Modality Effects on Emotion Perception in English by Chinese L2 English Users: An Eye-tracking Study

Xiazhen Liu[1], Weijing Zhou[2], Francis Nolan[3]
[1,2]Yangzhou University, [3]University of Cambridge
[1]xiazhenl1998@gmail.com, [2]zhouwj@yzu.edu.cn, [3]fjn1@cam.ac.uk

ABSTRACT

This eye-tracing study probes into the extent to which 3 modalities (Audio-only, AO; Visual-only, VO; Audio plus Visual, AV) influence Chinese L2 English users' (N=32) judgement of 6 emotions (happiness, sadness, anger, disgust, fear & neutral) expressed by native English speakers. Two parameters of data, i.e., eye movement (EM) and accuracy rate (AR) were measured to explore the relationship between presentation modality and emotion perception. The results show that for EM, all the emotions generated the same hierarchical rank in terms of Areas of Interest (AOI) under V and AV conditions (eyes > nose> mouth) in spite of salient variations between emotions, and that for AR, the participants in AV condition significantly outperformed themselves in both V and A conditions when judging all the emotions, although there were no significant differences in their distinguishing 3 emotions (anger, sadness and neutral) under 3 modalities, both demonstrating some universal and cultural-specific features.

Keywords: modality effects, emotion perception, Chinese L2 English users, eye-tracking

1. INTRODUCTION

Emotion perception has been a major topic in Natural Language Processing and human-computer interaction. Previous studies of emotion perception revealed that the presentation modality of emotions is a key relevant factor that unduly influences emotion processing [1, 2, 3], i.e., perceivers have different emotion perception performance when the emotions are presented via different modalities (visually, aurally or audio-visually). Furthermore, perceivers could identify the types of emotion in their native language most accurately when emotion is displayed through visual rather than auditory stimuli, i.e., native perceivers generally have a preference or bias to visual modality in emotion recognition [4, 5, 6, 7]. Yet, such a modality preference could be reshaped by perceivers' native linguistic or cultural background [1, 8, 9]. Therefore, to further investigate modality effects, it is necessary to take perceivers' linguistic or cultural background into consideration. In addition, emotion type has been considered as an effective factor in speech emotion perception, supported by the evidence that the identification performance differs between emotions of difference valence [10, 11, 12, 13]. In other words, emotion recognition is not only associated with how emotions are presented, but also correlated to which emotion is presented.

Universally speaking, the essential purpose of human communication is to convey ideas, or express emotions or both on the daily basis. With the increasing globalization and amazing advance of cross-space human communication science and technology, real or virtual, cross-cultural cooperation and collaboration has been a world-wide reality for all the countries. In this very international context, speech emotion in English as a Lingua Franca (ELF) is no doubt of significance both for the research on cross-cultural ELF communication and for the pedagogy of ELF speech learning and teaching in China.

Based on the potential roles of presentation modality and emotion type in processing speech emotions and the well-acknowledged advantages of eye-trackers in detecting human emotion, the present study adopted an eye-tracking diagram to explore the effects of modalities on the perception of speech emotions in English among Chinese L2 English users so as to meet the needs of research and instruction on cross-cultural ELF communication.

2. METHODOLOGY

2.1. Participants

Thirty-nine university students (graduates: 70%, undergraduates: 30%,) were recruited to participate the present study. They spoke Mandarin as their L1 and English as L2, and had normal or corrected-to-normal vision and no hearing impairments. One participant did not complete all the tasks and two other participants did not pass calibration before the experiment. Therefore, thirty-six of them in fact completed the experiment (18 females, 18 males). This sample size ensured proper counterbalancing of the experimental blocks. They used English frequently (all above 10 hours every week, M=24.89, SD=15.05). As their scores of National College English Tests and overall scores of self-reported English proficiency (including listening, speaking, reading, writing and translation) were excellent, they

could be assumed as advanced L2 English users among the population of college students in China.

