

Psychological Science

<http://pss.sagepub.com/>

I Feel Your Voice : Cultural Differences in the Multisensory Perception of Emotion

Akihiro Tanaka, Ai Koizumi, Hisato Imai, Saori Hiramatsu, Eriko Hiramoto and Beatrice de Gelder
Psychological Science 2010 21: 1259 originally published online 16 August 2010

DOI: 10.1177/0956797610380698

The online version of this article can be found at:

<http://pss.sagepub.com/content/21/9/1259>

Published by:



<http://www.sagepublications.com>

On behalf of:



[Association for Psychological Science](http://www.sagepub.com)

Additional services and information for *Psychological Science* can be found at:

Email Alerts: <http://pss.sagepub.com/cgi/alerts>

Subscriptions: <http://pss.sagepub.com/subscriptions>


Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

I Feel Your Voice: Cultural Differences in the Multisensory Perception of Emotion

Akihiro Tanaka^{1,2}, Ai Koizumi³, Hisato Imai⁴, Saori Hiramatsu⁴,
 Eriko Hiramoto⁴, and Beatrice de Gelder^{1,5}

¹Cognitive and Affective Neuroscience Laboratory, Tilburg University; ²Waseda Institute for Advanced Study; ³Department of Psychology, University of Tokyo; ⁴Department of Psychology, Tokyo Woman's Christian University; and ⁵Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School

Psychological Science
 21(9) 1259–1262
 © The Author(s) 2010
 Reprints and permission:
sagepub.com/journalsPermissions.nav
 DOI: 10.1177/0956797610380698
<http://pss.sagepub.com>


Abstract

Cultural differences in emotion perception have been reported mainly for facial expressions and to a lesser extent for vocal expressions. However, the way in which the perceiver combines auditory and visual cues may itself be subject to cultural variability. Our study investigated cultural differences between Japanese and Dutch participants in the multisensory perception of emotion. A face and a voice, expressing either congruent or incongruent emotions, were presented on each trial. Participants were instructed to judge the emotion expressed in one of the two sources. The effect of to-be-ignored voice information on facial judgments was larger in Japanese than in Dutch participants, whereas the effect of to-be-ignored face information on vocal judgments was smaller in Japanese than in Dutch participants. This result indicates that Japanese people are more attuned than Dutch people to vocal processing in the multisensory perception of emotion. Our findings provide the first evidence that multisensory integration of affective information is modulated by perceivers' cultural background.

Keywords

emotion, multisensory perception, cultural difference, audiovisual speech

Received 10/15/09; Revision accepted 3/29/10

Are expressions of emotion universal, or is their perception culture-specific? Classical investigations of how humans communicate emotions focused on the universality of facial expression across cultures (e.g., Ekman, 1972; Ekman & Friesen, 1971). More recent studies demonstrated considerable cultural differences in the appearance and the perception of facial expressions (e.g., Elfenbein & Ambady, 2002; Jack, Blais, Scheepers, Schyns, & Caldara, 2009; Yuki, Maddux, & Masuda, 2007). In the natural environment, however, a facial expression is not seen in isolation; emotion expressed in the voice is heard at the same time. Recent cross-cultural studies on the perception of emotions raise the possibility that there may also be cross-cultural differences in the way multiple emotional cues are combined.

Such differences in the way multiple signals are combined have already been reported within the visual modality. For example, East Asian observers rely more on context when perceiving emotion in faces than Westerners do (Masuda et al., 2008). When the emotion expressed by the central figure was incongruent with that of the surrounding figures, the level of emotion of the central person was underestimated by Japanese, but not by American, participants. Reliance on context by East Asians has also been reported for perception

of emotions within the auditory modality (Ishii, Reyes, & Kitayama, 2003; Kitayama & Ishii, 2002). Using the Stroop-type interference paradigm (Stroop, 1935), Ishii et al. (2003) showed that in Japanese participants, the interference effect of vocal affect on judgments of verbal meaning is larger than the interference effect of verbal meaning on judgments of vocal affect, whereas in American participants, the opposite is true—that is, the interference effect of verbal meaning on judgments of vocal affect is larger than the interference effect of vocal affect on judgments of verbal meaning.

These studies have shown that there are cultural differences in how multiple sources of information are combined within the same modality. However, social interactions involve information from multiple modalities, such as faces and voices (Campanella & Belin, 2007; de Gelder & Bertelson, 2003). Literature has shown that facial and vocal emotional expressions interact in emotion perception (de Gelder & Bertelson, 2003; de Gelder, Bocker, Tuomainen, Hensen, & Vroomen,

Corresponding Author:

Akihiro Tanaka, Waseda Institute for Advanced Study, 1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050, Japan
 E-mail: akih.tanaka@gmail.com

1999; de Gelder & Vroomen, 2000; Massaro & Egan, 1996). Thus, the very process of integrating emotional cues from different modalities may also be culture sensitive. No previous studies have investigated the integration of emotional cues from the face and voice in a cross-cultural setting.

