pa/) is played over lip movements for /ga/ (/ka/), /da/ (/ta/) can be perceived by a listener. This phenomenon reflects the fact that humans' perception of speech sound is affected by lip movements. Thus, phoneme also is perceived by the integration of information from facial and vocal cues.

Recent studies have suggested that the way of the audiovisual integration varies depending on the perceiver's cultural background. When people judge others' emotion from short movies, Japanese adults tend to be influenced by vocal expression, while Dutch adults are affected by facial expression [5]. These results suggest that Japanese people place weight on auditory information in emotion perception to a greater extent than Dutch people in audiovisual integration. Such a cultural difference has also been reported in the context of phoneme perception. For instance, Sekiyama and Tohkura's study [6] reported that the McGurk effect was induced more frequently in English speakers than in Japanese speakers[1]. That is, Japanese people are less affected by lip movement in audiovisual phoneme perception than English speakers. Moreover, another study revealed the McGurk effect in English speakers increases during childhood, while that of Japanese speakers does not [7].

Given these findings, it can be stated that Japanese people are less influenced by visual information from facial cues and that they place weight on auditory information from vocal cues in both emotion perception and phoneme perception. It is interesting that such a similar tendency can be observed in different types of perception. Why do Japanese people perceive both emotion and phoneme weighting with respect to auditory information? There are

[1] These previous studies [6][7] targeted two different language speakers, not people of different cultures. However, given the case of Japan, we consider people with a Japanese cultural background can be equated with Japanese native speakers.

two possible hypotheses. The first hypothesis is that audiovisual perception of affective and phonetic information share the same process in which visual and auditory information are integrated (Figure 1A). The alternative hypothesis is that there is an independent integration process for each process (Figure 1B), whereby other cultural factors may affect Japanese people's culturally specific way of audiovisual integration.

To examine which possibility holds true, we explored the commonality of audiovisual emotion perception and phoneme perception from the following three perspectives: First, we conducted both an audiovisual emotion perception task and a phoneme perception task with adult Japanese and Dutch people and calculated the correlation coefficient to determine a potential relation in these audiovisual perceptions in Study 1. Second, we focused on the acquisition process of audiovisual integration and compared the developmental patterns for these types of perception with Japanese children from 5 to 12 years old in Study 2. Finally, we conducted a more direct investigation in Study 3, where we examined whether the incongruency of audiovisual phoneme information can affect audiovisual emotion perception or vice versa in adult Japanese people.

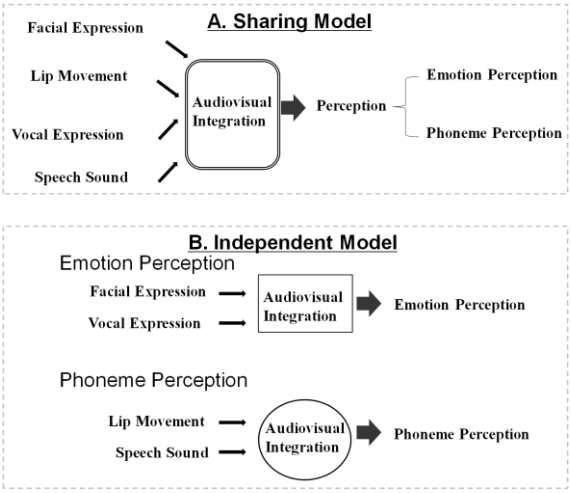

Figure 1 Two models of audiovisual integration in perception of affective and phonetic information

2. Study 1: Correlation between audiovisual emotion perception and phoneme perception

2.1 Purpose of Study 1

The aim of Study 1 is to examine whether audiovisual emotion perception and audiovisual phoneme perception are correlated.

2.2 Method

2.2.1 Participants

The participants composed Japanese participant and Dutch participant groups. The Japanese participants were 50 undergraduate students of Tokyo Woman's Christian University (age: *M*=20.0, *SD*=2.6), and the Dutch participants were 50 undergraduate and graduate school students of Leiden University (age: *M*=20.6, *SD*=2.9). All participants in the Japanese participant group and the Dutch participant group spoke Japanese and Dutch as their native language, respectively. They were all recruited in each university.

2.2.2 Apparatus

The experimenters used a 15-inch laptop to present and control audiovisual stimuli through Hot Soup Processor 3.4 (Onion software). The 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 resolution (12.2×16.3 cm size on the display). Sound stimuli were presented through headphones (HDA300, SENNHEISER) at approximately 70 dB SPL at most, which was adjusted using headphone amplifiers (DAC-HA200, ONKYO).

