JSLHR

Research Article

Prosody Dominates Over Semantics in Emotion Word Processing: Evidence From Cross-Channel and Cross-Modal Stroop Effects

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

Purpose: Emotional speech communication involves multisensory integration of linguistic (e.g., semantic content) and paralinguistic (e.g., prosody and facial expressions) messages. Previous studies on linguistic versus paralinguistic salience effects in emotional speech processing have produced inconsistent findings. In this study, we investigated the relative perceptual saliency of emotion cues in cross-channel auditory alone task (i.e., semantics–prosody Stroop task) and cross-modal audiovisual task (i.e., semantics–prosody–face Stroop task).

Method: Thirty normal Chinese adults participated in two Stroop experiments with spoken emotion adjectives in Mandarin Chinese. Experiment 1 manipulated auditory pairing of emotional prosody (happy or sad) and lexical semantic content in congruent and incongruent conditions. Experiment 2 extended the protocol to cross-modal integration by introducing visual facial expression during auditory stimulus presentation. Participants were asked to judge emotional information for each test trial according to the instruction of selective attention.

Results: Accuracy and reaction time data indicated that, despite an increase in cognitive demand and task complexity in Experiment 2, prosody was consistently more salient than semantic content for emotion word processing and did not take precedence over facial expression. While congruent stimuli enhanced performance in both experiments, the facilitatory effect was smaller in Experiment 2.

Conclusion: Together, the results demonstrate the salient role of paralinguistic prosodic cues in emotion word processing and congruence facilitation effect in multisensory integration. Our study contributes tonal language data on how linguistic and paralinguistic messages converge in multisensory speech processing and lays a foundation for further exploring the brain mechanisms of cross-channel/modal emotion integration with potential clinical applications.

Encoding, decoding, and interpreting emotions are, by default, a multisensory integration (MSI) process. In order to fully understand the emotional information, we tend to combine the available linguistic (i.e., verbal content) and paralinguistic signals (i.e., nonverbal messages conveyed by means of tone of voice, gesture, facial expression, and body movement) in multiple communication channels. Unlike unimodal emotion processing, co-occurring verbal and nonverbal emotional signals in MSI are not independent but interactive in natural communicative settings, yielding either enhancement or inhibition effects in behavioral tests (Barnhart et al., 2018).

A growing body of literature has adopted a Stroop-like paradigm (Stroop, 1935) to examine how different communication channels interact with one another in emotion processing (Filippi et al., 2017; Ishii et al., 2003; Kitayama & Ishii, 2002; Ovaysikia et al., 2011; Sutton et al., 2007; Zhu et al., 2010). In these cross-channel (i.e., semantics vs. prosody) and cross-modal (i.e., auditory [semantics vs. prosody] vs. visual [face]) Stroop experiments, participants are instructed to attend to the emotion conveyed through one communication channel or modality and ignore information in another. Participants generally make faster and more accurate responses when processing congruent information across channels/modalities, and incongruent stimuli tend to cause interference with emotion

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

Editor-in-Chief: Frederick (Erick) Gallun
Editor: Christian E. Stilp
Received October 1, 2019
Revision received December 16, 2019
Accepted January 15, 2020
https://doi.org/10.1044/2020_JSLHR-19-00258

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

perception (Liu et al., 2015; Niedenthal, 2007; Pell et al., 2011; Schwartz & Pell, 2012). In short, the Stroop effects in these emotion processing experiments are represented as a congruence-induced facilitation effect in parallel with an incongruence-induced interference effect.

One account for why incongruent stimuli engender interferences lies in the competition between linguistic and paralinguistic processing over their relative salience (or dominance) in emotion cognition. In this account of emotional speech processing, one sensory channel or modality is so perceptually salient/dominant that it can inhibit the perception of another (Spence et al., 2012). However, the empirical findings are mixed. Some researchers have discovered a processing bias toward linguistic content (e.g., semantic meaning; Kitayama & Ishii, 2002; Ovaysikia et al., 2011; Pell et al., 2011). Others have claimed the perceptual saliency of paralinguistic vocal cues such as emotional prosody (i.e., the change of a series of acoustic cues, including pitch, duration, and stress, to convey an emotion; Ben-David et al., 2016; Filippi et al., 2017; Kim & Sumner, 2017; Schirmer & Kotz, 2003). In addition, still others have identified a visual dominance effect supporting the salient role of paralinguistic signals in emotion processing (Beall & Herbert, 2008; Colavita & Weisberg, 1979; Egeth & Sager, 1977; Schmid et al., 2011; see also Spence, 2009, for a review).

The inconsistent findings on the perceptual salience of linguistic versus paralinguistic channels are likely to arise from participant-related factors and experiment-related factors. For instance, a number of studies have demonstrated that emotion processing performances can be shaped by participants' language experience. Yow and Markman (2011) observed that bilingual children are more adept than their monolingual counterparts in judging emotional information in natural speech according to speakers' tone of voice, especially when the semantic content is ambiguous. While native speakers of English are inclined to be oriented by linguistic contents (Kitayama & Ishii, 2002; Pell et al., 2011), Japanese speakers are more attuned to emotional prosody in the understanding of emotions in spoken language (Ishii et al., 2003; Kitayama & Ishii, 2002; Tanaka et al., 2010). Similar findings have been reported on the influences of cultural backgrounds. In western countries such as Germany, the Netherlands, and the United States, emotional semantics or faces tend to override prosody. By contrast, prosody takes precedence among the participants from Asian countries such as Japan, Philippines, and China (Ishii et al., 2003; Liu et al., 2015; Paulmann & Kotz, 2008; Tanaka et al., 2010).

