Language or appearance? The trigger of the in-group effect in multisensory emotion perception

Misako Kawahara¹,², Hisako W. Yamamoto¹ and Akihiro Tanaka¹,

¹Department of Psychology, Tokyo Woman’s Christian University, 2–6–1 Zempukuji, Suginami-ku, Tokyo, 167–8585 Japan
²Japan Society for the Promotion of Science, Kojimachi Business Center Building, 5–3–1 Kojimachi, Chiyoda-ku, Tokyo, 102–0083 Japan

(Received 19 April 2019, Accepted for publication 26 April 2019)

1. Introduction

For successful social communication, humans need to accurately perceive linguistic and non-linguistic information from facial expressions and voices [1–3]. Recent studies have demonstrated that there are cultural differences in multisensory emotion perception from facial and vocal expressions (i.e., nonverbal information such as prosody) [1,2]. Tanaka et al. found cultural differences between Japanese and Dutch participants in how multisensory information (i.e., facial and vocal expressions) is integrated [1]. In their study, experimenters showed both Japanese and Dutch speakers’ emotional expressions of the face and voice to Japanese and Dutch participants and asked them to judge the speakers’ emotion. The results showed that Japanese individuals rely on vocal expressions more than Dutch individuals do, while Dutch individuals rely on facial expressions more than Japanese individuals do. In addition, our earlier experiments demonstrated that such cultural differences arise during childhood; both Japanese and Dutch preschoolers judge speakers’ emotion based mainly on facial expressions, whereas Japanese (but not Dutch) children come to recognize emotions by placing weight on vocal expressions as well as facial expressions for in-group speakers (the degree to which Japanese participants rely on Japanese speakers’ vocal expressions was 12.9% in 5–6-year olds, 25.4% in 11–12-year olds, and 33.9% in adults) [4]. These results suggest that multisensory emotion perception is dependent on whether the speaker is an in-group member or not.

What, then, is the precursor of Japanese children’s focus on in-group members’ vocal expressions? The face and voice contain not only nonverbal affective information (e.g., facial and vocal expressions) but also information on the cultural group to which the individual belongs (e.g., appearance and language). In previous studies [1,4], emotional expressions of Japanese-speaking Mongoloids and Dutch-speaking Caucasians were presented as audiovisual stimuli. It is unclear what cue is critical in judging in-group members in emotion perception.

There are several candidates that may affect Japanese vocal superiority for in-group speakers. First, speaker’s appearance (i.e., Mongoloid or Caucasian) may be an important factor (appearance-superiority hypothesis). Physical appearance is an important predictor of personal identity [5]. Although the task (i.e., to choose an image of the child of an adult) differed from that of the present study, it has been shown that children believe that race (i.e., skin color) is a more informative factor than physique or occupation in the preservation of identity [5]. On the basis of this importance of appearance, we hypothesized that Japanese children categorize same-appearance people as in-group members in the perception of emotion from the face and voice. Specifically, they focus more on vocal expressions for Japanese-appearance speakers than for Dutch-appearance speakers.

Second, language (i.e., whether the person is speaking the listener’s native language or not) may be another important factor (language-superiority hypothesis). Recent studies on social preferences have suggested the importance of language in children’s social preferences [6,7]. Kinzler and colleagues showed that children choose to be friends with native-language individuals rather than foreign-language individuals [6,7]. They also demonstrated that children prefer individuals who speak their native language with a native accent over those with a foreign accent. Importantly, native-accented speakers whose race is different from the children are preferred to foreign-accented speakers whose race matches that of the children [7]. These results suggest that children use the speakers’ language or accent as cues to distinguish in-group members. On the basis of these findings on social preferences, we hypothesize that Japanese children also categorize same-language people as in-group members in emotion perception, and thus focus more on vocal expressions for Japanese-language speakers than for Dutch-language speakers.

The present study was aimed at examining which of the candidate cues (i.e., speakers’ appearance, language, or both) leads East Asian children to focus on in-group speakers’ vocal expressions. We tested Japanese children aged 5–12 years in addition to adults and investigated how the degree to which the focus on speakers’ vocal expressions in Japanese participants varies depending on the speakers’ language or appearance. In our earlier experiments [4], we demonstrated that Japanese participants gradually focus on in-group speak-
ers’ vocal expressions during childhood (between 5 and 12 years). Because of those results, we examined children of the same ages.

2. Methods

2.1. Participants

The participants were 20 Japanese university students between the ages of 18 and 25 years (all female) and 168 children between the ages of 5 and 12 years. The children were divided into 4 groups in accordance with age: 39 aged 5–6 years (19 girls), 48 aged 7–8 years (26 girls), 49 aged 9–10 years (19 girls), and 32 aged 11–12 years (15 girls). All participants had normal hearing and normal or corrected-to-normal vision. Participants or parents of child participants provided written informed consent before the experiments.

