Emotional McGurk Effect? A Cross-Cultural Investigation on Emotion Expression under Vocal and Facial Conflict

Aijun Li[1], Qiang Fang[1], Yuan Jia[1], and Jianwu Dang[2]

[1]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
{liaj,fangqiang,jiayuan}@cass.org.cn
[2] Tianjin University, Tianjin, China
Dangjianwu@tju.edu.cn

Abstract. A multi-modal emotion perceptual experiment is conducted cross-culturally to investigate the difference of expressing and perceiving emotions across Chinese and Japanese. Focus is on the cultural effect on the interaction between the combination of vocal and facial expression and perception. In this paper, part of the perceptual results is reported for the AV-conflicting stimuli produced by a Chinese female speaker and perceived by Chinese and Japanese listeners. The results support the assumptions that (i) When listeners decoding the conflicting AV stimuli, they might rely on some modality more than another across different emotions. (ii) Although common psychological factor contributes to the emotional communication, the decoding of conflicting AV information will be affected by culture background, and (iii) the emotional McGurk effect exists, and it may also be related to cultural norms of the encoder/listener.

Keywords: Emotion, emotional McGurk effect, Cross culture, multimodalily.

1 Introduction

Emotion theories and emotion psychology are currently concerned with the universality and cultural relativity of emotional expression, which, in fact, is a key issue exploring 'what is the essence and function of emotion?' It is generally acknowledged that emotional expression is both psychobiologically and culturally controlled, but the respective effect imposed by psychobiology and culture on emotional expression remains unexplored. The earliest predecessors studying cross-cultural emotion include Charles Darwin[1], Ekman[2] and Izard[3]. They notice that listeners from one culture have the ability to decode the facial expression of an actor from another culture. They claim that like decoding facial expression of emotion, people from different culture can decode vocal expression of emotion. Therefore, from the psychobiological perspective, emotion decoding is universal. Cross-cultural studies on emotion encoding and decoding are needed to supply speech technology with culturally-relative emotional expressions. Erickson[14] made a review on cross-linguistic studies, recently, lot of research have been carrying on cross-cultural research on emotional speech[4-11], some of the results are consistent such as the perception of emotional expressions is more successful

M. Sun et al. (Eds.): CCL and NLP-NABD 2013, LNAI 8202, pp. 214–226, 2013.

when the stimuli are multimodal, facial expression plays a major role in the correct decoding of emotion; some are inconsistent such as the recognition of the emotions was not influenced by cultural differences[11], while some thought there exist cross-cultural difference[4-6]. Besides, they suggest "Anger joy and sad" may constitute three basic emotions[10], and listeners from different cultures show different sensitive degrees to the acoustic parameters when decoding emotions.

Speech communication is a physiological process, conveying both audio and visual information. Human's perception bases on the information transmitted by both channels. Generally, in speech communication, the information from two channels is complementary and coherent. But when the information is conflicting and nevertheless integrated then the percept in one of the modalities might be changed by the other modality. [12]

Fagel [12] claims that the stimulus with conflicting audio and visual content can be perceived as an emotion which is neither the emotion indicated by the audio information nor the emotion indicated by the visual information, which is called emotional McGurk Effect. It is assumed that the valence (positive or negative emotion) is primarily conveyed by the visual channel while the degree of arousal is reflected by the audio channel. A match of a positive facial expression with a negative voice will be perceived as joy. However, the identification of content emotion will be derived from the combination of sad voice with happy facial expression. Other mismatches between audio and visual information are only perceived as either the emotion indicated by audio channel or the emotion indicated by visual channel.

Since the encoding and decoding of emotion may depend on multiple modalities and language backgrounds, the purpose of the present study is to clarify the process of the encoding and decoding of emotion through a cross-cultural perceptual experiment for multi-modal emotions. The preliminary analysis on Chinese and Japanese listening to emotional speech of a Chinese speaker in three conditions of congruent audio-video, audio-only and video-only, reveals that language and culture will impose an influence on the identification of emotion. In the present paper, we will continue to explore the emotional speech communication but modulated in conflicting AV channels. Here, the issues concerned are as: (i) what is the interplay between the two conflicting AV channels in conveying emotional information? (ii) Does the emotional McGurk effect exist when the emotions are conveyed in conflicting channels? (iii) Are there any culture effects on perception on the conflicting AV emotions?

The assumptions are: (i) When listeners decode the conflicting AV stimuli, they might rely on some modality more than others across different emotions, i.e. one modality should have stronger emotional modulation for some emotions than that in another modality. (ii) Although the common psychological factor contributes to the emotional communication, the decoding of conflicting AV information will be affected by linguistic and cultural background and (iii) the emotional McGurk effect may also be related to culture norms of the encoder/listener.

