JSLHR

Research Article

Unisensory and Multisensory Stroop Effects Modulate Gender Differences in Verbal and Nonverbal Emotion Perception

Yi Lin,[a] Hongwei Ding,[a] and Yang Zhang[b]

Purpose: This study aimed to examine the Stroop effects of verbal and nonverbal cues and their relative impacts on gender differences in unisensory and multisensory emotion perception.
Method: Experiment 1 investigated how well 88 normal Chinese adults (43 women and 45 men) could identify emotions conveyed through face, prosody and semantics as three independent channels. Experiments 2 and 3 further explored gender differences during multisensory integration of emotion through a cross-channel (prosody-semantics) and a cross-modal (face-prosody-semantics) Stroop task, respectively, in which 78 participants (41 women and 37 men) were asked to selectively attend to one of the two or three communication channels.
Results: The integration of accuracy and reaction time data indicated that paralinguistic cues (i.e., face and prosody) of emotions were consistently more salient than linguistic ones (i.e., semantics) throughout the study. Additionally, women demonstrated advantages in processing all three types of emotional signals in the unisensory task, but only preserved their strengths in paralinguistic processing and showed greater Stroop effects of nonverbal cues on verbal ones during multisensory perception.
Conclusions: These findings demonstrate clear gender differences in verbal and nonverbal emotion perception that are modulated by sensory channels, which have important theoretical and practical implications.
Supplemental Material: https://doi.org/10.23641/asha.16435599

Verbal (e.g., semantic content) and nonverbal (e.g., facial expressions, prosody) cues of emotion play an important role in conveying the feelings and intentions in speech communication. They have been shown to have nonequal salience in emotion cognition, which is referred to as the sensory dominance effects (Colavita, 1974). The degree of salience among different informational channels constitutes a crucial part of the classic Stroop effect (Melara & Algom, 2003; Stroop, 1935). Empirical evidence also suggests that the ability in emotion cognition is not equally represented in males and females (see Thompson & Voyer, 2014, for a review). It remains unclear how the two factors of sensory channel and gender can be associated in influencing emotion perception in unisensory and multisensory situations. This study aimed to examine the Stroop effects of verbal and nonverbal cues and their relative impacts on gender differences in emotion perception in uni-channel, cross-channel, and cross-modal settings wherein auditory and visual signals are referred to as two sensory modalities and face, prosody, and semantics as three communication channels.

Stroop Effects in Emotion Perception

Stroop effects have a long history in the study of human cognition, dating back to the mid-1930s when Stroop (1935) first reported his findings. A typical Stroop setup involves two tasks with the use of color words and word colors as two perceptually congruent or incongruent dimensions. The baseline task tests perception in only one dimension by controlling the other, in which the participant is asked to read the color words printed in black or name the colors of squares. The orthogonal task manipulates the color words printed in congruent (e.g., the word "red" printed in red ink) or conflicting (e.g., "red" printed in blue, green, or purple) colors, in which the participant is asked to recognize one dimension and ignore the other. Findings

[a]Speech-Language-Hearing Center, School of Foreign Languages, Shanghai Jiao Tong University, China
[b]Department of Speech-Language-Hearing Sciences & Center for Neurobehavioral Development, University of Minnesota, Minneapolis
Correspondence to Hongwei Ding: hwding@sjtu.edu.cn; Yang Zhang: zhanglab@umn.edu

Editor-in-Chief: Frederick (Erick) Gallun
Editor: Alexander L. Francis
Received June 13, 2020
Revision received December 8, 2020
Accepted May 21, 2021
https://doi.org/10.1044/2021_JSLHR-20-00338

Disclosure: The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

of the classic Stroop studies can be characterized as two effects. First, task effects show the asymmetry of the two dimensions, indicating that color naming is more difficult than word reading. Second, congruity effects highlight the influences of semantic correspondence between the two dimensions, showing that congruent stimuli produce faster and/or more accurate responses than incongruent ones. Additional research on this topic further demonstrates that tasks containing more congruent trials exert larger interferences than those with more incongruent trials, which is referred to as the proportion congruent effect (Bugg & Crump, 2012). The general conclusion is that perception is driven by both dimensional imbalance and relatedness, which modulates the presence, magnitude, and direction of Stroop effects (Melara & Algom, 2003; Tzelgov et al., 1992).

Stroop effects on different communication dimensions have been a focal point of research in the context of emotion perception. While a wealth of studies converge on the congruence-induced facilitation effect in emotion processing (Barnhart et al., 2018; Lin et al., 2020; McGurk & MacDonald, 1976; Pell, 2005; Schirmer et al., 2005; Schwartz & Pell, 2012), there have been mixed findings in the literature concerning the sensory dominance of communication channels. Some researchers found a processing bias toward linguistic information (Kitayama & Ishii, 2002; Pell et al., 2011), whereas others claimed the predominance of paralinguistic signals, including visual (Lin & Ding, 2019; Spence et al., 2011) and auditory prosodic (Ben-David et al., 2016; Filippi et al., 2017; Kim & Sumner, 2017; Mehrabian & Wiener, 1967; Schirmer & Kotz, 2003) cues. These findings align with studies using nonaffective stimuli (Green & Barber, 1981), suggesting that the magnitude of Stroop effects varies across verbal and nonverbal contents. It is likely that the sensory dominance effect is meditated by participants' language and cultural backgrounds. In emotional speech processing, while a linguistic dominance effect is more often found in participants from the western culture background (e.g., Germany, the Netherlands, and the United States), paralinguistic cues tend to take precedence over linguistic ones in East Asian settings (e.g., Japan and Philippines; Ishii et al., 2003; Kitayama & Ishii, 2002; Kotz & Paulmann, 2007; Paulmann & Kotz, 2008; Pell et al., 2011). In addition to linguistic/cultural contexts, variations in performance including a reverse pattern may be observed due to the differences in experimental stimuli and measurement indices across studies (Ben-David et al., 2016).

Of particular note is the existence of perceptual asymmetries in emotion perception, whose interpretation may be complicated by several experimental design issues in previous research. First, different sensory modalities have not received an equal amount of attention in previous studies. As most research on emotion perception focuses on visual facial displays and written words, the perception of auditory prosodic cues is much understudied (Citron, 2012; Ethofer et al., 2006; Hawk et al., 2009; Lausen & Schacht, 2018; Vasconcelos et al., 2017). Second, a common experimental design is that linguistic and nonlinguistic information has largely been separated. For instance, many studies on spoken word recognition have been conducted with little consideration for the role of paralinguistic messages (Nygaard & Queen, 2008). These works generally assume that nonlinguistic information serves as a source of noise that may disrupt the retrieval of abstract linguistic units (Shankweiler et al., 1977). On the other hand, studies on voice emotion perception have often employed nonlinguistic vocalizations (e.g., interjections, laughter, cries, sighs, groans, screams) to avoid the potential bias of semantic content (Belin et al., 2008; Collignon et al., 2008; Juslin & Laukka, 2001; Koeda et al., 2013; Vasconcelos et al., 2017), thereby underrepresenting the complexity of the everyday running speech. Third, even though research efforts have delved into bi-channel/modal interactions (Ethofer et al., 2006; Garrido-Vasquez et al., 2018; Schwartz & Pell, 2012), the dynamic interplay among more than two channels has scarcely been addressed. In this regard, two recent investigations (Filippi et al., 2017; Lin et al., 2020) have examined sensory dominance by involving three communication channels (i.e., semantics, prosody, and face), which provided evidence for the salient role of nonverbal signals in cross-channel and cross-modal emotion tasks. However, emotional signals from different channels are simultaneously presented in these studies on multisensory processing with manipulation of selective attention. While traditional color–word Stroop studies involve baseline tasks testing perception in a single dimension, few emotion Stroop studies include unisensory tasks for comparison. It remains unclear whether paralinguistic signals dominate over linguistic ones due to bottom-up influences (i.e., the intrinsic physical properties of the stimuli), or top-down modulation (i.e., the requirements of the task, the facilitation of available cues in other channels, and the involvement of other higher-level cognitive skills such as attention and integration; Latinus et al., 2016). To disentangle the roles of the bottom-up and top-down psychological mechanisms in perceptual biases, it is of great significance to examine how emotions are perceived separately in single channels and integrated in multichannel situations.

Gender Differences in Emotion Perception

Gender has been repeatedly proposed to influence emotion perception performance (Fischer et al., 2018; Hall, 1978; Hall & Matsumoto, 2004; Kret & de Gelder, 2012; Lambrecht et al., 2014; Proverbio et al., 2006; Schirmer et al., 2002; Thompson & Voyer, 2014; Zupan et al., 2016). Women are often believed to be intuitively more sensitive to emotional information, but empirical findings do not show a consistent pattern of female superiority. While some researchers did not observe a significant effect of gender (Hawk et al., 2009; Martin & Altarriba, 2017; Paulmann et al., 2008; Sauter et al., 2013), there is mounting evidence that women are more competent in emotion perception than men (Belin et al., 2008; Collignon et al., 2010; Donges et al., 2012; Hall, 1978; Hall & Matsumoto, 2004; Hampson et al., 2006; Hoffmann et al., 2010; Koeda et al., 2013; Kret & de Gelder, 2012; Lausen & Schacht, 2018; Li et al., 2008; Menezes et al., 2017; Scherer & Scherer, 2011; Vasconcelos et al., 2017; Zupan et al., 2016). But the magnitude of gender differences remains unclear. Hall (1978) showed a moderate effect size of female advantages, while a recent meta-analysis conducted by Thompson and Voyer (2014) indicated that female advantages are by and large small (d = .19).

Some studies further delved into the differences between men and women using a Stroop paradigm. While the presence and extent of gender differences in traditional color–word Stroop effects are far from conclusive (Baroun & Alansari, 2006), studies on emotion perception tend to converge on the view that women show greater Stroop interference effects of nonverbal cues on verbal ones compared to men. For instance, Schirmer and Kotz (2003) provided electrophysiological evidence showing women's greater difficulty in preventing interferences by prosody in semantic processing in a prosody–word interference task, which is in agreement with their earlier priming study indicating an earlier use of prosodic information during word processing in women (Schirmer et al., 2002). Similarly, later neuroimaging studies using Stroop-like paradigms also suggested that women are more susceptible to influences from emotional prosody in word processing and may need less conscious inferences in emotional prosody processing (Imaizumi et al., 2004; Schirmer et al., 2004).

One noteworthy point is that most of the Stroop studies on gender differences in emotion perception were conducted with participants speaking a nontonal language and relying on an alphabetic writing system (e.g., German). It remains to be tested to what extent these gender differences in emotion perception can be generalized across languages and culture (Alansari, 2004; Alqarni & Dewaele, 2018; Chentsova-Dutton & Tsai, 2007; Hawk et al., 2009; Koeda et al., 2013; Sauter et al., 2010; Schirmer & Kotz, 2006; Schirmer et al., 2005). Emotion perception performances can also vary due to the effects of other participant-related factors (Thompson & Voyer, 2014), such as age (e.g., children vs. young adults vs. old adults; Lausen & Schacht, 2018; Paulmann et al., 2008; Sauter et al., 2013; Thompson & Voyer, 2014), health status (e.g., healthy vs. clinical [e.g., schizophrenia and dementia]; Horley et al., 2010; Scholten et al., 2008), and hearing status (Husain et al., 2014; Most & Michaelis, 2012; Picou et al., 2018). In addition, findings may diverge due to manipulations in task paradigms (e.g., unisensory vs. multisensory; Collignon et al., 2010; Hawk et al., 2009), target communication channels (e.g., facial vs. prosodic vs. semantic; Hall & Matsumoto, 2004; Hawk et al., 2009; Schirmer et al., 2005), and the number and types of emotional categories involved (Belin et al., 2008; Collignon et al., 2010; Deng et al., 2016; Kret & de Gelder, 2012; Schirmer & Kotz, 2003; Schirmer et al., 2002).