Moreover, the effects of modality salience depend on experimental manipulations, including stimulus design and task difficulty. de Silva et al. (1997) showed that sadness and fear displayed in videos are better recognized in the auditory modality whereas other emotions such as anger and happiness can be effortlessly identified in the visual modality. Similarly, Paulmann and Pell (2011) also did not identify a uniform visual dominance effect across emotional categories, especially for the recognition of sad, neutral, and disgust emotions. Further work by Paulmann et al. (2013) showed that, despite the influence of valence and arousal, no difference was found between implicit and explicit task instructions. Although many studies showed no impacts from task demands among healthy participants (Sander et al., 2005; Wildgruber et al., 2006), some studies on clinical populations have reported task-related attentional modulation during multimodal emotional processing (de Gelder et al., 2005; de Jong et al., 2010).

In view of these inconsistencies, it is essential to carefully reconsider some variations in language backgrounds, sociocultural characteristics, and experimental design before generalizing the existing findings. First, the language samples in previous work were limited to English, German and Japanese, which are either stress-timed or mora-timed nontonal languages. As most studies were conducted in western countries with a "low-context" culture that favors a more direct and explicit approach in communication, it remains to be tested whether the findings are applicable to a "high-context" culture like Chinese that relies heavily on contextual cues and implicit messages (Hall, 1976). Second, a number of Stroop experiments employed emotional cues that were dimensional (e.g., positive, negative) instead of categorical (e.g., happy, sad, angry, disgust). For instance, Schirmer and Kotz (2003) used positive, negative, or neutral verbal contents corresponding or not corresponding to the matched tone of voice. Sutton et al. (2007) adopted negative (e.g., nervous, panic) and neutral (e.g., seat, deck) words spoken with negative or neutral prosody. It remains unclear whether the relative salience effects are applicable to discrete emotional categories. Third, previous Stroop tasks were confined to binary contrasts of emotional stimuli (e.g., facial vs. prosodic, facial vs. semantic, prosodic vs. semantic). The experimental design needs to take into account the MSI complexity of emotional speech processing in naturalistic settings, where we often have access to more than two communication channels of emotional information. Moreover, few studies have investigated how task demand/difficulty, especially an increase in the number of communication channels, affects the interplay among different sensory modalities. Some recent studies have made tentative explorations by modulating the task demands, but they only narrowed the focus on the role of attentional shift by using task-relevant or irrelevant stimuli to examine facilitation effects of MSI rather than the inhibition posed by different sensory cues (Paulmann et al., 2013; Paulmann & Pell, 2011; Wildgruber et al., 2006). One study by Filippi et al. (2017) was able to demonstrate the perceptual saliency of prosody involving two-dimensional semantics–prosody and three-dimensional semantics–prosody–face Stroop tasks. However, group differences in the study could not be excluded as two panels of participants were respectively recruited for each task.

This study applied the experimental protocols of Filippi et al. (2017) and tested whether the perceptual salience/dominance effect of prosody could be generalized to a typical East Asian language and culture setting. Our target language was Mandarin Chinese, a syllable-timed, tonal language in a typical high-context culture. We selected an uncontroversial pair of discrete categories of basic emotions (i.e., happy and sad; Ekman, 1992; Ekman & Cordaro, 2011) and adopted a within-subject experimental design by inviting a group of 30 subjects to complete the two sets of Stroop experiments. Prosody and lexical semantic content were directly contrasted in a cross-channel (linguistic vs. paralinguistic) Stroop task in the auditory-alone Experiment 1. Based on the existing findings on the salient role of vocal emotions in Asian countries, it was hypothesized in the cross-channel Experiment 1 that prosody would be more perceptually salient than lexical semantic content during emotion perception. In the audio-visual Experiment 2, we included emotional facial expressions in a three-dimensional cross-modal Stroop task to further examine the relative salience of the three communication channels. The inclusion of visual facial cues as a third channel and a new modality required the participants to allocate more cognitive resources so that they could successfully disentangle cross-channel/modal interactions, thus rendering emotion recognition more complex and demanding compared to the semantics–prosody Stroop task. We hypothesized that prosody would continue to be more salient than lexical semantic content despite an increase in task difficulty and take precedence over facial expression, which might reflect an auditory salience effect.

By testing a new linguistic and sociocultural system with a multidimensional emotion Stroop protocol, we investigated whether the reliance on paralinguistic features (especially prosody) in emotion recognition tasks among Westerners can be extended to people in the East (with Chinese as a representative) and even serve as a universal pattern of MSI of emotion. We aimed to contribute to our understanding of the nature of modality salience effects and congruence-induced facilitation effects in emotional speech processing under different task difficulty levels. Such empirical efforts are necessary and important for advancing existing theories of emotion recognition by investigating the interactions among multiple communication channels and testing possible language-specific variations, which holds promise to reveal insights to guide communication practices for both healthy and clinical populations.