2.2. Stimuli

As the stimuli used in lots of previous studies were quite often static, e.g., just pictures of speech emotions [14], they were quite less informative or quite inconsistent with real-life human interaction situations, where both facial and vocal organs are simultaneously and dynamically involved in natural and authentic communication. Again, the majority of previous studies of speech perception focused only on two contrastive emotions, e.g., positive vs. negative or happy vs. sad [15], which limited the research scope and depth of speech perception research to a great extent, as emotions conveyed in human speech are quite various in its types to mirror complicated psychological temperatures and convey authentic speech functions. To keep pace with the progress of emotion perception, the present study used dynamic auditory and visual spoken expressions of 6 basic emotions, i.e., happiness, sadness, fear, anger, disgust, and neutral [1] taken from Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [16]. As is shown in Table 1, one Caucasian male and female speaker were selected from each of the three age groups respectively. Their videos included three modalities, audio-only (AO), visual-only (VO), and audio-visual (AV). The experiment thus included a total of 108 different stimuli for each participant (6 speakers×6 emotions×3 modalities). The semantically neutral sentence "*I wonder what this is about*" was used as the target sentence to convey 6 emotions so as to control the potential influence of semantic meanings.

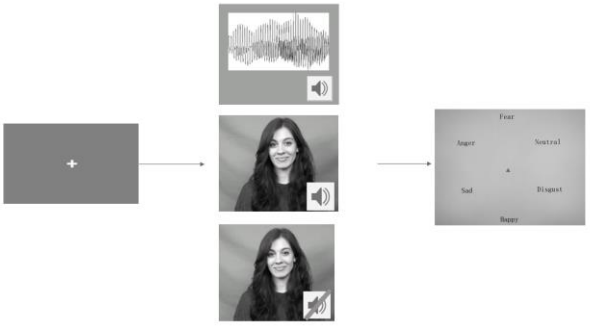
Speakers	Sentence	Modalities	Emotions
Speaker	I wonder what this is about.	AV	happiness, sadness, fear, anger, disgust, neutral
		VO	happiness, sadness, fear, anger, disgust, neutral
		AO	happiness, sadness, fear, anger, disgust, neutral

Table 1: The example of stimuli presented

2.3. Procedure

Experiments and data collection was conducted in conjunction with eye tracking. In each trial, the fixation point was presented first before the stimuli to be presented in either AV, VO, or AO modality. Then the response screen with labels indicating the emotion types appeared and remained on the screen until the participants made a response (See Figure 1). Participants were asked to identify the emotion type of the stimuli and click their mouse on the screen to make the choice. The order of all three blocks (AO, VO, AV) was counterbalanced among participants. Stimulus order was randomized in each block.

Figure 1: Order of experiment events.

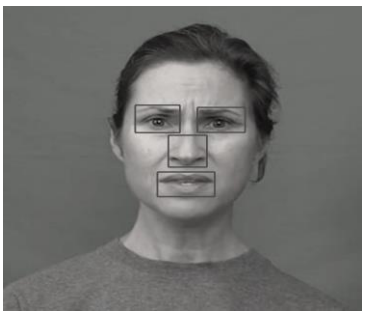


Calibration was performed before the formal experiment started and only the participants who had passed calibration continued to conduct the experiment. They were encouraged to take a self-paced break between each block, and calibration would be again performed after the break. A drift correction was performed before each trail.

2.4. Data collection and analysis

AOI-based fixation proportion (hereafter FP) was analyzed to investigate perceivers' weight of each facial cue in emotion perception [15] so as to test whether the participants had different preferences on certain face regions for different conditions. Since there were no visual stimuli during audio-only block, only data from visual block and audio-visual block were collected. Face regions including eyes, nose, and mouth are dynamic areas of interest (Figure 2). Only fixations located within these areas were recorded. In addition to FP values, accuracy rate (AR) of emotion identification were generated and compared as an indicator of modality preference.

Figure 2: An example for AOI.



Statistical analysis and visualization were conducted in R [17]. Mixed leaner models from lme4 [18] with emotion types, modality being fixed factors which were allowed to interact, and participants and task items being random factors. The final model was:

(1) FP~ Modality*Emotion type + (1+Modality|Participant) + (1+Modality | Item).

The likelihood-ratio test was used to get *p*-values to better illustrate significance. Repeated measures ANOVA from rstatix [19] were used with modality (6 levels) and emotion types (2 levels, i.e., AV vs. VO) being independent variables and AR being the dependent variable.

When there was a significant main effect, Tukey's post hoc tests were conducted to assess the difference between modality pairs. When significant interactions were observed, simple effect analysis was conducted at each level of experiment conditions, and at last, Bonferroni-adjusted pairwise comparisons were performed to further evaluate interactions between modality and emotion types for each modality when necessary.

.

