INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Effects of communication channels and actor's gender on emotion identification by native Mandarin speakers

Yi Lin, Hongwei Ding[*]

Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China

carol.y.lin@sjtu.edu.cn, hwding@sjtu.edu.cn

Abstract

Communication channels and actor's gender have been increasingly reported to influence emotion perception, but past literature exploring these two factors has largely been disassociated. The present study examined how emotions expressed by actors of the two genders are perceived in three different sensory channels (i.e. face, prosody, and semantics). Eighty-eight native Mandarin participants (43 females and 45 males) were asked to identify the emotion displayed visually through face, or auditorily through prosody or semantics in a fixed-choice format, in which accuracy and reaction time were recorded. Results revealed that visual facial expressions were more accurately and rapidly identified, particularly when posed by female actors. Additionally, emotion perception in the auditory modality was modulated by actor's gender to a greater extent: emotional prosody yielded more accurate and faster responses when expressed by female than male actors, while emotional semantics produced better performances when presented by males. To sum up, paralinguistic (i.e., visual and prosodic) dominance effects are more evident in emotions expressed by female than male actors.

Index Terms: communication channels, actor's gender, unisensory emotion perception, face, prosody, semantics

1. Introduction

The ability to identify the emotional states of others, and appropriately express one's own emotions forms a crucial socio-cognitive basis for everyday social interactions [1]. To successfully interpret the communicative intentions of others, interlocutors need to integrate emotional information signaled in different channels. In addition to linguistic content, emotions are also conveyed through various types of paralinguistic cues, such as facial expressions, body movements, gestures and tone of voice.

One focal point of research on emotion processing is whether some forms of emotional signals might hold certain communicative advantages over others. According to some studies [2, 3], generally, people rely on facial expressions compared to prosody or semantic content when recognizing multimodal emotional information, revealing a visual dominance effect. Emotional prosody has also been found to override semantics, which again supports the salient role of non-verbal cues in multisensory emotion perception [4, 5]. However, since information is simultaneously conveyed through different domains in multisensory integration, it remains unclear whether these non-verbal signals are intrinsically more unequivocal and perceptually salient, or they might be facilitated by the availability of other emotional cues in multichannel emotion processing [6]. Thus, it is necessary to disentangle the interplay among multisensory cues by examining how emotions are perceived in a single channel. Indeed, from time to time one may encounter real-life situations when they need to decipher emotional information through single channels (e.g. talking on the phone, listening to a news broadcasts, or looking at one's face when he/she is not speaking) [7]. According to some studies [8-10], one can also decode affective states highly above chance using just one channel, suggesting that using unisensory emotion recognition tasks for certain emotions may be sufficient to evaluate one's emotion perception skills.

Apart from modality dominance effects, gender has also been repeatedly recognized as a core factor that might affect emotion processing [11, 12]. An extensive body of research has provided evidence for the effect of decoder's gender, demonstrating female advantages in perceiving emotional information [9, 13-15]. By comparison, there has been a relatively smaller amount of research examining the effect of encoder's gender on emotion perception, and these existing studies have often produced inconsistent results: while some researchers did not observe a significant effect of actor's gender [7, 16-18], others proposed that emotions are more accurately decoded when presented by females in agreement with the general belief that women are more emotionally expressive [8, 9, 14, 15, 19-21]. There was also literature reporting better recognition of words with positive or negative semantics uttered by males [8]. Moreover, a vast majority of past research has selected accuracy as the only behavioral index in measurements. However, one possibility is that emotion recognition performance is sometimes at or nearly ceiling, thus underemphasizing significant gender differences that may have had a larger magnitude [22]. In this case, it has been recommended that more sensitive measurement methods, such as reaction time, be integrated with accuracy when interpreting the effect of gender in emotion recognition [21].