We investigated cultural differences between Japanese and Dutch people in the multisensory perception of emotion. We used the immediate cross-modal bias paradigm (Bertelson & de Gelder, 2004), widely used in the field of cross-modal perception. A face and a voice, which expressed either congruent or incongruent emotions, were presented on each trial (e.g., a happy face was presented with an angry voice, in the incongruent case). Participants were instructed to judge the emotion expressed in one of the two sources (face or voice) and to ignore the other source. The difference in accuracy between the congruent and the incongruent conditions was compared between Japanese and Dutch participants.

Given the findings that East Asians are more sensitive to context information than Westerners are (for a review, see Nisbett, Peng, Choi, & Norenzayan, 2001), we expected that Japanese participants would be more sensitive to context than Dutch participants. Specifically, we examined two possible hypotheses based on two different views of what is considered context. From one perspective, one can simply assume that task-relevant information is central and task-irrelevant information is context. On the basis of this assumption, one can hypothesize that Japanese people are more likely than Westerners to be sensitive to the context provided by task-irrelevant information during emotion perception (the *task-dependent-context hypothesis*). According to this hypothesis, then, the interference effect of voices on judgments of facial expressions would be larger in Japanese participants than in Dutch participants, as would the effect of faces on judgments of vocal expressions.

From a different perspective, one can assume that in emotion perception, one type of information is always central, and other types are context (e.g., Ishii et al., 2003; Kitayama & Ishii, 2002). Given the importance of the face, the face might always be central, whereas the voice might function as added information (i.e., as context). On the basis of this assumption, one can hypothesize that Japanese people would weight cues in voices more than Dutch people do, regardless of whether they are instructed to focus on or ignore the voices (the *face-central hypothesis*). According to this hypothesis, the interference effect of voices on judgments of facial expressions would be larger in Japanese participants than in Dutch participants, whereas the interference effect of faces on judgments of vocal expressions would be larger in Dutch participants than in Japanese participants.

Method

The audiovisual stimuli for this experiment 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 was uttered with happy or angry emotion. Happy and angry facial expressions were combined with happy and angry vocal expressions for each of the eight utterances in each language (two speakers \times four fragments), resulting in a total of 32 bimodal stimuli (16 congruent and 16 incongruent) in each language.

Participants were 20 native speakers of Japanese living in Japan (ages 21–29 years; 11 male, 9 female) and 16 native speakers of Dutch living in The Netherlands (ages 18–30 years; 3 male, 13 female). A trial consisted of a 1-s fixation point around the speaker's mouth followed by simultaneous presentation of dynamic face and voice. In the face task, participants were instructed to categorize the emotion of the faces as happy or angry and ignore the voices. In the voice task, participants were instructed to categorize the emotion of the voices and to look at the faces when the voices were being presented, but to ignore the faces when rating the voices. Participants responded by pressing one of two buttons. The experimenter instructed the participants that accuracy, rather than response speed, was important. The experiment began with four multisensory sessions (two repetitions of 32 stimuli in each session), followed by four unisensory sessions (in which only the faces or only the voices were presented). Each multisensory session and each unisensory session presented a different combination of task (face or voice) and speaker (Japanese or Dutch). Thus, both Japanese and Dutch participants observed both Japanese and Dutch targets.

In a preliminary experiment, the average overall performance in the face task was very high (98.0%). Consequently, we decreased the visibility of the face stimuli so that performance on unisensory trials was matched between face-only and voice-only trials. This enabled us to compare the differences between face and voice judgments without a ceiling effect. Visibility of the stimuli was reduced by adding dynamic noise to the face images (see, e.g., Collignon et al., 2008). (For more details about the method, see Section 1 in the Supplemental Material available online.)

Results

Data from 1 Japanese participant were excluded from the analysis because the participant misunderstood the task. Performance in all multisensory conditions is shown in Table 1. The difference between performance on the face-only trials ($M = 83.9\%$) and performance on the voice-only trials ($M = 87.2\%$) in the unisensory sessions was not significant, $t(34) = 1.57$, $p = .12$, confirming that difficulty was closely matched between the tasks.

