Cultural similarities and differences in the recognition of audio-visual speech stimuli

Sumi Shigeno

Kitasato University

Teacher Training Courses, 1-15-1, Kitasato, Sagamihara, 228-8555 Japan

ABSTRACT

Cultural similarities and differences were compared between Japanese and North American subjects in the recognition of emotion. Seven native Japanese and five native North Americans (four Americans and one Canadian) subjects participated in the experiments. The materials were five meaningful words or short-sentences in Japanese and American English. Japanese and American actors made vocal and facial expression in order to transmit six basic emotions—happiness, surprise, anger, disgust, fear, and sadness. Three presentation conditions were used—auditory, visual, and audio-visual. The audio-visual stimuli were made by dubbing the auditory stimuli on to the visual stimuli. The results show: (1) subjects can more easily recognize the vocal expression of a speaker who belongs to their own culture, (2) Japanese subjects are not good at recognizing "fear" in both the auditory-alone and visual-alone conditions, (3) and both Japanese and American subjects identify the audio-visually incongruent stimuli more often as a visual label rather than as an auditory label. These results suggest that it is difficult to identify the emotion of a speaker from a different culture and that people will predominantly use visual information to identify emotion.

1. INTRODUCTION

The investigation of cultural similarities and differences in the recognition of emotion have mainly focused on facial expression. For instance, the findings from the study by Ekman & Friesen (1971) provided evidence for cultural agreement about the universal facial expression of six basic emotions—happiness, surprise, anger, fear, disgust, and sadness. It was also reported that among these emotions there was high agreement for some emotions (e.g., in happiness) and low agreement for others (e.g., in fear and sadness) (Ekman & Friesen, 1975).

Face-to-face communication, however, is rather rare in our daily-life and we often don't see the speaker's face while he/she is speaking, as Ekman & Friesen (1975) noted. Instead we use the auditory (vocal) or audio-visual (vocal and facial) expression of emotion more often than facial expression to recognize the speaker's expression. Although research on the auditory and audio-visual expression of emotion is very important, little research has been done in this area.

Meanwhile those who are brought up in the culture where the expression of emotion is parsimonious are not good at perceiving others' emotion as well as at expressing their own. It is one possibility that those who don't use facial expression in communication might use other signals such as voice. For example, Japanese express strong emotion on their faces to a lesser extent than Americans, but they can usually interact with others without any difficulties. If one hides his/her true emotion and disguises it with the facial expression of other emotions, are there any differences between Japanese and Americans in perceiving or guessing the true emotion?

The present study aims at exploring cultural similarities and differences when the auditory and visual emotions are incongruent. Experiments were conducted to compare the performance in recognition of audio-visually incompatible emotions between Japanese and Americans and to discuss how one's influences this recognition.

2. EXPERIMENT

2.1. Method

Subjects. Seven native Japanese and five native North Americans (four Americans and one Canadian) undergraduate students participated in the experiment. None of them had ever lived outside of their countries for more than one year.

Stimuli. The materials were five meaningful words or short sentences (which express no feelings on their own) in Japanese and American English, which are listed in Table 1. Two professional actors in their thirties—one Japanese and the other American—made vocal or facial expressions so that they could transmit the six basic emotions as accurately and naturally as possible. Voices and faces were recorded separately—faces were first recorded and then voices. The six basic emotions were happiness, surprise, anger, disgust, fear, and sadness. The duration of utterance was about 693-1254 ms in Japanese and about 858-1650 ms in English. The utterances and facial expressions were recorded several times and two people including the experimenter selected the best voice and face for each stimulus. The best voice was then dubbed on to the best faces to make audio-visual stimulus. Thus 36 audio-visual stimuli were made (6 vocal expressions x 6 facial expressions).

2.2 Procedure

There were three presentation conditions—auditory-alone, visual-alone, and audio-visual. In the auditory-alone condition, subjects were instructed to listen to what the speaker said and to judge the emotion he expressed. In the visual-alone condition, subjects were required to carefully watch his facial movements without sound and to judge the emotion he expressed. In the audio-visual condition, subjects were required to watch the actor's face while listening to his voice and to judge the emotion he expressed. The subjects did not have to respond what he said in any of the conditions.