Despite considerable attention given to the effects of communication channels and actor's gender, the two domains of research have largely been disassociated. In other words, it remains to be clarified how emotions expressed by actors of the two genders are processed in different communicative channels. The current study attempts to explore how unisensory emotion perception is shaped by the interplay between communication channels and actors' gender. We tested how accuracy and reaction time were affected by these two factors in unisensory (i.e. face, prosody, and semantics) emotion identification tasks by 88 native Mandarin participants. In accordance with visual dominance effects, we expected to find face as the most salient

[*] Corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1498

channel in emotion perception (hypothesis 1). It was also predicted that actor's gender would exert greater influences on auditory emotion perception, with female presentation of emotion better recognized in the prosodic channel and male expressions more easily identified in the semantic one (hypothesis 2). We hope to deepen the understanding of how theoretical accounts of modality dominance effects and gender differences reconcile in shaping emotion perception.

2. Methods

2.1. Participants

The present study was conducted with ethical approval from the Institutional Review Boards (IRB) in accordance with the Declaration of Helsinki at Shanghai Jiao Tong University (SJTU). We recruited 88 participants (43 females and 45 males, mean age ± SD: 23.74 ± 2.69) through an online campus advertisement. All participants were undergraduate or graduate students at SJTU, and spoke Mandarin Chinese as their native language. All had normal or corrected-to-normal vision and normal hearing as verified by standard audiological screening [23]. None of them reported a history of speech, language, or hearing impairment or any cognitive difficulty. They completed written informed consent at the study onset, and were financially compensated for their time and involvement.

2.2. Stimuli

The stimuli used in the study consisted of emotions conveyed in three different communication channels, namely facial expressions, prosody (i.e. tone of voice) and words with semantic content. Each stimulus expressed one of the three basic emotions (i.e. happiness, sadness, and anger) [24] or neutrality, in which "happiness" and "sadness" were the target emotions in the current study, and "anger" and "neutrality" were included as emotional and non-emotional distractors respectively to mitigate the effect of chance level and confuse participants' judgments.

We selected the facial stimuli posed by eight actors (four females and four males) from Chinese Affective Picture System [25], a well-established database with a standardized set of black-and-white photographs of Chinese actors displaying facial expressions of emotions. The prosodic and semantic stimuli contained vocalizations portrayed by four amateur actors (two females and two males) in a quiet laboratory setting, and digitized at a sampling rate of 44.100 kHz with a 16-bit resolution. For the prosodic stimuli, the speakers enunciated semantically neutral disyllabic concrete words (e.g. 报纸 (newspaper in Chinese), 电话 (telephone in Chinese)) in happy, sad, angry and neutral tone of voice. For the semantic stimuli, the speakers uttered the disyllabic words indicating happy (e.g. 愉快 (joyful in Chinese)), sad (e.g. 悲痛 (sorrowful in Chinese)), angry (e.g. 狂怒 (furious in Chinese)) emotions and neutrality (e.g. 平凡 (ordinary in Chinese)) in a neutral prosody. The duration measures of the target auditory stimuli used for emotional prosody and semantics identification are shown in Tables 1 and 2 respectively.

Sixty-four stimuli were included in the facial, prosodic, and semantic channels respectively. The number of stimuli in each of the three channels was balanced between four emotional or non-emotional categories (16 stimuli for each category), and between actors and actresses (8 females and 8 males for each category in each channel). An earlier norming study was conducted to perceptually validate the experimental stimuli by 23 Chinese university students (11 females and 12 males, mean age ± SD: 23.70 ± 2.59) who did not participate in the current experiment. All included emotional stimuli received over 90% identification accuracy for emotional categories, and an average rating of >3 for emotional intensity on a 7-point Likert scale (0 = not intense, 6 = very intense). Table 3 presents the identification accuracy of emotional category and rating of emotional intensity for the target stimuli adopted in the experiment.

Table 1: *Duration (milliseconds) of the auditory stimuli used for emotional prosody identification*

Emotion / Actors	Happy		Sad		Mean / SD	
	Mean	SD	Mean	SD		
Females	1066	94	1913	165	1489	444
Males	1149	132	1770	211	1459	357
Mean / SD	1108	122	1841	203	1474	403

Table 2: *Duration (milliseconds) of the auditory stimuli used for emotional semantics identification*

Emotion / Actors	Happy		Sad		Mean / SD	
	Mean	SD	Mean	SD		
Females	1007	78	1001	88	1004	83
Males	1076	95	1091	107	1084	101
Mean / SD	1041	93	1046	108	1044	101

