Journal of Cross-Cultural Psychology
2021, Vol. 52(3) 275–294

DOI: 10.1177/0022022121990897
journals.sagepub.com/home/jcc

Unattended Emotional Prosody Affects Visual Processing of Facial Expressions in Mandarin-Speaking Chinese: A Comparison With English-Speaking Canadians

Pan Liu[1,2], Simon Rigoulot[1,3], Xiaoming Jiang[1,4], Shuyi Zhang[1], and Marc D. Pell[1]

Abstract

Emotional cues from different modalities have to be integrated during communication, a process that can be shaped by an individual's cultural background. We explored this issue in 25 Chinese participants by examining how listening to emotional prosody in Mandarin influenced participants' gazes at emotional faces in a modified visual search task. We also conducted a cross-cultural comparison between data of this study and that of our previous work in English-speaking Canadians using analogous methodology. In both studies, eye movements were recorded as participants scanned an array of four faces portraying fear, anger, happy, and neutral expressions, while passively listening to a pseudo-utterance expressing one of the four emotions (Mandarin utterance in this study; English utterance in our previous study). The frequency and duration of fixations to each face were analyzed during 5 seconds after the onset of faces, both during the presence of the speech (early time window) and after the utterance ended (late time window). During the late window, Chinese participants looked more frequently and longer at faces conveying congruent emotions as the speech, consistent with findings from English-speaking Canadians. Cross-cultural comparison further showed that Chinese, but not Canadians, looked more frequently and longer at angry faces, which may signal potential conflicts and social threats. We hypothesize that the socio-cultural norms related to harmony maintenance in the Eastern culture promoted Chinese participants' heightened sensitivity to, and deeper processing of, angry cues, highlighting culture-specific patterns in how individuals scan their social environment during emotion processing.

Keywords

eye-tracking, pseudo-speech prosody, facial emotion, cross-sensory, Mandarin Chinese, cross-cultural

[1]McGill University, Montréal, QC, Canada
[2]Western University, London, ON, Canada
[3]Université du Québec à Trois-Rivières, QC, Canada
[4]Tongji University, Shanghai, China

Corresponding Author:
Pan Liu, Department of Psychology, Brain & Mind Institute, Western University, Western Interdisciplinary Research Building, Room 2172, London, ON N6A 5B7, Canada.
Email: pliu261@gmail.com

In daily interactions, cues signifying our emotional state, motivations, and intentions are derived simultaneously from several communication sources, including language, facial expressions, and our tone of voice while speaking, or speech prosody (variation in supra-segmental acoustic features such as pitch or duration of speech elements). Given the high salience of non-linguistic cues for evaluating basic emotions (Ekman, 1992), recent studies have focused on how emotional details in the face and voice are simultaneously processed, which has provided new insights about the nature and time course for integrating emotional information across sensory channels based on behavioral, eye-tracking, and electrophysiological evidence (de Gelder & Vroomen, 2000; Liu et al., 2012, 2015a, 2015b; Paulmann et al., 2012; Paulmann & Pell, 2010; Pell, 2005; Vroomen et al., 2001). The current study extends this literature by investigating how emotional cues in the voice and face interact to influence gaze behavior, while testing recent claims that these processes may be shaped by cultural learning.

Cross-channel bias in the perception of emotional prosody and facial expressions was first reported in the behavioral literature. de Gelder and Vroomen (2000) demonstrated that when an ambiguous facial expression, morphed between sadness and happiness, was accompanied by a vocal stimulus (in Dutch) expressing either sadness or happiness, participants were more likely to identify the face as expressing the same emotion as the voice, even when they were ignoring the voice. Similar effects were replicated when a morphed vocal utterance was recognized in the presence of a happy or sad face (de Gelder & Vroomen, 2000), and when participants diverted attention from emotional features of the stimuli by performing an extra task (Vroomen et al., 2001). This work suggests that emotional information encountered in different sensory channels is registered and processed involuntarily, such that information from one channel tends to facilitate the processing of congruent emotional cues of the other. Supportive evidence for emotional congruence effects was also reported in other studies (Jaywant & Pell, 2012; Pell, 2005; Pell & Skorup, 2008; Schwartz & Pell, 2012): when listening to vocal cues that express an emotion congruent with a face, evaluative judgments of the face tend to be more accurate and/or quicker due to the activation of associative knowledge about emotions shared across channels (Massaro & Egan, 1996; Pell et al., 2011).

Among the approaches for advancing knowledge of how vocal and facial emotions interact in communication, eye-tracking can supply a direct, real-time measure of attention allocation to faces as participants are exposed to emotional cues in the auditory modality. Recent eye-tracking studies corroborate the notion that emotional cues registered in the voice involuntarily bias how facial expressions are processed. Using a modified visual search task in which native Canadian English speakers viewed an array of six faces expressing different emotions, Paulmann et al. (2012) reported that participants dwell longer to expressions that match the emotional tone of a concurrent vocal instruction (e.g., *"Click on the happy face"*). These data underscore that emotional meanings encoded by prosody play a major role in directing attention to congruent features of an adjoined face stimulus (Paulmann et al., 2012).

Using a similar paradigm, Rigoulot and Pell (2012) further demonstrated cross-modal interactions by presenting emotionally-inflected "pseudo-utterances" in English (e.g., "Someone migged the pazing") while participants were viewing an array of four faces expressing different emotions. Participants were instructed to ignore the auditory stimulus and attend to the faces to make a recall judgment (whether a specific face had been presented or not). Results confirmed that participants looked longer and more frequently at faces expressing a congruent emotion with the prosody, an effect that endured even after the speech stimulus had ended (Rigoulot & Pell, 2012). In a follow-up study focusing on eye movements within a face (e.g., upper vs. lower regions), they found that prosody of congruent emotion as the face is likely to guide eye fixations toward facial regions that bear the most salient cues for the specific emotion (Rigoulot & Pell, 2014). Similar data highlighting voice-face congruence effects on fixation patterns have also been reported in English-speaking pre-schoolers (Berman et al., 2016). Taken together, these findings suggest that emotional prosody activates conceptual knowledge about emotion categories in an unconscious manner and promotes more efficient visual attention towards relevant (i.e., congruent) information in the social environment. Such tendencies could facilitate operations for emotion recognition, perceptual decisions, and the ability to generate coherent social inferences based on representations of another's emotional state (Kamachi et al., 2003; Noppeney et al., 2010; Paulmann & Pell, 2011; Rigoulot & Pell, 2012).

Most work to date has focused on individuals from Western societies (e.g., North American, European). However, differences in how individuals from Western versus Eastern cultures visually process emotional faces, including distinct looking strategies and fixation patterns to facial features have been reported by a number of studies: while Western Caucasians tend to look between the eyes and mouth for diagnostic information, East Asians mostly weight cues shown in the eyes for emotion recognition (Jack et al., 2009; Mai et al., 2011; Tan et al., 2015). This cultural difference might be related with social rules concerning emotional displays. Compared to Western individualistic cultures where the expression of individual feelings is encouraged, in Eastern collectivistic cultures, people learn to constrain their emotions to maintain group harmony (Matsumoto et al., 2002; Park et al., 2018). Since muscles around the eyes are more difficult to control compared to muscles around the mouth and thus easier to reveal the true feelings, the eye regions are thought to be the main diagnostic source for Easterners to understand the emotional states of other people (Mai et al., 2011; Yuki et al., 2007).

Cultural differences also emerge during cross-channel emotion processing, where Eastern Asian participants tend to show higher sensitivity to vocal cues than Western participants. In a cross-cultural study, Liu et al. (2015a) compared English Canadians and Chinese Mandarin speakers as they performed an emotional Stroop task, composed of paired facial expressions and emotional pseudo-utterances expressing sadness or fear. Each group judged voice-face displays expressed by members of their own language/culture, forming emotionally congruent and incongruent trials that were evaluated in two conditions: one in which participants focused on the face while ignoring the voice; and one in which they ignored the face while judging the voice. Behavioral results showed that when judging emotional prosody, the accuracy of Chinese participants was influenced less by the to-be-ignored faces than for Canadian participants (Liu et al., 2015a). This parallels earlier reports showing that Japanese participants are less susceptible to behavioral interference from emotional faces, and more sensitive to emotional prosody, compared to Dutch participants (Tanaka et al., 2010). The idea that Eastern (Chinese) individuals are presumably less sensitive to facial cues than Western (Canadian) individuals, and potentially more sensitive to emotional prosody, can also be inferred from related event-related potential (ERP) data (Liu et al., 2015a, 2015b). Based on the evidence from both behavioral and neural measures, culture seems to play an important role in modulating attention allocation to emotional cues at both controlled and automatic levels.