The Current Study

This study aimed to investigate the gender differences in Stroop effects of verbal and nonverbal emotion perception in a Mandarin Chinese context. We followed up our recent study (Lin et al., 2020) with a focus on happiness and sadness as our target emotion categories. These two represent a universal, uncontroversial, and highly distinctive pair of emotions that are more accurate and stable in recognition from an early age when compared with other basic emotions (e.g., fear, disgust, and surprise). The basic emotion categories of happiness and sadness play a crucial role in social interactions throughout developmental stages (Ekman & Cordaro, 2011; Flom et al., 2008; Garcia & Tully, 2020; Lawrence et al., 2015). Emotional stimuli were presented from three communication channels in two modalities: visual cues included facial expressions, and auditory signals were spoken words with variations in emotional prosody including neutral tone of voice.

We extended our previous study by increasing the sample size for each experiment by systematically testing unisensory and multisensory emotion perception in a unidimensional identification task and a two-dimensional and three-dimensional (3D) Stroop task. Specifically, Experiment 1 compared how men and women processed emotional face, emotional prosody, and semantics as three independent channels. Experiments 2 and 3 required participants to selectively attend to one of the concurrently presented channels in a cross-channel (prosody–semantics) and a cross-modal (face–prosody–semantics) emotion identification task, respectively. Based on previous studies showing paralinguistic dominance effects in East Asian contexts and female advantages in emotion processing, we predicted that during unisensory and multisensory emotion perception, paralinguistic signals of emotions would be more perceptually salient than linguistic ones in the Chinese participants (Hypothesis 1). In addition, women would demonstrate greater sensitivity to the emotional signals, and be more susceptible to the influences of nonverbal signals on verbal ones in multisensory emotion perception (Hypothesis 2). It is possible that the two hypotheses are only partially supported or even not supported in the three experimental conditions featuring unisensory and multisensory processing of emotion, which provides an opportunity to examine under what conditions/contexts the hypotheses are not supported. The findings would contribute to our understanding of how theoretical accounts of Stroop effects of verbal (semantic) and nonverbal (facial and prosodic) channels and gender differences converge in shaping uni- and multisensory emotion perception. The results can also provide insights for further efforts in exploring the underlying brain mechanisms and developing appropriate diagnosis and intervention methods that may benefit patients with hearing impairments and gender-related psychiatric disorders.

Experiment 1

Method

Participants

The study was approved by the institutional review board (IRB) at Shanghai Jiao Tong University (SJTU). Eighty-eight volunteers (43 women and 45 men[1]) were

[1]The biological sex of the participants was recorded and entered for gender in data analysis for all three experiments.

recruited to take part in this unisensory emotion identification test through an online campus advertisement on the SJTU website. Women were on average 23.8 years old (SD = 2.4), and men had a mean age of 23.6 years (SD = 2.9). Men and women were balanced in terms of formal school education (M_{women} = 17.3, SD = 2.4; M_{men} = 17.1, SD = 2.9). All participants were native speakers of Mandarin Chinese, and were studying at SJTU as undergraduate or graduate students at the time of testing. All were nonmusicians, who had received no formal musical training within the last 5 years and less than 2 years of musical training prior to that (Wu et al., 2015). All had normal or corrected-to-normal vision and normal hearing as verified by a standard audiometric assessment for pure tones from 0.25–8 kHz (≤ 20 dB HL; Koerner & Zhang, 2018), and were without any medical history of speech, language, and hearing disorders, or neurological problems. All participants gave their written informed consent prior to inclusion in the experiment, and were financially compensated for their time.

Stimuli

The stimuli in the unisensory emotion identification test consisted of emotions conveyed in three different communication channels, namely facial expressions, prosody, and semantic content. Each stimulus expressed one of the three basic emotions (i.e., happiness, sadness, and anger; Ekman, 1992) or neutrality. The target emotions of this experiment were "happiness" and "sadness," which were consistently investigated in Experiments 2 and 3. Stimuli of "anger" and "neutrality" were inserted as emotional and nonemotional distractors respectively to make the judgment process more challenging and lower the effect of chance-level binary selection, and did not enter statistical evaluation of the experimental conditions.

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 categories of expressions and between actors and actresses, yielding eight emotional stimuli per category for each gender in each channel. All stimuli were presented to the participants once. An earlier norming study was conducted to perceptually validate the experimental stimuli. Twenty-three normal adult native speakers of Mandarin Chinese (11 women and 12 men) who did not participate in the current research were invited to judge the reliability of the materials. All included target stimuli received over 90% identification accuracy for emotional categories, and an average rating > 3 on a 7-point Likert scale (0 = *not intense*, 6 = *very intense*) for emotion intensity within each emotional category. Only words with an average rating of > 3 for familiarity on a 7-point Likert scale (0 = *not familiar*, 6 = *very familiar*) were included as the prosodic and semantic stimuli in the auditory modality.

Specifically, *the facial stimuli* came from the Chinese Affective Picture System (Bai et al., 2005), a well-established database with a standardized set of black-and-white photographs of Chinese actors and actresses posing emotional or neutral facial expressions. The facial expressions of happiness, sadness, anger, and neutrality produced by eight actors and eight actresses were selected based on the identification accuracy of emotional category and ratings of emotional intensity in the norming study.

The prosodic and semantic stimuli included words uttered by four Chinese professional speakers (two actors and two actresses) in a quiet laboratory setting, and digitized into a MacBook Pro computer with AVID Mbox Mini at a sampling rate of 44.100 kHz with a 16-bit resolution. A word familiarity test was administered in the norming study to ensure that all the selected words were common to native speakers of Mandarin Chinese. Each word was portrayed 3 times by the four speakers, and the best ones were selected according to the results of the norming study. For *the prosodic stimuli*, the speakers enunciated semantically neutral disyllabic concrete words (e.g., 风扇, "fan" in Chinese; 面包, "bread" 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., 喜悦, "delighted" in Chinese), sad (e.g., 沮丧, "upset" in Chinese), angry (e.g. 愤怒, "indignant" in Chinese) emotions, and neutrality (e.g., 普通, "common" in Chinese) in a neutral prosody. See Supplemental Tables S1 and S2 for the list of included words for semantic and prosodic stimuli, respectively.

Procedure

The experiment was carried out in a sound booth with the participant seated in a comfortable chair at around 70 cm from an LCD monitor. E-Prime (Version 2.0.8.22; Psychology Software Tools, 2012) was used as a stimulus presentation program. 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.

This unisensory emotion identification test comprised three subtests, namely face emotion identification task, voice emotion identification task, and emotional semantics identification task. There were altogether 192 trials of three separate blocks, with each block containing 64 trials of emotional stimuli presented in one of the facial, prosodic, or semantic channels. Each block started with a familiarization phase in which there was a practice test with eight trials. Participants needed to reach 100% accuracy in the practice test before entering the test session. Each trial began with a fixation cross in the center of the screen for 1,100 milliseconds (ms). Then emotional stimuli were presented, at the onset of which participants were required to identify which emotion was being portrayed by the actor or actress in a four-choice task. They responded by pressing one of the four emotion-coded keys on a keyboard ("v" for happy, "b" for sad, "n" for angry, and "m" for neutral) as quickly as possible. Accuracy and response time were measured from stimulus onset. After the response, a blank screen was displayed for 1,000 ms before the next trial began. The presentation order of the blocks was fully randomized across participants, and the trial order in each block was pseudorandomized so that none of the two consecutive trials presented emotion of the same category. Participants were allowed to have a short break for up to 30 s after running every 32 trials.

Statistical Analyses

We applied a series of generalized linear mixed-effects models in R (Version 3.6.1) with the lme4 package (Bates et al., 2015) for data analysis. We excluded the filler trials so that only trials of the target happy and sad emotions were included in the analyses. Accuracy and reaction time data were entered as dependent variables in the mixed-effects models respectively. A binomial distribution with a logit link function was applied for the analysis of accuracy data, and a gamma distribution with a log-link function was employed for the analysis of reaction time data (Lo & Andrews, 2015). Between-subject variable participants' gender (two conditions: female and male) and within-subject variable identification task (three conditions: facial, prosodic, and semantic) were entered as categorical fixed factors, in which female participants and the facial channel were set as the baseline level, respectively. When a pairwise comparison between the prosodic and semantic channels was conducted, prosody was used as the baseline level. Individual viewer (listener) participants and test items were entered as random factors for intercepts. Tukey's post hoc tests in the emmeans package (Lenth, 2020) were implemented for pairwise comparison in case of a significant main effect or interaction. The full models for accuracy and reaction time analyses in Experiment 1 are specified in formulas (1) and (2) in Supplemental Material S1.

Results

The table in Appendix presents a summary of the significant effects in all three experiments. The detailed results of the generalized linear mixed-effects models for accuracy and reaction time for target emotions (happiness and sadness) in Experiment 1 are summarized in Supplemental Tables S3 and S4.[2]

Accuracy

Overall, participants identified the emotional stimuli with substantially high accuracy (M = 93.9%, SD = 10.7%). Figure 1(a) illustrates the mean proportion correct in three emotion tasks by men and women.

Generalized linear mixed-effects analyses showed significant main effects of gender, $\chi^2(1) = 6.37$, $p = .012$, Cohen's $w = .27$, and task, $\chi^2(2) = 74.81$, $p < .001$, $w = .92$. Compared with men, women achieved 1.2% ± 0.8% higher

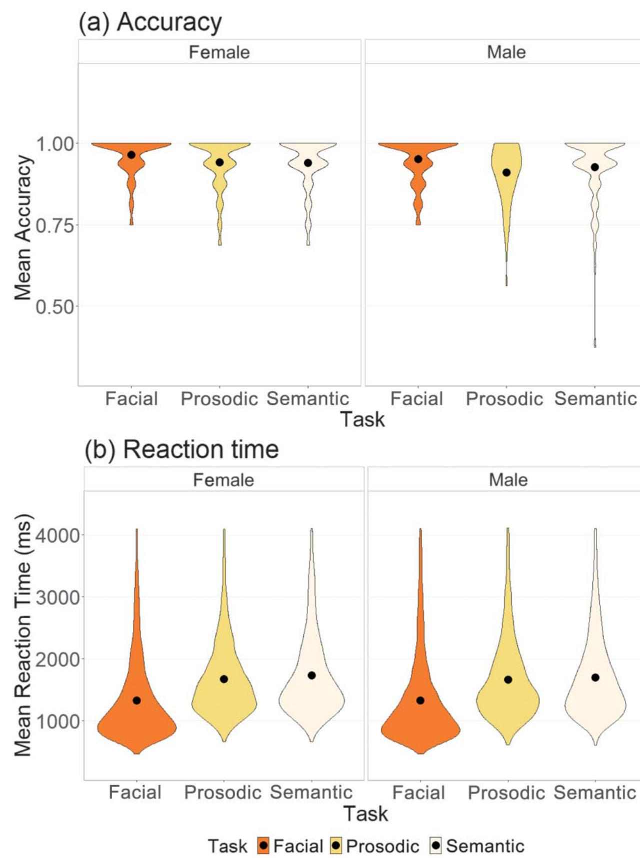


Figure 1. Mean (a) accuracy and (b) reaction time in the three emotion identification tasks by men and women in Experiment 1. Mean accuracy and reaction time are displayed in the violin plots with data distribution shape indicated by the density plots, and mean values represented by the black dots.

identification accuracy regardless of task type ($\hat{\beta}_1 = .41$, $SE = .16$, $z = 2.58$, $p = .010$, Cohen's $d = .28$). Women tended to outperform men in all three channels with the greatest gender difference manifested in the prosodic task ($\hat{\beta}_3 = .56$, $SE = .20$, $z = 2.83$, $p = .005$, $d = .30$), though no significant interaction between task and gender was found, $\chi^2(2) = 1.81$, $p = .404$, $w = .14$. When collapsed across gender, the facial task elicited 3.3% ± 0.8% more accurate responses than the prosodic one ($\hat{\beta}_2 = 1.05$, $SE = .15$, $z = 7.13$, $p < .001$, $d = .76$), and 2.6% ± 0.8% more than the semantic one ($\hat{\beta}_2 = 1.17$, $SE = .15$, $z = 7.63$, $p < .001$, $d =$ .81). No significant difference was found between the prosodic and semantic tasks ($p = .638$).

Reaction Time

When analyzing reaction time data, we excluded incorrect responses (3.9%) and responses over 2 SDs from the mean (3.6%; Baayen & Milin, 2010; Chien et al., 2017). Figure 1(b) illustrates the mean reaction time in three emotion tasks by men and women.