To examine the general cross-modal bias, we performed a Task (face or voice) \times Group (Japanese or Dutch) \times Speaker (in-group or out-group) mixed-factor analysis of variance (ANOVA) on congruency effects, which were calculated by subtracting mean accuracy in the incongruent condition from

Table 1. Mean Percentage Accuracy in the Multisensory Conditions

Stimuli	Face task				Voice task			
	Japanese group		Dutch group		Japanese group		Dutch group	
	Japanese speaker	Dutch speaker	Japanese speaker	Dutch speaker	Japanese speaker	Dutch speaker	Japanese speaker	Dutch speaker
Angry face								
Angry voice	79.9 (2.7)	76.6 (3.2)	84.8 (3.3)	85.2 (2.8)	92.8 (2.2)	85.9 (3.4)	84.8 (2.4)	91.8 (3.0)
Happy voice	59.2 (4.8)	55.9 (4.9)	78.5 (2.8)	66.4 (4.2)	94.1 (2.2)	79.9 (3.3)	75.8 (3.9)	83.6 (3.3)
Happy face								
Angry voice	66.4 (4.3)	59.5 (3.9)	82.4 (2.3)	70.7 (4.3)	92.1 (1.9)	78.9 (3.5)	74.2 (3.5)	84.0 (4.4)
Happy voice	88.8 (2.1)	79.6 (2.9)	88.7 (2.2)	88.7 (2.5)	94.7 (1.4)	82.2 (3.2)	92.2 (1.7)	88.7 (2.9)

Note: Standard errors are given in parentheses.

mean accuracy in the congruent condition. The congruency effect was stronger in the face task than in the voice task, $F(1, 33) = 19.19, p < .001, \eta_p^2 = .37$. Effects of group, $F(1, 33) = 0.09, p = .76, \eta_p^2 = .003$, and speaker, $F(1, 33) = 0.16, p = .69, \eta_p^2 = .005$, were not significant. The absence of a main effect of group does not support the task-dependent-context hypothesis. Instead, results were consistent with the face-central hypothesis, as the two-way interaction between task and group was significant, $F(1, 33) = 11.48, p = .002, \eta_p^2 = .26$ (Fig. 1). The two-way interaction between task and speaker, $F(1, 33) = 18.11, p < .001, \eta_p^2 = .35$, and the three-way interaction, $F(1, 33) = 6.08, p = .02, \eta_p^2 = .16$, were also significant.

To examine the cultural difference in the cross-modal bias for each task, we conducted separate Group \times Speaker ANOVAs. Instead of conducting a one-way ANOVA on the group factor, we built the speaker factor into the analysis, because the cross-modal bias was different for in-group and out-group stimuli. For the face task, the main effect of group was significant,

$F(1, 33) = 4.88, p = .03, \eta_p^2 = .13$. Consistent with the face-central hypothesis, the congruency effect was larger in Japanese than in Dutch participants. The main effect of speaker, $F(1, 33) = 10.44, p = .003, \eta_p^2 = .24$, and the interaction, $F(1, 33) = 7.13, p = .01, \eta_p^2 = .18$, were also significant. Simple main-effects analyses showed that the congruency effect was not different between in-group and out-group stimuli in Japanese participants, $F(1, 33) = 0.16, p = .69, \eta_p^2 = .005$, but was larger for in-group than for out-group stimuli in Dutch participants, $F(1, 33) = 17.42, p < .001, \eta_p^2 = .35$. For the voice task, the main effect of group was also significant, $F(1, 33) = 10.27, p = .003, \eta_p^2 = .24$. Again, consistent with the face-central hypothesis, the congruency effect was larger in Dutch than in Japanese participants. The congruency effect was also larger for out-group than for in-group stimuli, $F(1, 33) = 7.58, p = .01, \eta_p^2 = .19$. The interaction was not significant, $F(1, 33) = 0.60, p = .44, \eta_p^2 = .02$. (For more details about participants' accuracy on the tasks, see Section 2 in the Supplemental Material.)

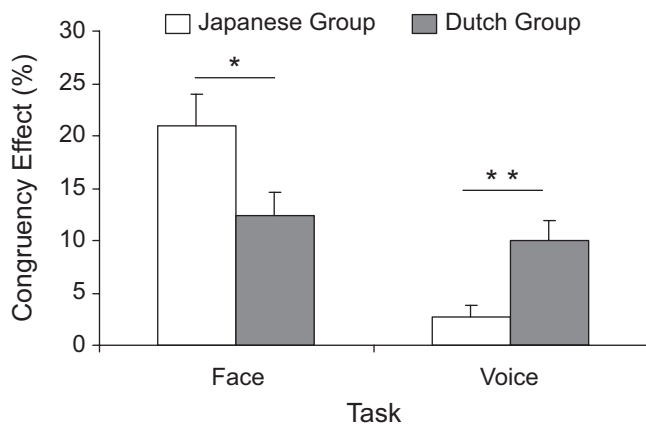


Fig. 1. Congruency effects (mean accuracy in the congruent condition minus mean accuracy in the incongruent condition) in the face task and the voice task among Japanese and Dutch participants. Error bars represent standard errors. Asterisks indicate significant differences between groups (* $p < .05$, ** $p < .01$).