Arguably, observed differences in cross-channel emotion processing are mediated by culture-specific norms that shape emotion communication in social settings (Gorodnichenko & Roland, 2012; Oyserman et al., 2002). East Asian cultures are considered higher in "interdependence" (i.e., the intention and preparedness to be socially connected with others), as opposed to more independent Western cultures which emphasize autonomy and individual thoughts (Kitayama et al., 2007; Markus & Kitayama, 1991). Interdependent cultures place greater importance on harmonious social/interpersonal relations and group interests (Hall & Hall, 1990; Kitayama et al., 2007; Scollon & Scollon, 1995). Members of these groups, therefore, learn and practice social norms to maintain harmony and avoid social conflicts, for example, less eye contacts (Hawrysh & Zaichkowsky, 1991; McCarthy et al., 2006, 2008), restrained facial expressivity (Ekman, 1971; Markus & Kitayama, 1991; Matsumoto et al., 1998, 2008), and indirect speech that uses prosodic cues to convey negative intentions beyond the literal meaning of words (Bilbow, 1997). As a result, Easterners may develop a greater sensitivity to social cues that signal potential conflicts within the group, such as negative facial expressions (Goto et al., 2013), mismatched eye gaze directions (Cohen et al., 2017), and deviant social behaviors (Mu et al., 2015).

Further, as East Asian cultures encourage emotion constraints that may result in less salient visual emotional cues (e.g., less eye contact, less expressive faces), Easterners may also learn to broaden their attentional span to make use of information from different resources when engaging in the processing of cross-channel emotions. As a result, they may show distinct patterns in attention orientation when processing multi-channel emotions. Specifically, their less reliance on linguistic (indirect speech) and visual cues (less eye contact, restrained facial expressions) may lead to a greater reliance on, and sensitivity to, prosodic cues (Engelmann & Pogosyan, 2013; Tanaka et al., 2010).

This literature motivates our current investigation on the processing strategies of cross-modality emotional cues in a Chinese sample. Specifically, we asked two questions: (1) how do individuals from Eastern cultures visually process facial displays of emotion in the context of emotional prosody? (2) how is this process modulated by culture? To address these questions, we examined the on-line eye-movement patterns of a group of Mandarin-speaking Chinese participants by using a visual search paradigm analogous to that used in our previous work on a group of English-speaking Canadian participants (Rigoulot & Pell, 2012) as reviewed earlier. Specifically, we presented visual arrays of various facial expressions using the same facial stimuli as in Rigoulot and Pell (2012), each array accompanied by Mandarin-like emotional pseudo-utterances, where the participants were instructed to ignore the speech stimuli and focus on facial expressions. The only methodological difference between the current study and our previous work was that the previous study used English pseudo-utterances as the speech stimuli, which were cut to consistent length across trials (see details in the Method section). By mirroring real-life situations when someone hears a person speaking while scanning the faces of a group of interlocutors, our paradigm allowed processing of emotional cues to be assessed in real time. Using an analogous paradigm as our previous work (Rigoulot & Pell, 2012) also allowed direct cultural comparison between the current Chinese group and our previous data from an English-speaking Canadian group.

Based on the literature, we hypothesized that Chinese participants' visual processing of facial expressions would be impacted by the accompanying emotional prosody, depending on the emotional congruence between each face and the utterance. In particular, given that our paradigm required explicit attention to faces (but not the voice), we expected that the Chinese participants might show greater sensitivity to the prosodic context compared to the Canadian group (Goto et al., 2013; Ishii et al., 2010; Liu et al. 2015a, 2015b; Mu et al., 2015). We also expected that the Chinese participants might be more sensitive to emotional cues that signal potential conflicts and social threats, regardless of the information channel (e.g., angry faces and/or angry speech); this may be especially salient when angry faces were accompanied by angry speech. Indeed, the general emotion recognition literature has indicated a facilitating effect in processing angry faces (Hansen & Hansen, 1988; LoBue, 2009; Öhman et al., 2001) and angry speech (Paulmann & Uskul, 2014), supporting the more efficient processing of threatening signals. This effect may be particularly salient in the context of an interdependent culture that places greater values on conflict avoidance and harmony maintenance.

Method

A Priori Power Analysis

We estimated required sample size based on findings of our previous work that used a highly analogous study paradigm (Rigoulot & Pell, 2012). In that study, English-speaking participants looked longer and more frequently at faces expressing congruent emotions as the prosody, with medium to large effect sizes reported ($r=0.33$–0.51). In the current study, to achieve a statistical power of 0.95, with an expected effect size of 0.4 and α of 0.05 (two-tailed), a minimum of 19 participants was required (G*Power 3.1.9.2).[1]

Participants

Twenty-five Chinese participants (12 females/12 males; $M_{\text{age}}=22.42$ years, $SD=2.81$) were recruited through campus advertisements at McGill University. All participants were native Mandarin speakers, were born and lived in mainland China until at least 18 years of age, and none had lived in Canada for more than 2 years at the time of participation. All participants were right-handed and reported normal hearing and normal/corrected-to-normal vision. Informed written consent was obtained from each participant prior to the study. Participants completed a demographic questionnaire prior to the experiment, as well as the State-Trait Anxiety Inventory (STAI; Spielberger et al., 1983), given data showing that individual anxiety levels have an important impact on emotion processing (e.g., Pell et al., 2015; Peschard et al., 2014). The results of STAI yielded a mean score of 35.75 across all participants for State-Anxiety ($SD=6.14$) and a mean score of 44.67 for Trait-Anxiety ($SD=8.90$). These results fall within the normal range of STAI scores for young Chinese adults reported by a normalization study administered in China (Li & Qian, 1995).

Stimuli

Stimuli consisted of emotional pseudo-utterances and static facial expressions, both selected from standardized databases. Specifically, vocal stimuli consisted of 64 emotionally inflected pseudo-utterances in Mandarin Chinese, produced by two native Mandarin speakers (one male) expressing four emotions, fear, anger, happiness, and neutrality. These stimuli were validated in a different group of native Mandarin-Chinese speakers who decided which emotion was being expressed by each stimulus in a seven-option forced choice task; stimuli that reached acceptable recognition rates across listeners (three times chance performance, 42.86%) were selected for the database (see Liu & Pell, 2012 for details). Pseudo-utterances have been successfully used to explore the processing of emotional prosody in the absence of linguistic-semantic information; they are composed of pseudo content words conjoined by real function words, rendering the utterances meaningless but possessing natural segmental and supra-segmental properties of the target language (Banse & Scherer, 1996; Pell et al., 2009). In order to better approach speech perception in real life, complete pseudo-utterances (duration ranging from 930 ms to 1,900 ms) were presented as the vocal stimulus here. Similarly, our previous work (Rigoulot & Pell, 2012) also employed 64 English pseudo-utterances, produced by two male and two female native English speakers conveying the same four emotions (fear, anger, happiness, and neutrality). However, rather than using complete pseudo-utterances, this previous study cut all utterances to a consistent duration of 1,250 ms. Realizing that the cut utterances might sound unnatural to participants, in the current study, we decided to use full Chinese pseudo-utterances with varied duration. The varied duration of vocal stimuli, a natural property of vocal emotion expressions, was then statistically controlled as a random factor. See Table 1 for an overview of acoustic properties of the vocal stimuli by emotion type and associated recognition rates for the selected tokens.

Facial stimuli were identical to those used by Rigoulot and Pell (2012), consisting of 24 color pictures (8.5 × 11 cm) of faces expressing the four target emotions (fear, anger, happiness, neutrality). The faces were posed by three female and three male actors of different racial backgrounds (Caucasian, East Asian, Black), cropped to show only facial features (Table 1). Major

Table 1. Mean (Standard Deviation) of the Recognition Rates and Acoustic Parameters of Vocal Stimuli, and Recognition Rates of the Facial Stimuli (for Acoustics, f_0 and Amplitude Values were Normalized; Speech Rate is in Number of Syllables Per Second; Duration is in Millisecond. See Liu and Pell (2012) for details).