[2]Filler trials (anger and neutrality) were originally included in the mixed-effects analyses, which did not change the gender differences reported here in meaningful ways. See Supplemental Table S5 for summary of the generalized linear mixed-effects models with both target (happiness and sadness) and filler (anger and neutrality) trials in Experiment 1.

Generalized linear mixed-effects analyses revealed a significant main effect of task, $\chi^2(2) = 1160.19$, $p < .001$, $w = 3.63$. Compared with the facial task, response time was increased by 440.0 ± 20.6 ms in the prosodic task ($\hat{\beta}_2 = -.30$, $SE = .01$, $z = -25.4$, $p < .001$, $d = -2.71$), and by 575.0 ± 20.6 ms in the semantic task ($\hat{\beta}_2 = -.39$, $SE = .01$, $z = -32.83$, $p < .001$, $d = -3.50$). There was also a significant increase by 135.0 ± 22.7 ms in the semantic task relative to the prosodic task ($\hat{\beta}_2 = -.09$, $SE = .01$, $z = -6.79$, $p < .001$, $d = -.72$). There was no main effect of gender, $\chi^2(1) = .01$, $p = .935$, $w = .01$. Neither of the gender produced significant interaction with the factor task, $\chi^2(2) = 1.05$, $p = .592$, $w = .11$.

The integration of accuracy and reaction time data demonstrated that while face served as the most dominant channel in unisensory emotion perception, prosody had greater perceptual salience than semantics. In addition, women showed greater advantages in decoding emotions from all three communication channels.

Experiment 2

In Experiment 1, emotional information was presented in isolation without reciprocal interactions that could happen in natural communication situations. This raises the question of to what extent the observed gender differences can be generalized to more complex and authentic interaction settings, in which gender effects might be confounded by higher-level cognitive demands. We thus tested whether the reported gender differences in the unisensory experiment could be preserved during multisensory integration of emotions by implementing a cross-channel (prosody–semantics) emotion perception task.

Method

Participants

A separate group of seventy-eight native Chinese students at SJTU (41 women and 37 men) was recruited for the multisensory Stroop tests (Experiment 2 and 3). We informed the volunteers that the experiment contained two phases, and participants would be reinvited 2 weeks later. Women had a mean age of 22.8 years ($SD = 2.5$), and men were on average 23.2 years old ($SD = 3.5$). Men and women were balanced in terms of formal school education ($M_{\text{women}} = 16.2$, $SD = 2.3$; $M_{\text{men}} = 16.7$, $SD = 3.5$). Subject inclusion/exclusion criteria were the same as Experiment 1. All reported no history of speech, language, or hearing disorders, and neurological impairments, and had normal or corrected-to-normal vision and normal hearing as verified by the audiometric screening as implemented in Experiment 1. They completed written consent forms as required and approved by the IRB.

Stimuli

In this cross-channel test, emotional information was expressed auditorily and concurrently through two communication channels, namely prosody and semantic content. The stimuli included 32 disyllabic spoken words in Mandarin Chinese, each of which conveyed either happiness or sadness in emotional prosody and semantics. Half of these auditory stimuli were congruent in emotional prosody and semantics, and the other half conveyed incongruent emotions in the two channels, yielding eight happy semantics with happy prosody, eight happy semantics with sad prosody, eight sad semantics with happy prosody, and eight sad semantics with sad prosody. Each spoken word was portrayed 3 times by the six speakers (three women and three men). As in Experiment 1, the best exemplars were selected based on the results of a norming study to ensure that words were relatively familiarized by the native speakers, and received high identification accuracy of emotional categories and similar rating of emotional intensity. A brief description of stimuli in the two channels is provided below. See Supplemental Table S6 for the spoken words used in Experiments 2 and 3.

(a) Prosodic channel. The 32 words were randomly chosen to be enunciated with either a happy or sad prosody by six professional speakers (three actors and three actresses). There was an equal number of words uttered by female and male actors (16 for each gender). The lexical tone combination of the disyllabic words was matched between the happy and sad prosodic sets so as to avoid the potential confound of lexical tones on emotion expression.

(b) Semantic channel. The 32 spoken words expressed a happy or sad emotion in semantic content. Sixteen words were synonyms of "happy," and the other 16 were synonyms of "sad," most of which were selected from Chinese Affective Words System (Wang et al., 2008). The number of spoken words was balanced between happy and sad emotional semantics (16 for each emotion), and between female and male actors (16 for each gender).

Procedure

The settings of this cross-channel Stroop test were similar to Experiment 1, and we followed the exact same procedures used in our previous study on prosody–semantics emotion perception (Lin et al., 2020). The test required the participants to manipulate their selective attention to either the emotional prosody or semantics according to the auditory instruction given at the beginning of each trial. They were asked to identify the emotion conveyed through tone of voice while ignoring the semantic meaning of the emotional words when the auditory instruction "emotional prosody" preceded the trial (prosodic task). When the instruction "emotional semantics" was provided (semantic task), participants needed to identify the emotion indicated by word meaning while ignoring emotional prosody.

After a familiarization phase with four practice trials consecutively reaching 100% accuracy, the test phase started. There were altogether 64 trials: each of the 32 auditory items with an equal number of congruent and incongruent pairs (16 each) was presented twice. Each trial began with an auditory instruction of either "emotional prosody" or "emotional semantics" for 1,100 ms as an instructional prompt. Then emotional prosody and semantics were binaurally and simultaneously presented over the headphones. Meanwhile on the screen, participants were asked to identify the emotion conveyed in the attended channel as quickly as possible by pressing the corresponding emotion-coded keys on a keyboard ("f" for happy and "j" for sad). The positioning of happy and sad response buttons was counterbalanced across participants. Accuracy and response time were measured from stimulus onset. After each response, a blank screen was displayed for 1,000 ms before the next trial began. The presentation order of the test trials was pseudorandomized so that none of the three consecutive trials presented emotion through the same channel. In this case, it is unlikely for the participants to anticipate the target information to attend to in the upcoming trial. There was a short break of up to 30 s after completing the first 32 trials.

Statistical Analyses

As in Experiment 1, a series of generalized linear mixed-effects models with accuracy and reaction time as separate dependent variables were then implemented. A binomial distribution with a logit link function was applied for the analysis of accuracy data, and a gamma distribution with a log-link function was employed for the analysis of reaction time data (Lo & Andrews, 2015). The categorical fixed factors included between-subject variable gender (two conditions: female and male), and within-subject variables task (two conditions: prosodic and semantic) and congruency (two conditions: congruent and incongruent), in which female participants, the prosodic task, and congruent condition were used as the baseline level, respectively. The random factors were listener participants and test items for intercepts. In case of a significant main effect or a significant interaction effect, Tukey's post hoc tests were performed, using the emmeans package (Lenth, 2020). The full models for accuracy and reaction time analyses in Experiment 2 are specified in formulas (3) and (4) in Supplemental Material S1.

Results

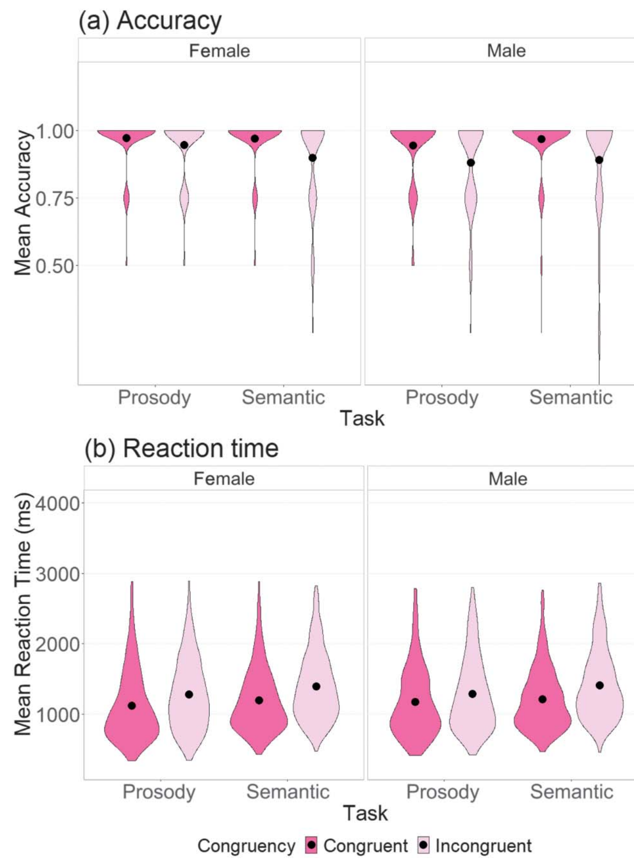
The detailed results of the generalized linear mixed-effects models for accuracy and reaction time in Experiment 2 are summarized in Supplemental Tables S7 and S8.

Accuracy

Overall, participants identified the emotional stimuli with substantially high accuracy (M = 93.4%, SD = 14.6%). Figure 2(a) illustrates the mean proportion correct in the prosody–semantics Stroop task by men and women.

Generalized linear mixed-effects analyses showed no main effect of task, $\chi^2(1) = .03$, $p = .857$, $w = .02$, but a significant main effect of gender, $\chi^2(1) = 4.50$, $p = .034$, $w = .24$, and a significant interaction between gender and task, $\chi^2(1) = 10.09$, $p = .001$, $w = .36$. Within the prosodic task,

Figure 2. Mean (a) accuracy and (b) reaction time in the prosody–semantics emotion identification tasks by men and women in Experiment 2. Mean accuracy and reaction time are displayed in the violin plots with data distribution shape indicated by the density plots, and mean values represented by the black dots.



women made emotional judgments with 5.1% ± 1.4% higher accuracy than men ($\hat{\beta}_4 = .80$, $SE = .22$, $z = 3.60$, $p < .001$, $d = .41$). Within the semantic task, there was no significant difference between participants of the two genders ($p = .849$). The Stroop interference effect of emotional prosody on semantics was significant for women ($\hat{\beta}_4 = .47$, $SE = .22$, $z = 2.12$, $p = .034$, $d = .24$) but not for men ($\hat{\beta}_4 = -.29$, $SE = .21$, $z = -1.43$, $p = .152$, $d = -.16$). A main effect of congruence was also found, $\chi^2(1) = 30.17$, $p < .001$, $w = .62$. Congruent stimuli elicited 6.0% ± 0.7% higher accuracy than the incongruent ones ($\hat{\beta}_3 = 1.09$, $SE = .18$, $z = 6.02$, $p < .001$, $d = .68$). There was no significant interaction between gender and congruence, $\chi^2(1) = .18$, $p = .672$, $w = .05$, between task and congruence, $\chi^2(1) = 1.71$, $p = .191$, $w = .15$, and among gender, task, and congruence, $\chi^2(1) = .03$, $p = .873$, $w = .02$.

Reaction Time

When analyzing reaction time data, we excluded incorrect responses (5.0%) and responses over 2 SDs from the mean (3.5%; Baayen & Milin, 2010; Chien et al., 2017). Figure 2(b) illustrates the mean reaction time in the prosody–semantics Stroop task by men and women.

Generalized linear mixed-effects analyses showed a significant main effect of task, $\chi^2(1) = 6.28$, $p = .012$, $w = .28$, and congruence, $\chi^2(1) = 27.05$, $p < .001$, $w = .59$. Participants responded at 92.6 ± 29.5 ms faster to the prosodic task than to the semantic task, $\hat{\beta}_2 = -.08$, $SE = .03$, $z = -2.60$, $p = .009$, $d = -.29$, and 169.0 ± 20.5 ms faster to the congruent stimuli than to the incongruent ones, $\hat{\beta}_3 = -.14$, $SE = .02$, $z = -6.27$, $p < .001$, $d = -.71$. The analyses revealed no main effect of gender, and no significant interaction between any of the two fixed factors, and among the three (all p values > .05).

In this cross-channel emotion Stroop experiment, participants showed better performances when selectively attending to prosody compared with semantics, and also when encountering stimuli with congruent information compared with incongruent ones. Additionally, female participants perceived emotional signals more successfully than males when attention was given to prosody irrespective of congruence conditions.