Experiment 1

Method

Participants

Thirty adult volunteers (15 men and 15 women) were recruited for this experiment through an online advertisement on Shanghai Jiao Tong University website. In the advertisement, we wrote that the experiment consisted of two phases, and volunteers would be re-invited for the test some days later. Only those who initially agreed to and completed the two experiments were counted as participants in the current study. They were financially compensated for their time. All participants were undergraduate or graduate students who were native speakers of Mandarin Chinese studying at Shanghai Jiao Tong University. Participants averaged 24.1 (*SD* = 2.6) years in age and had received an average of 17.5 (*SD* = 2.4) years of formal school education. Apart from their education backgrounds, participants were regarded as neurocognitively normal since none of them reported history of speech, language, or hearing impairment or any psychiatric disorder. All had normal or corrected-to-normal vision, and they had normal hearing as determined by standard audiometric screening (Koerner & Zhang, 2018). To verify their ability to follow instructions consistently, all participants passed the practice session with 100% accuracy prior to the experiment at a comfortable sound intensity level (approximately 70 dB SPL). They completed written informed consent at the study as required and approved by the institutional review board in accordance with the Declaration of Helsinki for research involving human subjects.

Stimuli

The stimuli included 32 disyllabic spoken words in Mandarin Chinese, each of which expressed either happiness or sadness in emotional semantic content and prosody (see Table A1 in Appendix A). Thus, emotional information in each auditory stimulus was conveyed simultaneously through two communication channels (see Appendix B with Tables B1 and B2 for procedures of stimulus construction and validation). There was an equal number of words with happy and sad meaning uttered by men and women in each prosodic set.

Semantic channel. The 32 spoken words in this study contained two sets of adjectives, most of which were selected from the Chinese Affective Words System (Wang et al., 2008). Sixteen words were synonyms of “happy,” and the other 16 were synonyms of “sad.” To ensure no rare or uncommon words were used, only words with an average rating of > 3 for familiarity on a 7-point Likert scale (0 = *not familiar*, 6 = *very familiar*) in a norming study were included.

Prosodic channel. The 32 words (16 happy and 16 sad) were randomly chosen to be enunciated with either a happy or sad prosody. Half of these words were spoken in a congruent prosody, and the other half in an incongruent prosody (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). The lexical tone combination of the disyllabic words was matched between the happy and sad word sets so as to control the influence of lexical tones on the expression of emotion. Each word was produced three times by six amateur actors (three men and three women), and the best ones were chosen to represent each voice based on the identification accuracy of emotional valence and ratings of intensity in the norming study.

The speech materials were first evaluated in a norming study with a group of 24 normal adult native speakers of

Mandarin Chinese who did not participate in either of the two Stroop experiments. The duration measures of the auditory stimuli are shown in Table 1. Emotions presented through both semantic and prosodic channels were recognized with over 90% accuracy for valence in a two-choice task (happy or sad) and received an average rating of > 3 for emotional intensity on a 7-point Likert scale (0 = *not intense*, 6 = *very intense*). In addition, we measured the ratio of semantics and prosody in terms of emotional intensity in the same trial, which indicated the extent to which emotional intensity was comparable across channels. We kept the ratio value as close to 1 as possible by including stimuli with a similar level of emotional intensity in semantics and prosody. The mean intensity ratio of the two channels for all included stimuli was 1.1 (*SD* = .2), suggesting a relatively balanced level of emotional intensity across communication channels and experiment trials. We intended to control the intra- and intertrial processing asymmetries in this way so that it is less likely for the intensity differences between semantics and prosody to serve as an extraneous factor affecting participants' performance.

Procedure

The experiment was administered in a sound booth with the participant seated in a comfortable chair at around 70 cm from an LCD monitor. Stimulus presentation program was written with E-Prime (Version 2.0.8.22; Psychology Software Tools, 2012). The auditory stimuli were presented binaurally over Sennheiser HD280 PRO headphones at 70 dB SPL. Prior to the experiment, instructions explaining the experimental procedures and task requirements were provided both visually in text on the screen and aurally through headphones. The experiment started with a familiarization phase in which there was a practice test with four trials. Participants needed to reach 100% accuracy in the practice test session before entering the test phase. Seventeen of 30 (56.7%) subjects passed the practice session in the first try, and the rest of the participants needed more than one (up to three) practice session. The test required the participants to selectively attend to either the semantic meaning or emotional prosody according to the auditory instruction provided at the beginning of each trial. They were asked to identify the emotional valence of word meaning while ignoring emotional prosody when the auditory instruction "emotional semantics" preceded the trial (semantic task). When the instruction "emotional prosody" was given (prosodic task), participants needed to identify the emotional valence conveyed through tone of voice while ignoring the semantic meaning of the emotional words.

Table 1. Duration (in milliseconds) of the auditory stimuli used in both experiments.

	Prosody					
	Happy		Sad			
Semantics	*M*	*SD*	*M*	*SD*	*M/SD*	
Happy	773	112	905	140	839	143
Sad	723	55	852	158	787	137
M/SD	750	92	879	153	813	142

