

## Cross-Cultural Differences in the Multisensory Perception of Emotion

Akihiro Tanaka<sup>1,4\*</sup>, Ai Koizumi<sup>2</sup>, Hisato Imai<sup>3</sup>, Saori Hiramatsu<sup>3</sup>, Eriko Hiramoto<sup>3</sup>, Beatrice de Gelder<sup>4</sup>

<sup>1</sup> Waseda Institute for Advanced Study, Waseda University, Japan

<sup>2</sup> Department of Psychology, The University of Tokyo, Japan

<sup>3</sup> Department of Psychology, Tokyo Woman's Christian University, Japan

<sup>4</sup> Cognitive and Affective Neuroscience Laboratory, Tilburg University, The Netherlands

\* a.tanaka@aoni.waseda.jp

### Abstract

Our recent study [1] showed that culture modulates the manner of the multisensory integration of affective information. Specifically, Japanese are more tuned to vocal processing than Dutch in the multisensory perception of emotion. The current study aimed to extend the findings by adding an experiment and conducting further analyses on the results of their study. In the experiments, pairs of affective faces and voices, expressing either congruent or incongruent emotion, were presented simultaneously. In Experiment 1, although the performance on the facial judgment was very high, a selective interference that angry voice interfered with judgment of happy face was found for the in-group judgment of Japanese participants. Also, the interference effect of to-be-ignored face on the judgment of vocal expressions was smaller in the Japanese group than in the Dutch group. The results suggest that Japanese are more tuned to vocal processing in the multisensory perception of emotion, regardless of whether the difficulty is matched between judgments of facial and vocal expressions. In Experiment 2, the effect of to-be-ignored voice was larger for the in-group than out-group speakers when judging the facial expression, whereas the opposite result was found in the vocal judgment. The results suggest that the relative weight of facial and vocal expressions can be modulated by the familiarity with the speakers.

**Index Terms:** emotion perception, cultural difference

### 1. 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., [2,3]). More recent studies observed considerable cultural differences in the appearance and the perception of facial expressions (e.g., [4-6]).

However, our social interactions involve information from multiple modalities such as faces and voices [7,8]. Literature has shown cross-modal interaction between facial and vocal emotional expressions [8-11]. Thus the very process of integrating emotional cues from different modalities may also be culture sensitive.

Our recent study [1] investigated cultural differences in the multisensory perception of emotion between Japanese and Dutch, using the immediate cross-modal bias paradigm [12]. A pair of dynamic face and voice, which expresses either congruent or incongruent emotions, was presented in the trials (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. 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 results indicated that Japanese are more tuned to vocal processing in the multisensory perception of emotion. Their results suggest that culture modulates the manner of the multisensory integration of affective information.

In the experiment of Tanaka et al. [1], both Japanese and Dutch participants observed both Japanese and Dutch targets. In other words, the participants observed emotional expressions of in-group and out-group speakers. Since Tanaka et al. [1] focused on the differences in the judgments of facial and vocal expressions between Japanese and Dutch participants, the differences in emotion perception between the stimuli of in-group and out-group speakers were not fully analyzed. Thus it remains unsolved whether the influence of the to-be-ignored modality differs between the in-group and the out-group speakers.

Tanaka et al. [1] decreased the visibility of the face stimuli so that unisensory performance was matched between face-only and voice-only trials. Specifically, they added a dynamic noise to the face images in order to decrease the visibility (e.g., [13]). This setting resembled the emotion perception in the darkness. This enabled them to compare the differences between facial and vocal judgments without ceiling effect. However, in the ideal situation, both face and voice are clearly perceivable. Besides matching the difficulty between the judgments of facial and vocal expressions, it is also important to examine the crossmodal bias under more naturalistic situation (i.e., without noise on the face).

The current study aimed to extend the findings of Tanaka et al. [1] by adding an experiment and conducting further analyses on the results of their study. In Experiment 1, clear face, not the face in noise, and voice were presented in order to examine the crossmodal effect under more naturalistic environment. We also investigated whether the influence of the to-be-ignored modality differs between the in-group and the out-group speakers.

### 2. Experiment 1

#### 2.1. Methods

Participants were 24 native speakers of Japanese living in Japan (ages 19–22, 12 female) and 20 native speakers of Dutch living in The Netherlands (ages 18–24, 16 female). They have never lived abroad for more than one year. All of them reported normal hearing and normal or corrected-to-normal visual acuity.