Table 3: *Identification accuracy of emotional category and rating of emotional intensity for the target stimuli adopted in the experiment*

Stimulus type	Emotion category	Accuracy		Intensity	
		Mean	SD	Mean	SD
Face	Happy	93.29%	2.03%	4.50	.60
	Sad	93.06%	1.96%	4.52	.62
Prosody	Happy	92.51%	3.49%	4.23	.38
	Sad	93.72%	3.89%	4.40	.24
Semantics	Happy	96.74%	3.61%	4.59	.70
	Sad	95.92%	3.91%	4.42	.47

2.3. Procedure

We carried out the study in a sound booth using E-Prime for stimulus presentation. The visual stimuli (i.e. emotional faces) were displayed in the center of an LCD screen over a constant white background, and the auditory ones (i.e. emotional prosody and words with semantic content) were presented binaurally over Sennheiser HD280 PRO headphones at 70 dB SPL. There were altogether 192 trials of three separate blocks in the experiment, with each block containing 64 trials of emotional stimuli presented in one of the facial, prosodic or semantic channels. Each trial started with a fixation cross in the center of the screen for 1100 milliseconds (ms). Then emotional stimuli were presented, at the onset of which participants were asked to identify which emotion was being portrayed by the actor or actress by pressing one of the 4 emotion-coded keys on a keyboard ("v" for happy, "b" for sad, "n" for angry, and "m" for neutral) as quickly as possible. There was no limit to the response time, and after the response was made, a blank screen was presented for 1000 ms before the next trial began. The

presentation order of the blocks and trials in each block was randomized across participants. Participants entered the experiment after familiarizing themselves with the experimental procedure by completing eight practice trials in each block with 100% accuracy. They were allowed to have a short rest after running every 32 trials.

2.4. Statistical analyses

A series of linear mixed-effects models in R (version 3.6.1) with the lme4 package were applied for data analysis [26]. We excluded the distractor trials so that only trials of happy and sad emotions were included in the analyses. Accuracy data were transformed into rationalized arcsine unit (RAU) to undermine the ceiling effects [27], and reaction time data were log-transformed to mitigate the influence of positive skewness [28]. The transformed accuracy and reaction time data were entered as dependent variables in the mixed-effects models. The fixed categorial factors consisted of actors' (speakers') gender (i.e. female and male), and communication channel (i.e. facial, prosodic and semantic), in which stimuli produced by females and the facial channel were set as the baseline level respectively. When the pairwise comparison between the prosodic and semantic channels was conducted, prosody was used as the baseline level. Random factors for intercepts included observer (listener) participants and test items. Tukey's post hoc tests in the lsmeans package [29] were employed for pairwise comparison in case of a significant main effect or interaction. The full models with intercepts, coefficients, and error terms for accuracy and reaction time analyses are represented as follows:

$$Accuracy\ (RAU)_{ij} = \beta_0 + (\beta_1 \times actor's\ gender) + (\beta_2 \times communication\ channel) + (\beta_3 \times communication\ channe \times actor's\ gender) + b_{0i} + b_{1j} + \varepsilon_{ij} \quad (1)$$

$$log(Reaction\ time)_{ij} = \beta_0 + (\beta_1 \times actor's\ gender) + (\beta_2 \times communication\ channel) + (\beta_3 \times communication\ channe * actor's\ gender) + b_{0i} + b_{1j} + \varepsilon_{ij} \quad (2)$$

In these models, β_0 represented the intercept, which was the predicted outcome when all other predictors were equal to 0. β_1, β_2, and β_3 represented the coefficients for actor's gender, communication channel and their interactions respectively. These coefficients reflected how much the outcome variable changed relative to a unit of change in the corresponding predictors. The random intercepts were represented as b_{0i} and b_{1j}, where i varied according to observer (listener) participants and j varied according to test items. An error term (ε) was also included in both models to account for the distance between the predicted value and the actual data point (i.e. residual).

3. Results

Mean accuracy and reaction time for emotional stimuli produced by males and females in three different communication channels are illustrated in Figures 1 and 2.