		Anger	Fear	Happy	Neutral
Vocal expression	Recognition rates	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)	0.9 (0.1)
	F0 mean	1.6 (0.5)	1.0 (0.5)	1.6 (0.5)	0.4 (0.1)
	F0 range	2.1 (0.6)	1.2 (0.5)	2.2 (0.7)	1.1 (0.5)
	Amplitude mean	0.6 (0.2)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)
	Amplitude range	0.9 (0.2)	0.9 (0.1)	0.9 (0.2)	0.8 (0.1)
	HNR mean	9.6 (1.6)	11.4 (2.2)	11.8 (1.8)	10.8 (1.9)
	Speech rate	7.4 (1.0)	5.9 (0.5)	6.2 (0.6)	6.5 (0.7)
	Duration (ms)	1,306 (187)	1,610 (211)	1,545 (180)	1,508 (208)
Face	Recognition rates	0.87 (0.11)	0.9 (0.1)	1.0 (0.0)	0.9 (0.0)

Note. F0: fundamental frequency; HNR: harmonics-noise ratio; ms: millisecond.

physical parameters of the selected pictures (luminance, contrast for gray and RGB layers, kurtosis, and skewness) were matched across the emotional categories by using ImageJ software to control the potential confounding effect of low-level physical features. A series of facial arrays was then constructed, each composed of four faces posed by the same actor expressing the four different emotions (Figure 1). The center of the four facial pictures was equidistant and localized at 11 cm from central fixation. A four-face array, when controlled for spatial arrangement of the faces, resulted in 24 spatially distinct arrays for each of the six actors (144 arrays in total), and were counterbalanced across participants during the task.

Experimental Procedure

The experimental procedure, including the size of facial display, the eye-tracking equipment and parameters, was exactly the same as what was used in Rigoulot and Pell (2012). In each trial, one pseudo-utterance was paired with a facial array posed by an actor of the same sex, although the identity of the speaker/actor across trials was not predictable. Each of the 64 pseudo-utterances (16 items × 4 emotions) was paired with each of the 24 facial arrays, for a total of 1,536 trials. To avoid excessive repetition of stimuli for each participant and ensure that the 24 possible spatial arrangements were fully counterbalanced across participants, each participant encountered all possible voice-face pairings but only a third of the unique spatial arrangements (randomly selected from the full set; $n=512$ trials/participant). In addition to the 512 trials in which combined vocal-facial stimuli were presented, 120 facial arrays without concurrent vocal stimuli were randomly inserted during the sequence of trials as fillers. As a result, a total of 632 trials were administered to each participant.

During the experiment, participants were seated in a quiet, dimly lit room at a 75 cm distance from the computer screen. Stimuli were presented by Experiment Builder software (SR Research) on a ViewSonic P95f monitor/PC computer. Eye-movements were recorded by an EyeLink II system (head mounted video-based; SR Research, Mississauga, Ontario, Canada) connected to a separate PC. With a 500 Hz sampling rate, the eye-tracker was calibrated at the onset of experiment and whenever needed during the testing. The calibration was accepted if the average error was less than 0.5 μ in pupil-tracking mode. Each trial began with a centrally located circle that participants were asked to fixate, allowing for drift-correction of the eye-tracker. The fixation was presented for a random duration of 100 to 300 ms to prevent anticipatory saccades, after which the facial array appeared on a gray background for 5,000 ms, at the same time as an

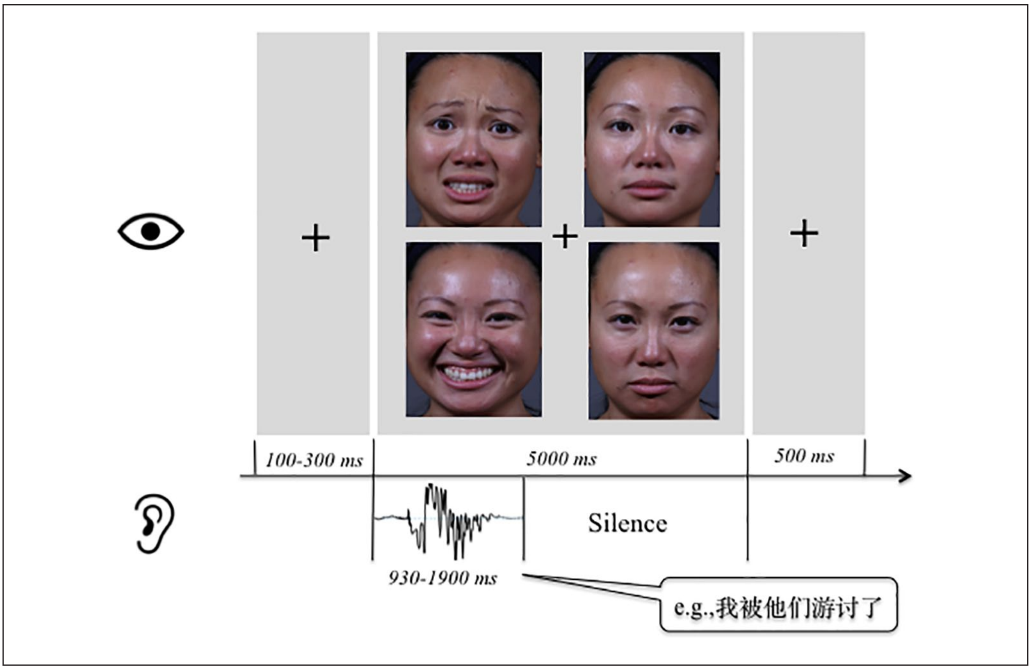

Figure 1. Illustration of the trial procedure of the visual search task.

emotional pseudo-utterance was presented binaurally over headphones. Looks with a minimum length of 100 ms were defined as fixations, and the onset of the vocal and facial stimuli in each trial was precisely synchronized (Figure 1).

Participants were informed that they would see an array of four faces and hear a meaningless sentence in Chinese in most cases; they were instructed to familiarize themselves with the faces in order to make a recall judgment following certain trials (1/3 of all trials). On recall trials, a single face was presented at the center of the computer screen and the participants had to indicate whether he/she had seen it during the preceding array by button pressing (yes/no). This task ensured that participants attended carefully to the facial arrays but did not explicitly orient their attention to underlying emotional features. Half of the recall faces yielded a “yes” response (i.e., the face was presented in the preceding array; an equal number of each of the four faces were presented) and half the trials yielded a “no” response (i.e., a facial expression posed by the same actor conveying emotions other than the four target emotions). The assignment of the yes and no response buttons was counterbalanced across participants. At the end of each trial (with or without a recall task), a blank screen appeared for 1,000 ms before the next trial started. Participants completed 11 practice trials before each recording session. The eye-tracking experiment lasted approximately 2 hours. We administered the study at two different sessions of 60 minutes each and scheduled 2 days in a row. After the experiment, the participants were compensated for their participation ($25 CAD). This study was reviewed and ethically approved by the Institutional Review Board of Faculty of Medicine at McGill University.

Statistical Analysis for the Chinese Group

Data for all 25 participants were included in statistical analyses. The recognition rates of the face recall task in one-third trials yielded a mean value of 85.43% (*SD* = 0.12), suggesting that participants were properly attending to the facial stimuli during the procedure. Eye-tracking data showed that participants looked at all four faces within each array for 97% of all trials. Data analysis concentrated on trials with emotional speech paired with the face array, considering four target cells defined as rectangles of the same size and location as the four faces in each array. Two sets of eye movement measures were examined: (1) the frequency and duration of the first fixations directed to different faces during each trial; (2) the frequency and duration of fixations directed to different faces during two time windows of interest in the 5,000 ms stimulus presentation period: early time window, from the onset (0 ms) to the offset of the pseudo-utterance when voice-face information was jointly available; late time window, from the offset of the pseudo-utterance to the offset of the facial array (5,000 ms) when only the facial array was present. These two time windows allow investigation of both the early and late biasing effects of emotional prosody on eye movement patterns to facial expressions.

A series of linear mixed effects models (LMM) were computed using the lme4 package (Bates, 2007; Bates & Sarkar, 2006) of R (version 2.13.1; Baayen, 2008; Baayen et al., 2008; R Development Core Team, 2010) to fit the eye movement measures and evaluate differences between conditions. In all models, fixed factors included Face (happy, angry, fearful, neutral) and Congruence between each face and the prosody (congruent, incongruent); control factors included each participant's sex (female, male) and years of education. Random factors (intercept only) differed depending on which eye-tracking measure served as the dependent variable (DV) in the model. For first fixations, as they occurred only once per trial, the frequency was calculated as the sum across trials for each of the four faces per participant; models with frequency of first fixations as the DV included subject as the random factor. When duration of first fixations served as the DV, subject, the specific item of speech stimuli, length of each speech item, and number of repetition of each speech item were included as four random factors.

For fixations during the early and late time windows that occurred multiple times per trial, frequency was summed across fixations within each trial for each of the four faces. Models with frequency of fixations during the two time windows as DV included subject, speech item, speech length, and number of repetition of each speech item as random factors; when duration of fixations during the two time windows was treated as DV, number of fixation of each face was added as the fifth random factor. For these DVs, we also included the total number of fixations across trials as a control factor in addition to sex and years of education, to account for the individual differences in this factor.

