



I feel your voice: Cultural differences in the multisensory perception of emotion

Journal:	<i>Psychological Science</i>
Manuscript ID:	PSCI-09-1731.R4
Manuscript Type:	Research report
Date Submitted by the Author:	27-Mar-2010
Complete List of Authors:	Tanaka, Akihiro; Tilburg University, Psychology Koizumi, Ai; University of Tokyo, Psychology Imai, Hisato; Tokyo Woman's Christian University Hiramatsu, Saori; Tokyo Woman's Christian University Hiramoto, Eriko; Tokyo Woman's Christian University de Gelder, Beatrice; Tilburg U, Cog Neurosci
Keywords:	Cross Cultural Differences, Facial Expressions, Speech Perception, Auditory Perception, Emotions

Running head: CULTURE AND MULTISENSORY PERCEPTION

I feel your voice:

Cultural differences in the multisensory perception of emotion

Akihiro Tanaka¹, Ai Koizumi², Hisato Imai³,

Saori Hiramatsu³, Eriko Hiramoto³, Beatrice de Gelder^{1,4}

¹ Cognitive and Affective Neuroscience Laboratory,

Tilburg University, The Netherlands

² Department of Psychology, the University of Tokyo, Japan

³ Department of Psychology, Tokyo Woman's Christian University, Japan

⁴ Martinos Center for Biomedical Imaging,

Massachusetts General Hospital and Harvard Medical School, USA

Correspondence author:

Akihiro Tanaka, Cognitive and affective neuroscience laboratory, Tilburg

University, PO Box 90153, 5000 LE Tilburg, The Netherlands. Tel:

+31-13-466-3644, Fax: +31-13-466-2067, E-mail: akih.tanaka@gmail.com.

Abstract

Cultural differences in emotion perception have mainly been reported for facial expressions and to a lesser extent for vocal expressions. But the way in which the perceiver combines auditory and visual cues may itself be subject to cultural variability. Our study investigated cultural differences in the multisensory perception of emotion between Japanese and Dutch. A pair of dynamic face and voice, which expresses either congruent or incongruent emotions, was presented in the experiment. Participants were instructed to judge the emotion expressed in one of the two sources. The effect of to-be-ignored voice on the facial judgment was larger in Japanese than Dutch participants whereas the effect of to-be-ignored face on the vocal judgment was smaller. The result indicates that Japanese are more tuned to vocal processing in the multisensory perception of emotion. Our findings provided the first evidence that culture modulates the manner of the multisensory integration of affective information.

Keywords: emotion; multisensory perception; cultural difference; audio-visual speech;

I feel your voice:

Cultural differences in the multisensory perception of emotion

Introduction

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 observed 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, we do not just see an isolated facial expression. We see the face and at the same time we hear emotion expressed in the voice. 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.

Cultural differences in the way multiple emotional signals are combined were already reported within the visual modality. For example, East Asian observers rely more on the context when perceiving emotion in faces (Masuda et al., 2008). When the emotion expressed by the central figure

1
2
3
4
5 is incongruent with that by the surrounding persons, the level of emotion of
6
7
8 the central person was underestimated by Japanese but not by American
9
10 participants. Reliance on the context by East Asians has also been reported
11
12 when perceiving emotions in the auditory modality (Ishii, Reyes, & Kitayama,
13
14 2003; Kitayama & Ishii, 2002). Using the Stroop-type interference paradigm
15
16 (Stroop, 1935), Ishii et al. (2003) showed that the interference effect of the
17
18 vocal affect on the judgment of verbal meaning is larger in Japanese
19
20 participants whereas that of verbal meaning on the judgment of vocal affect
21
22 is larger in American participants.
23
24
25
26
27
28
29
30

31 These studies have shown that there are cultural differences in how
32
33 multiple sources of information are combined within the same modality.
34
35 However, our social interactions involve information from multiple modalities
36
37 such as faces and voices (Campanella & Belin, 2007; de Gelder & Bertelson,
38
39 2003). Literature has shown cross-modal interaction between facial and vocal
40
41 emotional expressions (de Gelder & Bertelson, 2003; de Gelder, Bocker,
42
43 Tuomainen, Hensen, & Vroomen, 1999; de Gelder & Vroomen, 2000; Massaro
44
45 & Egan, 1996). Thus the very process of integrating emotional cues from
46
47 different modalities may also be culture sensitive. So far, integration of
48
49 emotional cues from the face and the voice has not been investigated in a
50
51 cross-cultural setting.
52
53
54
55
56
57
58
59
60

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

In line with the findings that East Asians are more sensitive to context information than Westerners (for a review, Nisbett, Peng, Choi, & Norenzayan, 2001), we expect that Japanese are more sensitive to the *context* than Dutch. Specifically, we examine two possible hypotheses from two different views on what is *contextualized*. From one perspective, one can simply assume that the task-relevant information is central and the task-irrelevant information is context. Based on this assumption, one can hypothesize that Japanese are more likely to contextualize task-relevant information during emotion perception (*task-in-context hypothesis*). This hypothesis predicts that the interference effect of the voices on the judgment

of facial expressions is larger in Japanese participants and the same applies to that of the faces on the judgment of vocal expressions.