Overall, participants achieved considerably high mean accuracy (Mean ± SD = 93.90% ± 10.68%). Mixed-effects models of accuracy data in RAU suggested no main effect of actors' (speakers') gender ($\chi2$ (1) = 1.27, $p > .05$), but a main effect of communication channel ($\chi2$ (2) = 19.89, $p < .001$), and a significant interaction between gender and channel ($\chi2$ (2) = 72.87, $p < .001$). To parse out the interaction effect, we compared how emotional information conveyed by females and males was processed in three different communication channels. Emotional signals presented by females elicited significantly higher identification accuracy in the facial ($\hat{\beta}_3 = .07$, $SE = .01$, $t = 7.00$, $p < .001$) and prosodic ($\hat{\beta}_3 = .05$, $SE = .01$, $t = 5.12$, $p < .001$) channels than the semantic one. No significant difference was found between the facial and prosodic channels for emotions expressed by the actresses ($\hat{\beta}_3 = .02$, $SE = .01$, $t = 1.89$, $p > .05$). However, male actors did not produce a consistent pattern of identification accuracy with their female counterparts: higher accuracy ($\hat{\beta}_3 = -.07$, $SE = .01$, $t = -6.53$, $p < .001$) was obtained in the semantic and facial ($\hat{\beta}_3 = .05$, $SE = .01$, $t = 4.38$, $p < .001$) channel than the prosodic one, and similarly, no significant difference was found between the two channels (i.e., emotional faces and semantics) that manifested more perceptual advantages ($\hat{\beta}_3 = -.02$, $SE = .01$, $t = -2.16$, $p > .05$).

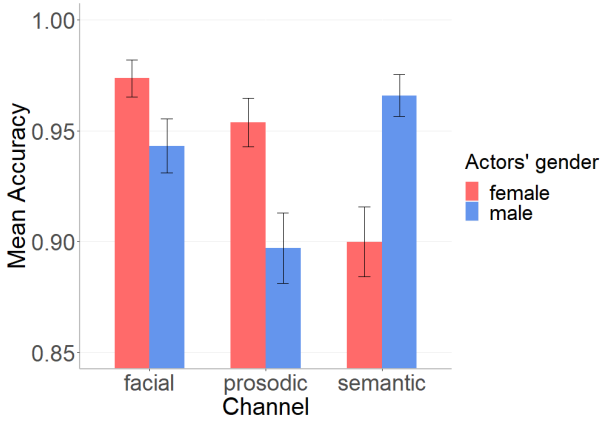

Figure 1: *Mean identification accuracy for stimuli produced by males and females in three different communication channels*

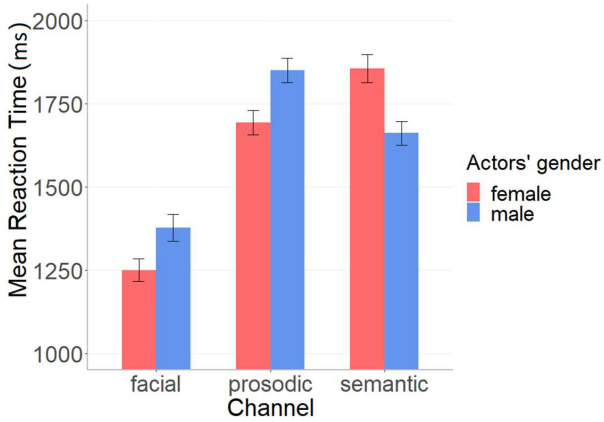

Figure 2: *Mean reaction time for stimuli produced by males and females in three different communication channels*