2.2.3 Stimuli and Procedure

The experiment was conducted in an experimental laboratory at Tokyo Woman's Christian University or Leiden University.

[Emotion perception task] Audiovisual stimuli were short movies in which a female actor (one of two native Japanese and two Dutch speakers) expressed her emotions. In each movie, the actor expressed happiness or anger in her face and voice. The linguistic vocal information was emotionally neutral ("*Hai, moshimoshi*" in Japanese and "*Hallo, daar ben je*" in Dutch (Hello) / "*Sayonara*" in Japanese and "*En goede dag*" in Dutch (Good-bye) / "*Korenani*" in Japanese and "*Hey, wat is dit*?" in Dutch (What is this?), or "*Sounandesuka*" in Japanese and "*Eh, is dat zo*?" in Dutch (Is that so?)). A total of 64 movies (two languages × two actors × four emotions (congruent movies: AngryFace / AngryVoice, HappyFace / HappyVoice; incongruent movies: AngryFace / HappyVoice; HappyFace / AngryVoice)) × four utterances) were used as test stimuli (Figure 2). In addition,

two other movies in which a Japanese actor expressed emotion were used as practice stimuli to check whether participants pressed the key properly.

In each trial, a fixation point was displayed at the center of the monitor for 500 ms, and a signal sound (a 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 to respond by pressing a key (D or K). After 500 ms from the participant's response, the next test trial began. A total of 64 test trials, including 32 congruent trials and as many incongruent trials, were conducted, following two practice trials. The order of test trials was randomized.

Figure 2 An example of audiovisual stimuli in the emotion perception task

[Phoneme perception task]

Audiovisual stimuli were short movies in which an actor (one of six native Japanese and six Dutch speakers) pronounced one syllable (/ka/, /pa/ or /ta/). The 24 audiovisual stimuli were congruent in lip movement and sound (two languages × six actors × two syllables[2]). The twelve movies were incongruent stimuli in which a sound /pa/ was combined with a lip movement /ka/ (Figure 3).

In each trial, a fixation point was displayed at the center of the monitor for 800 ms, and a signal sound (a 440 Hz pure tone lasting 100 ms) was played simultaneously; then, a blank display was presented 500 ms. After 1300 ms 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 to respond by pressing a key (Z, V, or M). After 500 ms from the participant's response, the next test trial began. Each congruent movie was presented once (24 congruent trials), while each McGurk type movie was presented twice (24 McGurk trials), resulting in a total of 48 test trials. Congruent trials were included in order to avoid response bias among participants. The order of test trials was randomized.

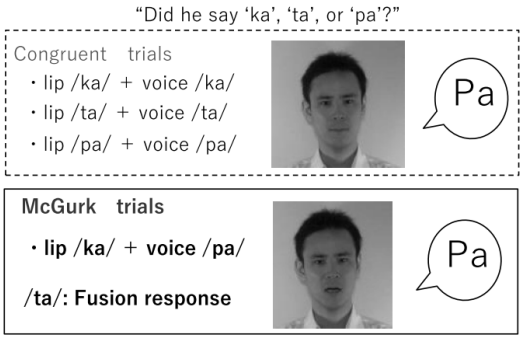

Figure 3 An example of audiovisual stimuli in the phoneme perception task

2.3 Results and Discussion

[Emotion perception task] The voice responses, which indicate the rates of participants' emotion judgement based on the speaker's voice, are shown in Figure 4. To examine cultural differences, we performed a Group (Japanese participant, Dutch participant) × Language (Japanese stimuli, Dutch stimuli) × Emotion (congruent, incongruent) three-way ANOVA on the voice responses. Since the second-order interaction ($F(1, 98)=100.11$, $p<.001$, $\eta_p^2=.51$) was significant, we conducted a Group × Language ANOVA on each emotion trial. For incongruent trials, the simple interaction between Group and Language ($F(1, 98)=148.13$, $p<.001$, $\eta_p^2=.60$) was significant. Shaffer's post hoc t-tests revealed that Japanese participants' voice responses to Japanese incongruent stimuli were higher than those of Dutch participants ($p<.001$). By contrast, the voice responses to Dutch incongruent stimuli did not differ between Japanese and Dutch participants ($p=.313$). Such a cultural difference was not observed for the congruent stimuli for both the Japanese and Dutch stimuli ($F(1, 98)=0.19$, $p=.663$, $\eta_p^2=.00$). Table 1 shows the voice responses in each combination of audiovisual information. Thus, the results demonstrated that Japanese participants judged emotion weighing on voice more than Dutch participants, replicating results in previous

[2] To equalize the number of congruent trials with that of incongruent trials, we adopted two of three syllables for each speaker. As a result, each syllable was pronounced four times in the whole congruent trials.

studies [5].