Experiment 3

In the present experiment, we were interested in determining whether participants' performances would continue to be influenced by attentional allocation when involved in cognitively more demanding cross-modal (face–prosody–semantics) emotion Stroop tasks. We explored whether gender differences in the Stroop effects of emotional prosody on semantics would still be present when visual facial expressions, a more perceptually salient channel/modality, were simultaneously presented. We involved the same panel of participants as in Experiment 2 to investigate whether or not and under what circumstances men's and women's performances in prosodic and semantic identification might be facilitated or interfered by emotional faces.

Method

Participants

As stated in Experiment 2, the same group of 78 students (41 women and 37 men) took part in the multisensory emotion identification tests (Experiments 2 and 3). For each participant, the second phase was administered at least 2 weeks apart from the first one to undermine carryover effects. Participants completed IRB-approved consent forms, and were paid for their participation in Experiments 2 and 3.

Stimuli

Happy and sad emotions were simultaneously expressed in three communication channels through two sensory modalities, which constituted the auditory (i.e., prosodic and semantic) and visual (i.e., facial) stimuli in this cross-modal test. The prosodic and semantic stimuli included the same set of 32 spoken words used in Experiment 2. The visual facial stimuli contained 32 black-and-white photographs of static facial expressions of emotion chosen from the Chinese Affective Picture System (Bai et al., 2005). Similar to the auditory stimuli, these facial expressions were subject to a norming study and adopted in our cross-modal emotion perception research (Lin et al., 2020), and also identical to the set of happy and sad faces employed in the block of face emotion identification in Experiment 1. The number of faces was balanced between happy and sad emotions, and between male and female actors, thus yielding eight faces per emotion category for each gender.

For each auditory stimulus, four emotional faces (two happy and two sad) portrayed by four actors of the same sex were paired in a trial. Among the three channels, two were congruent in the emotional category while the remaining one was incongruent against the other two, yielding three congruence conditions: prosody–semantic content congruent (facial incongruent), prosody–face congruent (semantic incongruent), and face–semantic content congruent (prosodic incongruent). In a fourth condition (cross-channel congruent condition), a congruent emotional category was expressed through three channels simultaneously, which served as a cross-congruent control for comparison.

Procedure

Like Experiment 2, we adopted the same protocol from our previous study (Lin et al., 2020) on face–prosody–semantics emotion perception for Experiment 3. The test required the participants to selectively attend to one of the three simultaneously presented channels (i.e., facial, prosodic, and semantic) based on the auditory prompt provided at the beginning of each trial. When the instruction "emotional face" was given (facial task), participants needed to identify the emotion displayed by the facial expression while ignoring the emotions conveyed through the auditory modality (i.e., prosody and semantics). When the instruction changed to "emotional prosody" (prosodic task), participants needed to concentrate on the emotion conveyed through tone of voice while ignoring information in the facial and semantic channels. When the instruction was to attend to "emotional semantics" (semantic task), they were told to identify the emotional category of word meaning while ignoring the information in the other two paralinguistic communication channels.

Participants entered the experiment after familiarizing themselves with the task requirements by running eight practice trials with 100% accuracy. The test session consisted of 96 trials: 32 pairs of audiovisual items were presented 3 times. There were a total of 32 trials targeting facial expression, 32 targeting prosody, and 32 targeting semantic content. In each target channel, there was an equal number of trials for the four congruence conditions (eight trials each). Each trial began with an auditory instruction prompt indicating the target channel to attend to. A facial expression was displayed in the center of the screen, together with emotional prosody and semantics delivered over headphones.

The onset of auditory and visual emotional cues was kept identical in each trial, from which accuracy and reaction time were measured. Participants were asked to identify the emotion conveyed in the attended channel as quickly as possible by pressing the corresponding emotion-coded keys on a computer keyboard ("f" for happy and "j" for sad). The positioning for "happy" and "sad" response buttons was counterbalanced across participants. After responses were made, a blank screen was displayed for 1,000 ms before the next trial began. As in Experiments 1 and 2, pseudorandomization of the test trials was implemented to ensure that none of the two consecutive trials presented emotion through the same channel. Thus, participants are unlikely to be able to make a strategic prediction of the target channel in the upcoming trials. Participants were allowed to take a short break of up to 30 s after every 32 trials.

Statistical Analyses

As in Experiments 1 and 2, accuracy and reaction time were entered as dependent variables respectively in generalized linear mixed-effects models. A binomial distribution with a logit link function was applied for the analysis of accuracy data, and a gamma distribution with a log-link function was employed for the analysis of reaction time data (Lo & Andrews, 2015). The categorical fixed factors included between-subject variable gender (two conditions: female and male), and within-subject variables task (three conditions: facial, prosodic, and semantic) and congruency (four conditions: facial incongruent, prosodic incongruent, semantic incongruent, and cross-channel congruent), in which female participants, the prosodic task, and cross-channel congruent condition were used as the baseline level, respectively. When conducting a pairwise comparison between prosodic and semantic tasks, prosody was set as the baseline level. The random factors were viewer (listener) participants and test items for intercepts. In case of a significant main effect or a significant interaction effect, Tukey's post hoc tests were performed, using the emmeans package (Lenth, 2020). The full models for accuracy and reaction time analyses in Experiment 3 are specified in formulas (3) and (4) in Supplemental Material S1.

In addition, we conducted within-participant analyses to compare performance between Experiments 2 and 3. A series of paired-sample t tests were performed to examine how accuracy and reaction time of men and women were affected by the inclusion of emotional faces under the congruent and incongruent conditions in the prosodic and semantic tasks.

Results

Generalized Linear Mixed-Effects Analyses

The detailed results of the generalized linear mixed-effects models for accuracy and reaction time in Experiment 3 are summarized in Supplemental Tables S9 and S10.

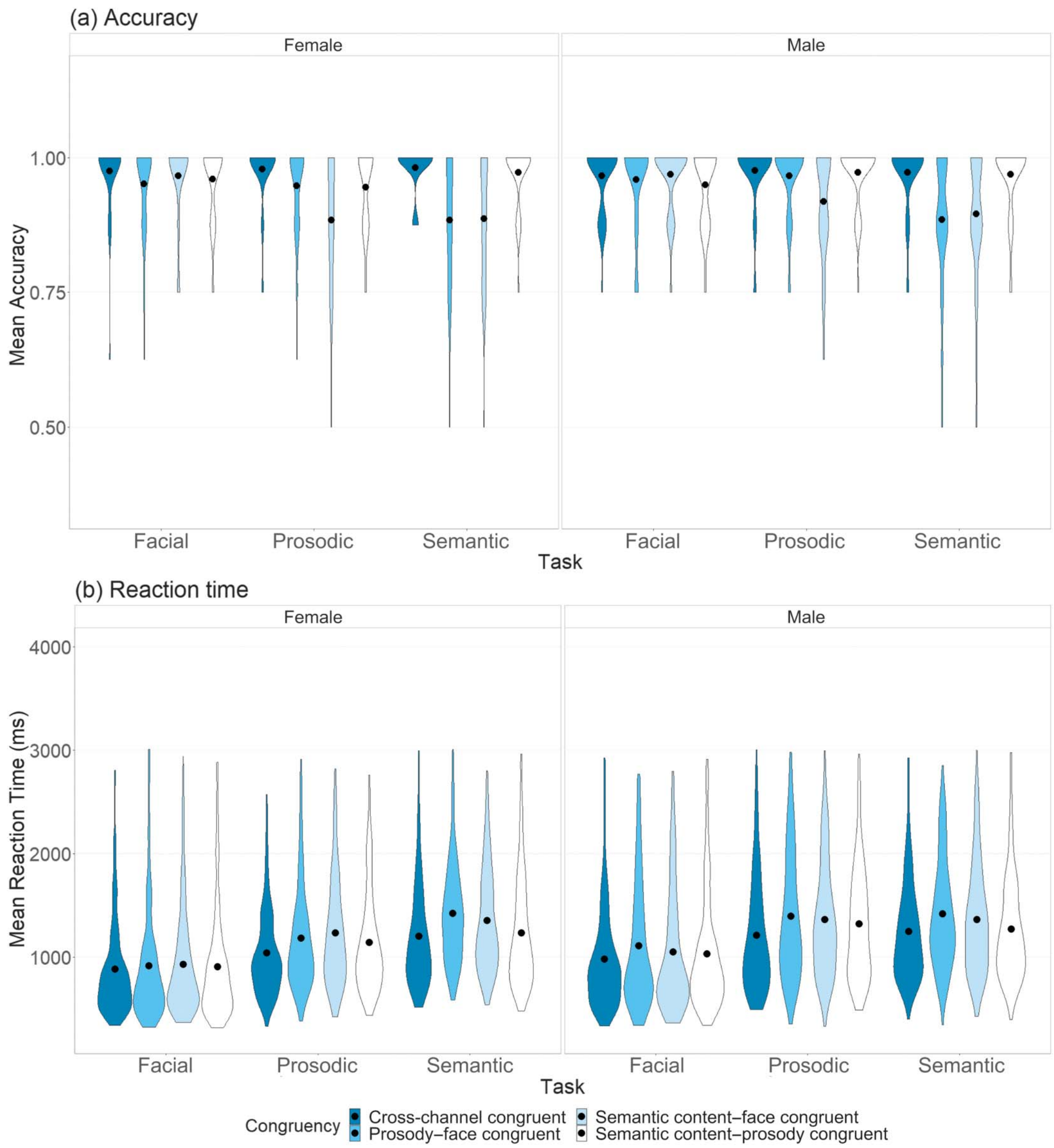
(a) Accuracy. Overall, participants identified the emotional stimuli with substantially high accuracy (M = 94.7%, SD = 9.1%). Figure 3(a) illustrates the mean proportion correct in the face–prosody–semantics Stroop task by men and women.

Generalized linear mixed-effects models suggested main effects of task, $\chi^2(2) = 7.25$, $p = .027$, $w = .30$, and congruence, $\chi^2(3) = 27.07$, $p < .001$, $w = .59$, and a significant interaction between task and congruence, $\chi^2(6) = 21.82$, $p = .001$, $w = .53$. To parse out the interaction effect, we compared the cross-channel congruent control condition with the other three conditions in which two communication channels were congruent while the remaining one was incongruent in each task. For the facial expression–oriented task, there was no significant difference in accuracy between the cross-channel congruent condition and any of the three incongruent conditions ($p > .05$). For the prosody-oriented task, the prosodic incongruent (semantic–facial congruent) condition was 7.7% ± 1.3% less accurately identified than the cross-channel congruent condition ($\hat{\beta}_5 = 1.61$, $SE = .39$, $z = 4.16$, $p < .001$, $d = .47$). The semantic incongruent (prosodic–facial congruent) condition ($p = .360$) and the facial incongruent (semantic–prosodic congruent) condition ($p = .559$) did not differ significantly from the baseline condition. For the semantics–oriented task, the semantic incongruent condition showed the poorest performance (9.3% ± 1.3% lower than the control condition, $\hat{\beta}_5 = 1.82$, $SE = 0.38$, $z = 4.75$, $p < .001$, $d = .54$), and the prosodic incongruent condition also showed a significant effect (8.7% ± 1.3% lower than the control condition, $\hat{\beta}_5 = 1.75$, $SE = .39$, $z = 4.55$, $p < .001$, $d = .52$). But facial incongruency (semantic–prosodic congruent) did not have such a significant influence on emotional identification ($p = .911$). The analyses revealed no main effect of gender, and no significant interaction between gender and task, between gender and congruence, and among the three fixed factors. The analyses revealed no main effect of gender, $\chi^2(1) = .50$, $p = .480$, $w = .08$, and no significant interaction between gender and task, $\chi^2(2) = 4.07$, $p = .131$, $w = .23$, between gender and congruence, $\chi^2(3) = 2.03$, $p = .566$, $w = .16$, and among the three fixed factors, $\chi^2(6) = 1.40$, $p = .966$, $w = .13$.

(b) Reaction time. When analyzing reaction time data, we excluded incorrect responses (4.8%) and responses over 2 SDs from the mean (3.8%; Baayen & Milin, 2010; Chien et al., 2017). Figure 3(b) illustrates the mean reaction time in the face–prosody–semantics Stroop task by men and women.

