Culture Shapes Emotion Perception from Faces and Voices: Changes Over Development

Kawahara, M.; Sauter, D.A.; Tanaka, A.

DOI
10.1080/02699931.2021.1922361

Publication date
2021

Document Version
Final published version

Published in
Cognition and Emotion

License
CC BY-NC-ND

Citation for published version (APA):
Culture shapes emotion perception from faces and voices: changes over development

Misako Kawahara\textsuperscript{a,b}, Disa A. Sauter\textsuperscript{c} and Akihiro Tanaka\textsuperscript{a}

\textsuperscript{a}Department of Psychology, Tokyo Woman's Christian University, Tokyo, Japan; \textsuperscript{b}Kojimachi Business Center Building, Japan Society for the Promotion of Science, Tokyo, Japan; \textsuperscript{c}Department of Psychology, University of Amsterdam, Amsterdam, Netherlands

ABSTRACT
The perception of multisensory emotion cues is affected by culture. For example, East Asians rely more on vocal, as compared to facial, affective cues compared to Westerners. However, it is unknown whether these cultural differences exist in childhood, and if not, which processing style is exhibited in children. The present study tested East Asian and Western children, as well as adults from both cultural backgrounds, to probe cross-cultural similarities and differences at different ages, and to establish the weighting of each modality at different ages. Participants were simultaneously shown a face and a voice expressing either congruent or incongruent emotions, and were asked to judge whether the person was happy or angry. Replicating previous research, East Asian adults relied more on vocal cues than did Western adults. Young children from both cultural groups, however, behaved like Western adults, relying primarily on visual information. The proportion of responses based on vocal cues increased with age in East Asian, but not Western, participants. These results suggest that culture is an important factor in developmental changes in the perception of facial and vocal affective information.

ARTICLE HISTORY
Received 3 June 2020
Revised 1 April 2021
Accepted 16 April 2021

KEYWORDS
Emotion perception; cultural differences; development

Humans need to accurately infer the intentions and feelings of others to enable successful social interactions. Faces and voices are important sources of information for understanding how others feel (de Gelder & Vroomen, 2000). The majority of research to date has been conducted in Western culture, but recent studies have demonstrated cultural differences in the cues that perceivers use to infer others’ emotional states. When perceiving emotions from facial expressions, Easterners focus more on the eyes than the mouth, whereas Westerners focus more on the mouth than the eyes (Yuki et al., 2007). In emotion perception from speech, Easterners are more attuned to vocal tone, while Westerners rely more on verbal meaning (Ishii et al., 2003; Kitayama & Ishii, 2002). These studies point to cultural differences in how different sources of information are weighed within a given modality. However, a growing body of research is starting to go beyond research within a single modality. That work shows that processing of facial and vocal emotional cues interacts in emotion perception (e.g. de Gelder & Vroomen, 2000; Massaro & Egan, 1996; Vroomen et al., 2001). Similar to the McGurk effect in auditory-visual speech perception (McGurk & MacDonald, 1976), audio-visual integration occurs also in emotion perception. Recent studies have examined cultural differences in multisensory emotion perception (Liu et al., 2015a, 2015b; Tanaka et al., 2010). For example, Tanaka and colleagues (Tanaka et al., 2010) found cultural differences when comparing emotion perception by East Asians (Japanese participants).
and Westerners (Dutch participants). Specifically, East Asians relied on voice cues more than Westerners did, while Westerners relied on facial cues more than East Asians did. Subsequent studies have found the same pattern of results when comparing Chinese and Canadian perceivers (Liu et al., 2015a, 2015b). This work shows that multisensory integration of affective information is influenced by adult perceivers’ cultural background. However, no research to date has compared children’s development of multisensory emotion perception from face and voice across cultures.

Many studies with Western children have examined developmental changes of emotion perception within a single modality. For faces, preschoolers can match expressions with emotion words for some emotion such as happiness, anger, and sadness (Bullock & Russell, 1984; Harrigan, 1984; Reichenbach & Masters, 1983; Russell & Widen, 2002). Previous studies have also demonstrated that children’s recognition performance improve with age during childhood, although performance also depends on the task and emotion categories examined (Brechet, 2017; Herba et al., 2006; Mancini et al., 2013; Montirosso et al., 2010). In terms of emotion perception from vocal cues, Sauter and colleagues demonstrated that 5-year-old Western children can already recognize emotion from voice at above-chance-levels, and that performances improve with age (Sauter et al., 2013). More recent studies have found similar results (Allgood & Heaton, 2015; Chronaki et al., 2015, 2018; Grosbras et al., 2018).