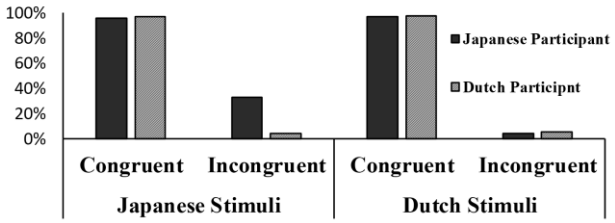

Figure 4 Voice response rates in the emotion perception task

[Phoneme perception task] Regarding the responses to the McGurk trials, we regarded participants' /ka/, /ta/, /pa/ responses as visual, fusion, and auditory responses, respectively. We present participants' rates for each type of response in Figure 5. To examine cultural differences in weighting auditory information, we performed a Group × Language mixed-factor ANOVA on auditory responses. The main effect of Group was marginally significant (F(1, 98)=3.69, p=.057, η^2=.04), suggesting that Japanese participants' auditory responses were slightly higher than those of Dutch participants. A two-way ANOVA on visual responses revealed the main effects of Group (F(1, 98)=5.44, p=.021, η_p^2=.05) and Language (F(1, 98)=76.18, p<.001, η_p^2=.44). A two-way ANOVA on fusion responses revealed the main effect of Language (F(1, 98)=50.23, p<.001, η_p^2=.34). However, fusion responses did not differ between Groups (F(1, 98)=0.59, p=.446, η_p^2=.01).

For congruent trials, we performed a Group × Language mixed-factor ANOVA on accuracy. Japanese participants' accuracy was higher than that of Dutch participants (F(1, 98)=6.57, p=.012, η_p^2.=04). The accuracy of participants in both groups was higher for Japanese stimuli than for Dutch stimuli (F(1, 98)=4.20, p=.043, η_p^2=.03).

Thus, the occurrence of fusion responses was not different between cultures. However, considering that Japanese participants responded based on visual information to a lesser extent and depended on auditory information to a greater extent than Dutch participants, the results of the present study show that Japanese participants place greater weight on auditory information than Dutch participants, as shown in previous studies [6][7].

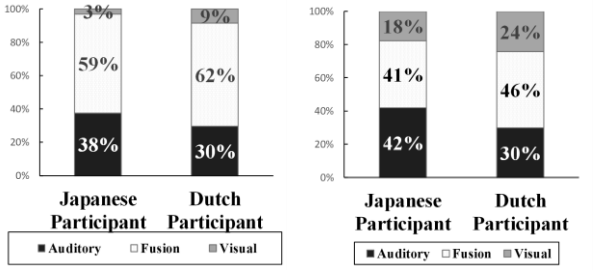

Figure 5 The rate of each response in McGurk trials (Left: Japanese Stimuli, Right: Dutch Stimuli)

[Correlation between two tasks] To explore the relation between two types of audiovisual perception, we calculated Pearson's correlation coefficients between the emotion perception task and the phoneme perception task (Table 2). For Dutch participants, there was no significant correlation between the voice responses in the emotion perception task and the auditory responses in the phoneme perception task. For Japanese participants, the correlation coefficient between voices responses between these two tasks was significant only when they viewed Dutch stimuli. This may suggest that Japanese people who tend to place greater weight on auditory information in emotion perception are more affected by visual information than phoneme perception information compared with Dutch people. However, considering that Japanese participants' voice responses to Dutch stimuli were very small (only 4%), this correlation coefficient is not reliable to support the hypothesis that audiovisual perception of affective and phonetic information share the same integration process.