2 Perceptual Experiment on Cross-Cultural Emotion

Table 1 lists the Chinese and the corresponding Japanese prompts. In order to control the time spent in the experiment, the prompts were divided into two sets. The sentences

were matched in the number of syllables, from 1 to 5, with different tonal combinations, in different grammatical structures. The contents of the texts were emotionally neutral.

The speech data used in the present paper is from a Chinese female student from Beijing Film Academy, who speaks Standard Chinese. Her emotion speech was videotaped with Canon Power Shot TX1 in the sound-proof room. She uttered the prompts in Table 1 in seven emotional states. The seven emotions are classified by valence (positive or negative emotions) : Happiness is positive and 'SAdness, ANger, DIsgust and Fear' are positive; while 'Surprise' can be positive or negative. In terms of the degree of arousal, 'Sadness and Fear' are being low arousal, while 'Happiness, Anger and Disgust' are being high arousal.

Table 1. Chinese and Japanese prompt

Set 1	Set 2
S1-1 妈 お母さん (mother) S2-1 大妈 おばさん (auntie) S3-1 吃拉面 ラーメンを食べる (to eat noodle) S4-1 打高尔夫 ゴルフをする (to play golf) S5-1 张雨吃拉面 張雨さんはラーメンを食べる (Zhangyu eats noodles.)	S1-4 骂 ののしる (to blame) S2-2 踢球 サッカーをする (to play football) S3-2 奥运会 オリンピック (Olympic Games) S4-2 足球比赛 サッカーの試合 (football match) S5-2 滑雪场教练 スキー場のスキーコーチ (coach of ski resort)

In order to explore the conflicting channel and the McGurk phenomenon, conflicting AV stimuli were obtained through dubbing a visual emotion with another vocal emotion for the same sentence. Then 5*7*7=245 dubbed stimuli were obtained for each set including 35 congruent AV tokens.

Listeners were 10 Chinese college students not knowing Japanese and 10 Japanese college students not knowing Chinese. They were recruited to identify the emotional states for all the dubbed stimuli and rate the expressive degrees on a 5-point scale (0-4), multiple choices are allowed. The higher the score the more expressive the stimulus is.

3 Results and Analysis

The perceived scores were averaged for each intended emotions to obtain the confusing pattern for Chinese and Japanese listeners respectively. To depict the perceptual patterns clearly, a kind of spider graphs representing the average perceptual scores from these10 Chinese and 10 Japanese listeners are plotted in Figures 1 to 4. Each ring in the graph represents the distribution of the rating scores of one perceived emotion for combinations of one facial (/vocal) expression (modality 1) and seven vocal (/facial) expressions (modality two). These rings are called here *Emotion Rings*. The changes in the shape and radius reflect the change of the perceptual patterns. If the

ring symmetrically distributes in all direction like a circle, then the perceived emotion is not related to the second modality. However, if the ring is in an unsymmetrical distribution, it means that the facial-vocal combination with higher scoring has a stronger tendency to be perceived as that emotion, on the contrary, the lower score the smaller chance. The variation of diameter size correlates with rating scores in various facial-vocal combinations.

3.1 Comparison on Perceptual Patterns

3.3.1 Perceptual Results for Chinese Listeners

(1) Fig. 1 (A) indicates that when *'Neutral'* facial expression is dubbed with the seven vocal emotions, the two primarily perceived emotions are *'Neutral' and 'Surprise'*. And the distribution patterns of the two emotion rings show a tendency to complement each other. The combinations of *'Neutral'* face with *'Neutral', 'Happy', 'Fear', 'Sad'* and '*Disgust'* voices tend to be perceived as *'Neutral'* while the combinations of *'Neutral'* face with *'Angry' and 'Surprise'* voices tend to be perceived as '*Surprise'*. Fig.2 (A) shows that the combinations of *'Neutral'* voice with varied facial expressions are mainly perceived as *'Neutral'*. Except for the combination of a *'Neutral'* voice with a *'Happy'* face, which is perceived as *'Happy'*, almost all combinations are perceived as *'Neutral'*. The combination of *'Neutral'* voice with *'Angry'* face is perceived as either '*Disgust'* or *'Neutral'* in equal probability, which is another exception.

(2) Fig.1 (B) indicates that the combinations of *'Happy'* facial expression with varied emotional voices tend to be perceived as *'Happy'*, which is illustrated by an evenly distributed emotion ring. It means that the perception of *'Happy'* depends more on visual information than audio information, although the facially *'Happy'* emotion also initiates the percept of '*Surprise' and 'Neutral'* as shown by the two small rings in the center. Fig.2 (B) reveals that the combinations of *'Happy'* voice with varied facial expressions (except for *'Happy'* face) could not be correctly perceived as *'Happy'*. The combinations of *'Happy'* voice with *'Sad', 'Surprise', or 'Neutral'* faces are perceived as *'Neutral'* emotion.