Developmental studies of multisensory emotion perception have primarily been conducted in infants, mainly in Western cultural contexts (e.g. Bahrick & Lickliter, 2012; Flom, 2013; Grossmann, 2010; Grossmann et al., 2006). Intermodal preference experiments have shown that infants can match facial emotion to vocal emotion by 7 months of age (Walker-Andrews, 1986). However, cross-modal emotional integration has been observed from middle childhood, but not in early childhood (Gil et al., 2016). To our knowledge, only one study has examined multisensory emotion perception from facial and vocal expressions in childhood (Gil et al., 2016). In their study, 5-, 7-, and 9-year-old French children and adults were simultaneously shown a face and a voice expressing either congruent or incongruent emotions, and were asked to judge the emotion of the stimuli. The results showed that the younger children judged the emotion mainly based on face.

Importantly, children then shifted into an adult-like pattern characterised by a high sensitivity to vocal emotion. At the age of nine, children were affected by the vocal emotion, although the degree of vocal influence was greater in adults than in 9-year-olds. These results suggest that multisensory emotion perception changes with age during childhood. However, no cross-cultural study to date has been conducted in the development of multisensory emotion perception from face and voice. Although there are cultural differences in multisensory integration in adults (Liu et al., 2015a, 2015b; Tanaka et al., 2010), it is thus not clear how these cultural differences appear during development. In the present study we sought to test three competing hypotheses concerning the onset and emergence of cultural differences in the differential weighting of facial and vocal cues in emotion perception. One possibility is that the cultural differences seen in adults exist already in childhood (H1, the early emergence hypothesis). In facial emotion perception, studies have showed that the tendency of East Asian perceivers to fixate more on the eye region compared to Westerners (Jack et al., 2009) is present already in infants (Geangu et al., 2016). It is thus possible that East Asians rely on voice cues more than Westerners do already early in ontogeny. A second possibility is that cultural differences emerge over the course of development. In audio-visual speech perception, the degree of visual influence is greater in English speakers than in Japanese speakers (Sekiyama & Tohkura, 1991), with auditory information driving Japanese perception more. Children across languages preferentially rely on vocal cues, with inter-language differences emerging over the course of childhood, (Sekiyama & Burnham, 2008). If there is an early developmental period during which there is no cultural difference, there are two possible scenarios: One possibility is that children in both cultures rely mainly on vocal expressions and that the Western children shift into a Western adult-like pattern with age (H2, voice-first hypothesis). This prediction for emotion perception would fit with the results from auditory-visual speech perception. Another possibility is that children in both cultures rely mainly on visual cues, and that the East Asian children then shift into an East Asian adult-like pattern with age (H3, face-first hypothesis). Children may rely more on facial cues because emotion perception from faces is generally easier than from emotional speech prosody (Collignon et al., 2008; Paulmann & Pell, 2011; Scherer, 2003; Takagi et al., 2015). A developmental study
also has demonstrated that preschoolers’ recognition performance in facial expressions is higher than in vocal expressions (Nelson & Russell, 2011). It is thus possible that children start out like adult Westerners (relying mainly on facial expressions) and that East Asian children shift into an East Asian adult-like pattern with age. It should be noted that we did not test children younger than the age of five. We thus could not examine whether infants or preschoolers start out like adult Westerners (relying mainly on visual cues) or East Asians (relying mainly on vocal cues). We called our hypothesis with “face-first hypothesis” or “voice-first hypothesis” as indicating the starting point in multisensory emotion perception style during childhood (5–12-year-olds), not from birth.