Cross-cultural Analysis

To further investigate the potential cultural differences in processing multi-sensory emotion, we directly compared the current data from our Chinese group and data from an English Canadian group adopted from our previous work (Rigoulot & Pell, 2012). The same LMMs were used as those described above, with Culture (Chinese, Canadian) included as a third fixed factor. We were particularly interested in whether any effects of Face and Congruence observed in the Chinese participants would be found in the Canadian participants; thus, we focused on the two interaction terms, Face × Culture and Congruence × Culture in testing these models.

Results

Frequency and Duration of First Fixations

Descriptives for the frequency and duration of first fixations are presented in Table 2. We found a significant effect of Face on the frequency of first fixations, $F(3, 365) = 15.67$, $p < .001$. Pairwise comparisons showed that neutral expressions elicited fewer first fixations than happy ($b = -4.8$, $SE = 1.57$, $t = -3.07$, $p < .001$), fearful ($b = -4.6$, $SE = 1.57$, $t = -2.94$, $p < .001$), and

Table 2. Mean Frequency and Duration of First Fixations for Each Experimental Condition.

		Anger		Fear		Happy		Neutral	
		Frequency of first fixation							
	Prosody	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Face	Anger	32.88	4.80	32.92	5.01	30.04	5.59	32.52	5.27
	Fear	32.80	5.68	32.24	6.81	33.64	5.57	31.20	4.30
	Happy	32.44	4.78	33.64	5.49	33.24	6.27	34.04	6.20
	Neutral	28.04	6.37	27.64	5.54	29.60	5.08	28.20	5.07
		Duration of first fixation (ms)							
Face	Anger	236.35	119.01	243.38	154.42	244.98	158.34	235.67	126.09
	Fear	248.97	202.56	243.02	163.33	235.18	172.40	251.79	226.22
	Happy	241.05	139.04	248.21	137.16	251.42	171.98	238.84	206.23
	Neutral	236.43	176.49	234.99	136.05	239.24	159.24	239.71	136.93

Note. SD: standard deviation; ms: millisecond; shaded cells: emotionally congruent face-voice pairings.

angry expression ($b=-5.24$, $SE=1.57$, $t=-3.35$, $p<.001$; Figure 2).[2] No significant effect of Congruence or Face × Congruence interaction were observed for first fixation frequency ($ps>.50$). No significant effect was found for the duration of first fixations ($ps>.20$).

Frequency of fixations during early and late time windows

Descriptives of the fixation frequency during the early and late time windows are presented in Table 3. The main effect of Congruence was significant during the late time window only, $F(1, 45{,}377)=13.93$, $p<.001$. In particular, eye fixations were more often directed to faces that were congruent versus incongruent with the paired emotional prosody ($b=-0.09$, $SE=0.08$, $t=-1.01$, $p<.001$; congruent as baseline). The main effect of Face was significant during both time windows (early, $F(3, 29{,}657)=7.51$, $p<.001$; late, $F(3, 45{,}377)=225.43$, $p<.001$). The Face × Congruence interaction was not significant in either time window ($ps>.20$). Figure 3 illustrates the effects of Face and Congruence.

Pairwise comparison between face types showed that in both time windows, more frequent fixations were directed to fearful faces than all other faces, including neutral (early, $b=0.44$, $SE=0.12$, $t=3.59$, $p<.001$; late, $b=1.32$, $SE=0.10$, $t=12.63$, $p<.001$; neutral as baseline), angry (early, $b=0.37$, $SE=0.13$, $t=2.96$, $p<.001$; late, $b=0.60$, $SE=0.10$, $t=5.82$, $p<.001$; angry as baseline), and happy (early, $b=0.35$, $SE=0.12$, $t=2.90$, $p<.001$; late, $b=0.90$, $SE=0.10$, $t=8.62$, $p<.001$; happy as baseline), regardless of the accompanying emotional voice. During the late time window, happy and angry faces also elicited more frequent fixations than neutral faces (happy, $b=0.42$, $SE=0.11$, $t=3.98$, $p<.001$; angry, $b=0.71$, $SE=0.10$, $t=6.85$, $p<.001$; neutral as baseline), indicating a broader emotionality effect during the later processing stage[3].

Duration of Fixations During Early and Late Time Windows

Descriptives of the fixation duration during the two time windows are presented in Table 4. During the early time window, no significant effect of Face, Congruence, or their interaction was observed on the duration of fixations ($ps>.20$). During the late time window, main effects of both fixed factors were significant: Face, $F(3, 118{,}520)=16.65$, $p<.001$;

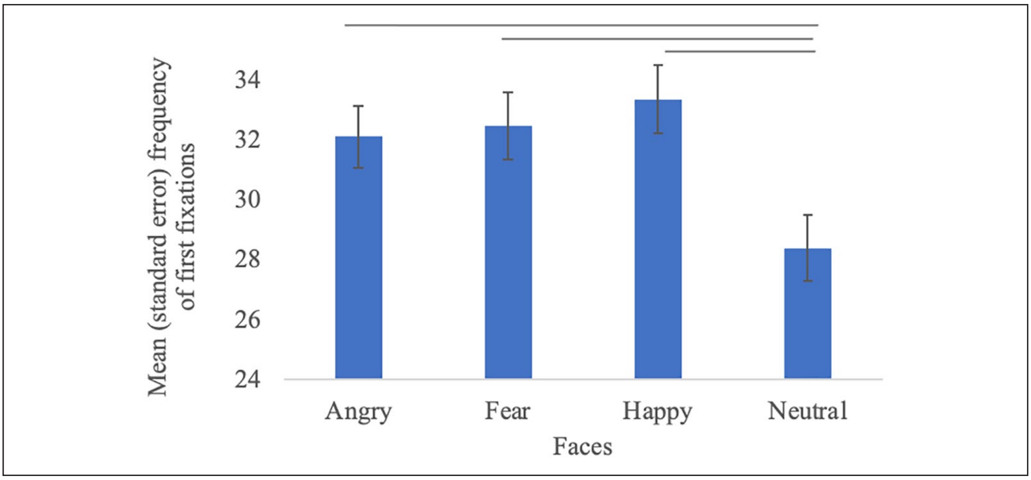

Figure 2. Frequency of first fixations by facial emotions. —: significant pair-wise differences found in LMMs.

Table 3. Mean Frequency of Fixations During the Early and Late Time Windows for Each Experimental Condition.

		Anger		Fear		Happy		Neutral	
		Early time window (0 ms to offset of speech)							
	Prosody	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Face	Anger	1.40	0.61	1.51	0.71	1.47	0.67	1.46	0.65
	Fear	1.44	0.65	1.55	0.74	1.52	0.71	1.52	0.71
	Happy	1.42	0.64	1.49	0.69	1.47	0.70	1.51	0.69
	Neutral	1.40	0.64	1.49	0.70	1.47	0.67	1.44	0.65
		Late time window (offset of speech to 5,000 ms)							
Face	Anger	2.78	1.57	2.57	1.48	2.61	1.50	2.69	1.53
	Fear	3.01	1.66	2.89	1.64	2.81	1.60	2.84	1.59
	Happy	2.52	1.32	2.32	1.32	2.53	1.51	2.48	1.43
	Neutral	2.41	1.39	2.29	1.33	2.34	1.34	2.41	1.43

Note. SD: standard deviation; ms: millisecond; shaded cells: emotionally congruent face-voice pairings.

Congruence, $F(1, 118{,}520) = 7.41$, $p = .01$, without any significant Face × Congruence interaction, $p = .31$ (Figure 4). Pairwise comparisons showed that angry faces were fixated longer than neutral ($b = 7.87$, $SE = 2.64$, $t = 2.98$, $p < .001$; neutral as baseline) and fearful faces ($b = 6.10$, $SE = 2.48$, $t = 2.46$, $p = .01$; fear as baseline). In addition, facial expressions that were congruent with the preceding voice were fixated longer than faces incongruent with the voice ($b = -1.96$, $SE = 2.06$, $t = -0.95$, $p = .01$; congruent as baseline).[4]

Cross-cultural Comparison

In cross-cultural analysis that included Culture as a third fixed factor, we found significant Face × Culture interactions for the frequency of fixations in both time windows (early, F(3,