From a different perspective, one can assume that one type of information is always central and others are context (e.g., Ishii et al., 2003; Kitayama & Ishii, 2002). Considering the importance of the face, the face can always be central whereas the voice functions as added information in the context of the face. Based on this assumption, one can hypothesize that Japanese weight cues in the voices more than Dutch despite task instructions (*voice-in-face-context hypothesis*). This hypothesis predicts that the interference effect of the voices on the judgment of facial expressions is larger in Japanese participants whereas that of the faces on the judgment of vocal expressions is larger in Dutch participants.

Method

The audiovisual stimuli were created from simultaneous audio and video recordings of Japanese and Dutch speakers' emotional utterances. Four short fragments were uttered by two Japanese and two Dutch female speakers in their native language. Each fragment with neutral linguistic meaning was uttered with happy or angry emotion. Happy and angry facial expressions were combined with happy and angry vocal expressions for each

1
2
3
4
5 of the eight utterances (two speakers \times four fragments), resulting in a total of
6
7
8 32 bimodal stimuli (16 congruent and 16 incongruent) in each language.
9

10
11 Participants were 20 native speakers of Japanese living in Japan
12
13 (ages 21-29, 9 female) and 16 native speakers of Dutch living in the
14
15 Netherlands (ages 18-30, 13 female).
16
17

18
19 A trial consisted of a 1-s fixation point around the speakers' mouth
20
21 and a simultaneous presentation of dynamic face and voice. In the face task,
22
23 participants were instructed to categorize the emotion of the faces into happy
24
25 or angry and ignore the voice. In the voice task, participants were instructed
26
27 to categorize the emotion of the voices, to look at the face when the voice was
28
29 being presented but to ignore the face when rating the voice. Participants
30
31 responded by pressing either of the two buttons. The experimenter instructed
32
33 to the participants that the accuracy rather than response speed is important
34
35 in this experiment. The experiment began with four multisensory sessions
36
37 (two repetition of 32 stimuli in each session), followed by four unisensory
38
39 sessions in which only the face or the voice was presented. The task (face or
40
41 voice) and speaker (Japanese and Dutch) were different across sessions. Thus,
42
43 both Japanese and Dutch participants observed both Japanese and Dutch
44
45 targets.
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

In a preliminary experiment, the overall performance in the face task

1
2
3
4
5 was very high (98.0% in average). This may mask some possible differences
6
7
8 due to a ceiling effect in the face task. Thus, we decreased the visibility of the
9
10
11 face stimuli so that unisensory performance is matched between face-only
12
13
14 and voice-only tasks. This enables us to compare the differences between face
15
16
17 and voice judgments without ceiling effect. Specifically, we added a dynamic
18
19
20 noise onto the face images in order to decrease the visibility (e.g., Collignon et
21
22
23 al., 2008). For more detail about the method, see the supporting information
24
25
26 available on-line.

31 Results

32
33
34 Data from one Japanese participant were excluded from the analysis
35
36
37 since the task was misunderstood. Performance in all multisensory conditions
38
39
40 is shown in Table 1. The difference between the performance of the face-only
41
42
43 ($M = 83.9\%$) and the voice-only ($M = 87.2\%$) tasks in unisensory sessions was
44
45
46 not significant, $t(34) = 1.57$, $p = .12$, confirming that the difficulty was closely
47
48
49 matched between tasks.