As no differences were found among the five words or short sentences, the results were summed. In the auditory-alone and visual-alone conditions, 20 responses were obtained per each subject for each of the six emotions. In the audio-visual condition, 10 responses were obtained.

The experiment was conducted in a sound-proof room. The stimuli were replayed by an S-VHS video deck (Victor, HR-Z1) and presented to the subjects on a 29-intch monitor and through the speaker attached to the monitor. Speech stimuli were presented to the subjects at 68-75 dB SPL. The distance from the monitor to the subjects was about 1 m73 cm. During the experiment, the experimenter sat next to the subject and kept checking whether the subject was watching the monitor.

2.3. Results and discussion

The averaged results from all subjects are shown in Figure 1.

Auditory-alone condition (Figure 1 (a)). The percent of correct responses of the six emotions ranged from 64% to 86%. Japanese subjects recognized the vocal expression of Japanese better than that of English, while the Americans recognized English vocal expression better than that of Japanese. That is, subjects can recognize the vocal expressions of a speaker who belongs to their own culture better than one who belongs to another culture (This is referred to here as "speaker effect").

Visual-alone condition (Figure 1 (b)). The percent of correct responses ranged from 90% to 99%. Overall the percentages were higher than those in the auditory-alone condition and both Japanese and American subjects show slightly higher performances in the recognition of the American speaker than of the Japanese speaker. The tendencies observed in the auditory-alone condition (speaker effect) were not obtained here.

Figures 2 (auditory-alone condition) and Figure 3 (visual-alone condition) represent the results for each emotion. Both in the auditory-alone and visual-alone conditions, Japanese subjects show lower percentages in the case of "fear". American subjects don't show this tendency. The results suggest that the Japanese are not good at recognizing "fear" for both vocal and facial expression.

Table 1: The words and short sentences used for the experiments.

Japanese	English
Tookyoo	New York
Kawarazaki-san	Rio de Janeiro
Juuichiji-han	Margaret
sayoonara	Saturday
soodesuka	Is that so?

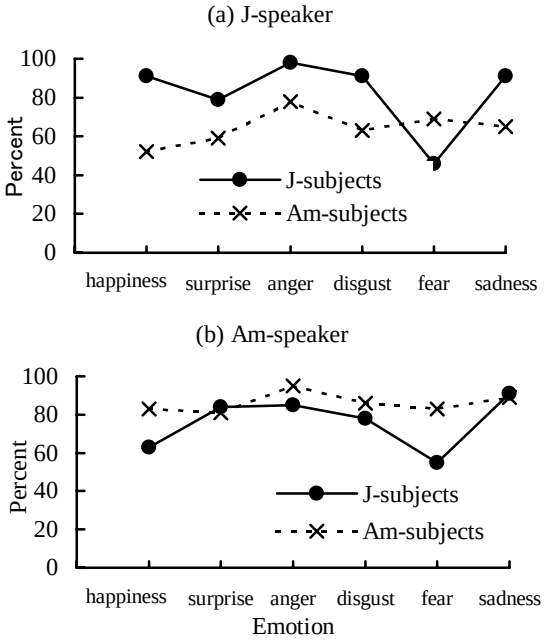

Figure 2: Percent of identification for each emotion by the (a) Japanese and (b) American subjects in the auditory-alone condition.

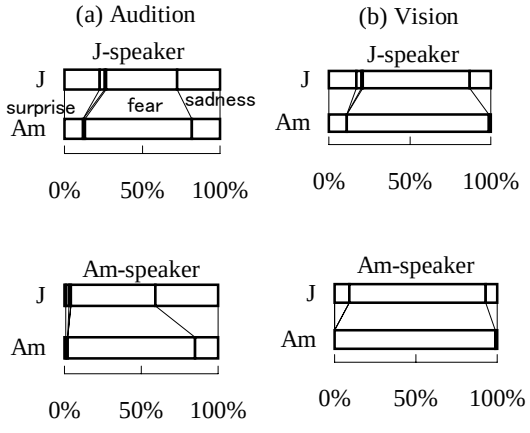

Figure 4: Percent of identification for "fear" in the (a) auditory-alone and (b) visual-alone condition.