In the current study, we employed an adapted version of the procedure used by Tanaka and colleagues (Tanaka et al., 2010) with East Asian and Western adults and children to test the three hypotheses outlined above. We used incongruent emotional stimuli, in which the emotions expressed from facial and vocal expressions are different. This type of incongruent stimuli has been employed in experiments applying the cross-modal bias paradigm (Bertelson & De Gelder, 2004), and is widely used in the field of cross-modal perception. In cross-modal bias experiments, participants are presented with video clips containing an emotional face and voice, which were either congruent or incongruent in terms of emotional content (e.g. a happy face paired with an angry voice on an incongruent trial). Participants are instructed to judge the emotion of the model as either happiness or anger, and are instructed to judge the emotion expressed in one of the two sources (face or voice) and to ignore the other source. Hence, this paradigm allows researchers to examine the degree of facial or vocal superiority in multisensory emotion perception, by calculating the difference in accuracy between the congruent and the incongruent conditions. However, there is a possibility that these instructions are difficult for young children. In fact, one developmental study used incongruent stimuli without attentional instructions to 5-, 7-, and 9-year-old children (Gil et al., 2016). Consistent with this approach, we did not instruct participants to focus on or ignore one of the two modalities. Although there is no “correct answer” in this task, we can examine the participants’ facial or vocal superiority by calculating the proportion of responses based on vocal and facial cues from their responses to incongruent stimuli. For example, when a participant saw a happy face paired with an angry voice and judged it as anger, we regarded that the judgment as based on the vocal expression. For this reason, we focused on the incongruent trials and compared the proportion of response based on vocal cues between East Asian and Western children and adults.

**Experiment 1**

**Materials and methods**

**Participants**

East Asian and Western adults and two age groups of children participated (5-6 and 11-12-year-olds). Participants were 33 East Asian adults (ages 18–32 years: 16 male, 17 female) and 104 children (5-6 years: 28 boys, 29 girls; 11–12 years: 28 boys, 19 girls) from Japan, and 38 Western adults (ages 19–30 years: 19 male, 19 female) and 46 children (5-6 years: 11 boys, 20 girls; 11–12 years: 7 boys, 8 girls) from the Netherlands. There were no deviations of distribution in sex and age between East Asians and Westerns in each age-group. Based on a work by Sekiyama and Burnham, which found that there are inter-language differences around 8 years in audio-visual speech perception (Sekiyama & Burnham, 2008), we set the lower age to 5–6 years and the higher age to 11–12 years. We chose Dutch and Japanese participants because they represent Western and East Asian cultures, respectively. East Asian participants were native Japanese speakers who lived in Japan. The East Asian adults were students at Tokyo woman’s Christian University, Waseda University, or Meiji University. East Asian children were recruited from the visitors to the National Museum of Emerging Science and Innovation (Miraikan). The Western participants were native Dutch speakers living in the Netherlands, with the Western adults being students at University of Amsterdam. The Western children were recruited via the University of Amsterdam and Dutch schools. All participants had normal hearing and normal or corrected-to-normal vision. The experiment was approved by the Tokyo Woman’s Christian University’s Research Ethics Board. Informed consent was collected from the adult participants and from the parents of the children who took part.

**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 for the Japanese and Dutch translations, and was spoken with happy or angry emotional intonation. 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. The audio was recorded at a sampling frequency of 48000 Hz. The recordings of the speakers’ visual expressions included their head and shoulders. The video frame rate was 29.97 frames per second.

For vocal stimuli, we calculated the average fundamental frequency (f0) for each vocal stimulus and compared it between Japanese and Dutch speakers (see Table S1 in the Supplemental Material). The f0 was higher in Japanese speakers than in Dutch speakers for the happy voice stimuli (z = −3.36, p < .001), but not for the angry voice stimuli. For facial stimuli, we used the Facial Action Coding System (FACS: Ekman and Friesen, 1978), which objectively describes visible facial movements based on anatomy, and a certified FACS coder coded all activated AUs from the onset to the offset of each stimulus. We compared the frequency of occurrence of AUs between Japanese and Dutch speakers. There were no cultural differences for any other AUs except AU17 in angry faces (z = −3.00, p = .01). The Japanese and Dutch stimuli were thus quite well matched. Moreover, our interest in this study was not in the culture of the individuals producing the emotional expressions. Nevertheless, for completeness, we report results of analyses that include the factor of the culture of stimuli in the supplemental material (Table S2).