3. RESULTS AND DISCUSSION

3.1. Eye movements

As shown in Table 2 and Figure 3 show, the overall FPs of three facial regions. Generally speaking, the eye preference appeared to be obvious with or without auditory cues, and emotion-specific analysis reported the same preference. There was a hierarchical rank with significant difference in terms of perceivers' gaze allocation to identify emotions, eyes > nose > mouth ($F_{(2, 8205)} = 288.9$, $p<0.0001$).

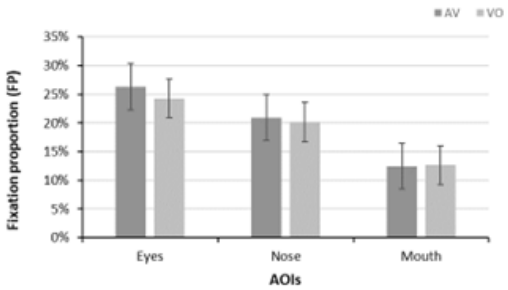


Figure 3: FP of each face region for each modality.

AOI	AV	VO
Eyes	26.31%	24.28%
Nose	20.96%	20.17%
Mouth	12.48%	12.62%

Table 2: The overall FP to three facial regions

As shown in Table 3, the mixed linear model showed a significant main effect of modality on FP ($\chi^2_{(1)} = 4.52$, $p<0.05$), a significant main effect of emotion type as well ($\chi^2_{(5)} = 81.798$, $p<0.001$). No significant interactions or inter-dependence between modality and emotion type was observed ($\chi^2_{(5)} = 1.48$, $p = 0.92$).

Effect	χ^2	df	*p*
Modality	4.523	1	<0.05
Emotion type	81.798	5	<0.001
Modality × Emotion type	1.481	5	0.915

Table 3: The effect of each main and interaction fixed effect.

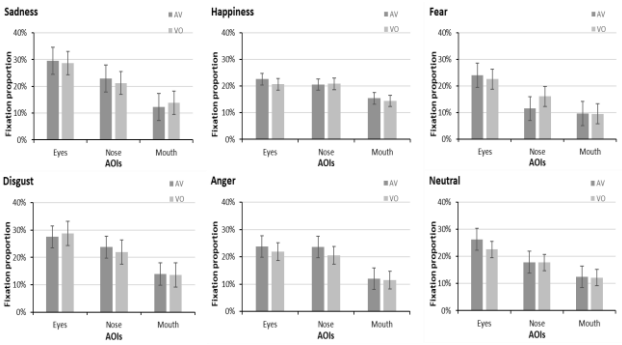


Figure 4: FP under all 12 conditions

As shown in Figure 4, the sub-categorical FPs in 6 emotions displayed the same hierarchical rank with the general pattern above. When emotions were presented in silent environment, in which the degree of processing difficulty increased, FP decreased for eyes ($F_{(1, 2734)} = 2.7$, $p=0.1$) and nose ($F_{(1, 2886)}=3.428$, $p=0.0642$), while increased for mouth region ($F_{(1, 2582)}=0.051$, $p=0.839$). When considering effects of emotion type, when identifying sadness, participants' attention devoted to mouth increased in VO (12.23% →13.85%, $F_{(1, 378)}=1.18$, $p=0.278$); when processing fear, perceivers' fixation on nose increased when the audio was removed (11.42% → 16.04%, $F_{(1, 454)}=1.381$, $p=0.241$); when processing disgust, fixation on eyes increased in silent condition (27.49% → 28.75%, $F_{(1, 454)}=0.3$, $p=0.584$). Although no significance was observed, this implies that perceivers would change their perceptual primacy according to modality condition.

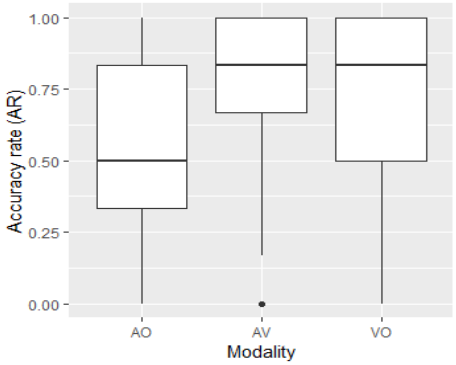
3.2. Accuracy rate

As shown in Figure 5, the overall ARs averaged over all the participants and experiment blocks clarify that AR was the highest for AV, which confirmed the existence of audio-visual integration in non-native context. Also, performance for VO was higher than that for AO; the differences of AR among the three modalities were overall significant, ($F_{(2, 645)}=33.8$, $p<0.0001$). This demonstrates a visual modality preference, which is consistent with the findings of