(3) Fig.1 (C) displays the complicated perceptual patterns activated by dubbing *'Angry'* face with varied emotional voices. The integrations of *'Angry'* face with *'Happy', 'Disgust' and 'Angry'* voices can be perceived as '*Anger', 'Surprise'* or '*Disgust'* with almost equal scores. Fig.2 (C) shows that the combinations of *'Angry'* voice with varied facial expressions are primarily perceived as '*Surprise'*. Only the combination of *'Angry'* voice with *'Happy'* face is perceived as *'Happy'*.

(4) Fig.1 (D) shows that the combinations of '*Disgust'* face with varied emotional voices could not be correctly perceived as '*Disgust'*. When '*Disgust'* face goes with '*Disgust', 'Surprise', 'Angry' and 'Happy'* voice, the percept of *'Surprise'* is induced. When '*Disgust'* facial expression is combined with 'Neutral' voice, the percept of *'Neutral'* emotion is initiated. The combination of '*Disgust'* facial expression with *'Sad'* voice is perceived as *'Sad'* with very low rating scores. Fig.2 (D) specifies that

the perceptual pattern of combinations of '*Disgust'* voice with varied facial expressions is similar to that shown in Fig.1 (D), with most combinations being perceived as '*Surprise'* except that the combination of '*Disgust'* voice with *'Happy'* facial expression is perceived as *'Happy'* and the combination of '*Disgust'* voice with *'Neutral'* face is perceived as *'Neutral'*.

(5) Fig.1 (E) reveals that most combinations of *'Fear'* expression with varied emotional voices could not be correctly perceived as *'Fear'*; instead, two obvious rings of '*Surprise' and 'Neutral'* emotion are displayed. The percept of '*Surprise* is induced when *'Fear'* expression is combined with '*Disgusted', 'Surprise'* or *'Angry'* voices. The percept of *'Neutral'* emotion is initiated when *'Fear'* expression is dubbed with *'Neutral'* voice. However, the result is vague when *'Fear'* expression is dubbed with *'Sad'*, *'Happy'* or *'Fear'* voices. Fig.2 (E) shows that the perception of *'Fear'* voice with varied facial expressions is ambiguous with rating scores lower than two points except that the combination of *'Fear'* voice with *'Happy'* face brings a percept of *'Happy'*.

(6) Fig.1 (F) shows that when *'Sad'* face is dubbed with varied emotional voices, two emotion rings of '*Surprise' and 'Neutral'* are exhibited in a symmetrical pattern. Specifically, the combinations of *'Sad'* face with *'Angry', 'Surprise' and 'Disgust'* voices lead to the percept of '*Surprise'* and the combinations of *'Sad'* face with *'Neutral' and 'Happy'* voices are perceived as *'Neutral'* emotion. When *'Sad'* face goes along with *'Fear' or 'Sad'* voice, either *'Neutral' or 'Sad'* is perceived with almost equal scores. Fig.2 (F) reveals that two overlapped emotion rings are formed when *'Sad'* voice is combined with varied facial expressions (except for *'Happy'* face), namely, *'Sad' and 'Neutral'* emotion rings. However, their rating scores are very low, which are less than 2 points. The combination of *'Sad'* voice with *'Happy'* face is more likely to be perceived as *'Happy'*.

(7) The two symmetrically distributed emotion rings in Fig.1 (G) display the patterns of the *'Surprise'* face dubbed with varied emotional voices: a *'Surprise'* ring derived from the combinations of *'Surprise'* face with *'Angry', 'Surprise'* and *'Disgust'* voices; and a *'Neutral'* emotion ring derived from the combinations of *'Surprise'* face with *'Neutral', 'Happy', 'Sad'* and *'Fear'* voices.

In Fig.2 (G), the perceptual pattern of *surprised* voice with varied facial expressions is represented by a dominant *'Surprise'* emotion ring. But the combination of *'Surprise'* voice with *'Happy'* face is inclined to be perceived as *'Happy'*.