50
51
52
53
54 ----Insert Table 1 about here----

55
56
57
58
59
60 In order to examine the general cross-modal bias, a Task (face or

voice) \times Group (Japanese or Dutch) \times Speaker (in-group or out-group) mixed-factor analysis of variance (ANOVA) was performed on congruency effects, which were calculated by subtracting the mean accuracy in the incongruent condition from that in the congruent condition. The congruency effect was stronger in the face task than in the voice task, $F(1,33) = 19.19$, $p < .001$, $\eta^2 = .37$. Effects of Group, $F(1,33) = 0.09$, $p = .76$, $\eta^2 = .003$, and Speaker, $F(1,33) = 0.16$, $p = .69$, $\eta^2 = .005$, were not significant. The absence of the main effect of Group does not support the task-in-context hypothesis. Instead, consistent with the voice-in-face-context hypothesis, a two-way interaction between Task and Group was significant, $F(1,33) = 11.48$, $p = .002$, $\eta^2 = .26$ (Fig. 1). A two-way interaction between Task and Speaker, $F(1,33) = 18.11$, $p < .001$, $\eta^2 = .35$, and a three-way interaction, $F(1,33) = 6.08$, $p = .02$, $\eta^2 = .16$, were also significant.

In order to compare the cultural difference in the cross-modal bias for each task, a Group \times Speaker two-way ANOVA was conducted separately for the face task and the voice task. Instead of conducting one-way ANOVA on the Group factor, the Speaker factor was built into the analysis, since the cross-modal bias was different when judging the in-group and the out-group stimuli. For the face task, the main effect of Group was significant, $F(1,33) = 4.88$, $p = .03$, $\eta^2 = .13$. Consistent with the voice-in-face-context hypothesis,

1
2
3
4
5 the congruency effect was larger in Japanese participants. The main effect of
6
7
8 Speaker, $F(1,33) = 10.44$, $p = .003$, $\eta^2 = .24$, and the interaction, $F(1,33) =$
9
10
11 7.13, $p = .01$, $\eta^2 = .18$, were also significant. Simple main effect analyses
12
13
14 showed that the congruency effect was not different between in-group and
15
16
17 out-group stimuli in Japanese participants, $F(1,33) = 0.16$, $p = .69$, $\eta^2 = .005$,
18
19
20 whereas it was larger for in-group stimuli in Dutch participants, $F(1,33) =$
21
22
23 17.42, $p < .001$, $\eta^2 = .35$. For the voice task, the main effect of Group was
24
25
26 significant, $F(1,33) = 10.27$, $p = .003$, $\eta^2 = .24$. Again, consistent with the
27
28
29 voice-in-face-context hypothesis, the congruency effect was larger in Dutch
30
31
32 participants. Congruency effect was larger for out-group than for in-group
33
34
35 stimuli, $F(1,33) = 7.58$, $p = .01$, $\eta^2 = .19$. The interaction was not significant,
36
37
38 $F(1,33) = 0.60$, $p = .44$, $\eta^2 = .02$.

39
40
41
42 ----Insert Figure 1 about here----

43 44 45 46 47 48 49 Discussion

50
51 Our findings provide the first evidence that culture modulates
52
53
54 multisensory integration of affective information. Supporting the
55
56
57 voice-in-face-context hypothesis, our results demonstrate that Japanese
58
59
60 participants weighted cues in the voices more than Dutch participants.

1
2
3
4
5 Despite instructions to focus on the faces, Japanese participants pay
6
7 attention to the voice, which is in the context of the face in everyday life. This
8
9 notion was further supported by the resistance to facial expressions in the
10
11 voice task among Japanese participants, compared with Dutch participants.
12
13
14
15
16
17 The task-in-context hypothesis was not supported, suggesting that the
18
19 cultural differences observed here are not due to high susceptibility to any
20
21 kinds of irrelevant stimuli in Japanese participants.
22
23
24

25
26 Our results are consistent but not mutually exclusive with several
27
28 lines of evidence that reported cultural differences. It has been shown that
29
30 East Asians tend to use a different strategy to judge the facial expression
31
32 (Jack et al., 2009; Yuki et al., 2007) and that they have a different attentional
33
34 bias to a different type of facial (Masuda et al., 2008) and vocal (Ishii et al.,
35
36 2003) information. Our results extend the cultural differences in strategy and
37
38 attentional bias to the multisensory information.
39
40
41
42
43
44

45
46 Our results are also in line with the finding that Japanese speakers
47
48 use visual information less than English speakers in the identification of
49
50 audiovisual speech (Sekiyama & Tohkura, 1991). The similarity between
51
52 their results and our results (i.e., lower reliance on the face and higher
53
54 reliance on the voice in Japanese) may be related to the fact that Japanese
55
56 people control the display of their own feelings in the face (Ekman, 1972;
57
58
59
60

Matsumoto, Takeuchi, Andayani, Kouznetsova, & Krupp, 1998).