Generalized linear mixed-effects models revealed no main effect of gender, $\chi^2(1) = 2.66$, $p = .103$, $w = .18$, but main effects of task, $\chi^2(2) = 77.78$, $p < .001$, $w = 1.00$, and congruence, $\chi^2(3) = 21.72$, $p < .001$, $w = .53$, and a significant interaction between gender and task, $\chi^2(2) = 41.45$, $p < .001$, $w = .73$. No significant interaction between task and congruence, $\chi^2(6) = 4.49$, $p = .611$, $w = .24$, between gender and congruence, $\chi^2(3) = 2.73$, $p = .435$, $w = .19$, and among the three fixed factors were found, $\chi^2(6) = 6.22$, $p = .398$, $w = .28$. When attending to emotional faces, participants responded 263.0 ± 26.3 ms faster than prosodic tasks ($\hat{\beta}_2 = -.26$, $SE = .03$, $z = -9.14$, $p < .001$, $d = -1.03$), and 346.0 ± 26.4 ms faster than semantic ones ($\hat{\beta}_2 = -.33$, $SE = .03$,

Figure 3. Mean (a) accuracy and (b) reaction time in the face–prosody–semantics emotion identification tasks by men and women in Experiment 3. Mean accuracy and reaction time are displayed in the violin plots with data distribution shape indicated by the density plots, and mean values represented by the black dots.



$z = -11.54$, $p < .001$, $d = -1.31$). Prosodic tasks also triggered 83.0 ± 26.4 ms faster responses than semantic ones ($\hat{\beta}_2 = -.06$, $SE = .03$, $z = -2.41$, $p = .042$, $d = -.27$). Pairwise comparison of the incongruent conditions and the baseline level showed that participants responded 155.4 ± 25.0 ms slower in the semantic incongruent condition ($\hat{\beta}_3 = -.13$, $SE = .02$, $z = -4.61$, $p < .001$, $d = -.52$), and 129.9 ± 25.0 ms slower in the prosodic incongruent condition ($\hat{\beta}_3 = -.11$, $SE = .02$, $z = -3.88$, $p = .001$, $d = -.44$), but the facial incongruent condition did not have such a significant influence ($p = .261$). To parse out the interaction between gender and task, we compared the response time by men and women within each of the three selectively attended channels. Compared with male participants, women responded 143.1 ± 65.9 ms faster in the facial expression–oriented task ($\hat{\beta}_4 = -.15$, $SE = .06$, $z = -2.31$, $p = .021$, $d = -.26$) and 181.5 ± 66.0 ms faster in the prosodic-oriented task ($\hat{\beta}_4 = -.14$, $SE = .06$, $z = -2.08$, $p = .037$, $d = -.24$). For women, there were significant interference effects of face on prosody ($\hat{\beta}_4 = -.27$, $SE = .03$, $z = -10.39$, $p < .001$, $d = -1.18$), of face on semantics ($\hat{\beta}_4 = -.39$, $SE = .03$, $z = -15.06$, $p < .001$, $d = -1.71$), and of prosody on semantics ($\hat{\beta}_4 = -.12$, $SE = .03$, $z = -4.66$, $p < .001$, $d = -.53$). For men, there were significant interference effects of face on prosody ($\hat{\beta}_4 = -.26$,

$SE = .03$, $z = -9.67$, $p < .001$, $d = -1.09$) and of face on semantics ($\hat{\beta}_4 = -.27$, $SE = .03$, $z = -10.01$, $p < .001$, $d = -1.13$), but no significant interference effect was found of prosody on semantics ($\hat{\beta}_4 = -.01$, $SE = .03$, $z = -.37$, $p = .926$, $d = -.04$). No significant difference between the two genders was found in the semantics-oriented task ($p = .717$).

Within-Participant Analyses

The summary of within-participant analyses for accuracy and reaction time between Experiments 2 and 3 is presented in Supplemental Table S11.

(a) Accuracy. Mean accuracy in Experiments 2 and 3 was calculated by subject under each congruence condition in the prosodic and semantic tasks respectively. Paired-sample *t* tests revealed that when the two auditory channels presented incongruent information, women showed significantly reduced accuracy in identifying emotional prosody with the included faces biased toward semantics (i.e., the prosodic incongruent condition), $t(40) = 2.23$, $p = .027$, $d = .36$. Male participants, however, showed significantly increased accuracy in prosodic identification when the included faces were congruent with prosody (i.e., the semantic incongruent condition), $t(36) = -3.31$, $p = .002$, $d = -.56$. No significant changes in accuracy were founded with the inclusion of faces in other paired contrasts (all *p* values > .05).

(b) Reaction time. For reaction time analyses, we excluded the trials with an inaccurate response in either of Experiments 2 and 3. Mean response time in both experiments was computed by subject under each congruence condition in the prosodic and semantic tasks respectively. Paired-sample *t* tests indicated that women spent more time in identifying emotional semantics that was congruent with prosody, when the included facial expressions were incongruent against the two auditory channels (i.e., the facial incongruent condition), $t(40) = -2.06$, $p = .046$, $d = -.33$. No significant changes in response time were founded with the inclusion of faces in other paired contrasts (all *p* values > .05).

This 3D cross-modal emotion Stroop experiment demonstrated that prosody continued to dominate over semantics when face, a more perceptually salient channel, was included. While both men and women made better identification performances in the congruent than incongruent conditions, women tended to be more sensitive to nonverbal emotional signals and show greater Stroop effects of nonverbal cues on verbal ones compared with men.

General Discussion

The current study investigated the relative perceptual salience of linguistic and paralinguistic cues in unisensory and multisensory emotion perception among men and women. We compared how men and women perceived emotional face, prosody, and semantics as three isolated channels in Experiment 1, and further examined gender differences in multisensory integration of emotion through a cross-channel (prosody–semantics) and a cross-modal (face–prosody–semantics) Stroop task in Experiments 2 and 3, respectively. We tested the paralinguistic dominance effect in unisensory and multisensory emotion perception (Hypothesis 1), and women's superiority over men in emotion perception with extra nonverbal sensitivity in multisensory emotion word processing (Hypothesis 2). The integration of accuracy and reaction time data in the three experiments has revealed that paralinguistic signals were more perceptually salient than linguistic cues in all emotion identification tasks, which has bolstered our first hypothesis. Our second hypothesis has also been supported: Though women demonstrated better performances in identifying both linguistic and paralinguistic cues of emotion in unisensory tasks, they preserved their advantages in processing paralinguistic cues with greater Stroop effects in multisensory tasks either in the accuracy or the reaction time measure. In other words, the paralinguistic dominance effect was found to be more pronounced in women than men in both unisensory and multisensory emotion perception.

Perceptual Asymmetries of Verbal and Nonverbal Cues in Emotion Cognition

We observed that paralinguistic signals of emotion (i.e., face and prosody) were consistently more salient than linguistic (i.e., semantic) cues in unisensory and multisensory experiments. Specifically, when processed separately as a unisensory signal of emotion (Experiment 1), facial expressions elicited the most accurate and fastest emotion identification performances among the three channels. More rapid responses were also triggered when participants were asked to recognize emotional prosody compared to word meaning. In the cross-channel and cross-modal Stroop tasks (Experiments 2 and 3), participants made faster responses when selectively attending to emotional face and prosody compared with semantics. Additionally, face showed the least interference in accuracy with the other two channels, and prosody demonstrated fewer interferences from semantics than the other way around. As predicted in Hypothesis 1, these results confirmed the communicative advantages of nonverbal signals (Beall & Herbert, 2008; Ben-David et al., 2016; Brazo et al., 2014; Filippi et al., 2017; Kim & Sumner, 2017; Kitayama & Ishii, 2002; Lin et al., 2018; Schirmer & Kotz, 2003), and replicated the Stroop effects of paralinguistic emotional cues on linguistic ones in our preceding work on multisensory integration with a larger sample size (i.e., Lin & Ding, 2019; Lin et al., 2020). While participants from a western society (e.g., Germany, the Netherlands, and the United States) tend to demonstrate a linguistic advantage in emotional speech perception (Ishii et al., 2003; Kitayama & Ishii, 2002; Kotz & Paulmann, 2007; Paulmann & Kotz, 2008; Pell et al., 2011), our study aligns with previous studies conducted among Asian participants (e.g., Japanese and Filipinos) by revealing a reverse paralinguistic advantage among Mandarin-speaking Chinese (Ishii et al., 2003; Kitayama & Ishii, 2002). The different profile observed in western and East Asian countries corroborates the idea that the perceptual asymmetries of verbal and nonverbal cues in emotion perception are shaped by perceivers' language experience and cultural backgrounds.

The perceptual asymmetries of verbal and nonverbal channels have often been interpreted by two different types of mechanisms (Latinus et al., 2016): top-down influences that involve voluntary, strategic, and supervisory cognitive processes with intentional control of attention based on task requirements; and bottom-up influences that refer to automatic and effortless processing driven by stimulus properties (Bugg & Crump, 2012). Top-down modulation with manipulation of selective attention may take some effects in our second and third Stroop experiments, in which nonverbal channels (i.e., prosody and face) are more perceptually salient when attended. However, if one sensory channel captures attention automatically, especially when stimuli in that channel are far more salient than those of the other channel, the processing of that channel would show more perceptual advantages even without directed attention (Andersen et al., 2004; Vroomen et al., 2001). In our cross-channel Stroop task, emotional prosody tended to interfere with semantic processing to a larger extent (with larger effect sizes) in incongruent conditions than congruent ones. In the cross-modal experiment, an interaction effect of task and congruence was found with greater interferences of prosody in accurate semantic processing. While semantic content alone was not perceptually salient enough to inhibit the processing of prosodic tasks, incongruent information conveyed through prosody could bias the responses when it was the unattended channel. Notably, the paralinguistic dominance effects not only occur during multisensory integration of emotion where selective attention is manipulated, but also hold true in unisensory emotion identification tasks where task instructions have no effect. Thus, it is likely that the automatic processing of nonverbal emotional signals would reduce the processing of verbal content when selective attention was required for emotional semantics in the cross-channel/modal Stroop tasks.

Compared with accounts based on top-down attentional mechanisms, bottom-up influences appear to better accommodate the findings in our study. Paralinguistic signals hold more communicative advantages in both unisensory and multisensory processing possibly due to intrinsic stimulus properties that provide more reliable information for the efficient realization of the tasks (Andersen et al., 2004; Latinus et al., 2016). In this study, the visual modality tends to be more reliable than the auditory one, since facial features could be more easily and readily extracted from standardized static images whereas emotions were communicated through dynamic speech streams in prosody and semantics. A line of studies also suggests that visual cues are highly informative and more discriminable than auditory ones (Biehl et al., 1997; Elfenbein & Ambady, 2002; Hawk et al., 2009; Scherer et al., 1991; Yuval-Greenberg & Deouell, 2009), and can affect vocal emotion processing at an early point in time, as revealed by early auditory event-related potentials (e.g., N100 and P200; Garrido-Vasquez et al., 2018; Pourtois et al., 2002). With regard to the predominance of prosody over semantics, it might be the case that prosodic judgments could be made based on available acoustic signals from the stimulus onset, while identifying the verbally expressed emotions would require waiting until the second syllable is heard in our tasks (Lin et al., 2020; Schirmer & Kotz, 2003). It is also plausible that listeners might differentiate word meanings by referring to variations in a multitude of prosodic cues (e.g., pitch, stress, duration; Nygaard et al., 2009; Nygaard & Lunders, 2002). As demonstrated in electrophysiological literature, emotional prosody contributes to the linguistic interpretation of an utterance similar to semantics, which is reflected by a similar sensitivity of the N400 component to prosodic and semantic contexts (Schirmer et al., 2002). Furthermore, such prosodic dominance is conceivable as the semantic processing of a word is more sensitive to and more easily influenced by task-irrelevant information than emotional prosody (Schirmer & Kotz, 2003).