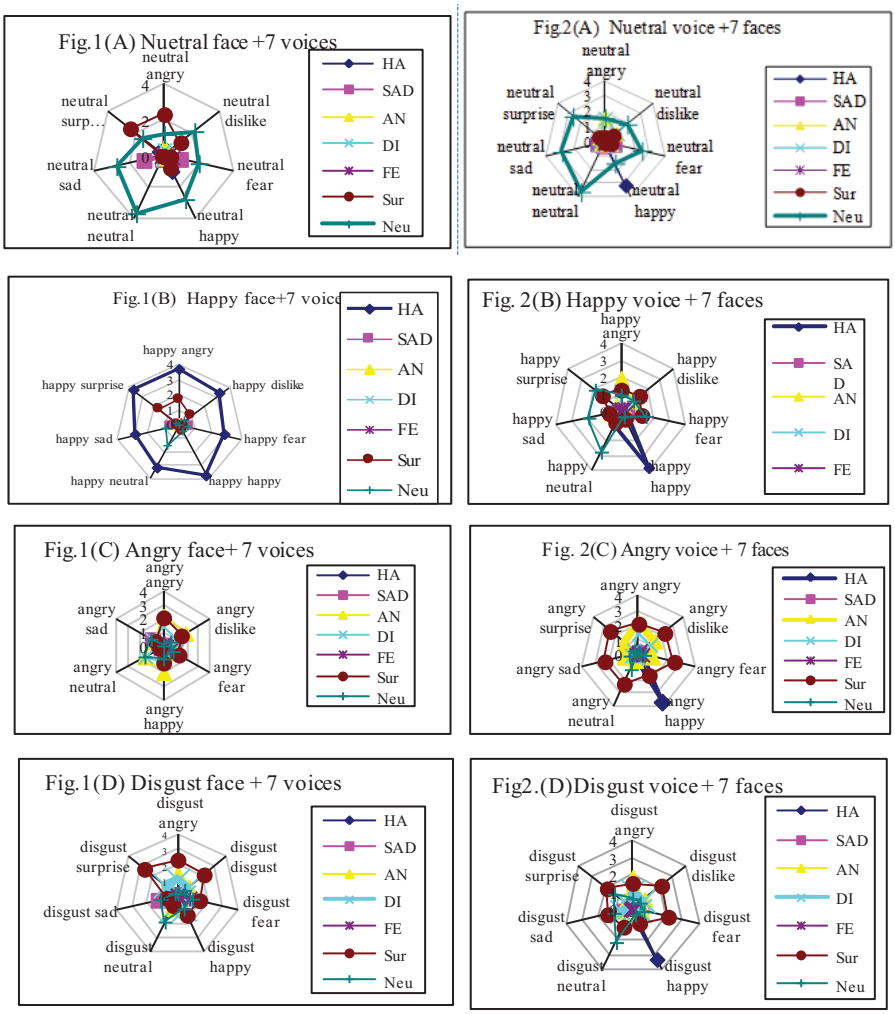
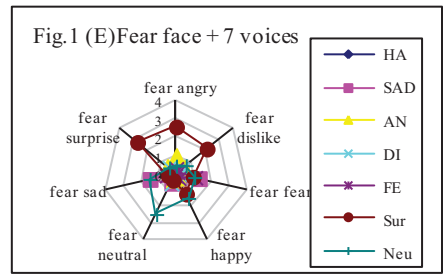
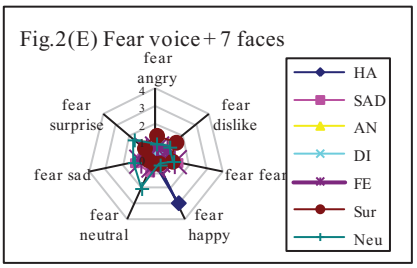
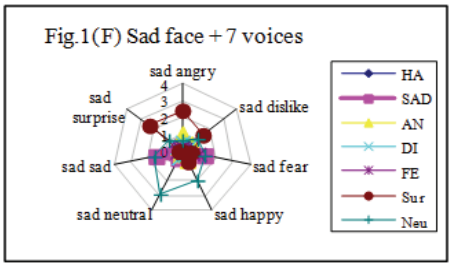


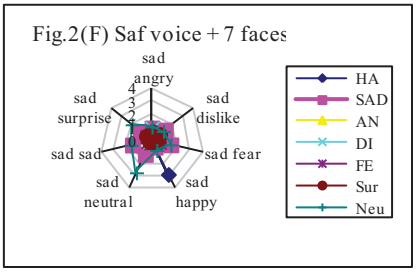
Fig. 1. Graphs from (A) to (G) reveal the perception patterns for the combinations of each facial expression with seven vocal expressions(10 Chinese listeners)

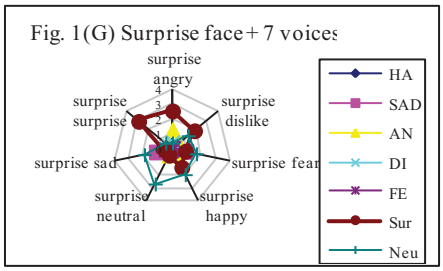
Fig. 2. Graphs from (A) to (G) reveal the perception patterns for the combinations of each emotional voice with seven facial expressions (10 Chinese listeners)











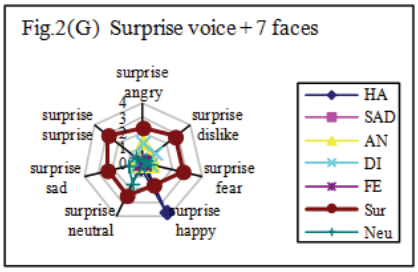


Fig. 1. (*continued*)

Fig. 2. (*continued*)

3.1.2 Perceptual Results for Japanese Listeners

(1) Fig.3 (A) indicates that when *'Neutral'* face is dubbed with non-*'Neutral'* voices, the emotion stimuli are primarily perceived as *'Neutral'*. Fig.4 (A) shows that the combinations of *'Neutral'* voice with varied facial expressions could not lead to the dominance of any emotion ring, except that the combination of *'Neutral'* voice with *'Happy'* face is regarded as *'Happy'*; and neutral voice with *'Neutral'* face is regarded as neutral. The combination of *'Neutral'* voice with an *'Angry'* or *'Disgust'* face is perceived as either *'Disgust'* or *'Anger'*, with almost equal scores less than 2 points.