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. Each fragment had an equivalent meaning between Japanese and Dutch. 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. Loudness (root mean square values of the amplitude) and the number of syllables were matched between languages.

The recorded materials were validated before the experiment. In a validation test, Japanese and Dutch speakers' emotional utterances were presented to four Japanese and four Dutch speakers, respectively. Three tokens from the combination of each speaker, fragment, and emotion were presented. Stimuli were presented unimodally (auditorily and visually). Participants categorized the emotion of the unimodal stimuli to either happy or angry. The token, whose auditory and visual recordings were correctly perceived by all the four raters, was used in the experiment.

Congruent and incongruent stimuli were created from the original audiovisual fragments. The unchanged fragments served as congruent stimuli. In order to make incongruent stimuli, the audio angry (happy) token was dubbed onto the audio happy (angry) token. The duration of spoken fragments was matched between happy ( $M = 771.1$  ms,  $SD = 88.5$  ms) and angry ( $M = 767.4$  ms,  $SD = 95.7$  ms) tokens [ $t(15) = 0.34$ ,  $p = .74$ ]. Since the amount of the difference in the duration of the corresponding fragments between happy and angry tokens ( $M = 27.3$  ms,  $SD = 31.9$  ms) was within the range of the temporal window of integration in audiovisual speech (around 200 ms for audio delay and around 100 ms for audio lead [14,15]), it was almost impossible for participants to notice the audiovisual asynchrony and use the asynchrony as a cue to their responses.

Happy and angry facial expressions were combined with happy and angry vocal expressions for each of the eight utterances (two speakers  $\times$  four fragments), resulting in a total of 32 bimodal stimuli in each language. A trial consisted of a 1-s fixation point around the speakers' mouth and a simultaneous presentation of dynamic face (happy or anger) and voice (happy or anger) on CRT and through loudspeaker, respectively. In the face task, participants were instructed to categorize the emotion of the faces into happy or angry and ignore the voice. In the voice task, participants were instructed to categorize the emotion of the voices, to look at the face when the voice was being presented but to ignore the face when rating the voice. Participants responded by pressing either of the two buttons, which were counterbalanced between participants. The experiment consisted of four sessions, in which the task (face or voice) and speaker (Japanese and Dutch) were different. Each session had 64 trials (two repetitions of the 32 bimodal stimuli).

## 2.2. Results and Discussion

The accuracy scores in all conditions are shown in Table 1. Overall, performance on the face task was very high. However, there was a selective interference that angry voices affected the perception of happy faces when Japanese participants judged the stimuli of Japanese speakers (happy vs. angry voice:  $t(23) = 2.41$ ,  $p < .05$ ).

In order to examine the 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 (Figure 1). The results of the main effect revealed that the congruency effect was much stronger in the voice task than in the face task,  $F(1,42) = 208.82$ ,  $p < .001$ . The asymmetric congruency effect suggests that when both face and voice are clearly perceivable, face is more dominant than voice in the multisensory perception of emotion.

There was a two-way interaction between Task and Group [ $F(1,42) = 5.74$ ,  $p = .02$ ]. Analyses of the simple main effect revealed that the congruency effect was smaller in the Japanese group than in the Dutch group in the voice task [ $F(1,84) = 5.84$ ,  $p = .02$ ], suggesting that Japanese people are better at focusing on the voice and ignoring the face. These results suggest that Japanese are more tuned to vocal processing than Dutch even when the difficulty is not matched between facial and vocal judgements.

Importantly, a three-way interaction,  $F(1,42) = 77.50$ ,  $p < .001$ , was significant. In order to compare the cultural difference in the cross-modal bias for each task, a Group  $\times$  Speaker two-way ANOVAs were conducted separately for the voice task and the face task. For the voice task, congruency effect was larger for out-group than for in-group stimuli,  $F(1,42) = 21.92$ ,  $p < .001$ . The interaction was also significant,  $F(1,42) = 70.35$ ,  $p < .001$ . For the face task, the main effect of Group was only marginally significant,  $F(1,42) = 3.77$ ,  $p = .06$ . The main effect of Speaker,  $F(1,42) = 3.45$ ,  $p = .07$ , and the interaction,  $F(1,42) = 0.75$ ,  $p = .39$ , were not significant.