2.2. Stimuli

The stimuli were created from simultaneous audio and video recordings of Japanese (Mongoloid) and Dutch (Caucasian) 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, the fragment “korenani?” (“What is this?” in English) was uttered by Japanese speakers while the equivalent fragment “Hey, wat is dit?” was uttered by Dutch speakers. The audio was recorded at a sampling frequency of 48,000 Hz. The recordings of the speakers’ visual expressions included their head and shoulders. The video frame rate was 29.97 frames per second.

We manipulated the culture (Japanese and Dutch) and emotion (happy and angry) of speakers’ faces and voices independently using the above original audiovisual fragments. Therefore, there were congruent and incongruent stimuli in the speakers’ appearances and languages (cultural congruency in Fig. 1) and in the speakers’ facial and vocal emotions (affective congruency in Fig. 1).

In the manipulation of cultural congruency, Japanese and Dutch speakers’ facial expressions were combined with Dutch and Japanese speakers’ vocal expressions, respectively. Thus, there were culturally incongruent [Figs. 1(c) and 1(d)] and culturally congruent stimuli [Figs. 1(a) and 1(b)]. In culturally incongruent stimuli, we combined Japanese facial expressions with Dutch vocal expressions and Dutch facial expressions with Japanese vocal expressions. Here, we combined audio stimuli with visual stimuli so that the meaning of the phrases was equivalent in the two languages. For the culturally congruent stimuli, we always combined the speaker’s facial expression with another speaker’s vocal expression in each culture (e.g., the face of Japanese speaker A was combined with the voice of Japanese speaker B). The phrases of faces and voices were different in culturally incongruent stimuli. Similarly, we combined the speakers’ faces with voices speaking different phrases in culturally congruent stimuli. To minimize the asynchrony between lip movement and voice, we selected and combined different phrases with a similar utterance duration (Ud) in culturally congruent stimuli [in Japanese: “Korenani?” (Ud = 679.8 ms, SD = 97.8 ms) and “Sayoonara” (Ud = 631.5 ms, SD = 38.0 ms) pairs, or “Soonandesuka” (Ud = 826.8 ms, SD = 9.7 ms) and “Hai moshimoshi” (Ud = 803.85 ms, SD = 64.3 ms) pairs; in Dutch, “Hey, wat is dit?” (Ud = 695.8 ms, SD = 72.7 ms) and “Een goede dag” (Ud = 705.0 ms, SD = 44.9 ms) pairs, or “Oh, is dat zo?” (Ud = 790.8 ms, SD = 29.9 ms) and “Hallo, dat ben ja” (Ud = 802.3 ms, SD = 65.2 ms) pairs].

In the manipulation of affective congruency, happy and angry facial expressions were combined with happy and happy vocal expressions, respectively. Therefore, there were affectively incongruent stimuli replacing the facial and vocal emotions [Figs. 1(b) and 1(d)] and affectively congruent stimuli not replacing the emotions [Figs. 1(a) and 1(c)]. In affectively incongruent stimuli, we combined angry facial expressions with happy vocal expressions (and happy facial expressions with angry vocal expressions).

Thus, four types of stimuli were created: culturally congruent–affectively congruent [Fig. 1(a)], culturally congruent–affectively incongruent [Fig. 1(b)], culturally incongruent–affectively congruent [Fig. 1(c)], and culturally incongruent–affectively incongruent [Fig. 1(d)] stimuli. In the experiments, the culturally congruent–affectively congruent stimuli [Fig. 1(a)] were excluded as we could not detect whether the participants judged the emotion from the face or voice when using this combination. However, we used the culturally incongruent–affectively congruent stimuli to reduce the overall sense of incongruity caused by presenting only the affectively incongruent stimuli. In total, 96 stimuli (32 culturally congruent–affectively incongruent, 32 culturally incongruent–affectively congruent, and 32 culturally incongruent–affectively incongruent stimuli) were used in the experiments. Thus, speakers’ lip movements were asynchronous with the voice in all stimuli.

2.3. Procedure

Participants were tested individually in quiet rooms. In each trial, a fixation point was displayed, and then a dynamic face (happy or angry) and voice (happy or angry) were presented simultaneously. The fixation point was displayed
at the position of the speaker’s mouth. The face was displayed on a PC monitor (Latitude 3540, Dell) and the voice was presented binaurally via headphones (HDA300, SENNHEISER) at a comfortable listening level, which was adjusted using headphone amplifiers (DAC-HA200, ONKYO). The order of trials was randomized. Participants were instructed to categorize the emotion of the speaker in the movies as happiness or anger. No instruction was given on which modality participants should pay attention to. Participants responded by pressing one of two keys, which were counterbalanced across participants. All participants completed the main session after a practice session of four trials. Adults took about 20 min to complete the experiment, young children took about 30 min.