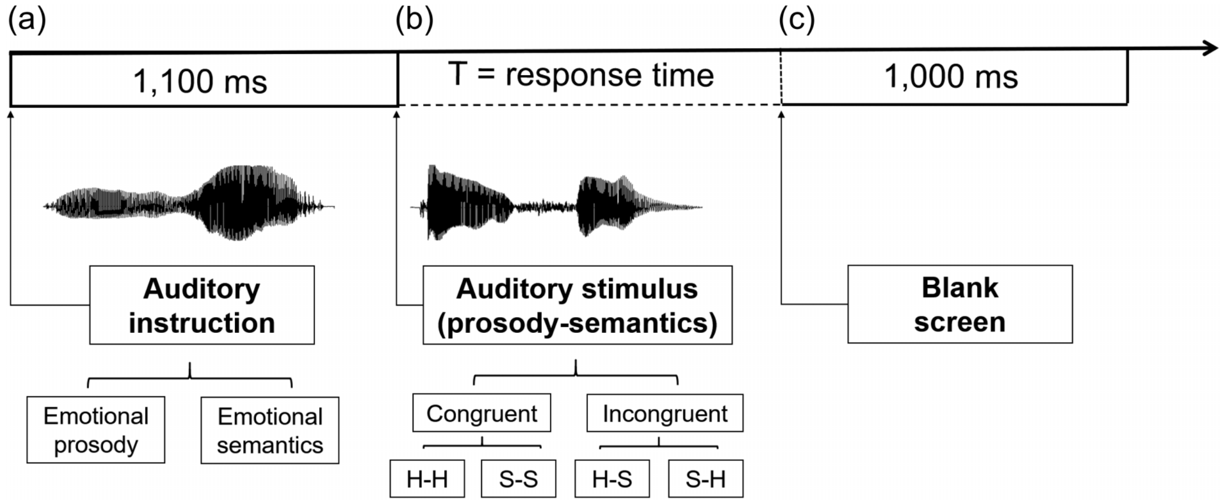
The experiment included 64 test trials with each of the 32 auditory items presented twice. Participants were guided to attend to either the semantic or prosodic channels for each given test trial based on the instructional prompt. The emotion pairing of semantic and prosodic channels was either congruent (happy–happy or sad–sad) or incongruent (happy–sad or sad–happy) in each trial. The presentation order of test trials was randomized across participants. As the emotional signals conveyed by each communication channel varied from trial to trial in a random fashion, participants were unlikely to be able to anticipate the pattern of emotional combination of semantic content and prosody for the upcoming trial. There was a short break of 15 s after completing the first 32 trials.

Each trial began with an auditory instruction of either "emotional semantics" or "emotional prosody" for 1,100 ms. Then simultaneous presentation of emotional semantic content and prosody were binaurally delivered over the headphones. Meanwhile on the screen, participants were asked to indicate the emotion conveyed in the attended channel as quickly as possible without sacrificing accuracy. They responded 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. We measured accuracy and response time from stimulus onset. After responses were made, a blank screen was displayed for 1,000 ms before the next trial began. The prosody–semantic content Stroop protocol is illustrated in Figure 1.

Statistical Analyses

A series of linear mixed-effects models were performed to analyze both types of data using R (Version 3.4.4; R Core Team, 2018) with the lme4 package (Bates et al., 2015). Considering that many participants obtained near-perfect scores, we transformed the percent correct scores into rationalized arcsine unit (rau) to decompress the ceiling effects (Studebaker, 1985). We also performed a log transformation to reaction time data since, in many perceptual experiments, response time exhibits positive skewness (Baayen & Milin, 2010), which is equivalent to implementing the logistic link function in generalized linear mixed-effects models (Lo & Andrews, 2015). Accuracy in raus and the logarithm of reaction time were respectively entered as dependent variables. Within-subject variables task (semantic vs. prosodic) and congruence (congruent vs. incongruent) were entered as categorical fixed factors, in which the prosodic task and the congruent condition were used as the default level, respectively. Listener subjects and test items were entered as random factors for intercepts. When there was a significant main effect or a significant interaction effect, Tukey's post hoc tests were conducted, using the lsmeans package

Figure 1. Schematic illustration of the prosody–semantic content cross-channel Stroop protocol. (a) In each trial, the auditory instruction guided the participants to attend to emotion conveyed by either prosody or semantic content. (b) Response could be made as soon as the auditory presentation of prosody and semantic content was delivered. Participants pressed the key buttons to indicate the emotional valence conveyed by the attended channel (e.g., "f" for happy and "s" for sad), the positions of which were counterbalanced across subjects, whenever they could make a judgment. (c) The two letters in "H-S" indicated the emotional valence of prosody and semantic content, respectively, with "H" standing for happy and "S" standing for sad. For instance, "H-S" represented a word with sad semantic meaning spoken with a happy prosody.



(Lenth, 2016). Satterthwaite approximation was used to obtain *p* values.

The full models with intercepts, coefficients, and error terms for accuracy and reaction time analyses in Experiments 1 and 2 are represented as follows:

$$\begin{aligned}\text{Accuracy(rau)}_{ij} = {} & \beta_0 + (\beta_1 \times \text{Task}) + (\beta_2 \times \text{Congruence}) \\ & + (\beta_3 \times \text{Task} \times \text{Congruence}) + b_{0i} \\ & + b_{1j} + \varepsilon_{ij,}\end{aligned} \tag{1}$$

$$\begin{aligned}\log(\text{Reaction time})_{ij} = {} & \beta_0 + (\beta_1 \times \text{Task}) + (\beta_2 \times \text{Congruence}) \\ & + (\beta_3 \times \text{Task} \times \text{Congruence}) \\ & + b_{0i} + b_{1j} + \varepsilon_{ij.}\end{aligned} \tag{2}$$

Results and Discussion

Tables 2 and 3 summarize the results of the linear mixed-effects models for accuracy and reaction time data in Experiment 1.

Accuracy

Overall, the participants responded to the auditory stimuli with substantially high accuracy (*M* = 94.3%, *SD* = 6.8%). Figure 2(a) illustrates the accuracy data in raus in Experiment 1, and the reported mixed-effects analyses were conducted based on data transformed into raus.

Linear mixed-effects analyses revealed a significant main effect of congruency condition, $\chi^2(1) = 23.88$, $p < .001$. Congruent stimuli elicited more (5.3% ± 1.1%) accurate responses than incongruent ones ($\beta_2 = 10.10$, $SE = 1.97$, $t = 5.14$, $p < .001$). There was no main effect of task, $\chi^2(1) = 0.38$, $p > .05$. Neither of the congruency condition produced significant interaction with the factor task, $\chi^2(1) = 0.52$, $p > .05$.