(2) In Fig.3 (B), the large and symmetrically distributed emotion ring shows that the combinations of *'Happy'* face with varied emotional voices are perceived as *'Happy'*, signifying that the visual modality contributes more than the audio modality in the identification of *'Happy'*. In other words, the perception of *'Happy'* can be independent of the audio modality. Fig.4 (B) reveals that the combinations of *'Happy'*

voice with varied facial expressions (except for 'Happy' face) could not be perceived as *'Happy'*. The combinations of *'Happy'* voice with *'Angry'* or *'Disgust'* faces are most likely to be perceived as *'Angry'* and then as *'Disgust'*. The combination of *'Happy'* voice with *'Neutral'* face is mainly perceived as *'Neutral'*.

(3) Fig.3 (C) presents two dominant rings, with the *'Disgust'* emotion ring embedded in the *'Angry'* emotion ring. This perceptual pattern is triggered by integrating *'Angry'* facial expression with varied emotional voices. Fig.4 (C) shows that the integrations of *'Angry'* voice with varied facial expressions are primarily perceived as *'Angry'*. Only the integration of *'Angry'* voice with *'Happy'* facial expression is perceived as *'Happy'*.

(4) Fig.3 (D) presents two similar evenly distributed emotion rings: *'Disgust'* ring and *'Angry'* ring, which are resulted from the combinations of *'Disgust'* facial expression with varied emotional voices. Fear can be perceived when *'Disgust'* facial expression is dubbed with *'Fear'* or *'Sad'* voice. Fig.4 (D) specifies that the combinations of *'Disgust'* voice with *'Angry'*, *'Happy'* and *'Neutral'* facial expressions are perceived as *'Angry'*, *'Happy'* and *'Neutral'* respectively, while the perceptual scores of other combinations are very low and show no obvious tendencies.

(5) Fig.3 (E) shows no obvious perceptual tendency for the combinations of *'Fear'* face dubbed with varied emotional voices (scores < 2 points). Fig.4 (E) reveals that except that the perception of the combination of *'Fear'* voice with *'Happy'* face is identified as *'Happy'* and the combination of *'Fear'* voice with *'Neutral'* face is identified as *'Neutral'*, all the scores of the combinations of *'Fear'* voice with other facial expressions are lower than 2 points.

(6) Fig.3 (F) shows no obvious perceptual tendency under conditions of *'Sad'* facial expression with varied emotional voices. Fig.4 (F) also reveals that there is no obvious perceptual tendency (scores < 2 points) when *'Fear'* voice is combined with varied facial expressions except for *'Happy'* and *'Neutral'* facial expressions. The perception scores for *'Fear'* and *'Sad'* are equal. The combinations of *'Fear'* voice with *'Happy'* and *'Neutral'* facial expressions tend to be perceived as the emotion implied in the facial expression.

(7) Fig.3 (G) shows no obvious perceptual tendencies under conditions of *'Surprise'* face with varied emotional voices. Each perceptual score is less than 2 points. It can be concluded form Fig.4 (G) that the combinations of surprised voice with varied facial expressions cannot be recognized as *'Surprise'*; the combination of surprised voice with *'Happy'* facial expression is inclined to be perceived as *'Happy'*; the combinations of surprised voice with 'Angry' and 'Disgust' facial expressions tend to be perceived as 'Angry'; and the combination of surprised voice with 'Neutral' facial expression is perceived as 'Neutral'.