Emotionally congruent and incongruent stimuli were created from the original audio-visual fragments. The unchanged fragments served as congruent stimuli (i.e. angry face with angry voice, happy face with happy voice). In order to make incongruent stimuli, happy and angry facial expressions were combined with angry and happy vocal expressions of the same speaker respectively (i.e. angry face with happy voice, happy face with angry voice). The composition of facial and vocal expressions was conducted for each of the eight utterances in each language (two speakers x four fragments), resulting in a total of 32 bimodal stimuli (16 congruent and 16 incongruent) in each language. The responses for the incongruent stimuli were the focus of the investigation, but congruent stimuli were included and presented to avoid participants noticing consistent incongruence and applying conscious strategies in the experiment. The unimodal stimuli were also created from the audio or video recordings of speakers’ utterances, resulting in a total of 16 unimodal stimuli (two speaker x four fragments x two language) in each sensory modality (audio only or visual only). Therefore, the total of 96 stimuli (32 Japanese bimodal, 32 Dutch bimodal, 16 Japanese unimodal, 16 Dutch unimodal stimuli) were presented in the experiment. All stimuli were presented once with no repetitions.

Procedure

Participants were tested in quiet rooms, either individually or in small groups of up to six participants. The experiment consisted of four sessions (Figure 1), beginning with two multisensory sessions, in which the affective faces and voices were presented simultaneously (i.e. bimodal stimuli were presented). The Japanese and Dutch bimodal stimuli were presented in separate sessions, one with Japanese and one with Dutch session. The order of the two multisensory sessions was counterbalanced between participants. These were followed by two uni-sensory sessions, in which either only faces or only voices (i.e. unimodal stimuli) were presented. In the uni-sensory sessions, Japanese and Dutch stimuli were mixed and presented within the same session. Although the multi-modal session was the focus of the investigation, a uni-sensory session was employed to check emotion recognition performance from unimodal cues. Thus, all participants judged all Japanese and Dutch targets.

In each trial, a fixation point was displayed, and then a dynamic face (happiness or anger) and/or voice (happiness or anger) was presented. The fixation point was displayed at a position of the mouth of speakers of stimuli. 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. Participants were instructed to categorise the emotion of the individual as either happiness or anger. Participants responded by pressing one of two keys, which were counterbalanced between participants. No instruction was given about which modality participants were to pay attention to. Therefore, there was no correct answer for the incongruent stimuli. For example, if a participant saw a happy face paired
with an angry voice and their response was anger, we regarded their answer as a response focusing on the speaker’s voice (voice choice). On the other hand, if a participant saw a happy face paired with an angry voice and their response was happy, we regarded their answer as a response focusing on the speaker’s face (face choice). Adults took about 20 min to complete the experiment, children took about 30 min.

Results

The percentages of voice choices for incongruent stimuli in all multisensory sessions are shown in Figure 2. To test for cultural differences and developmental changes in the proportion of responses based on vocal cues, participants’ responses for incongruent stimuli were analysed using a generalised linear mixed-effects model (GLMM) in the R software
interface (R Development Core Team, 2010), using the glmer function in the lme4 library (Bates et al., 2011). We treated the Perceiver Group (East Asians and Westerners) and Age Group (5-6 years, 11-12 years, and adult) as fixed factors, and participant as a crossed random factor. The binomial family call function was used because responses were coded in a binary fashion as voice choices (VC) or face choice (i.e. 1 or 0). For example, if a participant saw a happy face paired with an angry voice and their response was anger, this response would be categorised as a VC.

We used the Akaike information criteria (AIC) to establish the best fit model among regression models consisting of these variables and the interaction terms. The model including all single variables and interactions was the best (model 3 in Table 1). In this model, the main effects of Age Group (p < .001) and Perceiver Group were significant (p < .001). Most importantly, however, the interaction between Perceiver Group and Age was also significant (p < .001). To examine the power of model 3, we conducted a post-hoc power analysis using G*Power. The power (1-β) was equal to 0.8 when α = 0.05 and analysis parameters were set to a F test, hence confirming adequate power with the sample size. Also, we conducted a logistic regression analysis to check the overdispersion. The estimators of the dispersion parameter were calculated by dividing the value of residual deviance by degrees of freedom. The dispersion parameter was below 1.0 and hence confirmed that overdispersion was not an issue in our data.