Discussion

Our findings provide the first evidence that culture modulates multisensory integration of affective information. Supporting the face-central hypothesis, our results demonstrate that Japanese participants weighted cues in voices more than Dutch participants did. Despite instructions to focus on faces in the face task, Japanese participants paid attention to voices, which are, of course, present in the context of faces in everyday life. Further proof of this tendency to weight cues in voices was found in the voice task, in which Japanese participants demonstrated greater resistance to facial expressions compared with Dutch participants. The task-dependent-context hypothesis was not supported, which suggests that the cultural differences we observed were not due to high susceptibility to irrelevant stimuli in general in Japanese participants.

Our results are consistent with several lines of evidence for cultural differences. It has been shown that East Asians

(compared with Westerners) tend to use a different strategy to judge facial expressions (Jack et al., 2009; Yuki et al., 2007) and have a different attentional bias to different types of facial (Masuda et al., 2008) and vocal (Ishii et al., 2003) information. Our results extend the cultural differences in strategy and attentional bias to multisensory integration of affective information.

Our results are also in line with the finding that Japanese speakers use visual information less than English speakers do in interpreting audiovisual speech (Sekiyama & Tohkura, 1991). The similarity between Sekiyama and Tohkura's results and our results (i.e., less reliance on the face and greater reliance on the voice in Japanese) may be related to the fact that Japanese people control the display of their own feelings in the face (Ekman, 1972; Matsumoto, Takeuchi, Andayani, Kouznetsova, & Krupp, 1998).

Several issues should be examined in future studies. First, we used happy and angry emotions in our experiment. At the moment, it is not clear whether our findings can be replicated using other discrete emotions (e.g., sadness, fear), or if they can be observed only when hedonic valence is incongruent (i.e., pleasant vs. unpleasant stimuli—in our experiment, happy vs. angry expressions). Second, participants in our experiment were instructed to look at the mouth area. Given that there is cultural variation in the interpretation of facial expressions (Jack et al., 2009), it is worth investigating whether the differences found in our experiment persist when the instructions do not specify focusing on the mouth area. Third, it is noteworthy that there was no neutral condition in our experiment. Using neutral faces and voices would make it possible to separate facilitation from interference effects, which might yield another interesting cultural difference.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported in part by Grant-in-Aid for Specially Promoted Research No. 19001004 from the Ministry of Education, Culture, Sports, Science and Technology, Japan; by Postdoctoral Fellowships for Research Abroad from the Japan Society for the Promotion of Science; and by the European Commission (COBOL FP6-NEST-043403).

Supplemental Material

Additional supporting information may be found at <http://pss.sagepub.com/content/by/supplemental-data>

References

- Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence & J. Driver (Eds.), *Crossmodal space and crossmodal attention* (pp. 151–177). Oxford, England: Oxford University Press.
- Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*, 535–543.
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., et al. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242*, 126–135.
- de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*, 460–467.
- de Gelder, B., Bocker, K.B., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters, 260*, 133–136.
- de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion, 14*, 289–311.
- Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. Cole (Ed.), *Nebraska symposium on motivation, 1971* (pp. 207–282). Lincoln: University of Nebraska Press.
- Ekman, P., & Friesen, W.V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*, 124–129.
- Elfenbein, H.A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203–235.
- Ishii, K., Reyes, J.A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science, 14*, 39–46.
- Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology, 19*, 1–6.
- Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion, 16*, 29–60.
- Massaro, D.W., & Egan, P.B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review, 3*, 215–221.
- Masuda, T., Ellsworth, P., Mesquita, B., Leu, J., Tanida, S., & van de Veerdonk, E. (2008). Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology, 94*, 365–381.
- Matsumoto, D., Takeuchi, S., Andayani, S., Kouznetsova, N., & Krupp, D. (1998). The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology, 1*, 147–165.
- Nisbett, R.E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*, 291–310.
- 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 the Acoustical Society of America, 90*, 1797–1805.
- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 28*, 643–662.
- Yuki, M., Maddux, W.W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology, 43*, 303–311.