In regard to response time data, we excluded responses over two standard deviations from the mean (3.92%) and incorrect responses (3.62%) [28, 30]. Linear mixed-effects

analyses on the logarithm of reaction time revealed no main effect of actors' (speakers') gender ($\chi 2$ (1) = .17, $p > .05$), but a main effect of communication channel ($\chi 2$ (2) = 1441.40, $p < .001$), and a significant interaction between gender and channel ($\chi 2$ (2) = 126.77, $p < .001$). We parsed out the interaction effect by comparing how emotional information conveyed by females and males was processed in three different communication channels. Emotional cues presented by females yielded faster responses in the facial channel than the prosodic ($\hat{\beta}_3$ = -.33, *SE* = .02, t = -19.12, $p < .001$) and semantic ($\hat{\beta}_3$ = -.55, *SE* = .02, t = -31.37, $p < .001$) ones. Emotional prosody was also identified faster than emotional semantics expressed by the actresses ($\hat{\beta}_3$ = -.22, *SE* = .02, t = -11.53, $p < .001$). As for the signals displayed by males, faster responses were found in the facial channel than the prosodic ($\hat{\beta}_3$ = -.34, *SE* = .02, t = -21.73, $p < .001$) and semantic ($\hat{\beta}_3$ = -.28, *SE* = .02, t = -15.62, $p < .001$) ones. Emotional semantics was also identified faster than emotional prosody conveyed by the actors ($\hat{\beta}_3$ = .06, *SE* = .02, t = 3.40, p = .002).

4. Discussion

The current research investigated the effects of communication channels and actor's gender on emotion identification performances by native speakers of Mandarin Chinese. The integration of accuracy and reaction time data has demonstrated the dynamic interactions between the two factors. Specifically, visual facial expressions, especially those posed by female actors, gain more perceptual advantages than prosody and semantics in emotion identification. The perception of auditory emotional signals appears to depend more on the gender of the actors: emotional prosody dominates over semantics when produced by females, whereas male vocalizations produce a reverse processing pattern. Together, the salient role of paralinguistic (i.e. facial and prosodic) cues is more pronounced for emotional signals presented by female than male actors.

In line with our first hypothesis, the present study has provided evidence that participants tend to be visually dominant irrespective of the gender of the display, though female facial expressions exhibit greater perceptual advantages in both accuracy and reaction time performances. This indicates that such a dominance effect tends to take precedence over the role of actor's gender in the visual channel. Our confirmation of the communicative advantage of visual facial expressions in unisensory emotion identification provides a coherent picture with previous investigations on multisensory integration, in which face serves as a more dominant channel when simultaneously presented with either one or both of the prosodic and semantic channels [3, 5, 31, 32]. Thus, one can speculate that our sensitivity to visual facial cues in multimodal emotion integration might be underpinned by low-level visual dominance in unisensory perception [33].

Compared with visual emotion identification, actor's gender exerts greater influences on auditory emotion perception. As predicted in our second hypothesis, emotions portrayed by female actors are more perceptually salient in the prosodic channel, whereas those expressed by their male counterparts are more easily identified in the semantic channel. Our findings align with previous studies advocating better detectability of non-verbal affect bursts and speech-embedded vocal prosody uttered by females [8, 9, 14, 15, 19, 20]. We have also replicated the study demonstrating better recognition of words spoken by males [8], and extended male advantages in expressing semantic content from emotion-laden words (i.e. words with emotional connotations) to emotion-labelled words (i.e. words denoting emotions). Interestingly, neuroimaging research has also implicated that semantic processing in men is not as susceptible to influences from emotional prosody as is semantic processing in women in emotional speech perception [34]. Emotional semantics has been reported to elicit bilateral activation of inferior frontal gyrus in women but only left hemisphere activation in men. Though it is far from conclusive concerning why the two genders differ in how well their expressions of emotion can be processed, and to what extent studies on emotional expressions and perception in this field might be associated, these gender-related differential predispositions might emerge through a combination of various attributes, such as innate biological differences, social norms, and situational constraints (consult Chaplin [35] and Fischer and LaFrance [36] for reviews).

There are several limitations that might hinder the interpretation of findings in the current study. First, there was a small number of actors (especially for the auditory stimuli) involved in the experiment. Findings might also be limited since the sample of observer/listener participants were about the same age and all attended the same university, and we restricted our purview on their performances only from the perspective of actor's gender. Future researchers are highly recommended to examine how the gender of actors/speakers and observers/listeners interacts in shaping emotion perception with a larger size and wider range (e.g. the elderly people, second/foreign language learners, clinical populations) of sample. Another topic worthy of investigation in future behavioral and neurological studies, as mentioned above, is the interrelationship between emotion expression and perception in the generation of gender differences. It is also worthwhile to examine to what extent differences in encoder's and decoder's gender can be generalized to other categories of emotion (e.g. anger, disgust, fear).