To examine the onset and emergence of cultural differences (H1, H2 or H3), we ran a GLMM separately for each Age group. We treated the Perceiver Group as a fixed factor, and participant as a crossed random factor. The main effect of Perceiver Group was significant in 11–12 year-olds (p = .01, z = −2.55) and adults (p < .001, z = −3.25), but not in 5–6 year-olds (p = .51, z = −0.66). These results indicate that the cultural differences found for adults and 11-12-year-olds were not present in 5-6-year-olds. Therefore, our results do not support the early emergence hypothesis (H1).

Then, we ran a GLMM separately on each Perceiver group to determine which hypothesis (H2 or H3) would be supported by the data. We treated Age Group as a fixed factor, and participant as a crossed random factor. The main effect of Age Group was significant in East Asian participants (p < .001, z = 4.47), but not in Western participants (p = .26, z = 1.14). In East Asian participants, VC was different between 5-6-year-olds vs 11-12-year-olds (p = .002, z = 3.07), and 5–6 year-olds vs adults (p < .001, z = 4.40), but not between 11-12-year-olds vs adults (p = .12, z = 1.57). These results demonstrate that over the course of ontogeny, East Asians come to rely more on vocal cues, while Westerners do not. These findings support the face-first hypothesis (H3).

There is a possibility that the developmental changes of VC could be due to the development of uni-sensory emotion recognition. If this holds true, there should be cultural differences in the developmental trajectories of emotion perception from uni-sensory facial or vocal cues. To examine whether the developmental changes of VC could be due to the East Asian children becoming better at emotion recognition in the voice over development than the Western children, we calculated d-prime (Figure 3) and Criterion (i.e. C score, see Table 2) in the uni-sensory Audio only condition. d-prime and C score are signal detection theory (SDT) parameters. Techniques derived from SDT allow for the separation of participants’ abilities to recognise stimuli and their systematic strategies or biases inherent to recognition decisions. Therefore, we used d-prime as an index of participants’ abilities to recognise stimuli (instead of accuracy) and used C score as an index of participants’ systematic strategies or biases inherent to recognition decisions. We did not calculate the d-prime for the multimodal data since there is no correct answer in that condition. D-prime was calculated by subtracting the z-transform of the hit rate and false alarm rate. We defined the hit rate as the proportion of the trials presenting angry voice stimuli on which participants responded anger, and the false alarm rate as the proportion of the trials presenting happy voice stimuli on

### Table 1. The comparison among three models.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Age group</th>
<th>Perceiver group</th>
<th>Age group × Perceiver group</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>model 1</td>
<td>−3.19 (p &lt; .001)</td>
<td>0.38 (p &lt; .001)</td>
<td>—</td>
<td>4767</td>
<td>4787</td>
</tr>
<tr>
<td>model 2</td>
<td>−1.66 (p &lt; .001)</td>
<td>—</td>
<td>−0.59 (p &lt; .001)</td>
<td>4770</td>
<td>4790</td>
</tr>
<tr>
<td>model 3</td>
<td>−3.45 (p &lt; .001)</td>
<td>0.98 (p &lt; .001)</td>
<td>−0.11 (p &lt; .001)</td>
<td>4755</td>
<td>4789</td>
</tr>
</tbody>
</table>

Note: The values show the fixed effects estimates.
which participants responded anger. Since z-scores are not defined for 0 and 1 of hit rates or false alarm rates, we followed Macmillan and Creelman (1996) to add or subtract the equivalent of half of one response (i.e. 1/2 x n) from each score of zero or one. C score was calculated using the following equation: −0.5 * (zHits + zFalse Alarms).