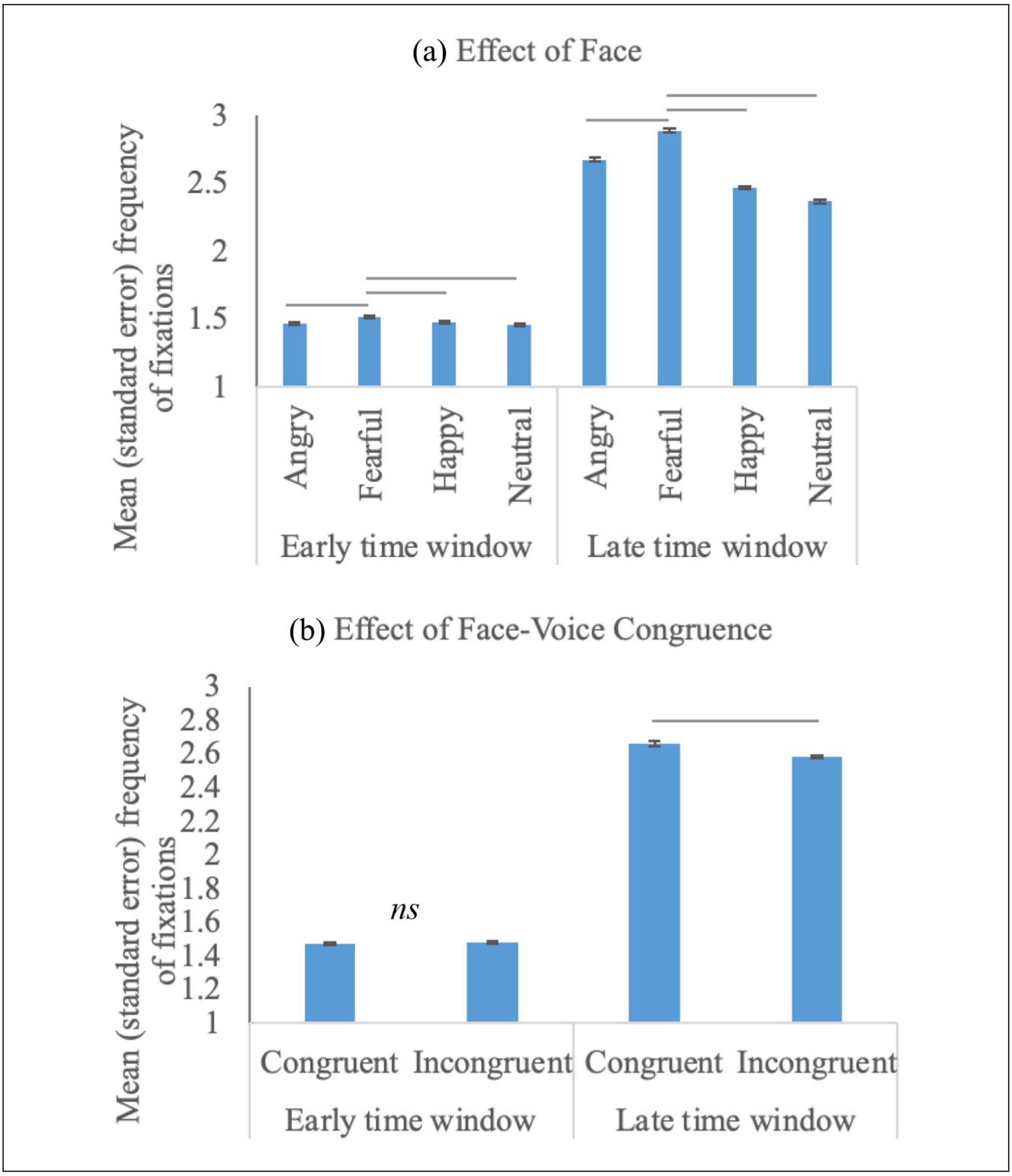

Figure 3. Frequency of fixations for the early and late time windows as a function of face (a) and face-voice congruence (b). —: significant differences found in LMMs; ns: non-significant.

59,876)=8.81, $p < .001$; late, $F(3,\ 103{,}499)=36.72$, $p < .001$), and the duration of fixations in the late time window ($F(3,\ 265{,}578)=5.28$, $p < .001$).[5] Decomposing the interactions did not show any significant simple effects of either factor in the early time window (*p*s > .18). During the late time window, we found that the significant interaction was driven by cultural differences in processing the angry versus other faces (Figure 5): Chinese participants fixated at angry faces more frequently than happy faces regardless of the emotion of the preceding speech prosody ($b=0.30$, $SE=0.10$, $t=2.84$, $p < .001$; happy faces as baseline), while Canadian participants did not ($b=0.09$, $SE=0.08$, $t=1.12$, $p > .20$). Chinese participants also looked longer at angry

Table 4. Mean Duration of Fixations During the Early and Late Time Windows for Each Experimental Condition.

		Anger		Fear		Happy		Neutral	
		Early time window (0 ms to offset of speech)							
	Prosody	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Face	Anger	228.40	105.22	233.86	115.78	233.18	113.64	232.03	109.37
	Fear	225.97	109.62	231.65	111.32	229.20	106.37	230.72	107.80
	Happy	227.85	108.39	236.70	111.15	235.11	111.67	228.66	109.66
	Neutral	226.48	103.15	235.53	111.22	229.43	111.12	230.63	105.44
		Late time window (offset of speech to 5,000 ms)							
Face	Anger	271.20	160.42	270.26	157.61	271.33	172.80	274.51	175.35
	Fear	267.36	167.81	275.34	182.07	273.06	171.69	269.15	179.66
	Happy	271.07	174.11	274.29	174.73	276.19	185.45	269.70	184.66
	Neutral	267.47	168.13	266.66	171.04	267.84	174.49	272.09	167.54

Note. SD: standard deviation; ms: millisecond; shaded cells: emotionally congruent face-voice pairings.

expressions than neutral faces regardless of the emotion of the preceding speech prosody ($b=7.87$, $SE=2.64$, $t=2.98$, $p<.001$; neutral faces as baseline), whereas Canadian participants did not show such a pattern ($b=2.26$, $SE=2.21$, $t=1.02$, $p>.30$). However, contrary to our expectation, we did not find any significant interactions between Culture and Congruence.

Discussion

In natural conversations, people typically are exposed to facial and vocal cues at the same time and tend to process the multisensory emotional information automatically, a process that could be modulated by cultural backgrounds (Campanella & Belin, 2007; Tanaka et al., 2010). Building on previous research on cross-sensory emotion processing, our study investigated how unattended emotional prosody expressed by Chinese pseudo-utterances affected the way Chinese participants scanned facial expressions that are emotionally (un)related to the voice. In cross-cultural analysis, we further explored the cultural differences in these processes by comparing our Chinese participants with a group of English-speaking Canadian participants adopted from our previous work (Rigoulot & Pell, 2012). As expected, we found that Chinese participants' eye gaze patterns were modulated by unattended emotional prosody, such that they looked more frequently, and with longer duration, at faces that were emotionally congruent (vs. incongruent) with the preceding prosody during the late time window. Finally, cross-cultural analysis showed that compared to the Canadian group, Chinese participants processed angry faces with more frequent fixations as well as longer fixation durations during the late time window, indicating a processing bias of this group toward emotional cues that potentially signal conflicts and social threats. However, we did not observe the expected cultural difference that Chinese participants would be impacted to a greater extent by unattended emotional prosody than Canadian participants.

First, LMMs conducted on Chinese participants showed that their eye gazing patterns were modulated by the emotions conveyed by faces. For first fixations, they looked more frequently at faces expressing an emotion (fear, happiness, anger) than neutral faces, regardless of the accompanying speech prosody, replicating a general emotionality effect of faces during the very early stage of emotion processing (e.g., Calvo et al., 2007; Nummenmaa et al., 2006). During the early