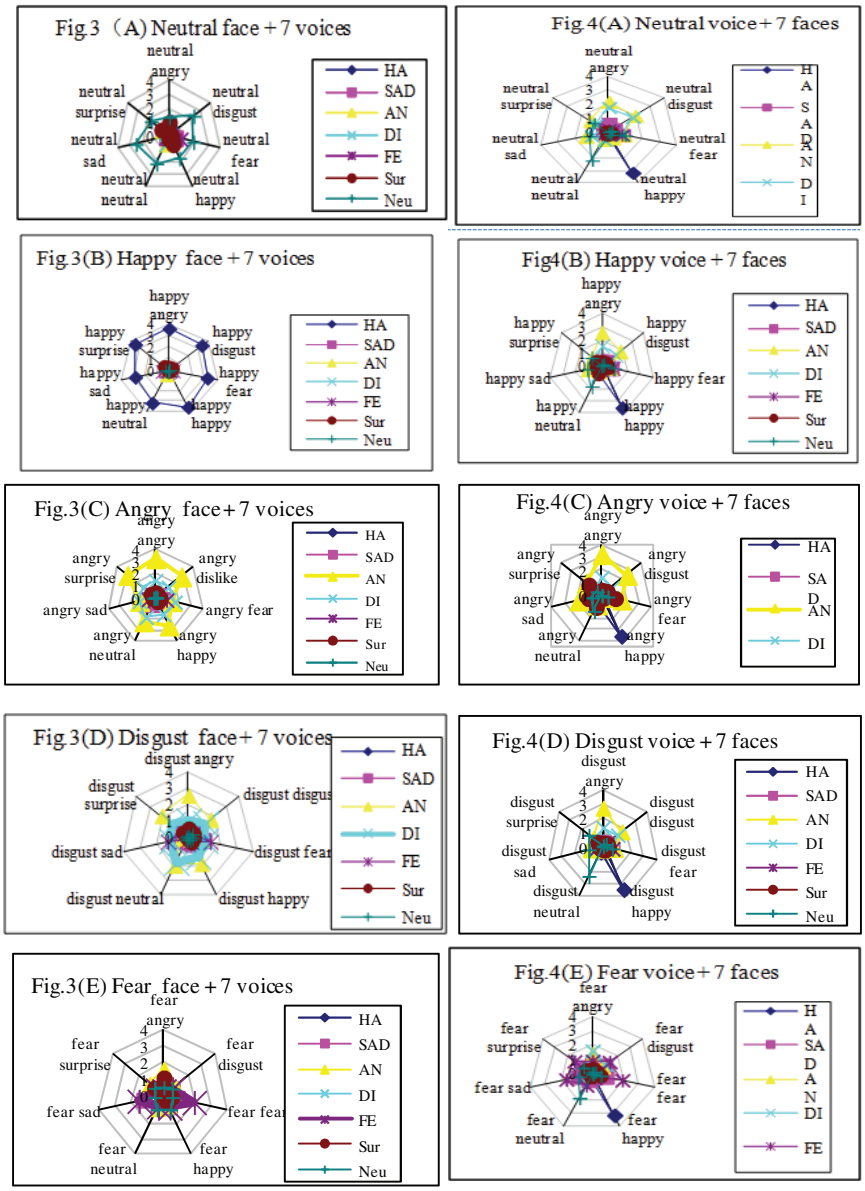
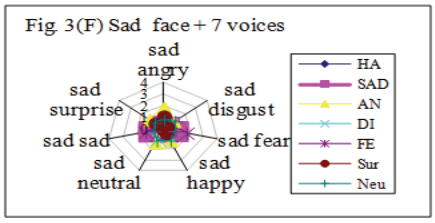
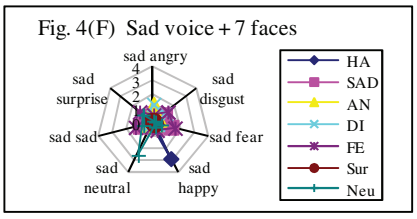
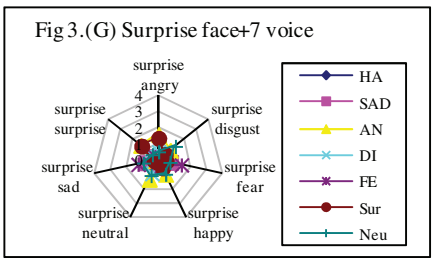


Fig. 3. Graphs from (A) to (G) reveal the perception modes for the combinations of each facial expression with seven voices under the AV-congruent and AV-conflicting condition (10 Japanese listeners)

Fig. 4. Graphs from (A) to (G) reveal the perception modes for the combinations of each emotional voice with seven facial expressions under the AV-congruent and AV-conflicting condition (10 Japanese listeners)







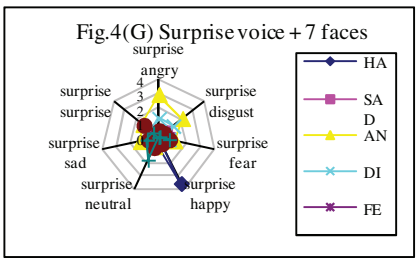


Fig. 3. (*continued*)

Fig. 4. (*continued*)