Table 2 Correlation coefficients between two tasks (Emotion: Voice response rate Phoneme: Auditory response rate)

Japanese Participant			Dutch Participant		
All	Japanese Stimuli	Dutch Stimuli	**All**	Japanese Stimuli	Dutch Stimuli
.30 *	.18	.49***	**-.19**	-.18	-.14

Table 1 Voice response rates in each combination in the emotion perception task

	Japanese Stimuli				Dutch Stimuli			
	AngryFace AngryVoice	HappyFace HappyVoice	AngryFace HappyVoice	HappyFace AngryVoice	AngryFace AngryVoice	HappyFace HappyVoice	AngryFace HappyVoice	HappyFace AngryVoice
Japanese Participant	100%	92%	6%	59%	99%	95%	2%	7%
Dutch Participant	98%	97%	4%	5%	99%	97%	6%	5%

We should note that previous studies have revealed that integration processes can be acquired during childhood [7][8]. Accordingly, it may be useful to consider the possibility that this relationship can be observed in development. Thus, Study 2 investigated this possibility by conducting similar tasks with children.

3. Study 2: The Development in audiovisual emotion perception and phoneme perception

3.1 Purpose of Study 2

The purpose of Study 2 is to compare a developmental path of audiovisual emotion perception with that of audiovisual phoneme perception and to investigate the correlation coefficient for each age group.

3.2 Method

3.2.1 Participants

The participants were 82 5- to 6-year-olds (M=5.6; SD=0.5), 92 7- to 8-year-olds (M=7.5; SD=0.5), 95 9- to 10-year-olds (M=9.5; SD=0.5), and 80 11- to 12-year-olds (M=11.4; SD=0.5). All participants were native Japanese speakers, recruited from the National Museum of Emerging Science and Innovation (Miraikan), Tokyo, Japan.

3.2.2 Stimuli and Procedure

The experiment was conducted in an experimental laboratory at Miraikan. The apparatus was the same as that in Study 1. We adopted half of the movies (movies in which Japanese actors appeared) used in Study 1 as audiovisual stimuli in Study 2. As a result, the procedure was the same as that in Study 1 except that the number of test trials was half (the emotion perception task: 32 trials, the phoneme perception task: 24 trials).

3.3 Results and Discussion

[Emotion perception task] As with Study 1, we used voice responses as an index of auditory influence. To examine age differences, we performed an Age (5-6 y, 7-8 y, 9-10 y, 11-12 y) × Emotion (congruent, incongruent) mixed-factor ANOVA on the voice responses. This revealed that the interaction between Age and Emotion (F(3, 345)=9.32, p<.001, η_p^2=.08) and the main effects of Age (F(3, 345)=18.81, p<.001, η_p^2=.14) and Emotion (F(1, 345)=3047.78, p<.001 , η_p^2=.90) were significant.

A simple main effect analysis showed that voice responses were different among ages for both incongruent (F(3, 345)=14.38, p<.001) and congruent (F(3, 345)=5.37, p=.001) trials. Shaffer's post hoc t-tests revealed that 5- to 6-year-olds' and 7- to 8-year-olds' voice responses were less than those of 9- to 10-year-olds and 11- to 12-year-olds and that 9- to 10-year-olds' voice responses were less than those of 11- to 12-year-olds in incongruent trials. Overall, voice responses increased with development during childhood (Figure 6). Table 3 shows voice responses for each combination of audiovisual stimuli. In congruent trials, 5- to 6-year-olds' voice responses were less than those of 7- to 8-year-olds, 9- to 10-year-olds, and 11- to 12-year-olds.

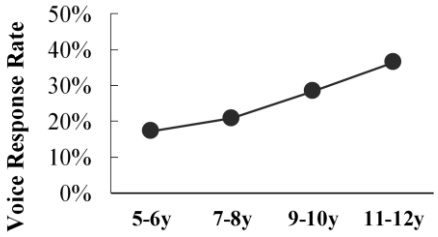

Figure 6 The developmental path of children's voice response rates in incongruent trials in the emotion perception task

Table 3 Children's voice response rates in each combination in the emotion perception task

	AngryFace AngryVoice	HappyFace HappyVoice	AngryFace HappyVoice	HappyFace AngryVoice
5-6y	93%	94%	15%	19%
7-8y	98%	97%	18%	24%
9-10y	99%	95%	19%	38%
11-12y	99%	93%	24%	49%

[Phoneme perception task] Regarding responses for McGurk trials, we regarded participants' /ka/, /ta/, /pa/ responses as visual, fusion, and auditory responses, respectively. We show participants' auditory responses in Figure 7 and 8. The one-way ANOVA on auditory responses showed no significant differences among age groups of children (F(3, 345)=1.05, p=.367, η^2=.01). This tendency was also observed for visual responses (F(3, 345)=0.1357, p=.939, η^2<.01) and fusion responses (F(3, 345)=1.35, p=.258, η^2=.01).