The main effect of the speaker in the voice task suggests that the relative weight of facial and vocal expressions in emotion perception is modulated by the familiarity with the speakers. The overall performance in the face task was very high in Experiment 1 (98.0% in average). Therefore, some possible differences may have been masked by a ceiling effect. In Experiment 2, the difficulty was matched between the face and voice stimuli by decreasing the visibility of the face stimuli. This enables us to compare the differences between face and voice judgments without ceiling effect in the face task. In line with the findings of Collignon et al. [13], which used both clear and noisy faces, we expect larger effect of the voice on the facial judgment when the face is degraded.

Table 1: Mean percentage accuracy in all conditions (Experiment 1). Standard errors are given in parentheses.

Voice task				
Stimuli	Japanese group		Dutch group	
	Japanese speaker	Dutch speaker	Japanese speaker	Dutch speaker
Angry face				
Angry voice	98.7 (0.6)	93.0 (2.0)	89.4 (1.7)	94.7 (1.7)
Happy voice	89.3 (4.6)	47.7 (4.3)	64.1 (3.8)	42.5 (3.5)
Happy face				
Angry voice	87.0 (3.0)	37.8 (3.9)	51.6 (4.9)	55.9 (5.4)
Happy voice	98.2 (0.9)	89.1 (1.5)	90.6 (2.2)	89.7 (2.8)
Face task				
Stimuli	Japanese group		Dutch group	
	Japanese speaker	Dutch speaker	Japanese speaker	Dutch speaker
Angry face				
Angry voice	99.2 (0.6)	99.5 (0.4)	99.7 (0.3)	99.4 (0.4)
Happy voice	99.0 (0.7)	97.7 (0.9)	99.7 (0.3)	97.5 (1.4)
Happy face				
Angry voice	90.6 (3.4)	99.0 (0.6)	98.4 (0.7)	98.4 (0.7)
Happy voice	99.0 (0.6)	99.7 (0.3)	98.4 (0.7)	98.8 (0.7)

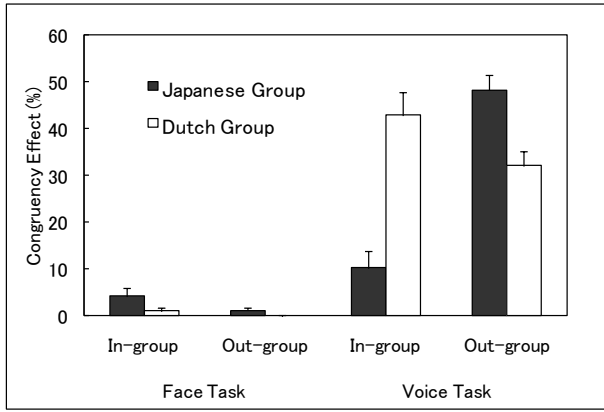


Figure 1: Congruency effects in the face task and the voice task among Japanese and Dutch participants in Experiment 1. Error bars represent standard errors.

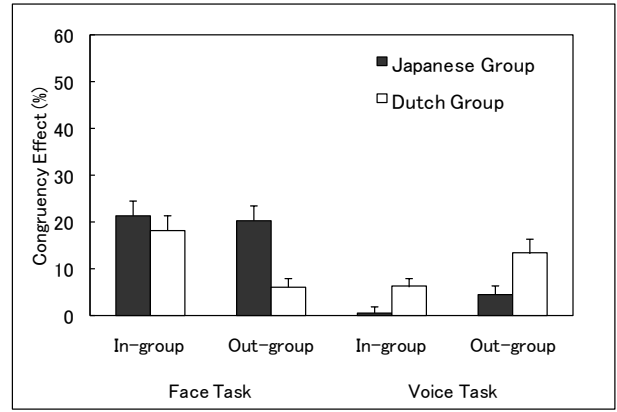


Figure 2: Congruency effects in the face task and the voice task among Japanese and Dutch participants in Experiment 2. Error bars represent standard errors.

### 3. Experiment 2

#### 3.1. Methods

Participants were 20 native speakers of Japanese living in Japan (ages 21-29, 9 female) and 16 native speakers of Dutch living in The Netherlands (ages 18-30, 13 female). Dynamic noise was added to the visual stimuli used in Experiment 1. The signal-to-noise ratio was adjusted so that the performance is matched between the face task and the voice task, based on the results of a preliminary experiment. In Experiment 2, unisensory control tasks, in which only the face or voice was presented, were also conducted. Namely, we conducted a face-only task (for Japanese and Dutch speakers) and a voice-only task (for Japanese and Dutch speakers). Apart from these differences, experimental procedure was exactly the same as in Experiment 1.