5. Conclusion

The present study examined how unisensory emotion identification performances by native Mandarin participants are influenced by communication channels and actor's gender. Results show that visual facial expressions are more perceptually salient than cues in the auditory channels (i.e., emotional prosody and semantics), particularly when presented by females. Auditory emotion perception has a greater propensity to be modulated by actor's gender: female vocalizations are better recognized in the prosodic channel, whereas male vocalizations produce better performances in the semantic channel. To conclude, female expressions of emotions tend to elicit a larger paralinguistic dominance effect in unisensory emotion perception. These findings may have significant implications on how theoretical perspectives concerning modality dominance effects and gender differences converge, and lay a foundation for future academic work to elucidate the behavioral and neural underpinnings of gender differences by associating emotion perception with expression.

6. Acknowledgements

This research was funded by the major Program of National Social Science Foundation of China (No. 18ZDA293).

7. References

[1] A. H. Fischer and A. S. R. Manstead, "Social functions of emotion.," in *Handbook of Emotions (3rd ed)*, M. Lewis, J. Haviland-Jones, and L. F. Barrett, Eds. New York: Guilford Press, 2008.

[2] O. Collignon *et al.*, "Audio-visual integration of emotion expression," *Brain Research,* vol. 1242, pp. 126-35, Nov 25 2008.

[3] P. M. Beall and A. M. Herbert, "The face wins: Stronger automatic processing of affect in facial expressions than words in a modified Stroop task," *Cognition and Emotion,* vol. 22, no. 8, pp. 1613-1642, 2008.

[4] P. Filippi *et al.*, "More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing," *Cognition and Emotion,* vol. 31, no. 5, pp. 879-891, Aug 2017.

[5] Y. Lin, H. Ding, and Y. Zhang, "Prosody dominates over semantics in emotion word processing: Evidence from cross-channel and cross-modal Stroop effects," *Journal of Speech, Language, and Hearing Research,* vol. 63, no. 3, pp. 896-912, 2020.

[6] W. R. Barnhart, S. Rivera, and C. W. Robinson, "Different patterns of modality dominance across development," *Acta Psychologica,* vol. 182, pp. 154-165, Jan 2018.

[7] S. T. Hawk, G. A. van Kleef, A. H. Fischer, and J. van der Schalk, "'Worth a thousand words': Absolute and relative decoding of nonlinguistic affect vocalizations," *Emotion,* vol. 9, no. 3, pp. 293-305, 2009.

[8] A. Lausen and A. Schacht, "Gender differences in the recognition of vocal emotions," (in English), *Frontiers in Psychology,* Article vol. 9, p. 22, Jun 2018, Art. no. 882.

[9] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior Research Methods,* vol. 40, no. 2, pp. 531-539, 2008.

[10] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion,* vol. 1, no. 4, pp. 381-412, 2001.

[11] S. Imaizumi, M. Honma, O. Y., M. Maruishi, and H. Muranaka, "Gender differences in emotional prosody processing - An fMRI study," *Psychologia,* vol. 47, no. 2, pp. 113-124, 2004.

[12] A. H. Fischer, M. E. Kret, and J. Broekens, "Gender differences in emotion perception and self-reported emotional intelligence: A test of the emotion sensitivity hypothesis," *PLoS One,* vol. 13, no. 1, p. e0190712, 2018.

[13] J. A. Hall, "Gender effects in decoding nonverbal cues," *Psychological Bulletin,* vol. 85, no. 4, pp. 845-857, 1978.

[14] M. Vasconcelos, M. Dias, A. P. Soares, and A. P. Pinheiro, "What is the melody of that voice? Probing unbiased recognition accuracy with the Montreal Affective Voices," *Journal of Nonverbal Behavior,* vol. 41, no. 3, pp. 239-267, 2017.

[15] O. Collignon, S. Girard, F. Gosselin, D. Saint-Amour, F. Lepore, and M. Lassonde, "Women process multisensory emotion expressions more efficiently than men," *Neuropsychologia,* vol. 48, no. 1, pp. 220-5, Jan 2010.