We conducted Perceiver Group (Westerners or East Asians) × Age (5-6 years, 11–12 years, and adults) ANOVAs on the d’ score. There were significant main effects of Age [F (2, 215) = 71.57, p < .001, $\eta_p^2 = .40$] and Perceiver Group [F (1, 215) = 41.52, p < .001, $\eta_p^2 = .16$]. However, the interaction between Perceiver Group and Age was not significant [F (2, 215) = 0.87, p = .42, $\eta_p^2 = .008$]. These results indicate that the ability to recognise emotion from vocal expressions improves at a similar rate in East Asians and Westerners. Therefore, the developmental change in VC cannot be explained by differential developmental improvements in the ability to perceive emotion from the voice.

There is also a possibility that the developmental changes of VC could be due to the Western children becoming better at emotion recognition in the face over development as compared to the East Asian children. To examine whether this prediction would be supported, we also calculated d’-prime (Figure 4) and C score (Table 3) in the uni-sensory Visual only condition. We conducted Perceiver Group (Westerners or East Asians) × Age (5-6 years, 11–12 years, and adults) ANOVAs on d’ score. There were significant main effects of Age [F (2, 215) = 8.46, p < .001, $\eta_p^2 = .07$] and Perceiver Group [F (1, 215) = 6.52, p = .01, $\eta_p^2 = .03$]. However, the interaction between Perceiver Group and Age was not significant [F (2, 215) = 2.36, p = .10, $\eta_p^2 = .02$]. These results indicate that the ability to recognise emotion from facial expressions improves at a similar rate in East Asians and Westerners. Therefore, the developmental change in VC cannot be explained by differential developmental improvements in the ability to perceive emotion from the face.

**Experiment 2**

The results of Experiment 1 clearly demonstrated that the proportion of responses based on vocal cues increased with age in East Asians, but not Westerners. In Experiment 2, we tested the developmental trajectories of East Asians’ multisensory emotion perception in detail. Specifically, we examined whether the VC increase would occur gradually during childhood in East Asians or whether there would be a sudden shift at a particular age. If the emotion perception style changes rapidly at a particular age, it raises the possibility that there is a relationship between multisensory emotion perception and other cognitive skills developing during the same period. To test for this possibility, we conducted the same experiment with East Asian children aged between 7 and 10 years. The results were compared with those of the East Asian adults and the 5-6-year-old and 11-12-year-old East Asian children included in Experiment 1.

**Materials and methods**

**Participants**

Seventy-five Japanese children between 7 and 10 years were combined with the Japanese participants from Experiment 1. They were native Japanese speakers who lived in Japan. We recruited them from the visitors to the National Museum of Emerging Science and Innovation (Miraikan). The children in Experiment 2 were divided into four groups, 5–6 years and 11–12 years of age, and adults. The participants were divided into four groups: East Asian or Western, and 5–6 years and 11–12 years of age. The results were compared with those of the East Asian adults and the 5-6-year-old and 11-12-year-old East Asian children included in Experiment 1.

<table>
<thead>
<tr>
<th>Perceiver group</th>
<th>East Asian</th>
<th>Western</th>
</tr>
</thead>
<tbody>
<tr>
<td>5–6 years</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>11–12 years</td>
<td>−0.04</td>
<td>0.13</td>
</tr>
<tr>
<td>adults</td>
<td>−0.03</td>
<td>−0.02</td>
</tr>
</tbody>
</table>

Note: The positive values of C signify a bias toward the happy response, whereas the negative values signify a bias toward the angry response.

Figure 3. d’-prime in AO session. Error bars represent standard errors.

Table 2. C-score in AO session.
years, 7–8 years (43 children: 29 boys, 14 girls), 9–10 years (32 children: 19 boys, 13 girls), and 11–12 years. There were no differences in the distribution of genders in each age-group. All participants had normal hearing and normal or corrected-to-normal vision.

**Stimuli and procedure**
The stimuli and procedure were identical to those in Experiment 1.

**Results**
Percentage of VC for incongruent stimuli in all multisensory sessions is shown in Figure 5. To examine developmental changes in the proportion of responses based on vocal cues, we ran a GLMM with Age Group (5-6 years, 7–8 years, 9–10 years, 11-12 years, or adults) as fixed factors, and participant as a crossed random factor. As in Experiment 1, the binomial family call function was used because responses were coded in a binary fashion as VC or face choice (i.e. 1 or 0). Results showed that the main effect of Age Group was significant ($p = .01$, $z = 4.38$). To examine the power of this model, we conducted a post-hoc power analysis using G*Power. The power (1-β) was equal to 0.7 when $\alpha = 0.05$ and analysis parameters were set to a F test. As in Study1, we conducted the logistic regression analysis to check for overdispersion. The dispersion parameter was below 1.0 and hence confirmed that overdispersion was not an issue in our data.