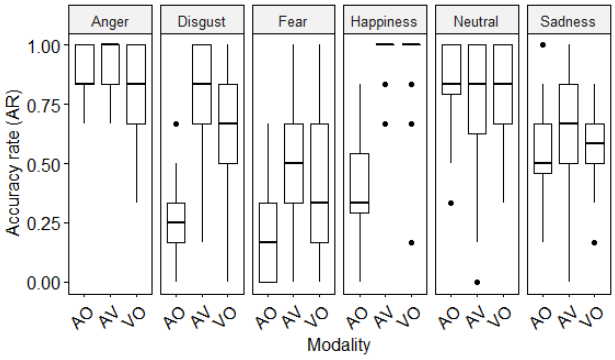
perception research focusing on native perceivers [4, 20, 21], but meanwhile is inconsistent with some studies indicating that people from Eastern cultures would rely more on visual cues compared with auditory ones [8].

Figure 5: AR for each modality.

Results of ANOVA showed a significant main



effect of modality, ($F_{(5, 175)}$ =66.639, p<0.05, ges=0.421), a significant main effect of emotion category, ($F_{(2, 70)}$ =106.643, p<0.05, ges=0.175), and a significant interaction between modality and emotion type, ($F_{(10, 350)}$ =24.263, p<0.05, ges=0.231). To assess the difference between modality pairs, Tukey's HSD test showed significant difference between AO and VO, as well as AV and AO, but no significant difference between AV and VO (AV-AO:



p<0.001; VO-AO: p<0.001; AV-VO: p=0.22), indicating a lower discriminability of AR between AV and VO. It is notable that accuracy results reported significant main effects of modality and emotion type, and a significant interaction between these two factors as well. However, no significant interactions or inter-dependence between modality and emotion type was observed from eye movements. This might imply that combination of task-free and spontaneous data such as eye-tracking is necessary.

Figure 6: AR under all 18 conditions

Emotion	AO-AV	AO-VO	AV-VO
Anger	0.441	0.819	0.069
Disgust	<0.001	<0.001	<0.001
Fear	<0.001	0.008	0.203
Happiness	<0.001	<0.001	1
Neutral	0.17	1	0.414
Sadness	1	1	0.78

Table 4: Results of simple main effect test.

As shown in Figure 6 and Table 4, happiness expressed with muted videos was easiest to be recognized, while fear portrayed through speech alone was the most difficult to be recognized; the effect of modality was significant for disgust, fear and happiness (p<0.05, ges=0.493, ges=0.165 and ges=0.703 respectively), but no significance was observed for anger (p=0.408), neutral (p=0.672) and sadness (p=1). The complete comparison is listed in Table 4 showing Bonferroni-adjusted p value.

4. DISCUSSION AND CONCLUSION

The present study found a visual dominance in L2 users. Research on native speakers suggested that Easterners showed an auditory preference, while Westerners did better in the visual modality [8, 24]. However, against that, the investigation on the modality effect on Mandarin ironic speech showed a better identification rate in the visual-only condition [25]. This inconsistency may imply that the modality effect is related to which emotion is presented and which language group we target.

In terms of the effect of emotion types, research on native speakers suggested that Easterners appear to be less accurate than Westerners in recognizing negative emotions, which could be explained by the cultural differences [26-27]. In this sense, participants in this study showed a combination of both Eastern and Western pattern, as they were less accurate in recognizing disgust and fear, which was a specific Eastern pattern, while their performance in response to anger was best. This could be explained by their long duration of exposure to the L2 culture, which leads to their similar processing behaviours to that of native speakers [9].

In conclusion, this eye-tracing study probes into the extent to which modalities influence Chinese L2 English users' judgement of emotions expressed by native English speakers. Results showed a consistent hierarchical rank of fixation on facial regions in both AV and VO conditions (eyes > nose> mouth), and that the visual cues play a dominant role for L2 learners' emotion perception, although there were no significant differences in their distinguishing anger, sadness and neutral. This study demonstrated some universal and cultural-specific features.