Regarding congruent trials, a one-way ANOVA revealed that the main effect of Age on correct responses was significant ($F(3, 345)=4.30$, $p=.005$, $\eta^2=.04$). Shaffer's post hoc t-tests revealed that 5- to 6-year-olds had fewer correct responses than other age groups.

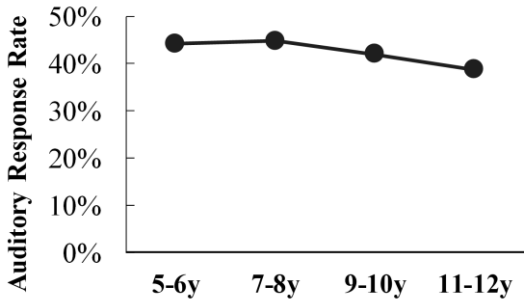

Figure 7 The developmental path of children's auditory response rates in McGurk trials in the phoneme perception task

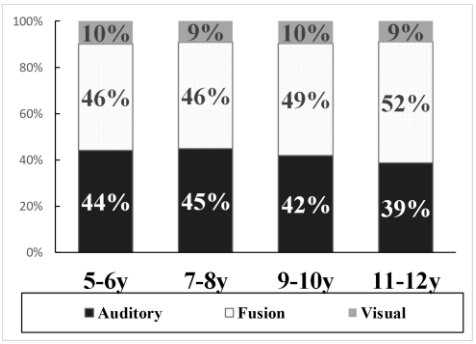

Figure 8 The rate of each response in McGurk trials

Thus, the results showed that the developmental path acquiring the manner of audiovisual integration differs between emotion perception and phoneme perception in two ways. First, the direction of development is different. In audiovisual emotion perception, the auditory influence increased. In contrast, the auditory influence decreased, and the visual influence increased in audiovisual phoneme perception together with the results of Study 1. Second, the age of development in audiovisual emotion perception differs from that in audiovisual emotion perception.

[Correlation between tasks] To explore the relation between two types of audiovisual perception, we calculated Pearson's correlation coefficients between the voice response rate in the emotion perception task and the auditory response rate in the phoneme perception task for each age group. A partial correlation was not significant when we controlled for each child's age ($r=.07$). Moreover, any significant correlations were not observed in each group (Table 4). Thus, as with Study 1, no significant correlation coefficients were observed even for children.

Table 4 Correlation coefficients between two tasks in each age group

5-6y	.08
7-8y	.01
9-10y	.06
11-12y	.13

To summarize, Study 2 demonstrated that each developmental path of audiovisual integration was "process-specific." Moreover, a correlation was not observed for either children or adults in Study 1 in the situation that they were shown Japanese stimuli. Thus, the results of Study 2 do not support the hypothesis that audiovisual perception of affective and phonetic information share the same integration process. However, the results are no more than an indirect suggestion about the integration process itself. It is desirable to investigate the independence of the mechanism of audiovisual integration in affective and phonetic information more directly. Additionally, in our daily life, we often perceive others' emotional expression and linguistic information simultaneously. Thus, we examine both audiovisual emotion perception and phoneme perception in a single trial in order to investigate the interaction between them in Study 3.

4. Study 3: Interaction between audiovisual emotion perception and phoneme perception

4.1 Purpose of Study 3

The purpose of Study 3 was to examine the interaction between audiovisual emotion perception and phoneme perception in order to more directly explore the commonality of audiovisual integration. For this examination, we investigated whether the congruency of audiovisual stimuli of emotion or phoneme affects the other types of perception. In each trial, participants were shown a short movie in which one speaker pronounced syllables (last phoneme was /ka/, /ta/, or /pa/) expressing emotion. They were asked to judge

her expressed emotion and pronounced phoneme simultaneously. If emotion perception and phoneme perception share the integration process, the congruency of one type perception would affect the other type of perception. For example, if an audiovisual phonemic incongruent movie inducing the McGurk effect is shown, a participant may be affected by visual information and perceive emotion weighing on facial expression (Figure 9 top). Moreover, if an emotional incongruent movie is shown, a participant would perceive phoneme weighing on lip movements (Figure 9 bottom).