Table 2: Mean percentage accuracy in the multisensory conditions (Experiment 2). Standard errors are given in parentheses.

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

#### 3.2. Results and Discussion

Data from one Japanese participant were excluded from the analysis since the task was misunderstood. In unisensory sessions, as the noise was added to the face images, the performance of the face-only task ( $M = 83.9\%$ ) in Experiment 2 was lower than in Experiment 1 (clear face). In line with previous studies which examined emotion perception from vocal expressions (e.g., [16]), the performance of the voice-only task ( $M = 87.2\%$ ) was not very high despite the fact that no noise was added, and comparable to that of the face-only task. The difference between the performance of the face-only and the voice-only tasks was not significant,  $t(34) = 1.57$ ,  $p = .12$ , confirming that the difficulty was closely matched between tasks.

Performance in all multisensory conditions is shown in Table 2. It is important to note that in Experiment 2, the interference in the face task, which was observed in Japanese group in Experiment 1, was not specific to the condition that angry voice was accompanied with happy face of Japanese speaker, but widely observed in any kinds of incongruent conditions and for both participant group.

In order to examine the cross-modal bias, a Task  $\times$  Group  $\times$  Speaker mixed-factor ANOVA was performed on congruency effects (Figure 2). The congruency effect was stronger in the face task than in the voice task,  $F(1,33) = 19.19$ ,  $p < .001$ . Effects of Group,  $F(1,33) = 0.09$ ,  $p = .76$ , and Speaker,  $F(1,33) = 0.16$ ,  $p = .69$ , were not significant. A two-way interaction between Task and Speaker,  $F(1,33) = 18.11$ ,  $p < .001$ , was significant. A two-way interaction between Task and Group was also significant,  $F(1,33) = 11.48$ ,  $p = .002$ , showing that congruency effect is larger in Japanese than in Dutch group when they judged facial expressions whereas the opposite tendency was found when they judged vocal expressions.

Importantly, a three-way interaction,  $F(1,33) = 6.08$ ,  $p = .02$ , was 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 voice task and the face task. For the voice task, the main effect of Group was significant,  $F(1,33) = 10.27$ ,  $p = .003$ , showing that congruency effect was larger in Dutch than in Japanese participants. Also congruency effect was larger for out-group than for in-group stimuli,  $F(1,33) = 7.58$ ,  $p = .01$ . The interaction was not significant,  $F(1,33) = 0.60$ ,  $p = .44$ . For the face task, the main effect of Group was significant,  $F(1,33) =$

4.88,  $p = .03$ , showing that the congruency effect was larger in Japanese than Dutch participants. The main effect of Speaker,  $F(1,33) = 10.44$ ,  $p = .003$ , and the interaction,  $F(1,33) = 7.13$ ,  $p = .01$ , were also significant. Simple main effect analyses showed that the congruency effect was not different between the in-group and out-group stimuli in Japanese participants,  $F(1,33) = 0.16$ ,  $p = .69$ , whereas it was larger for in-group than out-group stimuli in Dutch participants,  $F(1,33) = 17.42$ ,  $p < .001$ .

Results showed that the congruency effect was larger for the in-group than out-group speakers in the face task whereas the opposite result was found in the voice task. These results suggest that participants are more tuned to vocal expressions when they judge the multisensory expressions of in-group than out-group speakers.

The results in the voice-only task showed that vocal expressions of the in-group speakers were better recognized than those of the out-group [Japanese group:  $t(18) = 3.52$ ,  $p = .002$ ; Dutch group:  $t(15) = 2.43$ ,  $p = .03$ ]. In contrast, we did not observe any differences between in-group and out-group speakers in the face-only task [Japanese group:  $t(18) = 0.31$ ,  $p = .75$ ; Dutch group:  $t(15) = 1.24$ ,  $p = .23$ ].

#### 4. General Discussion

Following our recent study [1], the present study provided additional findings on how the multisensory perception of emotion is influenced by cultural background of the observers. First, Japanese are more tuned to vocal processing in the multisensory perception of emotion, regardless of whether the difficulty is matched between judgments of facial and vocal expressions. Second, the relative weight of facial and vocal expressions is modulated by the familiarity with the speakers.