5. REFERENCES

[1] Ekman, P. 1989. The argument and evidence about universals in facial expressions of emotion. In: Wiley, J. (eds). *Handbook of social psychophysiology*. Oxford, England, 143-164.

[2] Lorette, P. & Dewaele, J. M. 2019. The relationship between bi/multilingualism, nativeness, proficiency and multimodal emotion recognition ability. International Journal of Bilingualism, 23 (6), 1502-1516.

[3] Bhatara, A., Laukka, P., Boll-Avetisyan, N, Granjon, N., Elfenbein, H. A. & T. Banziger. 2016. Second Language Ability and Emotional Prosody Perception. *Plos One*. 11.

[4] de Boer, M. J., Baskent, D., & Cornelissen, F. W. 2020. on Emotion: Dynamic Gaze Allocation During Emotion Perception From Speech-Like Stimuli. *Multisensory Research*. 34, 17-47.

[5] Kim, J., & Davis, C. 2012. Perceiving emotion from a talker: How face and voice work together. *Visual Cognition*. 20, 902-921.

[6] Paulmann, S. & Pell, M. D. 2011. Is there an advantage for recognizing multi-modal emotional stimuli?. *Motivation and Emotion*. 35, 192-201.

[7] Vroomen, J. & de Gelder, B. 2000. Sound enhances visual perception: cross-modal effects of auditory organization on vision. *J Exp Psychol Hum Percept Perform*. 26, 83-90.

[8] Liu, P., & Rigoulot, S., & Pell, M. D. 2015. Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*. 67, 1-13.

[9] Chen, P., Chung-Fat-Yim, A. & Marian, V. 2022. Cultural Experience Influences Multisensory Emotion Perception in Bilinguals. *Languages*. 7.

[10] Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L. & Neel, R. 2011. The face in the crowd effect unconfounded: happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *J Exp Psychol Gen*, 140, 637-59.

[11] Becker, M. W. 2012. Negative emotional photographs are identified more slowly than positive photographs. *Attention, Perception, & Psychophysics*. 74,n1241-1251.

[12] Pinkham, A. E., Griffin, M., Baron, R., Sasson, N. J., & R. C. Gur. 2010. "The face in the crowd effect: anger superiority when using real faces and multiple identities. *Emotion*. 10, 141-6.

[13] Pitica, I., Susa, G., Benga, O, & Miclea, M. 2012. Visual search for real emotional faces: the advantage of anger. *Procedia - Social and Behavioral Sciences*. 33, 632-636.

[14] Henderson, J. M., Williams, C. C., & Falk, R. J. 2005. Eye movements are functional during face learning. *Memory & Cognition*. 33, 98-106.

[15] Schurgin, M. W., Nelson, J., Iida, S., Ohira, H., Franconeri, S. L. & Franconeri, S. L. 2014. Eye movements during emotion recognition in faces. *Journal of Vision*. 14.

[16] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A. & Verma, R. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. 2014. *Ieee Transactions on Affective Computing*. 5, 377-390.

[17] R Core Team, 2021. R: A Language and Environment for Statistical Computing (Version 4.0.5). Vienna, Austria. Retried from. https://www.R-project.org.

[18] Bates D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*. 67, 1-48.

[19] Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. https://CRAN.R-project.org/package=rstatix.

[20] Regel, S., Gunter, T. C., & Friederici, A. D. Isn't It Ironic? An Electrophysiological Exploration of Figurative Language Processing. 2011. *Journal of Cognitive Neuroscience*. 23, 277-293.

[21] Spotorno, N., Cheylus, A., Van Der Henst, J.-B., & Noveck, I. A. 2013. What's behind a P600? Integration operations during irony processing. *PloS one*. 8, e66839.

[22] Banse, R. & Scherer, K. R. 1996.Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*. 70, 614-636.

[23] Rottman, J. 2014. Evolution, Development, and the Emergence of Disgust. *Evolutionary Psychology*. 12.

[24] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17: 124-129.

[25] Li, S., Chen, A., Chen, Y., & Tang, P. (2022). The role of auditory and visual cues in the interpretation of Mandarin ironic speech. *Journal of Pragmatics*, 201: 3-14.

[26] Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: holistic versus analytic cognition. *Psychological Review*, 108: 291-310.

[27] Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability Data and Cross-National Differences. *Journal of Nonverbal Behavior*, 21: 3-21.