Gender Differences in Stroop Effects of Verbal and Nonverbal Emotional Cues

Another intriguing finding of this study is that the paralinguistic dominance effects are more pronounced in women than men. As predicted in Hypothesis 2 and widely confirmed in the extant literature, women outperformed men in unisensory and multisensory emotion identification tasks, especially for paralinguistic signals (Belin et al., 2008; Collignon et al., 2010; Demenescu et al., 2015; Donges et al., 2012; Hall, 1978; Kret & de Gelder, 2012; Lambrecht et al., 2014; Lausen & Schacht, 2018; Menezes et al., 2017; Scherer et al., 2001; Scherer & Scherer, 2011; Thayer & Johnsen, 2000; Zupan et al., 2016). One possible account is that men and women differ in the time course of emotion perception (Canli et al., 2002). For instance, women tend to make a faster use of facial and prosodic cues during emotion perception (Donges et al., 2012; Schirmer et al., 2002, 2004). From a socio-psychological perspective, women play a primary role as caretakers and are more strongly engaged in interpersonal interactions than men, which enables them to better detect subtle cues of emotion in regard to child-rearing and socialization (Fischer, 2000; Fischer et al., 2018; Waaramaa, 2017). The beliefs about a gendered division of social roles build on and drive a biological-essentialist view of gender differences (Saguy et al., 2021). According to this view, men and women are biologically different in various domains (e.g., brain structures, sex-related hormones; Kret & de Gelder, 2012), though the existence of these differences has not been fully established (Joel et al., 2015). Thus, to what extent these sociological and biological differences contribute to functional differences in emotion processing among men and women warrants further examination.

We also observed that women had higher accuracy in identifying emotional semantics, albeit only in the unisensory perception task, which is consistent with previous research showing female advantages in higher order semantic analysis (Baxter et al., 2003; Wirth et al., 2006). However, such advantages did not prevail in both multisensory tasks, which means that gender differences in sensory dominance effects can be mitigated by task design and attentional focus (Palmer et al., 2013; Schirmer et al., 2005; Shen & Itti, 2012). While participants only needed to focus on one sensory channel in Experiment 1, they were asked to selectively attend to one of the co-occurring interactive channels in Experiments 2 and 3, which might be one of the potential accounts for the gaps between our findings in unichannel and cross-channel/modal tasks. The inclusion of selective attention in multisensory Stroop tasks might facilitate women's perception of nonverbally expressed emotions to a greater extent than emotional semantics. Compared with men, women may implicitly allocate more attention to the expression of faces, and their semantic processing may be more susceptible to influences from prosody (Fulton et al., 2015; Herlitz & Lovén, 2013; Schirmer & Kotz, 2003; Schirmer et al., 2004). In this case, women might encounter greater difficulty to prevent influences from face and prosody when identifying the emotional meaning of a word in the multisensory Stroop tests. That women did not maintain their strengths in perceiving emotional semantics as face and prosody can also be stimulus-driven: The nature of linguistic and paralinguistic processing tends to be different. Word meaning identification generally relies on objective and global semantic judgments that happen at a late cognitive processing stage (Hoffmann et al., 2010; Schmid et al., 2011). By contrast, paralinguistic emotional processing favors women to a greater extent as they are more competent in perceiving the subtle changes in facial and prosodic features at an early perceptual stage (Fischer et al., 2018; Hoffmann et al., 2010).

Interestingly, emotional prosody interfered with the processing of semantic information, but this effect was only significant for women and produced a larger effect size in Experiment 3 than Experiment 2. This does not concord with the proportion congruent effect, which suggests that experimental paradigms with more congruent trials produce larger interference effects than paradigms with more incongruent trials (Bugg & Crump, 2012). One plausible explanation is that our two Stroop protocols differed not only in the proportion of congruent trials (50% congruent trials in Experiment 2 and 25% congruent trials in Experiment 3), but also in the number of communication channels, which changed dimensional uncertainty and imbalance at the same time (Melara & Algom, 2003). Though there appear to be robust congruence facilitation effects (i.e., congruent stimuli producing more accurate and rapid performances than incongruent ones) in the two Stroop experiments (Barnhart et al., 2018; McGurk & MacDonald, 1976; Pell, 2005; Schirmer et al., 2005; Schwartz & Pell, 2012), no significant interaction was found between gender and congruence. Thus, the observed gender differences may not depend much on congruence conditions, but instead more on channel dominance. Our within-participant analyses between Experiments 2 and 3 further showed that women's sensitivity in face emotion perception tended to inhibit their identification of prosody and semantics, while the inclusion of visual facial expressions appeared to be facilitatory in multisensory integration for male participants. This finding coincides with previous affective priming studies (Burton et al., 2004; Gohier et al., 2011), and implies that women are more likely to suffer perceptual loss whereas men tend to achieve multisensory facilitation in cross-channel/modal processing. The fact that women are more pervious to the interferences from nonverbal affective signals may underlie their susceptibility to emotional disorders such as depression (Gohier et al., 2011).

In our study, the observed gender differences were likely to be masked by the perceptual asymmetries among channels, thus producing smaller effect sizes than sensory dominance (Thompson & Voyer, 2014). There were also some discrepancies between the results of accuracy and reaction time among the three experiments. In Experiments 1 and 2, the gender effects were observed in the accuracy measure, whereas in Experiment 3, gender effects showed up only in response time. One potential account could be that different panels of participants and different numbers of emotions were involved in the unisensory and multisensory tasks. With regard to the two multisensory experiments, though we tried to mitigate the carryover effects by running Experiment 3 at least 2 weeks after Experiment 2, there remains the possibility that participants achieved better performances as they had familiarized themselves with the experimental stimuli and procedures, which is indicated by a slight increase in overall accuracy (from 93.4% in Experiment 2 to 94.6% in Experiment 3). As the identification performances were toward the ceiling and possibly influenced by the practice effect, response time might be a more sensitive measurement to indicate the subtle differences between the two genders in this case (Thompson & Voyer, 2014). It should be noted that accuracy and reaction time were not conceptually equivalent indices in our study as we excluded the trials with incorrect responses when analyzing reaction time data, which underlies the necessity to integrate them for a more comprehensive interpretation of the results. Future studies can examine the generalizability of the current findings with a more refined experimental design and integration of behavioral and neurological (e.g., event-related potentials with higher-temporal resolution) indices.

Implications, Limitations, and Future Research

This study investigated how sensory dominance effects are represented among men and women in unisensory and multisensory emotion processing. Theoretically, our findings suggest that a cohesive framework associating uni- and multisensory processing can be established to elucidate the interactions between Stroop effects of verbal and nonverbal channels and gender differences in emotion perception (Schreuder et al., 2016). This behavioral exploration lays a foundation for further unraveling the temporal dynamics and the corresponding cortical and subcortical substrates related to different levels of emotion cognition in future electrophysiological and neuroimaging research. As we focused on participants speaking a lexical tonal language that relies on a logographic writing system, this study can inspire further efforts to reveal cross-linguistic differences. In addition, our basic research investigating mutual influences of language and emotion processing skills could motivate new studies and intervention approaches for promoting both linguistic and socio-emotional functioning in clinical populations such as individuals with autism, developmental language disorder and hearing impairments who reportedly have deficits in emotion perception and multisensory integration (Bahn et al., 2021; Courtright & Courtright, 1983; Jiam et al., 2017; Most & Michaelis, 2012; St Clair et al., 2019). Furthermore, albeit gender similarity in selective attention, we found some evidence for small and yet significant gender differences in emotional processing in the different channels. Our results demonstrating a persevered advantage of women in perceiving paralinguistic signals during unisensory and multisensory processing can also have implications for investigating the psychopathological characteristics and social functioning abilities of patients with neurological and psychiatric conditions. Some mental disorders characterized by severe dysfunctions in emotion processing are found to be gender-related. For instance, schizophrenia and autism spectrum disorders affect men more than women, whereas depression reportedly disrupts women more often than men (Cahill, 2006; Feldman et al., 2018; Gohier et al., 2011; Lin et al., 2018). A larger proportion of men with traumatic brain injury is also reported to be impaired at vocal but not facial emotion recognition compared to their female counterparts (Zupan et al., 2016), whereas female advantages are absent in facial processing but preserved in prosody perception among schizophrenics (McBain et al., 2010; Scholten et al., 2008). It deserves to be explored to what extent gender differences in sensory dominance effects at different stages of emotion perception relate to the incidences and symptomatology of these diseases, which contributes to developing more efficient diagnosis and prevention techniques, and effective intervention and rehabilitation programs in clinical practice.

Some limitations need to be considered during the interpretation of the findings of this study. First, potential group differences in our study could not be excluded as two panels of participants were respectively recruited for unisensory and multisensory emotion tasks. However, given the sample size in our design, the gender differences in paralinguistic emotion processing are evident and generalizable across tasks despite potential individual variability in different measures. Second, we only focused on happy and sad emotions, disyllabic spoken words, static facial expressions, and native Mandarin Chinese university students as participants throughout the study. It remains to be tested whether the findings can be extended to other categories of emotion (e.g., angry, fearful), types of stimuli (e.g., words in text, spoken or written sentences, dynamic facial expressions), language and cultural settings (e.g., a nontonal language in a low-context culture), and population groups (e.g., young vs. old, healthy vs. clinical; Ben-David et al., 2019; Deng et al., 2016; Dupuis & Pichora-Fuller, 2010; Koeda et al., 2013; Thompson & Voyer, 2014; Vasconcelos et al., 2017). Third, findings might also be limited as we restricted our purview on the distinctions resulting from listeners/viewers' gender. Future researchers could explore whether speakers/actors' gender can generate different emotional responses (Thompson & Voyer, 2014; Vasconcelos et al., 2017), which is an interesting test bed to examine to what extent gender differences occur in the encoding and decoding processes for emotional speech.

Conclusions

The current research examined the perception of facial, prosodic, and semantic emotional signals among men and women with a unisensory emotion identification task, a cross-channel auditory-alone task, and a cross-modal audiovisual emotion Stroop task. Results indicated that paralinguistic signals, including face and prosody, consistently gain more perceptual salience than linguistic messages (i.e., semantic content) in unisensory and multisensory emotion perception. Though women are better able to identify signals in all three channels during unisensory emotion perception, they only preserve their advantages in paralinguistic emotion processing and are more susceptible to the influences of nonverbal channels on verbal ones during multisensory integration. Taken together, emotion cognition is modulated in accuracy or reaction time measures by the interplay between sensory dominance and gender differences in different contexts, which provides the basis for further examining the neural substrates of emotion perception as a dynamic complex and the clinical ramifications for individuals with affective processing deficits.

Acknowledgments

H. Ding and Y. Zhang were supported by the Major Program of National Social Science Foundation of China (No. 18ZDA293). Y. Zhang received additional support from University of Minnesota's Grand Challenges Exploratory Research Grant.