Our results showed that the crossmodal bias depends on the combination of the group and the stimuli. When the observers have the same cultural background as the speakers (i.e., in-group stimuli), observers are more tuned to vocal processing in both face and voice tasks. Previous studies on emotion perception of facial expression have shown the in-group effect that the emotion is well recognized when members of the same group that expressed the emotions also make the judgments (for a meta-analysis, see [4]). Our results revealed a larger in-group effect in the perception of vocal expressions than facial expressions in the unisensory tasks. Based on the findings from unisensory trials, our results in multisensory trials can be explained as follows. In the face task, the crossmodal bias was larger for the in-group than for the out-group stimuli, because the to-be-ignored vocal expressions of the in-group speakers were more automatically perceived than those of the out-group and therefore interfered with the processing of the faces. In contrast, in the voice task, it was larger for the out-group stimuli than for the in-group stimuli, because the vocal expressions of the out-group speakers were perceived less accurately than those of the in-group and thus participants were more susceptible to the accompanying face. These results are consistent with the notion of the “inverse effectiveness” that multisensory integration is stronger when the unisensory stimuli are relatively ambiguous when presented in isolation [17].

Experiment 1 showed that when both face and voice are clearly perceivable, the congruency effect was much stronger in the voice task than in the face task. However, this asymmetry disappeared in Experiment 2, in which the difficulty was matched between facial and vocal judgments by adding noise to the face. Our results extend previous findings that have shown the visual superiority in audiovisual

integration (e.g., McGurk effect and ventriloquist effect) to the perception of emotion. The visual superiority in emotion perception can also be explained by the principle of inverse effectiveness.

One interesting result from the accuracy scores is the selective interference found in Experiment 1. Angry voices affected the perception of happy faces in the Japanese participants only when the speaker was Japanese and when the face was clearly visible. This condition resembles the situation that a person hides his or her true (negative) feeling by making the face (positive) smiling. This kind of faked expression is based on the “display rule” [2], which depends on culture and individuals. For example, Japanese inhibit a direct expression of negative emotions [18]. Consequently, East Asian participants exhibit lower performances in perceiving some negative facial expressions compared to Western observers (see [4] for a meta-analysis). Similarly, this tendency in Japanese might lead to decreased performance in the Japanese participants when rating angry emotions expressed by multiple modalities. However, our data are not consistent with this straightforward prediction. Rather, our results can be interpreted as showing that Japanese participants rely on the vocal affect to judge the emotion of a speaker especially when the voice expresses negative emotion and the face expresses positive expression. When Japanese speakers get angry, they hide their facial expressions but it might be difficult to hide the emotion expressed by their voice. Thus, multisensory research on emotion can reveal the cultural differences in the display rule in ecologically valid situations. Of note, this selective interference was not specific to that condition, but widely observed in any kinds of incongruent conditions. It might be that the selectivity is most salient when both face and voice are clearly perceivable as in the naturalistic situation.

To conclude, the present study provided additional findings on how the multisensory perception of emotion is influenced by cultural background of the observers. Japanese are more tuned to vocal processing in the multisensory perception of emotion, regardless of whether the difficulty is matched between judgments of facial and vocal expressions. The relative weight of facial and vocal expressions is modulated by the familiarity with the speakers.

#### 5. Acknowledgements

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. The first author (A.T.) was supported by the Postdoctoral Fellowships for Research Abroad from the Japan Society for the Promotion of Science.

#### 6. References

- [1] Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., de Gelder, B. “I feel your voice: Cultural differences in the multisensory perception of emotion,” *Psychological Science*, in press.
- [2] 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.
- [3] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17, 124–129.
- [4] Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128, 203–235.

- [5] 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.
- [6] 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.
- [7] Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, 11, 535–543.
- [8] de Gelder, B. and Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Science*, 7, 460–467.
- [9] 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.
- [10] de Gelder, B., Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, 14, 289–311.
- [11] Massaro, D. W. & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, 3, 215–221.
- [12] 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.
- [13] 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.
- [14] Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk Effect. *Perception & Psychophysics*, 58, 351–362.
- [15] van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in bimodal speech. *Neuropsychologia*, 45, 598–607.
- [16] Scherer, K. R., Banse, R., & Wallbott, H. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32, 76-92.
- [17] Meredith, M. A. & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. *Science*, 221, 389–391.
- [18] Ekman, P. & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Englewood Cliffs, New Jersey: Prentice-Hall.