Alternatively, if each audiovisual integration is independent, emotion perception would not differ between phonemic congruent and incongruent stimuli, and phoneme perception would not differ between emotional congruent and incongruent stimuli (Figure 10).

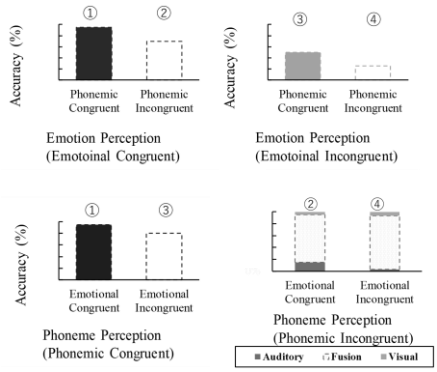

Figure 9 Expectation from the Sharing Model (numbers enclosed within a circle are corresponding to those in Figure 13 and 14)

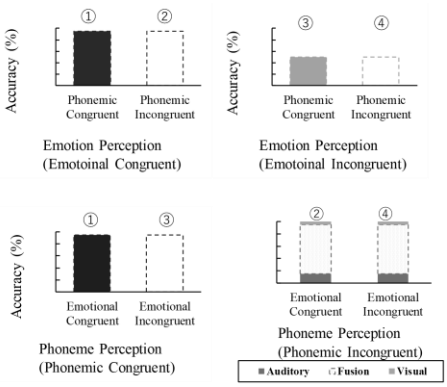

Figure 10 Expectation from the Independent Model

4.2 Method

4.2.1 Participants

The participants were 50 undergraduate and graduate students of Tokyo Woman's Christian University (age: *M*=18.8, *SD*=1.4). All participants spoke Japanese as their native language. They were all recruited in each university.

4.2.2 Stimuli

The audiovisual stimuli were short movies in which a female actor (four native Japanese speakers) said /aka/, /ata/, or /apa/ while expressing anger or happiness in her facial and vocal expression. Since it is difficult to express emotion in one syllable, we added /a/ to /ka/, /ta/, and /pa/. As in Study 1 and Study 2, the combination of facial and vocal emotion was congruent (AngryFace / AngryVoice, HappyFace / HappyVoice) or incongruent (AngryFace / HappyVoice, HappyFace / AngryVoice). As for phonemes, we used four types of combination (congruent: /aka/, /ata/, /apa/; incongruent (McGurk): visual /aka/ + auditory /apa/). That is, 64 (four actors × 4 emotions × 4 phonemes) movies were used as audiovisual stimuli (Figure 11). Each congruent phoneme stimulus was shown once (48 movies = 48 trials), while each McGurk stimulus was shown three times (16 movies × 3 = 48 trials) in order to present congruent and incongruent stimuli at the same time. Accordingly, a total of 96 trials were conducted, as described in the Procedure section.

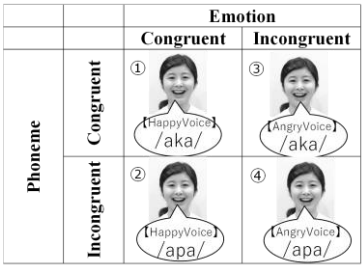

Figure 11 Audiovisual stimuli in Study 3 (In case of Angry Face)

4.2.3 Procedure

Study 3 was conducted in the experimental laboratory at Tokyo Woman's Christian University. The apparatus was the same as that in Study 1 and 2. In each trial, a fixation point was displayed at the center of the monitor for 800 ms, and a signal sound (a 440 Hz pure tone lasting 100 ms) was played simultaneously. After 1000 ms from the onset of the presentation of the fixation point, an audiovisual stimulus, the first instruction sentence, and the second instruction sentence display were presented successively (Figure 12).

For the half of participants, the first instruction was to judge whether the speaker's emotion was happy or angry and to press the corresponding key (E or U). Unlike Study 1 and 2, they were instructed to judge emotion based on the voice despite watching a face during trials in order to unified the attended modality between emotion perception and phoneme perception. After the response to the first instruction, as a second instruction, the participants were asked to judge which phoneme the speaker said, /aka/, /ata/, or /apa/ and to press the corresponding key (Z, V, or M). The other half of participants were asked to judge them in the reverse order (with the first instruction being phoneme judgement and the second instruction being emotion judgement). After the second response, the next test trial began. Thus, the participants judged both emotion perception and phoneme with respect to one movie simultaneously. In total, 96 test trials were conducted, and the order of test trials was randomized.