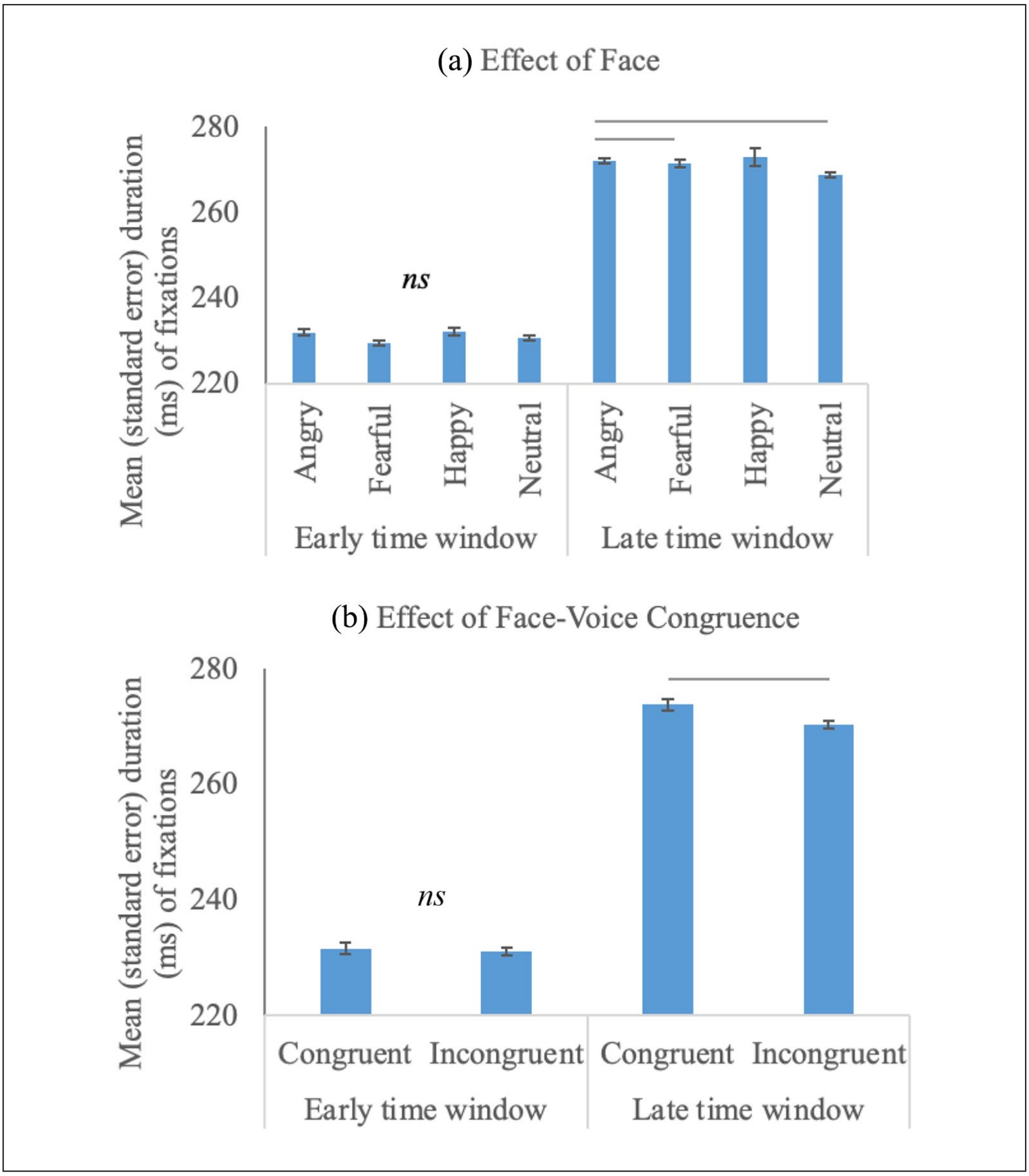

Figure 4. Duration of fixations for the early and late time windows as a function of face (a) and face-voice congruence (b). —: significant differences found in LMMs; ns: non-significant.

and late time windows, Chinese participants looked more frequently at fearful faces compared to most other faces; during the late time window, they looked longer at angry faces than other faces. In the eye-tracking literature, higher fixation frequency may indicate greater salience or noticeability of the target object (Eisenbarth & Alpers, 2011; Fitts et al., 1950; Gamer & Büchel, 2009); longer duration of fixations is thought to index deeper, more elaborate processing of the meaning of the target (e.g., Fischer et al., 2013; Loftus & Mackworth, 1978; Glöckner & Herbold, 2011). Hence, the observed patterns toward fearful and angry faces indicate that these faces bear particular salience and significance. Indeed, fearful faces inform the presence of immediate danger that may be life-threatening and thus can successfully capture early attentional resources,

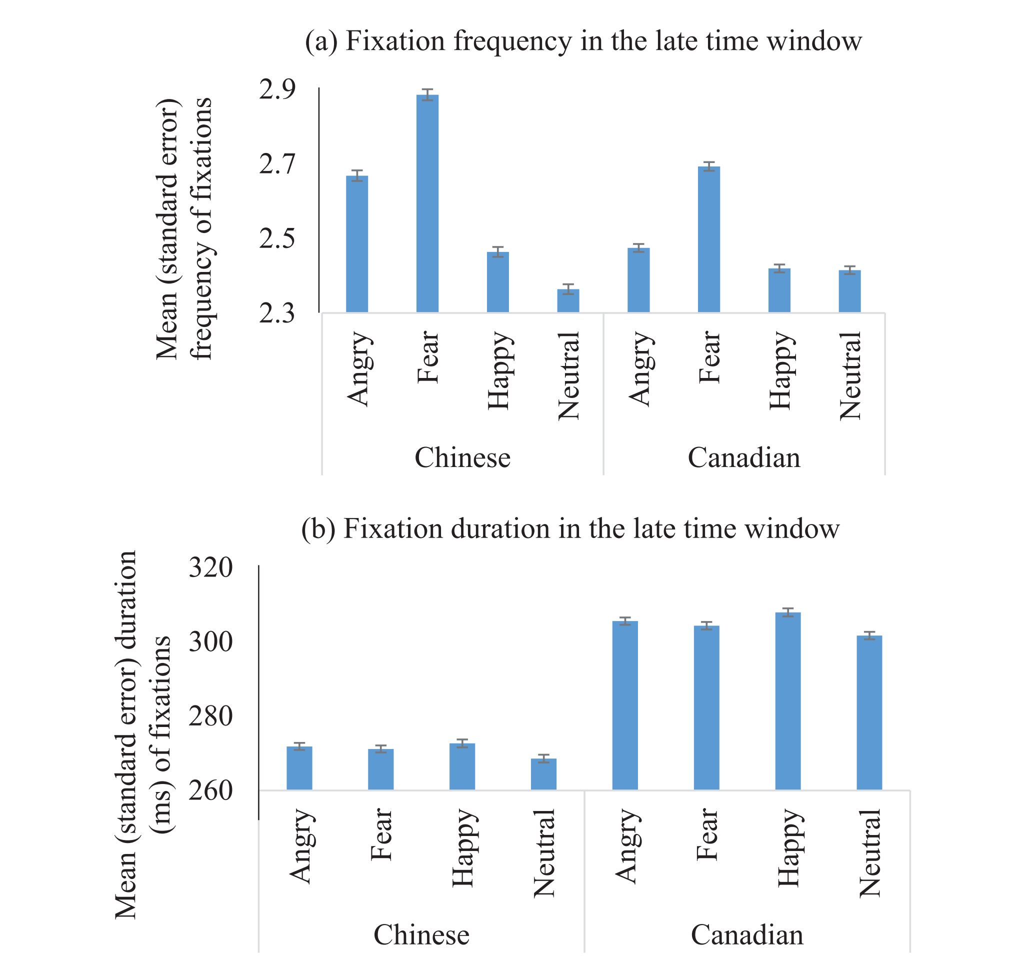

Figure 5. Cultural differences in processing different faces in the late time window across conditions of prosody. For the frequency of fixations (a), the group difference was driven by the angry versus happy faces; for the duration of fixations (b), the group difference was driven by the angry versus neural faces.

potentially facilitating adaptive behaviors (e.g., withdrawal actions; Chronaki et al., 2018; Frischen et al., 2008; Neuberg et al., 2011; Pourtois et al., 2004). A similar processing bias toward fearful faces was also reported in Rigoulot and Pell (2012). Angry faces, on the other hand, signal potential conflicts that may undermine interpersonal harmony; Chinese participants' longer duration at these angry faces during the late time window suggest that the signs of potential violation of interpersonal harmony elicited more elaborate processing in them. Interestingly, this pattern in angry faces was not reported in Rigoulot and Pell (2012), a cultural difference between the two groups confirmed by a significant Face × Culture interaction as detailed below.

Chinese participants also showed a significant main effect of face-voice congruence on both the frequency and duration of fixations during the late time window (after the utterance ended). The participants looked more frequently, and with longer duration, at faces that were emotionally congruent (vs. incongruent) with the preceding prosody, suggesting that the emotional congruence across modalities facilitated participants' visual processing of faces. This facilitation or priming effect across modalities is consistent with previous findings (Brosch et al., 2008;

Paulmann et al., 2012; Pell, 2005; Rigoulot & Pell, 2012; Vuilleumier et al., 2001). In particular, the preceding prosody activated the semantic knowledge of a certain emotion shared by emotionally congruent facial cues and therefore prioritized the processing of congruent faces, which is known to occur in the absence of voluntary deployment of attention or explicit meaning evaluation (Jaywant & Pell, 2012; Kitayama & Howard, 1994; Pell et al., 2011). In real-life communication, prioritizing emotionally congruent cues across modalities may also be beneficial for more efficient understanding of others' intentions, as oftentimes people tend to convey a particular emotion by using congruent multisensory cues (Attardo et al., 2003).