To examine whether there would be a sudden shift, we compared the proportion of VC responses between the Age Groups. There were significant differences between adults and all other ages, except 11–12 year-olds (vs 5–6 year-olds: $p < .001$, $z = 4.40$; vs 7–8 year-olds: $p = .02$, $z = 2.38$; vs 9–10 year-olds: $p = .02$, $z = 2.45$). There were also significant differences between 5–6 year-olds and 7–8 year-olds ($p = .04$, $z = 2.04$), and between 5–6 year-olds and 11–12 year-olds ($p = .002$, $z = 3.07$). Although there were no significant differences between 7–8 year-olds and 11–12 year-olds, the proportion of responses based on vocal cues was significantly lower in 7–8 year-olds than adults. As also can be seen from Figure 5, the VC continues to increase after 11–12 years, and does not appear to suddenly increase at only a particular age in East Asians.

**General discussion**
Our findings provide the first evidence of culture-specific developmental changes in multisensory emotion perception from face and voice. Our results demonstrate that young children focus on facial expressions in both cultures, but over the course of
development, East Asians increase the degree of weighting of vocal cues in multisensory emotion perception, whereas Westerners do not.

Several earlier studies have suggested that vocal expressions may be more potent signals for infants than facial expressions (Caron et al., 1988; Mumme et al., 1996; Vaish & Striano, 2004). For example, Mumme and colleagues examined the effects of facial and vocal cues on 12-month-old infants’ behaviour toward novel toys (Mumme et al., 1996). They found that the voice alone, but not the face alone, was sufficient to guide infants’ behaviour. This emphasis of vocal information disappears with age, however, and visual information is more important than auditory information for preschoolers (Nelson & Russell, 2011) and older children (Bugental et al., 1970). Quam and Swingley (2012) showed that preschoolers have trouble interpreting the pitch contours expressing happiness and sadness. They pointed out the possibility that preschoolers lost the iconicity of emotional prosody through reinterpretation during language acquisition despite the fact that infants are sensitive to pragmatic functions and emotions cued by prosody in maternal speech. Hence, the shift from the early vocal superiority to facial superiority might be explained by U-shaped decrements in the developmental function of children’s vocal emotion sensitivity, which might enhance language acquisition. In our study, East Asian and Westerner children aged five to six focused on facial emotion, and the face-first hypothesis (H3) was supported. However, we did not test children younger than the age of five. Therefore, we cannot rule out the possibility that infants or preschoolers start out from vocal superiority and then shift into a facial superiority by the age of five to six. If this possibility holds true, the vocal superiority of East Asian people may draw a U-shaped curve, declining until 5–6 years old and then increasing during childhood. However, we did not test children younger than the age of five and so we therefore cannot specify the age of the lowest vocal sensitivity with certainty. It is worth noting that the body of research showing a U-shaped developmental curve of vocal emotion perception is based exclusively on children in Western cultural contexts. Our results suggest that culture may be an important factor to consider when examining developmental changes in the use of facial and vocal cues. Further research is needed to establish whether the developmental change in emphasis of cues from different modalities that has been found in Western children will also be present in children from non-Western cultural contexts.

Our results demonstrated that the East Asian children increase the extent to which they rely on vocal cues in multisensory emotion perception during childhood, whereas Westerners do not. These findings differ from those of Gil et al. (2016), which showed that the degree of vocal influence in audiovisual emotion perception increased with age in Western children. One possible reason for the apparent discrepancy between the studies might be the difference in visual stimuli between the current studies and the previous work. We used dynamic faces (i.e. videos), synchronised with voices, while Gil and colleagues used static facial expressions (i.e. photos). In their study, temporal dynamic changes in vocal expressions might have attracted participants’ attention more than static facial expressions.