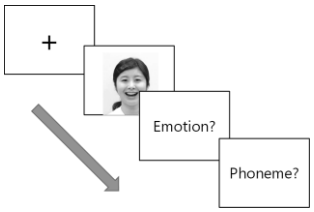

Figure 12 A flow of a trial

4.3 Results and Discussion

The effect of phonemic incongruency on emotion perception The results are shown in Figure 13. To investigate the impact of audiovisual phonemic information on audiovisual integration of emotional information, we conducted a Phonemic Congruency (phonemic congruent, phonemic incongruent) × Emotional Congruency (emotional congruent, emotional incongruent) within-subjects ANOVA (within subjects) on the accuracy of emotion perception. The main effect of emotional congruency was significant (F(1, 15)=66.35, p<.001, η_p^2=.82), suggesting that emotion perception from voice was inhibited by facial expression in emotional incongruent stimuli and that such perception was more difficult than congruent stimuli. However, the interaction (F(1, 15)=2.00, p=.178, η_p^2=.12) and the main effect of Phonemic Congruency (F(1, 15)=0.81, p=.383, η_p^2=.05) were not significant. Thus, audiovisual emotional perception was not affected by phonemic audiovisual incongruency in both the emotional congruent and incongruent trials.

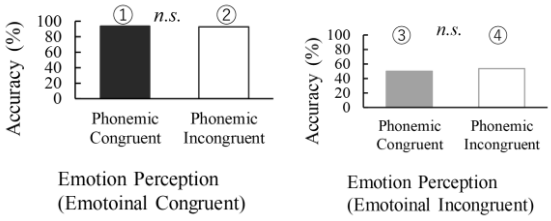

Figure 13 An accuracy in emotion perception

The effect of emotional incongruency on phoneme perception The results are shown in Figure 14. To investigate the impact of audiovisual emotional information on phonemic audiovisual integration, we compared the accuracy of phoneme perception in phonemic congruent trials and the occurrence of the McGurk effect in phonemic incongruent trials between emotional congruent and incongruent stimuli. The accuracy of phoneme perception in phonemic congruent trials did not significantly differ regardless of emotional congruency (t(15)=1.38, p=.188, d=.52). Regarding phonemic incongruent trials, the fusion responses (t(15)=0.53, p=.604 d=.09), auditory responses (t(15)=0.54, p=.596), and the responses (t(15)=0.14, p=.891 d=.07) were not significantly different between emotional congruent and incongruent stimuli. Thus, audiovisual emotional perception was not affected by phonemic audiovisual incongruency in both the emotional congruent and incongruent trials.

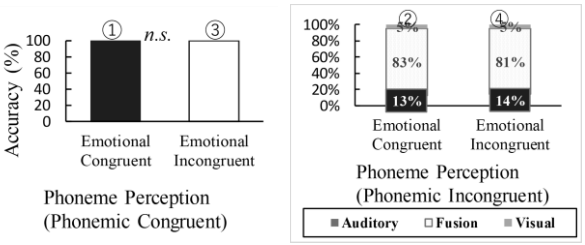

Figure 14 The results of phoneme perception

To summarize, the results of Study 3 suggest that the incongruency of audiovisual information does not affect other types of perception: there is no interaction between audiovisual emotion perception and phoneme perception.

5. General Discussion

This study aimed to examine the commonality of audiovisual emotion perception and phoneme perception in

three aspects: the characteristics of audiovisual integration in Japanese adult people, the developmental pattern, and the interaction of each perception. The results of Study 1 showed that there was no significant correlation between weighting on auditory information in audiovisual emotion perception and that in audiovisual phoneme perception in Dutch people. However, a significant correlation was observed only when Japanese participants were shown Dutch stimuli. Thus, the results in Study 1 were mixed. Study 2 demonstrated that the way of audiovisual integration in emotion perception develops during childhood in a different way from phoneme perception. Regarding emotion perception, the visual influence decreases and the auditory influence increase from the age of 5 to 12 years. In contrast, the rate of visual or auditory influence in audiovisual phoneme perception remains the same. Additionally, there was no significant correlation coefficient for each of the age groups. In Study 3, we examined whether the incongruency of phonemes affects audiovisual emotion perception and vice versa. The performance of audiovisual emotion perception did not differ between the congruency and incongruency (McGurk stimuli) of the actor's pronouncing phoneme. Moreover, the performance of audiovisual phoneme perception was the same regardless of the actor's emotional congruency. Overall, the results of the present study support the hypothesis that audiovisual perception of affective or phonetic information is processed independently (Figure 1B). Therefore, the reason why Japanese people place weight on auditory information can be explained by other mechanisms.