3.2 Comparison of Perceptual Patterns between Chinese and Japanese for Conflicting Stimuli

Figures 5~8 show the average perceptual score as a function of the intended emotion by vocal and facial expressions for Chinese and Japanese listeners in AV and CAV conditions. From the perspective of psychological dimension of emotion, Chinese and Japanese have similar perceptual patterns: for emotions with high arousal, the visual modality makes a major contribution to emotion decoding; while for emotions with low arousal, the audio modality makes a major contribution. In the AV-congruent setting, the perceptual scores of the Chinese for *'Neutral'*, *'Happy'*, and *'surprise'* are higher than those of the Japanese, signifying higher confidence for the Chinese; while the scores for *'Angry'*, *'Disgust'*, *'Sad'* and *'Fear'* are lower than those of the Japanese, signifying lower confidence for the Chinese. The comparison of Fig. 5 with Fig. 7 indicates that there is a sharper drop in the scores of the Japanese than the Chinese according to vocal emotions. Fig. 6 and Fig. 8 reveal that the degree of falling according to facial emotions between the Japanese and the Chinese is similar except *'Surprise'* and *'Neutral'* emotion. The results may imply that, for Japanese listeners, decoding Chinese emotion counts more on the visual modality than the audio modality, and their decoding for *'Neutral'* and *'Surprise'* facial expression is better than the Chinese. It confirms the results in the previous study that in cross-cultural communication the facial information could help non-native listeners in emotion decoding than only vocal information. By further comparing Fig.5 with Fig.7, Fig. 6 with Fig.8, we find the tendency that Japanese listeners are consistent with Chinese listeners where visual modality exists, while they are discrepant for vocal modality conditions. The result supports the assumption that cross-cultural effect also exists when decoding information transmitted in incongruent channels, and that this effect is greater in the vocal channel than the facial channel.

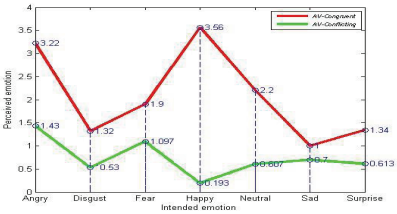


Fig. 5. Average perceptual score as a function of the intended emotion by vocal expressions for Japanese listeners in AV and CAV conditions

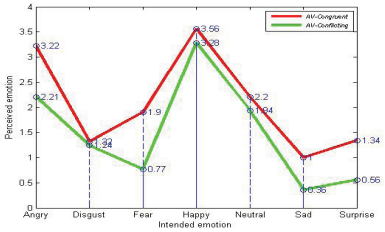


Fig. 6. Average perceptual score as a function of the intended emotion by facial expressions for Japanese listeners in AV-congruent and CAV conditions

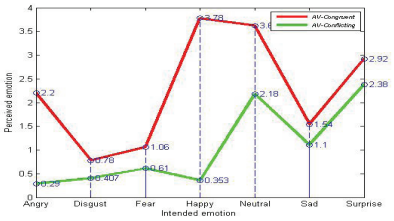


Fig. 7. Average perceptual score as a function of the intended emotion by vocal expressions for Chinese listeners in AV and CAV conditions

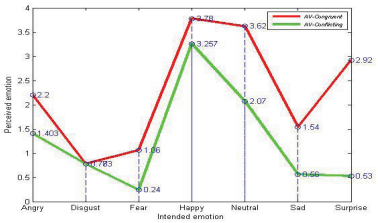


Fig. 8. Average perceptual score as a function of the intended emotion by facial expressions for Chinese listeners in AV and CAV conditions

3.3 Emotional McGurk Effect

Table 2 shows the cases of Emotional McGurk effect obtained for those cases with rating score>2. For an instance, *'Angry face+Disgust voice-> Surprise'* for Chinese listener; *'Happy voice + Angry face-> Disgust'* for Japanese listener.

It was shown that Emotional McGurk effect distributes differently between Chinese and Japanese. There is only one set of combination (surprise voice with various facial expressions) where McGurk effect is not observed. But there are four sets of combinations (neutral, fear, happy facial expression with various voice and angry voice with various facial expressions) where McGurk effect is not observed. The "third emotion" in McGurk effect is most likely to be surprise and more likely to be neutral for Chinese but most likely to be anger and more likely to be disgust for Japanese. These results demonstrate that culture has effect on the decoding of emotion. Taking a further look at the Chinese data, we find that the case where McGurk effect is observed is normally related to either a negative vocal or a negative visual expression (here *'Disgust, Sad and Surprise'* emotions as negative emotions. In fact, *'Surprise'* is an ambiguous emotion state concerning intended negative or positive one, here it is expressed more negative). Visual expression of *'Surprise, Fear and Sadness'* tend to be perceived as *'Neutral'* emotion. For Japanese, the combination where McGurk effect

is observed is also the one where either vocal expression or visual expression of emotion is negative. Of these combinations, as long as visual modality indicates the emotion of anger, the combination will be recognized as *'Disgust'*, most of the other cases will be decoded as *'Angry'*.