[16] L. Lambrecht, B. Kreifelts, and D. Wildgruber, "Gender differences in emotion recognition: Impact of sensory modality and emotional category," *Cognition & Emotion,* vol. 28, no. 3, pp. 452-69, Apr 2014.

[17] J. Thayer and B. H. Johnsen, "Sex differences in judgement of facial affect: A multivariate analysis of recognition errors," *Scandinavian Journal of Psychology,* vol. 41, no. 3, pp. 243-246, '2000/09/01 2000.

[18] M. T. Riviello and A. Esposito, *On the perception of dynamic emotional expressions: A cross-cultural comparison*. Springer, Dordrecht, 2016.

[19] M. Koeda, P. Belin, T. Hama, T. Masuda, M. Matsuura, and Y. Okubo, "Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners," *Frontiers in Psychology,* 10.3389/fpsyg.2013.00105 vol. 4, p. 105, 2013.

[20] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology,* vol. 32, no. 1, pp. 76-92, 2001.

[21] A. E. Thompson and D. Voyer, "Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis," *Cognition & Emotion,* vol. 28, no. 7, pp. 1164-95, 2014.

[22] E. Hampson, S. M. van Anders, and L. I. Mullin, "A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis," *Evolution and Human Behavior,* vol. 27, no. 6, pp. 401-416, 2006/11/01/ 2006.

[23] T. K. Koerner and Y. Zhang, "Differential effects of hearing impairment and age on electrophysiological and behavioral measures of speech in noise," *Hearing Research,* vol. 370, pp. 130-142, 2018.

[24] P. Ekman, "Are there basic emotions?," (in eng), *Psychological review,* vol. 99, no. 3, pp. 550-553, 1992/07/ 1992.

[25] L. Bai, H. Ma, Y. X. Huang, and Y. J. Luo, "The development of native Chinese affective picture system—A pretest in 46 college students," *Chinese Mental Health Journal,* vol. 19, no. 11, pp. 719-722, 2005.

[26] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software,* vol. 67, no. 1, pp. 1-48, 2015.

[27] G. A. Studebaker, "A 'rationalized' arcsine transform," *Journal of Speech and Hearing Research,* vol. 28, pp. 455-462, 1985.

[28] R. H. Baayen and P. Milin, "Analyzing reaction times," *International Journal of Psychological Research,* vol. 3, no. 2, pp. 12-28, 2010.

[29] R. V. Lenth, "Least-Squares Means: The R Package lsmeans," *Journal of Statistical Software,* vol. 1, no. 1, 2016.

[30] Y. F. Chien, J. A. Sereno, and J. Zhang, "What's in a Word: Observing the Contribution of Underlying and Surface Representations," *Language and Speech,* vol. 60, no. 4, pp. 643-657, Dec 2017.

[31] Y. Lin and H. Ding, "Multisensory integration of emotions in a face-prosody-semantics Stroop task," in *ICMI '19 NeuroManagement and Intelligent Computing Method on Multimodal Interaction*, Suzhou, China, 2019: ACM.

[32] C. Spence, "Explaining the Colavita visual dominance effect," *Progress in Brain Research,* vol. 176, pp. 245-258, 2009.

[33] M. M. Murray, A. F. Eardley, T. Edginton, R. Oyekan, E. Smyth, and P. J. Matusz, "Sensory dominance and multisensory integration as screening tools in aging," *Scientific Reports,* vol. 8, no. 1, p. 8901, Jun 2018.

[34] A. Schirmer, S. Zysset, S. A. Kotz, and D. Y. von Cramon, "Gender differences in the activation of inferior frontal cortex during emotional speech perception," *NeuroImage,* vol. 21, no. 3, pp. 1114-1123, 2004.

[35] T. M. Chaplin, "Gender and emotion expression: A developmental contextual perspective," *Emotion Review,* vol. 7, no. 1, pp. 14-21, Jan 2015.

[36] A. H. Fischer and M. LaFrance, "What drives the smile and the tear: Why women are more emotionally expressive than men," *Emotion Review,* vol. 7, no. 1, pp. 22-29, 2014.