In cross-cultural analysis, we found a significant Face × Culture interaction on frequency and duration of fixations during the late time window, which was driven by prioritized processing of angry faces in the Chinese group compared to the Canadian group. Specifically, the Chinese, but not the Canadians, looked longer at angry faces than neutral faces, and directed more frequent gazes toward angry faces than other faces in the late time window, irrespective of the paired emotional prosody. Similar early processing biases toward angry versus neutral cues have also been reported in Asian American participants, which occurred as early as 100 ms indicated by the P1 component of event-related potential data (Park et al., 2018). Across cultures, angry faces convey important social information indicating potential conflicts, disapproval, or violation of social rules or expectations (Averill, 1983). In the East Asian culture, which places greater value on group interests and social harmony (Hall & Hall, 1990), angry cues may bear particular importance as it may threaten harmony. Specifically, members from this culture engage in social learning and practices that help maintain their cultural values, for example, by indirect or restrained expression of feelings, especially those that may cause conflict or harm interpersonal harmony, such as anger. As a result, compared to their Western counterparts, individuals from the harmony-seeking culture may have less exposure to openly expressed anger, rendering angry cues more novel to them. This perceived novelty may elicit preferential visual processing, indicated by more frequent fixations and longer duration (e.g., Horstmann & Herwig, 2016; Yeung et al., 2016). Alternatively, when anger is present in the context of a harmony-seeking culture, it may signify violation of social norms that needs immediate attention and solution. Therefore, the presence of angry faces may be particularly concerning to East Asian individuals (Averill, 1983; Hall & Hall, 1990; Holtgraves, 1997), thus inducing deeper, more elaborate processing (Fischer et al., 2013; Glöckner & Herbold, 2011; Loftus & Mackworth, 1978).

In cross-cultural analysis, the Congruence × Culture interaction was not significant on any eye movement measures. We did not find evidence supporting our hypothesis that compared to Canadian participants, Chinese participants would be impacted to a greater extent by unattended emotional prosody. We suspect that this may be related to the facial stimuli that we used in the Chinese group, which were depicted by actors of different races (Caucasian, East Asians, and Black). We used the same multi-racial faces as those used in the Canadian group by Rigoulot and Pell (2012) to keep the two groups as methodologically comparable as possible. However, compared to Canadian participants, our Chinese participants may have less exposure to racial diversity given their experiences of growing up in China and having lived in Canada for a limited period of time, and the racial novelty of some of the faces might have impacted the results. This may be especially true when taking the paired prosody into consideration (i.e., Congruence effect), where the paring between Mandarin pseudo-utterances and non-Asian faces might have seemed unusual and odd, given that the Mandarin-speaking population is racially homogeneous in general. Interestingly, regardless of the racial novelty of the faces, we still observed significant cultural differences in processing angry faces, suggesting that for Chinese participants, the significance of angry cues might have outweighed the racial oddity of the faces, irrespective of the paired prosody. Future work using in-group stimuli for both facial and vocal modalities each cultural group is important to unravel this issue.

Other limitations of this study include the limited number of speakers for pseudo-utterances and certain methodological differences that still existed between the two groups (e.g., the length of the speech stimuli). Future research that tests cultural differences in a fully crossed study design using in-group facial and vocal stimuli will help evaluate and consolidate our arguments proposed here. Using vocal stimuli produced by a broader range of speakers is important for promoting the generalizability of our findings. Additionally, examining East Asian participants that have lived in North America for a wider range of length of time will help address the question of how the amount of exposure to a novel culture would modulate eye gaze behavior in emotion processing.

In summary, by comparing a group of Mandarin-speaking Chinese participants to a group of Canadian participants in a multi-modality eye-tracking task, we found that Chinese participants displayed a heightened sensitivity to, and deeper processing of, angry faces compared to Canadians. This processing bias in the Chinese group may be associated with the interdependent, harmony-seeking social norms in the Eastern culture. The absence of the expected result that the Chinese participants would be impacted to a greater extent by unattended speech prosody may be related to methodological factors, for example, the paring between multi-racial faces and Mandarin speech. Additionally, results from the Chinese group alone also extend the literature in cross-modality emotion processing: unattended emotional prosody affects visual processing of emotional faces in systematic ways, for example, a facilitation effect of emotional congruence during the later processing stage, consistent with those reported in Western participants (Paulmann et al., 2012; Rigoulot & Pell, 2012). To our knowledge, these findings provide the first evidence on the cultural differences and culture-specific patterns in cross-modality emotion processing. This study will also inform and motivate future research in this area, for example, using a fully crossed study design and in-group stimuli to further explore the cultural differences in multi-modality emotion processing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Pan Liu https://orcid.org/0000-0003-1278-2129

Notes

1. While the power analysis was conducted based on the traditional MANOVA approach, we used linear mixed-effects models (LMM) for data analysis in this study. Compared to MANOVA, LMM preserves as many data points as possible and further increases statistical power by accounting for random factors on the trial level (e.g., length of the speech stimulus; Baayen et al., 2008; Newman et al., 2012). See the method section for details.
2. We also examined whether positive and negative faces elicited different eye gaze patterns and found a significant main effect of valence (positive, negative) on first fixation frequency, $F(1, 369)=6.06$, $p=.01$, with negative faces (fear and anger) elicited more fixations than positive (happy) faces, $b=2.52$, $SE=1.16$, $t=2.18$, $p=.03$.
3. We examined the effect of valence of faces (positive, negative), which showed a significant main effect on the frequency of fixations during the late time window, $F(1, 45{,}381)=533.82$, $p<.001$, with negative faces (fear and anger) evoking more frequent fixations than positive (happy) faces, $b=0.80$, $SE=0.07$, $t=10.95$, $p<.001$.

4. No effect of the valence of faces (positive, negative) was found on the duration of fixations during the early and late time windows (*ps* > .23).
5. For the purpose of this study, we focused on results of the comparison between the two groups. For the results of the Canadian group only, please refer to Rigoulot and Pell (2012).

References

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243–260.

Averill, J. R. (1983). Studies on anger and aggression. Implications for theories of emotion. *American Psychologist*, *38*(11), 1145–1160.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introductionto statistics using R*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Bates, D. M. (2007). *Linear mixed model implementation in lme4* (Unpublished manuscript). University of Wisconsin.

Bates, D. M., & Sarkar, D. (2006). *lme4: Linear mixed-effects modeling using S4 classes R package* (Version 0.9975–10) [Computer software]. R Foundation for Statistical Computing.

Berman, J. M. J., Chambers, C. G., & Graham, S. A. (2016). Preschoolers' real-time coordination of vocal and facial emotional information. *Journal of Experimental Child Psychology*, *142*, 391–399.

Bilbow, G. T. (1997). Spoken discourse in the multicultural workplace in Hong Kong: Applying a model of discourse as 'impression management'. In F. Bargiela-Chiappini & S. Harris (Eds.), *The languages of business: An international perspective* (pp. 21–48). Edinburgh University Press.

Brosch, T., Grandjean, D., Sander, D., & Scherer, K. R. (2008). Behold the voice of wrath: Cross-modal modulation of visual attention by anger prosody. *Cognition*, *106*(3), 1497–1503.

Calvo, M. G., Nummenmaa, L., & Hyönä, J. (2007). Emotional and neutral scenes in competition: Orienting, efficiency, and identification. *The Quarterly Journal of Experimental Psychology*, *60*, 1585–1593.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535–543.

Chronaki, G., Wigelsworth, M., Pell, M. D., & Kotz, S. A. (2018). The development of cross-cultural recognition of vocal emotion during childhood and adolescence. *Scientific Reports*, *8*(1), 8659.

Cohen, A. S., Sasaki, J. Y., German, T. C., & Kim, H. S. (2017). Automatic mechanisms for social attention are culturally penetrable. *Cognitive Science*, *41*, 242–258.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*(3), 289–311.

Eisenbarth, H., & Alpers, G. (2011). Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion*, *11*, 860–865.

Ekman, P. (1971). *Universals and cultural differences in facial expressions of emotion*. In J. Cole (Ed.), *Nebraska symposium on motivation* (pp. 207–282). University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Engelmann, J. B., & Pogosyan, M. (2013). Emotion perception across cultures: The role of cognitive mechanisms. *Frontiers in Psychology*, *4*, 118.

Fischer, T., Graupner, S.-T., Velichkovsky, B. M., & Pannasch, S. (2013). Attentional dynamics during free picture viewing: Evidence from oculomotor behavior and electrocortical activity. *Frontiers in Systems Neuroscience*, *7*, 17.

Fitts, P. M., Jones, R. E., & Milton, J. L. (1950). Eye movement of aircraft pilots during instrument-landing approaches. *Aeronautical Engineering Review*, *9*, 24–29.

Frischen, A., Eastwood, J. D., & Smilek, D. (2008). Visual search for faces with emotional expressions. *Psychological Bulletin*, *134*, 662–676.

Gamer, M., & Büchel, C. (2009). Amygdala activation predicts gaze toward fearful eyes. *Journal of Neuroscience*, *15*, 9123–9126.

Glöckner, A., & Herbold, A. K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, *24*, 71–98.