Although both affective and phonetic information is conveyed through visual information from face and auditory information from voice, their integration processes are different and independent, as demonstrated in the present study. There can be two differences between these two types of perception. First, the way of perceiving and expressing emotion is variable depending on the individual's communication manner. For example, expression of negative emotion may be natural in one culture, while it may be considered impolite in another culture, such as Japanese culture [9]. However, the way of perceiving and pronunciation of the phoneme is not dependent on the manner of expression as perceivers do not need to determine whether it is true or false. Second, for perception of affective and phonetic information, other aspects within same modality may be required. Lipreading skill may be independent of the recognition of facial expression, considering that lipreading skills develop later than the perception of facial expression [10]. Moreover, phoneme is segmental information of speech sound, while vocal emotion is conveyed by suprasegmental information. As demonstrated in previous studies using fMRI, fNIRS, and EEG, it is possible that segmental and suprasegmental information are processed in different brain areas [12]. Considering these characteristics of audiovisual perception, some audiovisual integration can exist simultaneously along with readable information.

Nonetheless, it should be noted that the results of the present study do not confirm that Japanese tendency of weighting vocal expression and their characteristics of audiovisual phoneme perception are totally irrelevant. Therefore, future research should investigate factors which can induce cultural difference in audiovisual integration. For example, it is useful to examine the relationship between gaze patterns on faces and audiovisual integration. Cultural difference in such gaze patterns [13] may have affect audiovisual perception. Additionally, we should reveal whether cultural differences is caused by people's culture or their languages. To shed light on these mechanism, it may be important to investigate factors shaping the way of audiovisual integration.

Acknowledgement

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 JSPS KAKENHI (No. 15H02714), and Grant-in-Aid for Scientific Research on Innovative Areas No. 17H06345 "Construction of the Face-Body Studies in Transcultural Conditions".

References

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

[2] de Gelder, B., Pourtois, G., & Weiskrantz, L. (2002). "Fear recognition in the voice is modulated by unconsciously recognized facial expressions but not by unconsciously recognized affective pictures." Proceedings of the National Academy of Sciences of the United States of America, vol. 99,

no. 6, pp. 4121-4126.
[3] Tanaka, A., Takagi, S., Hiramatsu, S., Huis In't Veld, E., & de Gelder, B. (2015). "Towards the development of facial and vocal expression database in East Asian and Western cultures.". Proceedings of the International Conference on Facial Analysis, Animation, and Auditory-Visual Speech Proceeding 2015, pp.63-66.
[4] McGurk,H., & McDonald, J. (1976)."Hearing lips and seeing." Nature, vol. 264, pp.746–748.
[5] Tanaka, A., Koizumi., A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder., B. (2010). "I feel your voice: Cultural differences in the multisensory perception of emotion." Psychological Science, vol.21, no.9, pp.1259-1262.
[6] Sekiyama, K., & Tohkura, Y. (1991). "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.
[7] Sekiyama, K., & Burnham, D. (2008). "Impact of language on development of auditory-visual speech." Developmental Science, vol.11, no.2, pp.306-320.
[8] Kawahara, M., Sauter, D., & Tanaka, A. (2017). "Impact of Culture on the Development of Multisensory Emotion Perception." Proceedings of the 14th International Conference on Auditory-Visual Speech Processing, D2. S5. 2.
[9] Ekman, P., & Sorenson, E. R., & Friesen, W. V. (1969). "Pan-cultural elements in facial displays of emotion." Science, vol.164, no.3875, pp.86-88.
[10] Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., & Foxe, J. J. (2011). "The development of multisensory speech perception continues into the late childhood years." European Journal of Neuroscience, vol.33, pp.2329-2337.
[11] Grandjean, D., Bänziger, T., & Scherer, K. R. (2006). "Intonation as an interface between language and affect." Progress in Brain Research, vol.156, pp.235-247.
[12] Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). "Cultural confusions show that facial expressions are not universal." Current Biology, vol.19, pp.1543-1548.