Reaction Time

In analysis of reaction time data, incorrect responses and responses over 2 *SD*s from the mean, which respectively accounted for 5.4% and 3.4% of the overall data set, were excluded (Chien et al., 2017; Baayen & Milin, 2010). Reaction time data in the two tasks and conditions are displayed in Figure 2(b).

Linear mixed-effects analyses on the logarithm-transformed reaction time showed main effects of task, $\chi^2(1) = 13.64$, $p < .001$, and congruency, $\chi^2(1) = 14.42$, $p < .001$, with no significant interaction between task and congruency, $\chi^2(1) = 1.82$, $p > .05$. On average, participants responded 119.0 ± 32.1 ms faster to the prosody task than to the semantic task ($\beta_1 = -.11$, $SE = .03$, $t = -3.86$, $p < .001$), and 134.9 ± 32.1 ms faster to the congruent stimuli than to the incongruent ones ($\beta_2 = -.10$, $SE = .03$, $t = -3.96$, $p < .001$).

These data showed that emotions were identified faster in the prosodic task than in the semantic task regardless of congruency condition. Additionally, congruent trials yielded more accurate and rapid responses than incongruent trials in both tasks. Thus, prosody facilitated more rapid emotional processing than semantic content. There was also a consistent congruency facilitation effect across tasks in both accuracy and response time with no significant

Table 2. Linear mixed-effects model with task and congruence as the fixed effects and accuracy in rationalized arcsine units as the dependent variable in Experiment 1 (pairwise contrasts are indented).

Parameter	Any effect?	Estimate	*SE*	Test (*df*)	*p*
Task	No				
Prosody vs. semantics		–1.35	2.21	–0.61 (89)	.542
Congruence	Yes				
Congruent vs. incongruent	Yes	10.10	1.97	5.14 (88)	< .001
Task × Congruence	No				
Congruent (prosody vs. semantics)		–2.76	2.80	–0.99 (87)	.757
Incongruent (prosody vs. semantics)		0.06	2.80	0.02 (87)	1.000
Prosody (congruent vs. incongruent)		8.73	2.80	3.12 (87)	.013
Semantics (congruent vs. incongruent)		11.56	2.80	4.13 (87)	.001

Note. The prosodic task and the congruent condition were used as the default level of task and congruence, respectively. *df* = degrees of freedom.

interaction. These results are not surprising as human beings are more likely to encounter congruent emotional information conveyed by both prosodic and semantic channels in daily communication (Nygaard & Queen, 2008). Consistent with previous studies (Filippi et al., 2017; Nygaard & Queen, 2008; Wildgruber et al., 2006), the Mandarin Chinese speakers in the study also relied on affective prosody as an effective and primary cue for interpreting emotion.

Experiment 2

Apart from semantic content and tone of voice, visual cues, such as facial expressions, gestures, and body movements, also serve as an essential component of emotional expression and comprehension in real-life communication (Beall & Herbert, 2008). In Experiment 2, we further tested whether prosody would continue to be more salient than semantic content in emotion word processing with co-occurring facial expression for each test trial. We were interested in determining whether congruent stimuli would produce a facilitatory effect in an emotional identification task with a cross-modal Stroop task that is more demanding and closer to real-life experience.

Method

Participants

As stated in Experiment 1, the same group of 30 students (15 men and 15 women; $M_{\text{age}} = 24.1$, $SD = 2.6$) participated in the two experiments. For each participant, the second experiment was conducted at least 2 weeks apart from the first one to mitigate carryover effects. As in Experiment 1, participants completed institutional review board–approved consent form and were financially compensated for their time.

Stimuli

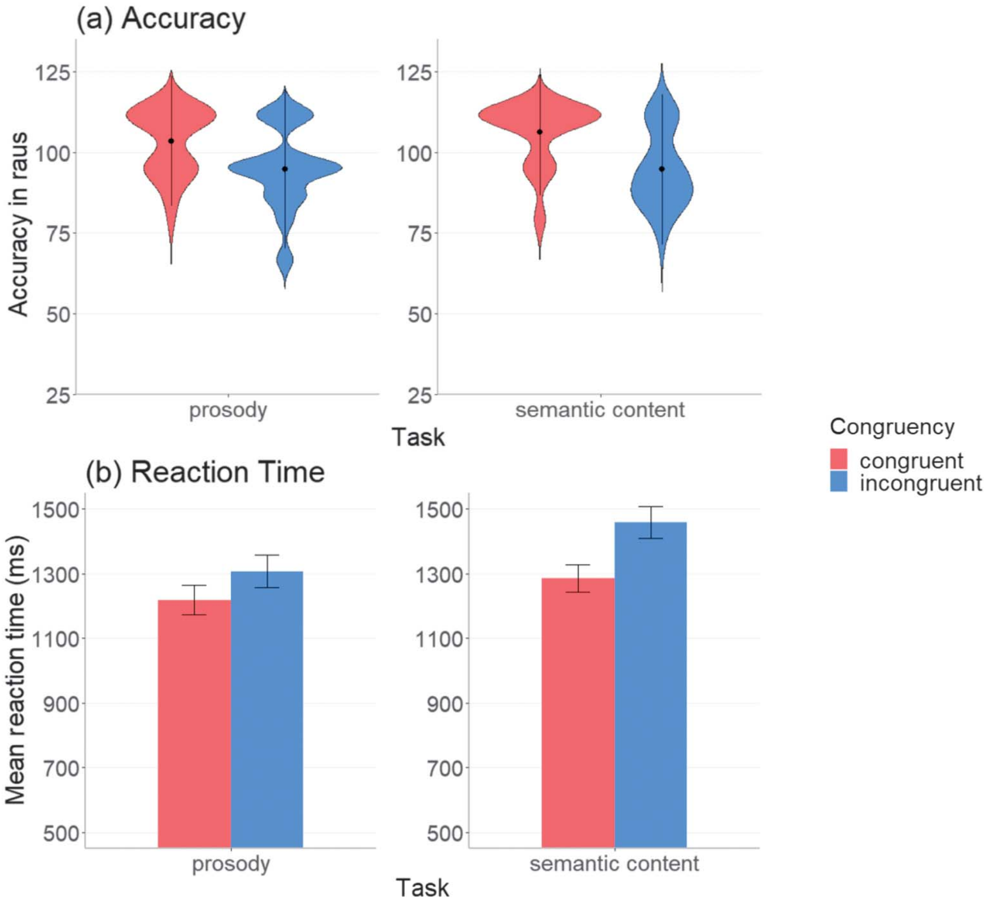
In this cross-modal task, emotional information was simultaneously conveyed through three communication channels, namely, semantic content, prosody, and facial expression (see Appendix B for procedures of stimulus construction and validation). The same set of 32 spoken word stimuli (see Appendix A) was used in this experiment. In addition, 32 faces expressing happy or sad emotions were employed. These visual stimuli contained black-and-white photographs of a facial expression from the Chinese Affective Picture System (Bai et al., 2005), which were posed