3. Results

We focused on the responses in affectively incongruent trials and calculated the percentage of voice choice responses (VCs) for incongruent expressions. For example, when participants observed a happy face paired with an angry voice and responded “anger,” this response was categorized as a VC. The mean VC under the affectively incongruent condition for each age group is presented in Fig. 2.

To examine which of the candidates (i.e., speakers’ appearance or language) lead Japanese children to focus on vocal expression, we performed Age (5–6 years, 7–8 years, 9–10 years, 11–12 years, or students) × Speaker condition (Japanese appearance + Japanese language, Dutch appearance + Dutch language, Japanese appearance + Dutch language, and Dutch appearance + Japanese language) ANOVAs on the mean VC under the affectively incongruent conditions. Results showed that the main effect of Speaker condition \([F(3, 549) = 102.75, p < 0.001, \eta_p^2 = 0.36]\) and the interaction between Speaker condition and Age \([F(12, 549) = 6.30, p < 0.001, \eta_p^2 = 0.12]\) were significant. Simple main effect analyses showed that the differences in VCs among various Speaker conditions were observed in children aged 7–8 years \([F(3, 181) = 11.92, p < 0.001, \eta_p^2 = 0.17]\), 9–10 years \([F(3, 181) = 20.80, p < 0.001, \eta_p^2 = 0.26]\), 11–12 years \([F(3, 181) = 15.42, p < 0.001, \eta_p^2 = 0.20]\), and students \([F(3, 181) = 18.78, p < 0.001, \eta_p^2 = 0.24]\), but not in 5–6 years \([F(3, 181) = 1.01, p = 0.39, \eta_p^2 = 0.02]\). Bonferroni post-hoc tests showed that VCs were higher for Japanese language speakers than for Dutch language speakers regardless of speaker’s appearance in each age group (all \(p < 0.001\)). These results show that children aged 5–6 years focus on facial expressions of speakers regardless of the speaker’s appearance and language, and that Japanese children increase their attention on vocal expressions when they are presented with Japanese language speakers’ emotional expressions. These results are consistent with those of our earlier study [4]. Therefore, our findings support the language-superiority hypothesis. Given that the speakers’ lip movement was asynchronous with the voice in all stimuli, the difference in VC among Speaker conditions cannot be explained by audiovisual asynchronization.

Next, we compared the VCs between Japanese and Dutch appearances under each Japanese and Dutch language condition to examine the effect of speaker’s appearance. The results showed that VCs were higher for Japanese-appearance speakers than for Dutch-appearance speakers, but only under the Japanese language condition. The difference in VC was marginally significant at age 11–12 years (\(p = 0.10\)) and was significant for university students (\(p = 0.007\)). The effect of speaker’s appearance was also observed by later childhood, although the effect was less pronounced than the effect of the speaker’s language.

4. Discussion

In this study, we found that children aged 5–6 years rely on facial expressions of speakers regardless of the speaker’s appearance and language, and that Japanese children increase their reliance on vocal expressions when they are presented with Japanese language speakers’ emotional expressions. These results support the language-superiority hypothesis and indicate that the speaker’s language is the main factor in leading Japanese children to focus on vocal expression in emotion perception. Our results are consistent with those of previous studies, indicating that language is important in children’s social preferences [6,7].

The effect of the speaker’s appearance in emotion perception was also observed from 11–12 years of age, although the effect was less than that of the speaker’s language. These results partly support the appearance-superiority hypothesis. The increase in the importance of visual information is consistent with a previous study demonstrating that preschool children perceive race (i.e., skin color) as an unchanging property when they infer an individual’s identity [5]. Our findings suggest that a speaker’s appearance also modulates the degree of focus on vocal expressions among in-group speakers.

It should be noted that we could not distinguish between the effect of linguistic information and paralinguistic information in the current study since we used only native-accented utterances as vocal stimuli. Vocal stimuli contain both linguistic and paralinguistic information. Therefore, paralinguistic information may lead participants to focus on vocal expression rather than linguistic information. Previous studies have suggested that the advantage of vocal emotion
perception in native language reflects the extent to which perceivers are familiar with conventions for expressing vocal emotions in a language, rather than with linguistic elements in the language [8,9]. On the basis of these findings, our results may also reflect the effect of paralinguistic information, such as culture-specific emotion expression shared within in-group members, in addition to linguistic information. In future studies, the effects of the linguistic and paralinguistic factors should be examined separately.

Acknowledgements
We thank all participants and staff at the National Museum of Emerging Science and Innovation (Miraikan). We also thank the student volunteers from Tokyo Woman’s Christian University for assistance with 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” and JSPS KAKENHI Grant Number JP18J23316.

References