Our results demonstrate culture-specific developmental changes in the multisensory integration of affective information. A point to consider is whether the developmental changes of VC could be due to the development of emotion recognition from only faces (e.g. Brechet, 2017; Harrigan, 1984) or voices (e.g. Sauter et al., 2013). If the East Asian’s developmental change of VC could be explained by the improvement of the ability to perceive emotion from a single type of cue, there should be cultural differences in the developmental trajectories of unisensory emotion perception. In particular, East Asian children should show more marked improvements in their emotion perception from vocal cues. However, the results indicate that age-related improvements were consistent across both East Asians and Westerners. Therefore, the developmental change in VC cannot be explained by the cultural differences of the developmental trajectory in unisensory emotion perception.

Unexpectedly, the performance on uni-sensory emotion perception was higher in East Asians than in Westerners in both Visual-only and Audio-only conditions, although performance increased with age in both cultures. The results on facial expressions are consistent with Markham and Wang (1996), which reported that 4–8 year-old East Asian children inferred emotion from facial expressions more correctly than did Western children of these ages, although performance improved with age in both cultures. As for vocal emotions, no study to date has compared children’s development of emotion perception from voice across cultures, although previous studies suggest
that East Asian adults are more sensitive to vocal information than are Westerners (Ishii et al., 2003; Tanaka et al., 2010). Cross-cultural studies are necessary on the development of vocal emotion perception.

Several issues should be examined in future studies. First, there was a large age gap between the 11–12 year-olds children and adults who participated in our experiments. We could not examine how East Asian children shift their perception style after the age of twelve. Conducting the experiment with adolescents would make it possible to investigate the developmental change in greater detail. Second, the sample size of Western 11–12 year-olds was small, although sufficiently large to detect cultural differences. Since the individual differences in children’s data tend to be large, it will be important to replicate these findings with a larger sample. Third, although we regarded participants’ response based on vocal expression as a vocal choice, we did not confirm that participant perceived emotion from each modality correctly. For example, when a participant was presented with a happy face paired with an angry voice, it is possible that the participant might perceive it as an angry face paired with a happy voice. In this case, it may be inappropriate to regard the participant’s response as “angry” as a vocal choice. In a future study, it would be preferable to integrate a confirmation of the source modality of participants’ response. Forth, we only used happy and angry expressions. Based on a previous study using the six basic emotion, which found that East Asian adults tended to focus on vocal emotion more than Westerners across different emotions (Takagi et al., 2015), we expect that our findings should hold also for other emotions, but further research will be needed to test this directly. Fifth, we reported the cultural differences of multisensory emotion perception regardless of the cultural origin of stimuli (i.e., East Asian and Western speakers). However, the vocal superiority might be influenced by the culture of stimuli; our exploratory analyses suggest that this may be the case (see the supplemental material). Future studies are needed to investigate the effect of congruency of culture between emotion perceivers and expressers, such as the in-group effect (Elfenbein & Ambady, 2003), in multisensory emotion perception.

In conclusion, we show culture-specific developmental changes in multisensory emotion perception. Our results demonstrate that while Western children and adults focus consistently on facial cues, East Asians gradually come to weigh vocal cues over the course of childhood. These results support the face-first hypothesis.

Acknowledgments
We thank the National Museum of Emerging Science and Innovation (Miraikan), and student volunteers from Tokyo Woman’s Christian University for assistance with data collection in Japan, and Onur Sahin, Maien Sachisthal, and Narda Schenk for assistance with data collection in the Netherlands.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Funding
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” (https://kaken.nii.ac.jp/en/grant/KAKENHI-PLANNED-17H06345/) and JSPS KAKENHI [grant number 15H02714] to A.T., and The Dutch Research Council (NWO) [grant number 275-70-033] and European Research Council (ERC) Starting grant, number 714977 to D.A.S.

References
Caron, A. J., Caron, R. F., & MacLean, D. J. (1988). Infant discrimination of naturalistic emotional expressions: The role of face


Takagi, S., Hiramatsu, S., Tabei, K., & Tanaka, A. (2015). Multisensory perception of the six basic emotions is modulated by attentional instruction and unattended modality.
Frontiers in Integrative Neuroscience, 9, 1. https://doi.org/10.3389/fnint.2015.00001