Table 3. Linear mixed-effects model with task (prosody as baseline) and congruence (congruent as baseline) as the fixed effects and the logarithm of reaction time as the dependent variable in Experiment 1 (pairwise contrasts are indented).

Parameter	Any effect?	Estimate	*SE*	Test (*df*)	*p*
Task	Yes				
Prosody vs. semantics	Yes	–0.11	0.03	–3.86 (62)	< .001
Congruence	Yes				
Congruent vs. incongruent	Yes	–0.10	0.03	–3.96 (61)	< .001
Task × Congruence	No				
Congruent (prosody vs. semantics)		–0.08	0.04	–2.14 (58.6)	.152
Incongruent (prosody vs. semantics)		–0.14	0.04	–3.99 (61.5)	.001
Prosody (congruent vs. incongruent)		–0.07	0.04	–1.88 (59.8)	.248
Semantics (congruent vs. incongruent)		0.01	0.04	0.25 (59.6)	.994

Note. The prosodic task and the congruent condition were used as the default level of task and congruence, respectively. *df* = degrees of freedom.

Figure 2. Accuracy in (a) rationalized arcsine units (raus) and (b) mean reaction time in the two tasks and congruence conditions in Experiment 1. In (a), accuracy in raus is displayed in the violin plots, with data distribution shape indicated by the density plots, mean values represented by the black dots, and 95% confidence intervals shown by the error bars. In (b), mean reaction time is displayed in the bar charts with error bars showing 95% confidence intervals.



by 16 male and 16 female actors. The number of faces was equal between happy and sad emotions and between male and female actors, thus yielding eight faces per emotion category for each gender.

For each voice, four emotional faces (two happy and two sad) posed by four actors of the same sex were matched in a trial. Similar to Experiment 1, the stimuli were subject to a norming study for evaluation by another group of 24 native speakers of Mandarin Chinese. Emotions presented through semantic, prosodic, and facial channels were all recognized with over 90% accuracy for valence in a two-choice task (happy or sad) and received an average rating of > 3 for emotional intensity on a 7-point Likert scale (0 = *not intense*, 6 = *very intense*) in the norming study. The mean ratios of emotional intensity ratings between every two channels were comparable (semantic content/prosody: 1.1, *SD* = 0.2; semantic content/face: 1.1, *SD* = 0.1; prosody/face: 1.1, *SD* = 0.2) so that participants' performances were less likely to be explained by the intensity asymmetry across channels.

Procedure

The basic protocols followed Experiment 1. Participants were told to identify the emotional valence of word meaning while ignoring emotional prosody and the co-occurring facial expression when the trial prompt before stimulus presentation was "emotional semantics" (semantic task). When the instruction changed to "emotional prosody" (prosodic task), participants needed to identify the emotional valence conveyed through tone of voice while ignoring emotional semantic content and facial expression on screen. When the instruction was to attend to "emotional face," participants needed to respond based on the facial expression while ignoring information in the other two communication channels. After a familiarization phase with eight practice trials consecutively reaching 100% accuracy, the test phase started. Twenty of 30 (66.7%) subjects passed the practice session in the first try, and the rest of the participants needed more than one practice session.

The test session included 96 trials: 32 items were presented in three blocks. Participants were guided to selectively attend to the semantic, prosodic, or facial channels according to the given instruction for each trial. There were a total of 32 trials targeting semantic content, 32 targeting prosody, and 32 targeting facial expression. In each trial, two channels were congruent, while the remaining one was incongruent with the other two, yielding the following congruency conditions: prosody–semantic content congruent, prosody–face congruent, and face–semantic content congruent. In a fourth condition, which served as cross-channel congruent control, emotional valence was congruent across all three. Our purpose was to test how emotionally congruent channels interacted with a third incongruent channel in multimodal processing for each selective attention task. Full randomization of trial order was implemented in order to prevent strategic prediction. A short break of up to 30 s was allowed for every 32 test trials.