Gorodnichenko, Y., & Roland, G. (2012). Understanding the individualism-collectivism cleavage and its effects: Lessons from cultural psychology. In International Economic Association (Eds.), *Institutions and Comparative Economic Development* (pp. 213–236). Palgrave Macmillan.

Goto, S. G., Yee, A., Lowenberg, K., & Lewis, R. S. (2013). Cultural differences in sensitivity to social context: Detecting affective incongruity using the N400. *Social Neuroscience*, *8*(1), 63–74.

Hall, E. T., & Hall, M. R. (1990). *Understanding cultural differences – Germans, French and Americans*. Intercultural Press.

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology*, *54*(6), 917–924.

Hawrysh, B. M., & Zaichkowsky, J. L. (1991). Cultural approaches to negotiations: Understanding the Japanese. *Asia Pacific Journal of Marketing and Logistics*, *3*(1), 40–54.

Holtgraves, T. (1997). Styles of language use: Individual and cultural variability in conversational indirectness. *Journal of Personality and Social Psychology*, *73*(3), 624.

Horstmann, G., & Herwig, A. (2016). Novelty biases attention and gaze in a surprise trial. *Attention, Perception, & Psychophysics*, *78*, 69–77. https://doi.org/10.3758/s13414-015-0995-1

Ishii, K., Kobayashi, Y., & Kitayama, S. (2010). Interdependence modulates the brain response to word-voice incongruity. *Social Cognitive and Affective Neuroscience*, *5*(2–3), 307–317.

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology*, *19*, 1543–1548.

Jaywant, A., & Pell, M. D. (2012). Categorical processing of negative emotions from speech prosody. *Speech Communication*, *54*, 1–10.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice. *Current Biology*, *13*(19), 1709–1714.

Kitayama, S., Duffy, S., & Uchida, Y. (2007). Self as cultural mode of being. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 136–174). Guilford Press.

Kitayama, S., & Howard, S. (1994). Affective regulation of perception and comprehension: Amplification and semantic priming. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influences in perception and attention* (pp. 41–65). Academic Press.

Li, W., & Qian, M. (1995). A validation of the State-Trait Anxiety Inventory in Chinese university students. *Universitatis Pekinensis (Acta Scientiarum Naturalium)*, *31*(1), 108–112.

Liu, P., & Pell, M. D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli. *Behavior Research Methods*, *44*(4), 1042–1051.

Liu, P., Rigoulot, S., & Pell, M. D. (2015a). Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*, *67*(1), 1–13.

Liu, P., Rigoulot, S., & Pell, M. D. (2015b). Cultural differences in on-line sensitivity to emotional voices: Comparing East and West. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00311

Liu, T., Pinheiro, A., Zhao, Z., Nestor, P. G., McCarley, R. W., & Niznikiewicz, M. (2012). Emotional Cues during simultaneous face and voice processing: Electrophysiological insights. *PLoS ONE*, *7*(2), e31001. https://doi.org/10.1371/journal.pone.0031001

LoBue, V. (2009). More than just another face in the crowd: Superior detection of threatening facial expressions in children and adults. *Developmental Science*, *12*(2), 305–313.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565–572.

Mai, X., Ge, Y., Tang, H., Liu, C., & Luo, Y.-J. (2011). Eyes are windows to the Chinese soul: Evidence from the detection of real and fake smiles. *PLoS ONE*, *6*(5), e19903.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*(2), 224–253.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review*, *3*, 215–221.

Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., Ray, R., & Uchida, H. (2002). American-Japanese cultural differences in judgements of emotional expressions of different intensities. *Cognition & Emotion*, *16*(6), 721–747.

Matsumoto, D., Takeuchi, S., Andayani, S., Kouznetsova, N., & Krupp, D. (1998). The contribution of individualism vs. collectivism to cross-national differences in display rules. *Asian Journal of Social Psychology*, *1*(2), 147–165.

Matsumoto, D., Yoo, S. H., & Fontaine, J. (2008). Mapping expressive differences around the world: the relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology*, *39*(1), 55–74.

McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2006). Cultural display rules drive eye gaze during thinking. *Journal of Cross-Cultural Psychology*, *37*(6), 717–722.

McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2008). Gaze display when thinking depends on culture and context. *Journal of Cross-Cultural Psychology*, *39*(6), 716–729.

Mu, Y., Kitayama, S., Han, S., & Gelfand, M. (2015). How culture gets embrained: Cultural differences in event-related potentials of social norm violations. *Proceedings of National Academy of Sciences*, *112*, 15348–15353.

Neuberg, S. L., Kenrick, D. T., & Schaller, M. (2011). Human threat management systems: Self-protection and disease avoidance. *Neuroscience & Biobehavioral Reviews*, *35*(4), 1042–1051.

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience*, *24*(5), 1205–1223.

Noppeney, U., Ostwald, D., & Werner, S. (2010). Perceptual decisions formed by accumulation of audiovisual evidence in prefrontal cortex. *Journal of Neuroscience*, *30*(21), 7434–7446.

Nummenmaa, L., Hyönä, J., & Calvo, M. G. (2006). Eye movement assessment of selective attentional capture by emotional pictures. *Emotion*, *6*, 257–268.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*(3), 381–396.

Oyserman, D., Coon, H. M., & Kemmelmeier, M. (2002). Rethinking individualism and collectivism: Evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin*, *128*(1), 3–72.

Park, G., Lewis, R. S., Wang, Y. C., Cho, H. J., & Goto, S. G. (2018). Are you mad at me? Social anxiety and early visual processing of anger and gaze among Asian American biculturals. *Culture and Brain*, *6*(2), 151–170.

Paulmann, S., & Pell, M. D. (2010). Contextual influences of emotional speech prosody on face processing: How much is enough? *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 230–241.

Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, *35*(2), 192–201.

Paulmann, S., Titone, D., & Pell, M. D. (2012). How emotional prosody guides your way: Evidence from eye movements. *Speech Communication*, *54*, 92–107.

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition and Emotion*, *28*(2), 230–244. https://doi.org/10.1080/02699931.2013.812033

Pell, M. D. (2005). Nonverbal emotion priming: Evidence from the 'Facial Affect Decision Task'. *Journal of Nonverbal Behavior*, *29*(1), 45–73.

Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition & Emotion*, *25*(5), 834–853.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120.

Pell, M. D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., & Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, *111*, 14–25. https://doi.org/10.1016/j.biopsycho.2015.08.008

Pell, M. D., & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, *50*(6), 519–530.

Peschard, V., Maurage, P., & Philippot, P. (2014). Towards a cross-modal perspective of emotional perception in social anxiety: Review and future directions. *Frontiers in Human Neuroscience*, *8*, 322.

Pourtois, G., Grandjean, D., Sander, D., & Vuilleumier, P. (2004). Electrophysiological correlates of rapid spatial orienting towards fearful faces. *Cerebral Cortex*, *14*, 619–633.

R Development Core Team. (2010). *R: A language and environment for statistical computing* (Version 2.13.1). R Foundation for Statistical Computing. http://www.R-project.org

Rigoulot, S., & Pell, M. D. (2012). Seeing emotion with your ears: Emotional prosody implicitly guides visual attention to faces. *PLoS ONE*, *7*(1), e30740. https://doi.org/10.1371/journal.pone.0030740

Rigoulot, S., & Pell, M. D. (2014). Emotion in the voice influences the way we scan emotional faces. *Speech Communication*, *65*, 36–49.

Schwartz, R., & Pell, M. D. (2012). Emotional speech processing at the intersection of prosody and semantics. *PLoS ONE*, *7*(10), e47279.

Scollon, R., & Scollon, S. W. (1995). *Intercultural communication: A discourse approach*. Blackwell.

Spielberger, C. D., Gorsuch, R. L., Lushene, P. R., Vagg, P. R., & Jacobs, A. G. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press, Inc.

Tan, C., Sheppard, E., & Stephen, I. (2015). A change in strategy: Static emotion recognition in Malaysian Chinese. *Cogent Psychology*, *2*, 1085941. https://doi.org/10.1080/23311908.2015.1085941

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I feel your voice: Cultural differences in the multisensory perception of emotion. *Psychological Science*, *21*(9), 1259–1262.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective and Behavioral Neuroscience*, *1*(4), 382–387.

Vuilleumier, P., Armony, J. L., Driver, J., & Dolan, R. J. (2001). Effects of attention and emotion on face processing in the human brain: An event-related fMRI study. *Neuron*, *30*, 829–841.

Yeung, H. H., Denison, S., & Johnson, S. P. (2016). Infants' looking to surprising events: When eye-tracking reveals more than looking time. *PLoS ONE*, *11*(12), e0164277. https://doi.org/10.1371/journal.pone.0164277

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology*, *43*(2), 303–311.