References

Alansari, B. (2004). Gender and cultural performance differences on the Stroop color and word test. *Social Behavior and Personality, 32*(3), 235–245. https://doi.org/10.2224/sbp.2004.32.3.233

Alqarni, N., & Dewaele, J.-M. (2018). A bilingual emotional advantage? An investigation into the effects of psychological factors in emotion perception in Arabic and in English of Arabic-English bilinguals and Arabic/English monolinguals. *International Journal of Bilingualism, 24*(2), 141–158. https://doi.org/10.1177/1367006918813597

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research, 21*(3), 301–308. https://doi.org/10.1016/j.cogbrainres.2004.06.004

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Bahn, D., Vesker, M., Schwarzer, G., & Kauschke, C. (2021). A multimodal comparison of emotion categorization abilities in children with developmental language disorder. *Journal of Speech, Language, and Hearing Research, 64*(3), 993–1007. https://doi.org/10.1044/2020_JSLHR-20-00413

Bai, L., Ma, H., Huang, Y. X., & Luo, Y. J. (2005). The development of native Chinese affective picture system: A pretest in 46 college students. *Chinese Mental Health Journal, 19*(11), 719–722.
Barnhart, W. R., Rivera, S., & Robinson, C. W. (2018). Different patterns of modality dominance across development. *Acta Psychologica, 182,* 154–165. https://doi.org/10.1016/j.actpsy.2017.11.017
Baroun, K., & Alansari, B. (2006). Gender differences in performance on the Stroop test. *Social Behavior and Personality: An International Journal, 34*(3), 309–318. https://doi.org/10.2224/sbp.2006.34.3.309
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01
Baxter, L. C., Saykin, A. J., Flashman, L. A., Johnson, S. C., Guerin, S. J., Babcock, D. R., & Wishart, H. A. (2003). Sex differences in semantic language processing: A functional MRI study. *Brain and Language, 84*(2), 264–272. https://doi.org/10.1016/S0093-934X(02)00549-7
Beall, P. M., & Herbert, A. M. (2008). The face wins: Stronger automatic processing of affect in facial expressions than words in a modified Stroop task. *Cognition and Emotion, 22*(8), 1613–1642. https://doi.org/10.1080/02699930801940370
Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods, 40*(2), 531–539. https://doi.org/10.3758/BRM.40.2.531
Ben-David, B. M., Gal-Rosenblum, S., van Lieshout, P. H. H. M., & Shakuf, V. (2019). Age-related differences in the perception of emotion in spoken language: The relative roles of prosody and semantics. *Journal of Speech, Language, and Hearing Research, 62*(4S), 1188–1202. https://doi.org/10.1044/2018_JSLHR-H-ASCC7-18-0166
Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. H. M. (2016). Prosody and semantics are separate but not separable channels in the perception of emotional speech: Test for rating of emotions in speech. *Journal of Speech, Language, and Hearing Research, 59*(1), 72–89. https://doi.org/10.1044/2015_JSLHR-H-14-0323
Biehl, M., Matsumoto, D., Ekman, P., Hearn, V., Heider, K., Kudoh, T., & Ton, V. (1997). Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences. *Journal of Nonverbal Behavior, 21*(1), 3–21. https://doi.org/10.1023/A:1024902500935
Bugg, J. M., & Crump, M. J. C. (2012). In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects. *Frontiers in Psychology, 3,* 367. https://doi.org/10.3389/fpsyg.2012.00367
Burton, L. A., Rabin, L., Vardy, S. B., Frohlich, J., Wyatt, G., Dimitri, D., & Guterman, E. (2004). Gender differences in implicit and explicit memory for affective passages. *Brain and Cognition, 54*(3), 218–224. https://doi.org/10.1016/j.bandc.2004.02.011
Cahill, L. (2006). Why sex matters for neuroscience. *Nature Reviews Neuroscience, 7*(6), 477–484. https://doi.org/10.1038/nrn1909
Canli, T., Desmond, J. E., Zhao, Z., & Gabrieli, J. D. E. (2002). Sex differences in the neural basis of emotional memories. *Proceedings of the National Academy of Sciences, 99*(16), 10789–10794. https://doi.org/10.1073/pnas.162356599
Chentsova-Dutton, Y. E., & Tsai, J. L. (2007). Gender differences in emotional response among European Americans and Hmong Americans. *Cognition & Emotion, 21*(1), 162–181. https://doi.org/10.1080/02699930600911333
Chien, Y. F., Sereno, J. A., & Zhang, J. (2017). What's in a word: Observing the contribution of underlying and surface representations. *Language and Speech, 60*(4), 643–657. https://doi.org/10.1177/0023830917690419
Citron, F. M. M. (2012). Neural correlates of written emotion word processing: A review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and Language, 122*(3), 211–226. https://doi.org/10.1016/j.bandl.2011.12.007
Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics, 16*(2), 409–412. https://doi.org/10.3758/BF03203962
Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research, 1242,* 126–135. https://doi.org/10.1016/j.brainres.2008.04.023
Collignon, O., Girard, S., Gosselin, F., Saint-Amour, D., Lepore, F., & Lassonde, M. (2010). Women process multisensory emotion expressions more efficiently than men. *Neuropsychologia, 48*(1), 220–225. https://doi.org/10.1016/j.neuropsychologia.2009.09.007
Courtright, J. A., & Courtright, I. C. (1983). The perception of nonverbal vocal cues of emotional meaning by language-disordered and normal children. *Journal of Speech and Hearing Research, 26*(3), 412–417. https://doi.org/10.1044/jshr.2603.412
Demenescu, L. R., Kato, Y., & Mathiak, K. (2015). Neural processing of emotional prosody across the adult lifespan. *BioMed Research International, 2015,* 1. https://doi.org/10.1155/2015/590216
Deng, Y., Chang, L., Yang, M., Huo, M., & Zhou, R. (2016). Gender differences in emotional response: Inconsistency between experience and expressivity. *PLOS ONE, 11*(6), Article e0158666. https://doi.org/10.1371/journal.pone.0158666
Donges, U.-S., Kersting, A., & Suslow, T. (2012). Women's greater ability to perceive happy facial emotion automatically: Gender differences in affective priming. *PLOS ONE, 7*(7), Article e41745. https://doi.org/10.1371/journal.pone.0041745
Dupuis, K., & Pichora-Fuller, M. K. (2010). Use of affective prosody by young and older adults. *Psychology and Aging, 25*(1), 16–29. https://doi.org/10.1037/a0018777
Ekman, P. (1992). Are there basic emotions? *Psychological Review, 99*(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550
Ekman, P., & Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review, 3*(4), 364–370. https://doi.org/10.1177/1754073911410740
Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203–235. https://doi.org/10.1037/0033-2909.128.2.203
Ethofer, T., Anders, S., Erb, M., Droll, C., Royen, L., Saur, R., Reiterer, S., Grodd, W., & Wildgruber, D. (2006). Impact of voice on emotional judgment of faces: An event-related fMRI study. *Human Brain Mapping, 27*(9), 707–714. https://doi.org/10.1002/hbm.20212
Feldman, J. I., Dunham, K., Cassidy, M., Wallace, M. T., Liu, Y., & Woynaroski, T. G. (2018). Audiovisual multisensory integration in individuals with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 95,* 220–234. https://doi.org/10.1016/j.neubiorev.2018.09.020
Filippi, P., Ocklenburg, S., Bowling, D. L., Heege, L., Güntürkün, O., Newen, A., & de Boer, B. (2017). More than words (and faces): Evidence for a Stroop effect of prosody in emotion word processing. *Cognition and Emotion, 31*(5), 879–891. https://doi.org/10.1080/02699931.2016.1177489
Fischer, A. H. (2000). *Gender and emotion: Social psychological perspectives.* Cambridge University Press. https://doi.org/10.1017/CBO9780511628191
Fischer, A. H., Kret, M. E., & Broekens, J. (2018). Gender differences in emotion perception and self-reported emotional intelligence:

A test of the emotion sensitivity hypothesis. *PLOS ONE, 13*(1), Article e0190712. https://doi.org/10.1371/journal.pone.0190712

Flom, R., Gentile, D. A., & Pick, A. D. (2008). Infants' discrimination of happy and sad music. *Infant Behavior & Development, 31*(4), 716–728. https://doi.org/10.1016/j.infbeh.2008.04.004

Fulton, E. K., Bulluck, M., & Hertzog, C. (2015). Orienting to face expression during encoding improves men's recognition of own gender faces. *Acta Psychologica, 161,* 18–24. https://doi.org/10.1016/j.actpsy.2015.08.005

Garcia, S. E., & Tully, E. C. (2020). Children's recognition of happy, sad, and angry facial expressions across emotive intensities. *Journal of Experimental Child Psychology, 197,* 104881. https://doi.org/10.1016/j.jecp.2020.104881

Garrido-Vasquez, P., Pell, M. D., Paulmann, S., & Kotz, S. A. (2018). Dynamic facial expressions prime the processing of emotional prosody. *Frontiers in Human Neuroscience, 12,* 244. https://doi.org/10.3389/fnhum.2018.00244

Gohier, B., Senior, C., Brittain, P. J., Lounes, N., El-Hage, W., Law, V., Philips, M. L., & Surguladze, S. A. (2011). Gender differences in the sensitivity to negative stimuli: Cross-modal affective priming study. *European Psychiatry, 28*(2), 74–80. https://doi.org/10.1016/j.eurpsy.2011.06.007

Green, E. J., & Barber, P. J. (1981). An auditory Stroop effect with judgments of speaker gender. *Perception & Psychophysics, 30*(5), 459–466. https://doi.org/10.3758/BF03204842

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85*(4), 845–857. https://doi.org/10.1037/0033-2909.85.4.845

Hall, J. A., & Matsumoto, D. (2004). Gender differences in judgments of multiple emotions from facial expressions. *Emotion, 4*(2), 201–206. https://doi.org/10.1037/1528-3542.4.2.201

Hampson, E., van Anders, S. M., & Mullin, L. I. (2006). A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis. *Evolution and Human Behavior, 27*(6), 401–416. https://doi.org/10.1016/j.evolhumbehav.2006.05.002

Hawk, S. T., van Kleef, G. A., Fischer, A. H., & van der Schalk, J. (2009). "Worth a thousand words": Absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293–305. https://doi.org/10.1037/a0015178

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. *Visual Cognition, 21*(9–10), 1306–1336. https://doi.org/10.1080/13506285.2013.823140

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., & Traue, H. C. (2010). Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. *Acta Psychologica, 135*(3), 278–283. https://doi.org/10.1016/j.actpsy.2010.07.012

Horley, K., Reid, A., & Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *Journal of Speech, Language, and Hearing Research, 53*(5), 1132–1146. https://doi.org/10.1044/1092-4388(2010/09-0030)

Husain, F. T., Carpenter-Thompson, J. R., & Schmidt, S. A. (2014). The effect of mild-to-moderate hearing loss on auditory and emotion processing networks. *Frontiers in Systems Neuroscience, 8,* 10. https://doi.org/10.3389/fnsys.2014.00010

Imaizumi, S., Honma, M., Ozawa, Y., Maruishi, M., & Muranaka, H. (2004). Gender differences in emotional prosody processing—An fMRI study. *Psychologia, 47*(2), 113–124. https://doi.org/10.2117/psysoc.2004.113

Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone: Differences among three cultures. *Psychological Science, 14*(1), 39–46. https://doi.org/10.1111/1467-9280.01416

Jiam, N. T., Caldwell, M., Deroche, M. L., Chatterjee, M., & Limb, C. J. (2017). Voice emotion perception and production in cochlear implant users. *Hearing Research, 352,* 30–39. https://doi.org/10.1016/j.heares.2017.01.006

Joel, D., Berman, Z., Tavor, I., Wexler, N., Gaber, O., Stein, Y., Shefi, N., Pool, J., Urchs, S., Margulies, D. S., Liem, F., Hänggi, J., Jäncke, L., & Assaf, Y. (2015). Sex beyond the genitalia: The human brain mosaic. *Proceedings of the National Academy of Sciences of the United States of America, 112*(50), 15468–15473. https://doi.org/10.1073/pnas.1509654112

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion, 1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381

Kim, S. K., & Sumner, M. (2017). Beyond lexical meaning: The effect of emotional prosody on spoken word recognition. *The Journal of the Acoustical Society of America, 142*(1), EL49–EL55. https://doi.org/10.1121/1.4991328

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition and Emotion, 16*(1), 29–59. https://doi.org/10.1080/0269993943000121

Koeda, M., Belin, P., Hama, T., Masuda, T., Matsuura, M., & Okubo, Y. (2013). Cross-cultural differences in the processing of non-verbal affective vocalizations by Japanese and Canadian listeners. *Frontiers in Psychology, 4,* 105. https://doi.org/10.3389/fpsyg.2013.00105

Koerner, T. K., & Zhang, Y. (2018). Differential effects of hearing impairment and age on electrophysiological and behavioral measures of speech in noise. *Hearing Research, 370,* 130–142. https://doi.org/10.1016/j.heares.2018.10.009

Kotz, S. A., & Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Research, 1151,* 107–118. https://doi.org/10.1016/j.brainres.2007.03.015

Kret, M. E., & de Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia, 50*(7), 1211–1221. https://doi.org/10.1016/j.neuropsychologia.2011.12.022

Lambrecht, L., Kreifelts, B., & Wildgruber, D. (2014). Gender differences in emotion recognition: Impact of sensory modality and emotional category. *Cognition & Emotion, 28*(3), 452–469. https://doi.org/10.1080/02699931.2013.837378

Latinus, M., Vanrullen, R., & Taylor, M. J. (2016). Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neuroscience, 11*(1), 36. https://doi.org/10.1186/1471-2202-11-36

Lausen, A., & Schacht, A. (2018). Gender differences in the recognition of vocal emotions. *Frontiers in Psychology, 9,* 822. https://doi.org/10.3389/fpsyg.2018.00882

Lawrence, K., Campbell, R., & Skuse, D. (2015). Age, gender, and puberty influence the development of facial emotion recognition. *Frontiers in Psychology, 6,* 761. https://doi.org/10.3389/fpsyg.2015.00761

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means.* R package.