Each trial began with an auditory instruction prompt of target information to attend to (see Figure 3). Then, emotional semantic content and prosody were binaurally presented over headphones, together with a facial expression shown in the center of the LCD screen. The onset of auditory and visual emotional cues were kept identical in each trial. Participants were asked to indicate the emotion conveyed in the attended channel as quickly as possible. They responded by pressing the corresponding emotion-coded keys on a computer keyboard ("f" for happy and "j" for sad). The positioning of "happy" and "sad" response buttons was counterbalanced across participants. After each response, a blank screen was displayed for 1,000 ms before the next trial began.

Statistical Analyses

As in Experiment 1, we transformed the accuracy data into rau to handle the ceiling effects and applied log-transform to the reaction time data due to the positive skewness. A series of mixed-effects models with the transformed data as separate dependent variables were then implemented. The fixed factors were task (three conditions: semantic, prosodic, and facial) and congruency (four conditions: prosodic incongruent, semantic incongruent, facial incongruent, and all congruent), in which the facial task and all-congruent condition were used as the baseline level, respectively. When conducting pairwise comparison between prosodic and semantic tasks, prosody was set as the baseline level. The random factors were listener subjects and test items for intercepts. When there was a significant main effect or a significant interaction effect, Tukey's post hoc tests were performed, using the lsmeans package (Lenth, 2016). Satterthwaite approximation was used to obtain p values. The full models (1) and (2), as specified in the Method section in Experiment 1, were applied in the current analyses.

Results and Discussion

Tables 4 and 5 summarize the results of the linear mixed-effects models for accuracy and reaction time data in Experiment 2. As in Experiment 1, accuracy data will be reported first, followed by the analysis of reaction time.

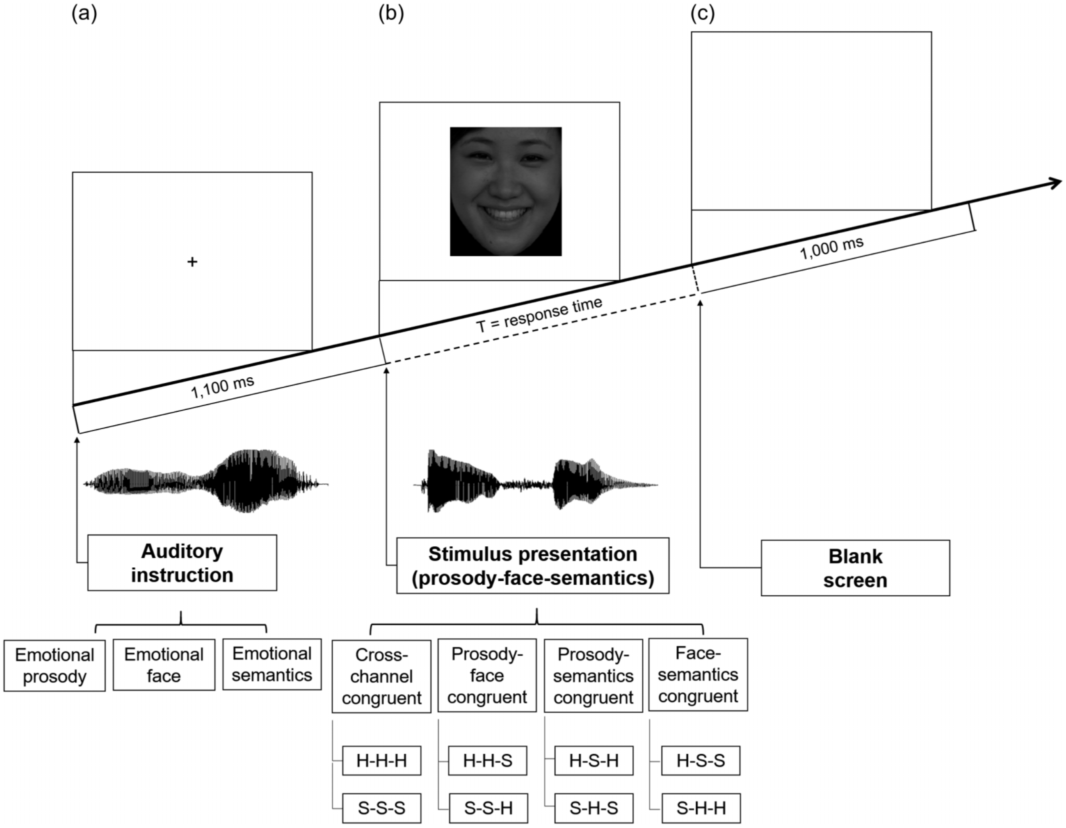
Accuracy

Participants achieved considerably high accuracy in this experiment (M = 93.5%, SD = 9.9%). Accuracy data in raus in different tasks and congruence conditions in Experiment 2 are illustrated in Figure 4(a).