Table 2. Emotional McGurk effect containing only the scores marked by 2 or 3 asterisks

Facial	Chinese		Japanese		Vocal	Chinese		Japanese	
	Vocal	Perceived	Vocal	perceived		Facial	perceived	Vocal	perceived
Ne	AN	SU***	----	----	Ne	AN	DI**	AN DI/SA/SU	DI** AN**
AN	Di Ne	SU** DI**	FE/HA/Neu/SA	DI**	AN	F/Neu/S DI HA	SU*** SU*** SU**	----	----
FE	An DI HA/SA	SU*** SU** Neu**	---	---	FE	SU	Neu**	AN	DI**
HA	AN	SU**	-----	-----	HA	FE/SA/SU	Neu**	DI AN	AN** DI**
DI	AN	SU***	HA/Neu SU	AN** AN***	DI	AN /FE/SA	SU**	SU	Ne**
SA	AN DI HA	SU*** SU** Neu**	Ne/SU	AN**	SA	FE/SU	Neu**	AN	DI**
SU	FE/HA/SA	Neu**	DI Ne	Ne** AN**	SU	--	--	DI SA	AN*** AN**

4 Conclusions

The main conclusion is that cultural background poses difference in the perceptual patterns between the Chinese and the Japanese. From the perspective of psychological dimension of emotion, in the AV-conflicting setting, the Chinese and the Japanese have similar patterns: for emotions with high arousal, the visual modality makes a major contribution to emotion decoding; while for emotions with low arousal, the audio modality makes a major contribution. Due to linguistic and cultural difference, the Chinese listeners make more use of the audio modality to decode emotion; for those Japanese who don't know Chinese, the emotion recognition counts more on the visual modality and their decoding of *'Neutral'* and *'Surprise'* facial expression is better than that of the Chinese. Regarding to the rating confidence, the Chinese give higher scores than the Japanese. The findings here are different from those in [12], which assumed that the visual modality mainly transmits valence (positive or negative emotion) and the audio modality mainly transmits arousal (the degree of excitement). One explanation for this discrepancy lies in the difference in the number of emotions: seven in our research and only four in [12].

The emotional McGurk effect is found in the AV–conflicting experiment. Though the occurrence of the McGurk effect relates highly to negative emotions, the perception patterns are different due to the culture effect. Inconsistent with Fagel [56], no

cases of the emotional McGurk effect relating to positive emotions are found. To some extent, the occurrence frequency of the McGurk effect shows that the Chinese listeners have a tendency to jump to a conclusion (*'Surprise'*) while the Japanese favor ambiguity (*'Anger', 'Disgust'* or *'Neutral'*).

The results support the assumptions that (1) When listeners decoding the conflicting AV stimuli, they might rely on some modality more than another across different emotions, as shown in Table 3-7. (2) Although common psychological factor contributes to the emotional communication, the decoding of conflicting AV information will be affected by culture background, and (3) the emotional McGurk effect exists, and it may also be related to cultural norms of the encoder/listener.

Future research will focus on more speakers from various cultures to verify the emotional McGurk effect patterns.

This work was supported by the National Basic Research Program (973Program) of China (No. 2013CB329301), NSFC Project with No. 60975081 and CASS innovation project.

References

1. Darwin, C.: The expression of the emotions in man and animals. John Murray, Oxford University Press, London, New York (1988, Original work published in 1872) (reprinted with introduction, afterword, and commentary by Ekman, P. (ed.))
2. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology 17, 124–129 (1971)
3. Izard, C.E.: Innate and universal facial expressions: Evidence from developmental and cross-cultural research. Psychological Bulletin 115, 288–299 (1994)
4. Scherer, K.R.: A Cross-Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology. In: ICSLP 2000 (2000)
5. Scherer, K.R., Banse, R., Wallbott, H.G.: Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. Journal of Cross-Cultural Psychology 32, 76 (2001)
6. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256 (2003)
7. Abelin, A., Allwood, J.: Cross Linguistic Interpretation of Emotional Prosody. In: Proc. ISCA Workshop on Speech and Emotion, Belfast (2000)
8. Yanushevskaya, I., Chasaide, A.N., Gobl, C.: Cross-Language Study of Vocal Correlates of Affective States. In: Interspeech 2008 (2008)
9. Abelin, Å.: Cross-Cultural Multimodal Interpretation of Emotional Expressions – An Experimental Study of Spanish and Swedish. In: SP 2004 (2004)
10. Huang, C.F., Akagi, M.: A three-layered model for expressive speech perception. Speech Communication 50, 810–828 (2008)
11. Barkhuysen, P., Krahmer, E., Swerts, M.: Incremental perception of acted and real emotional speech. In: ICPHS (2007)
12. Fagel, S.: Emotional McGurk effect. In: Speech Prosody 2006 (2006)
13. Grandjean, D., Scherer, K.R.: Unpacking the Cognitive Architecture of Emotion Processes. Emotion 8(3), 341–351 (2008)
14. Erickson, D.: Expressive speech: Production, perception and application to speech synthesis. Japan Acoust. Sci. & Tech. 26, 4 (2005)

[This paper was published in Springer Cs Proceedings,2013]