Li, H., Yuan, J., & Lin, C. (2008). The neural mechanism underlying the female advantage in identifying negative emotions: An event-related potential study. *NeuroImage, 40*(4), 1921–1929. https://doi.org/10.1016/j.neuroimage.2008.01.033

Lin, Y., & Ding, H. (2019). Multisensory integration of emotions in a face–prosody–semantics Stroop task. In *Proceedings of ICMI '19: neuromanagement and intelligent computing method on multimodal interaction* (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3357160.3357668

Lin, Y., Ding, H., & Zhang, Y. (2018). Emotional prosody processing in schizophrenic patients: A selective review and meta-analysis. *Journal of Clinical Medicine, 7*(10), 363. https://doi.org/10.3390/jcm7100363

Lin, Y., Ding, H., & Zhang, Y. (2020). Prosody dominates over semantics in emotion word processing: Evidence from cross-channel and cross-modal Stroop effects. *Journal of Speech, Language, and Hearing Research, 63*(3), 896–912. https://doi.org/10.1044/2020_JSLHR-19-00258

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6,* 1171. https://doi.org/10.3389/fpsyg.2015.01171

Martin, J. M., & Altarriba, J. (2017). Effects of valence on hemispheric specialization for emotion word processing. *Language and Speech, 60*(4), 597–613. https://doi.org/10.1177/0023830916686128

McBain, R., Norton, D., & Chen, Y. (2010). A female advantage in basic face recognition is absent in schizophrenia. *Psychiatry Research, 177*(1–2), 12–17. https://doi.org/10.1016/j.psychres.2009.02.005

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748. https://doi.org/10.1038/264746a0

Mehrabian, A., & Wiener, M. (1967). Decoding of inconsistent communications. *Journal of Personality and Social Psychology, 6*(1), 109–114. https://doi.org/10.1037/h0024532

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review, 110*(3), 422–471. https://doi.org/10.1037/0033-295X.110.3.422

Menezes, C. B., Hertzberg, J. C., Neves, F. E. D., Prates, P. F., Silveira, J. F., & Vasconcellos, S. J. L. (2017). Gender and the capacity to identify facial emotional expressions. *Estudos de Psicologia, 22*(1), 1–9. https://doi.org/10.22491/1678-4669.20170001

Most, T., & Michaelis, H. (2012). Auditory, visual, and auditory–visual perceptions of emotions by young children with hearing loss versus children with normal hearing. *Journal of Speech, Language, and Hearing Research, 55*(4), 1148–1162. https://doi.org/10.1044/1092-4388(2011/11-0060)

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science, 33*(1), 127–146. https://doi.org/10.1111/j.1551-6709.2008.01007.x

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory and Cognition, 30*(4), 583–593. https://doi.org/10.3758/BF03194959

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance, 34*(4), 1017–1030. https://doi.org/10.1037/0096-1523.34.4.1017

Palmer, M. A., Brewer, N., & Horry, R. (2013). Understanding gender bias in face recognition: Effects of divided attention at encoding. *Acta Psychologica, 142*(3), 362–369. https://doi.org/10.1016/j.actpsy.2013.01.009

Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language, 105*(1), 59–69. https://doi.org/10.1016/j.bandl.2007.11.005

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104*(3), 262–269. https://doi.org/10.1016/j.bandl.2007.03.002

Pell, M. D. (2005). Prosody–face interactions in emotional processing as revealed by the facial affect decision task. *Journal of Nonverbal Behavior, 29*(4), 193–215. https://doi.org/10.1007/s10919-005-7720-z

Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition and Emotion, 25*(5), 834–853. https://doi.org/10.1080/02699931.2010.516915

Picou, E. M., Singh, G., Goy, H., Russo, F., Hickson, L., Oxenham, A. J., Buono, G. H., Ricketts, T. A., & Launer, S. (2018). Hearing, emotion, amplification, research, and training workshop: Current understanding of hearing loss and emotion perception and priorities for future research. *Trends in Hearing, 22,* 2331216518803215. https://doi.org/10.1177/2331216518803215

Pourtois, G., Debatisse, D., Despland, P.-A., & de Gelder, B. (2002). Facial expressions modulate the time course of long latency auditory brain potentials. *Cognitive Brain Research, 14*(1), 99–105. https://doi.org/10.1016/S0926-6410(02)00064-2

Proverbio, A. M., Brignone, V., Matarazzo, S., Del Zotto, M., & Zani, A. (2006). Gender differences in hemispheric asymmetry for face processing. *BMC Neuroscience, 7*(1), 44. https://doi.org/10.1186/1471-2202-7-44

Psychology Software Tools. (2012). *E-Prime 2.0*. Author. https://www.pstnet.com

Saguy, T., Reifen-Tagar, M., & Joel, D. (2021). The gender-binary cycle: The perpetual relations between a biological-essentialist view of gender, gender ideology, and gender-labelling and sorting. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 376*(1822), 20200141. https://doi.org/10.1098/rstb.2020.0141

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences, 107*(6), 2408–2412. https://doi.org/10.1073/pnas.0908239106

Sauter, D. A., Panattoni, C., & Happe, F. (2013). Children's recognition of emotions from vocal cues. *British Journal of Developmental Psychology, 31*(1), 97–113. https://doi.org/10.1111/j.2044-835X.2012.02081.x

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology, 32*(1), 76–92. https://doi.org/10.1177/0022022101032001009

Scherer, K. R., Banse, R., Wallbott, H. G., & Goldbeck, T. (1991). Vocal cues in emotion encoding and decoding. *Motivation and Emotion, 15*(2), 123–148. https://doi.org/10.1007/BF00995674

Scherer, K. R., & Scherer, U. (2011). Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the emotion recognition index. *Journal of Nonverbal Behavior, 35*(4), 305–326. https://doi.org/10.1007/s10919-011-0115-4

Schirmer, A., & Kotz, S. A. (2003). ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience, 15*(8), 1135–1148. https://doi.org/10.1162/089892903322598102

Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences, 10*(1), 24–30. https://doi.org/10.1016/j.tics.2005.11.009

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research, 14*(2), 228–233. https://doi.org/10.1016/S0926-6410(02)00108-8

Schirmer, A., Kotz, S. A., & Friederici, A. D. (2005). On the role of attention for the processing of emotions in speech: Sex differences revisited. *Cognitive Brain Research, 24*(3), 442–452. https://doi.org/10.1016/j.cogbrainres.2005.02.022

Schirmer, A., Zysset, S., Kotz, S. A., & von Cramon, D. Y. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. *NeuroImage, 21*(3), 1114–1123. https://doi.org/10.1016/j.neuroimage.2003.10.048

Schmid, P. C., Mast, M. S., Bombari, D., & Mast, F. W. (2011). Gender effects in information processing on a nonverbal decoding task. *Sex Roles, 65*(1), 102–107. https://doi.org/10.1007/s11199-011-9979-3

Scholten, M. R. M., Aleman, A., & Kahn, R. S. (2008). The processing of emotional prosody and semantics in schizophrenia: Relationship to gender and IQ. *Psychological Medicine, 38*(6), 887–898. https://doi.org/10.1017/S0033291707001742

Schreuder, E., van Erp, J., Toet, A., & Kallen, V. L. (2016). Emotional responses to multisensory environmental stimuli. *SAGE Open, 6*(1), 2158244016630591. https://doi.org/10.1177/2158244016630591

Schwartz, R., & Pell, M. D. (2012). Emotional speech processing at the intersection of prosody and semantics. *PLOS ONE, 7*(10), Article e47279. https://doi.org/10.1371/journal.pone.0047279

Shankweiler, D., Strange, W., & Verbrugge, R. R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 315–345). Erlbaum.

Shen, J., & Itti, L. (2012). Top-down influences on visual attention during listening are modulated by observer sex. *Vision Research, 65,* 62–76. https://doi.org/10.1016/j.visres.2012.06.001

Spence, C., Parise, C., & Chen, Y. C. (2011). The Colavita visual dominance effect. In M. M. Murray & M. T. Wallace (Eds.), *Frontiers in the neural bases of multisensory processes* (pp. 523–550). CRC Press/Taylor & Francis. https://doi.org/10.1201/9781439812174-34

St Clair, M. C., Forrest, C. L., Yew, S. G. K., & Gibson, J. L. (2019). Early risk factors and emotional difficulties in children at risk of developmental language disorder: A population cohort study. *Journal of Speech, Language, and Hearing Research, 62*(8), 2750–2771. https://doi.org/10.1044/2018_JSLHR-L-18-0061

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Thayer, J., & Johnsen, B. H. (2000). Sex differences in judgement of facial affect: A multivariate analysis of recognition errors. *Scandinavian Journal of Psychology, 41*(3), 243–246. https://doi.org/10.1111/1467-9450.00193

Thompson, A. E., & Voyer, D. (2014). Sex differences in the ability to recognise non-verbal displays of emotion: A meta-analysis. *Cognition & Emotion, 28*(7), 1164–1195. https://doi.org/10.1080/02699931.2013.875889

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory and Cognition, 20*(6), 727–735. https://doi.org/10.3758/BF03202722

Vasconcelos, M., Dias, M., Soares, A. P., & Pinheiro, A. P. (2017). What is the melody of that voice? Probing unbiased recognition accuracy with the Montreal affective voices. *Journal of Nonverbal Behavior, 41*(3), 239–267. https://doi.org/10.1007/s10919-017-0253-4

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective, & Behavioral Neuroscience, 1*(4), 382–387. https://doi.org/10.3758/CABN.1.4.382

Waaramaa, T. (2017). Gender differences in identifying emotions from auditory and visual stimuli. *Logopedics Phoniatrics Vocology, 42*(4), 160–166. https://doi.org/10.1080/14015439.2016.1243725

Wang, Y. N., Zhou, L. M., & Luo, Y. J. (2008). The pilot establishment and evaluation of Chinese affective words system. *Chinese Mental Health Journal, 22*(8), 608–612.

Wirth, M., Horn, H., Koenig, T., Stein, M., Federspiel, A., Meier, B., Michel, C. M., & Strik, W. (2006). Sex differences in semantic processing: Event-related brain potentials distinguish between lower and higher order semantic analysis during word reading. *Cerebral Cortex, 17*(9), 1987–1997. https://doi.org/10.1093/cercor/bhl121

Wu, H., Ma, X., Zhang, L., Liu, Y., Zhang, Y., & Shu, H. (2015). Musical experience modulates categorical perception of lexical tones in native Chinese speakers. *Frontiers in Psychology, 6,* 436. https://doi.org/10.3389/fpsyg.2015.00436

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research, 193*(4), 603–614. https://doi.org/10.1007/s00221-008-1664-6

Zupan, B., Babbage, D., Neumann, D., & Willer, B. (2016). Sex differences in emotion recognition and emotional inferencing following severe traumatic brain injury. *Brain Impairment, 18*(1), 36–48. https://doi.org/10.1017/BrImp.2016.22

Appendix

Summary of Significant Effects in All Three Experiments

	Accuracy			Reaction time		
Experiment No.	**Gender**	**Task**	**Congruence**	**Gender**	**Task**	**Congruence**
Experiment 1	W > M	F > P F > S	-	n.s.	F > P > S	-
Experiment 2	P: W > M S: n.s		C > I	n.s.	P > S	C > I
Experiment 3	n.s	F: n.s; P: Cross > F-S; S: Cross > F-S; Cross > F-P		F: W > M P: W > M		Cross > F-S Cross > F-P

Note. (1) “>” indicates better performances (i.e., more accurate or rapid responses). (2) Abbreviations: W = Women; M = Man; F = Face; P = Prosody; S = Semantics; C = Congruent; I = Incongruent; Cross = Cross-channel congruent; F-S = Face-semantics congruent (prosodic incongruent); F-P = Face-prosody congruent (semantic incongruent); n.s. = No significant difference. (3) For a significant main effect or an interaction, pairwise comparisons between conditions are presented. If the main effect and interaction involving a single factor were both significant, pairwise comparisons under are presented for the interaction effect instead of the main effect.