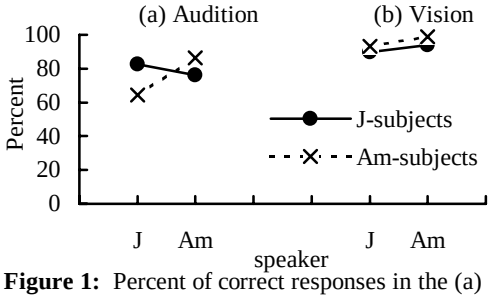

Figure 1: Percent of correct responses in the (a) auditory- and (b) visual-alone condition obtained form the Japanese and American subjects.

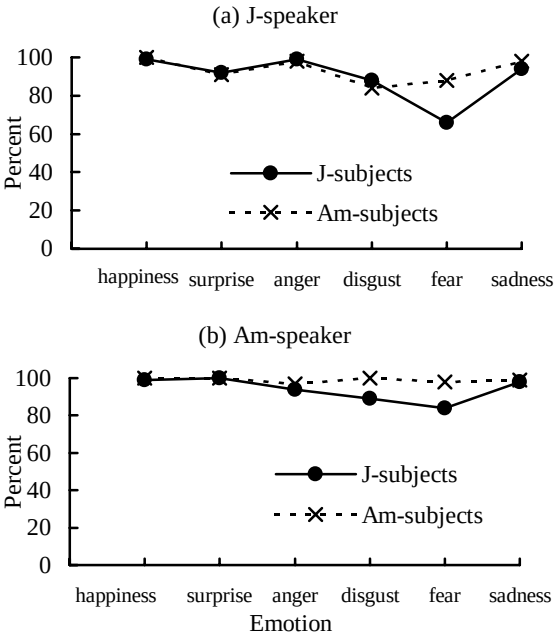

Figure 3: Percent of identification for each emotion by the (a) Japanese and (b) American subjects in the visual-alone condition.

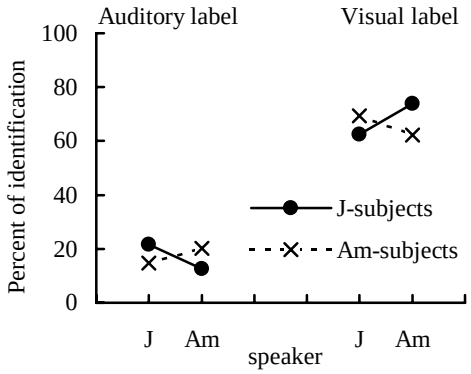

Figure 5: Percent of identification for emotion in audition and in vision in the audio-visually incongruent condition.

Figure 4 shows how the subjects identify “fear”. “Fear” is confused with surprise and sadness. Japanese subjects confused it with surprise (e.g., in the case of auditory-alone condition with a J-speaker: 23%) or sadness (46%). American subjects also confused it with surprise or sadness, although at a lower rate than the Japanese.

Audio-visual condition (Figure 5). Figure 5 represents the results from the audio-visually incongruent condition in which vocal and facial expressions are different. In Figure 5, the left graph shows the percent of identification labeled as the auditory emotion (voice) and right graph shows the percent labeled as the visual emotion (face). Both Japanese and American subjects identify the audio-visually incongruent stimuli more often as visual label than as auditory label. These results suggest that when we are required to identify the audio-visually incongruent emotion, visual information is used much more than auditory information. Furthermore, the “speaker effect” observed in the auditory-alone and visual-alone conditions is obtained only when subjects identify the stimuli as auditory label. When subjects identify the stimuli as visual label, the tendencies are reverse: the percentage is higher in the case of the speaker who belongs to a different culture from the subjects than one who belongs to the same culture. These results suggest that it is difficult to identify the emotion of a speaker from a different culture and we will often use visual information to identify the emotion. On the other hand, the emotion of a speaker is slightly easier to identify if the speaker comes from the same culture and we will use both auditory and visual information to guess his/her true feeling when it is disguised by another emotion.

3. ACKNOWLEDGMENTS

This research was supported by the Grant-in-Aid for Scientific Research on Priority Areas (No. 09207218) and the Grant from the Sound Technology Promotion Foundation (1997).

4. REFERENCES

1. Ekman, P., and Friesen, W. V. “Constants across cultures in the face and emotion,” J. Personal. Soc. Psychol., 17, 124-129, 1971.
2. Ekman, P., and Friesen, W. V. “Unmasking the face,” Englewood Cliffs, NJ: Prentice Hall, 1975.