Several issues should be examined in future studies. First, we used happy and angry emotions in the experiment. At the moment, it is not clear whether the findings can be replicated using other discrete emotions (e.g., sad and fear) or they can be observed only when the hedonic valence is incongruent (i.e., in our experiment, happy is pleasant while angry is unpleasant). Second, participants were instructed to look at the mouth area in our experiment. Since there are cultural differences in the diagnostic facial information when observers judge facial expressions (Jack et al., 2009), it is worth investigating whether this tendency persists when the instructions specify otherwise. Third, it is noteworthy that a neutral condition was not present. Using neutral faces and voices, one can separate facilitation from interference effects, which might yield another interesting cultural difference.

References

- Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In C. Spence and J. Driver (Ed.), *Crossmodal space and crossmodal attention* (pp. 151-177). Oxford: 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.,
& Lepore, F. (2008). Audio-visual integration of emotion expression.

Brain Research, 1242, 126–135.

de Gelder, B. and Bertelson, P. (2003). Multisensory integration, perception
and ecological validity. *Trends in Cognitive Science, 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 and 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-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*

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.*

Acknowledgments

A part of this work was supported by a Grant-in-Aid for Specially Promoted Research No. 19001004 from the Ministry of Education, Culture, Sports, Science and Technology, Japan and the Postdoctoral Fellowships for Research Abroad from the Japan Society for the Promotion of Science.

For Review Only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure Captions

Figure 1. The amount of the congruency effects in the Japanese and the Dutch groups obtained from the face task and the voice task. The error bars represent standard errors. * $p < .05$, ** $p < .01$.

For Review Only

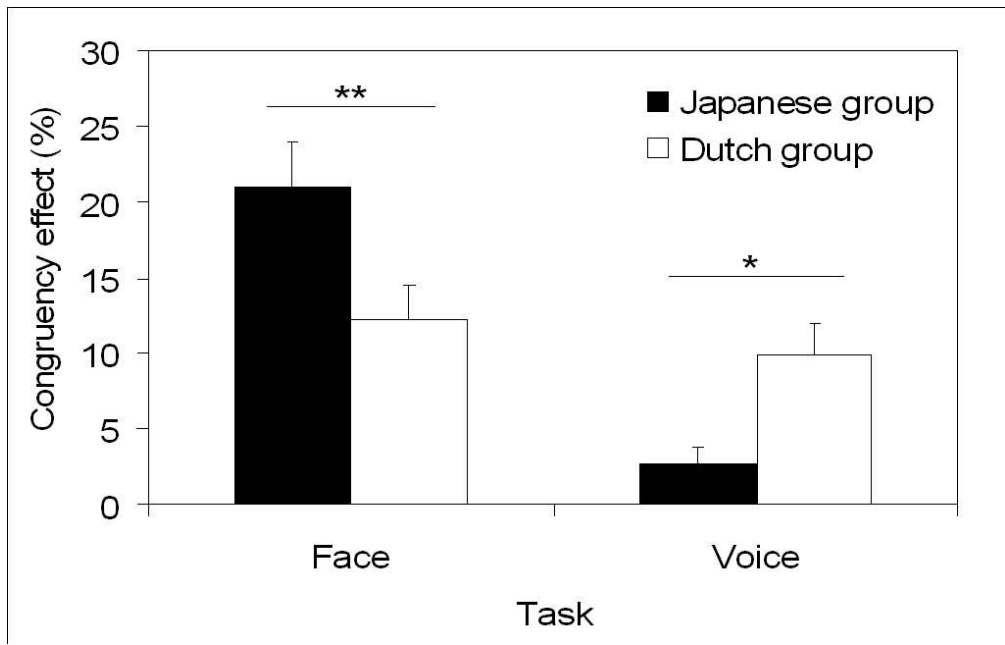
Table 1. Accuracy scores in all conditions.

Group	(a) Face task				Group	(b) Voice task			
	Japanese group		Dutch group			Japanese group		Dutch group	
Stimuli	Japanese	Dutch	Japanese	Dutch	Stimuli	Japanese	Dutch	Japanese	Dutch
Angry face					Angry face				
Angry voice	79.9 (2.7)	76.6 (3.2)	84.8 (3.3)	85.2 (2.8)	Angry voice	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)	Happy voice	94.1 (2.2)	79.9 (3.3)	75.8 (3.9)	83.6 (3.3)
Happy face					Happy face				
Angry voice	66.4 (4.3)	59.5 (3.9)	82.4 (2.3)	70.7 (4.3)	Angry voice	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)	Happy voice	94.7 (1.4)	82.2 (3.2)	92.2 (1.7)	88.7 (2.9)

Note. Mean accuracy is expressed as a percentage. Standard errors are given in parentheses.

For Review Only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



269x171mm (96 x 96 DPI)

ew Only