Linear mixed-effects analyses on accuracy in rau revealed main effects of task, $\chi^2(2) = 13.53$, $p = .001$, and congruence, $\chi^2(3) = 33.39$, $p < .001$. There was also a significant interaction between the two fixed factors, $\chi^2(6) = 34.34$, $p < .001$. To parse out the interaction effect, the all-congruent control condition was compared with the other three conditions in which two communication channels were congruent while the remaining one was incongruent. For the semantics-oriented task, the semantic incongruent (prosodic–facial congruent) condition showed the poorest performance (12.1% ± 2.7% lower than the all-congruent condition, $\beta_3 = 17.80$, $SE = 3.19$, $t = 5.59$, $p < .001$), and the prosodic incongruent (semantic–facial congruent) condition also showed a significant effect (10.8% ± 2.7% lower than the all-congruent condition, $\beta_3 = 15.83$, $SE = 3.19$, $t = 4.97$, $p < .001$). But facial incongruency (semantic–prosodic congruent) did not have such an effect ($p > .05$). For the prosody-oriented task, only the prosody-incongruent condition showed a significant effect with its accuracy rate 9.2% ± 2.0% lower than the baseline condition ($\beta_3 = 14.28$, $SE = 3.19$, $t = 4.48$, $p < .001$). The semantic-incongruent condition ($p > .05$) and the facial-incongruent condition ($p > .05$) did not have significant influences on the emotion identification. For the facial expression–oriented task, none of the three incongruent conditions showed significant differences from the all-congruent condition in emotional identification ($p > .05$).

Notably, identification accuracy largely depended on the interplay between information modalities and congruency conditions. In comparison with the all-congruent condition, the semantics-oriented task engendered less accurate responses not only when the semantic content (the attended channel) was incongruent with face and prosody (the other two congruent channels) but also when prosody (the unattended channel) was incongruent with both semantic content and face. These data suggest that incongruent prosody could bias the emotion identification responses even if prosody was not the attended channel. However, semantics did not appear to have the same kind of influential weight in emotion identification as the prosody-oriented task elicited less accurate performance only when prosody (the attended channel) was incongruent with both semantic content and facial expression for the speech material. In other words, when unattended in this cross-modal Stroop task, semantic content alone did not appear to carry the same perceptual weight as prosodic information to interfere with emotion identification. In the facial expression–oriented task, identification accuracy rates were comparable among all the experimental conditions, suggesting that visual facial

Figure 3. Schematic illustration of the prosody–face–semantic content Stroop protocol. (a) In each trial, the auditory instruction guided the participants to attend to emotion conveyed by prosody, facial expression, or semantic content. (b) Response could be made as soon as auditory and visual stimuli were simultaneously presented. Participants pressed the key buttons to indicate the emotional valence conveyed by the attended channel (e.g., "f" for happy and "s" for sad), the positions of which were counterbalanced across subjects, whenever they could make a judgment. (c) The three letters in "H-H-H" indicated the emotional valence of prosody, face, and semantic content, respectively, with "H" standing for happy and "S" standing for sad. For instance, "H-S-S" represented a word with sad semantic meaning spoken with a happy prosody and simultaneously presented with a sad facial expression.



expressions are the most unambiguous cues in emotion identification to avoid potential influences from other informational channels in this three-dimensional Stroop task.

Reaction Time

In the analysis of reaction time data, we excluded incorrect responses (5.1%) and responses over 2 *SD*s from the mean (3.7%; Chien et al., 2017; Baayen & Milin, 2010). Reaction time data across conditions are displayed in Figure 4(b). As the reaction time data were positively skewed, the data were log-transformed as in the analysis for Experiment 1.

Linear mixed-effects analyses revealed a main effect of task, $\chi^2(2) = 116.70$, $p < .001$; a main effect of congruency, $\chi^2(3) = 13.78$, $p = .003$; and no significant interaction between the two fixed factors, $\chi^2(6) = 12.53$, $p > .05$. Compared with the facial expression–oriented task, response time was increased by 212.4 ± 24.9 ms in the prosody-oriented task ($\beta_1 = -.24$, $SE = .02$, $t = -9.90$, $p < .001$) and by 329.0 ± 24.9 ms in the semantics-oriented task ($\beta_1 = -.35$, $SE = .02$, $t = -14.64$, $p < .001$). There was also a significant increase by 116.6 ± 25.0 ms in the semantics-oriented task relative to the prosody-oriented task ($\beta_1 = -.11$, $SE = .02$, $t = -4.73$, $p < .001$). When collapsed across tasks, cross-channel and cross-modality all-congruent stimuli elicited 94.6 ± 28.7 ms faster responses than prosody–face congruent stimuli ($\beta_2 = -.08$, $SE = .03$, $t = -3.23$, $p = .009$). No significant reaction time difference was found between all-congruent stimuli and semantics–face congruent stimuli and between the baseline and prosody–semantic content congruent stimuli ($p > .05$). Consistent with the accuracy data and previous work (Filippi et al., 2017), the reaction time results provide further evidence that the visual cues of facial expression were perceptually salient and unambiguous as emotions were identified significantly faster in the facial expression–oriented task than the other two tasks regardless of congruency conditions. Interestingly, despite the inclusion of facial expression in our cross-modality and cross-channel Stroop experiment for emotion identification, there existed significant differences between the semantics-oriented task and the prosody-oriented task. As in

Table 4. Linear mixed-effects model with task and congruence (cross-congruent as baseline) as the fixed effects and accuracy in rationalized arcsine units as the dependent variable in Experiment 2 (pairwise contrasts are indented).

Parameter	Estimate	*SE*	Test (*df*)	*p*
Task				
Face vs. prosody	2.51	1.74	1.44 (328)	.320
Face vs. semantics	6.40	1.74	3.68 (328)	.001
Prosody vs. semantics	3.89	1.74	2.23 (328)	.067
Congruence				
All congruent vs. prosodic incongruent	10.32	1.92	5.38 (325)	< .001
All congruent vs. semantic incongruent	7.77	1.92	4.05 (325)	< .001
All congruent vs. facial incongruent	3.22	1.92	1.68 (325)	.339
Task × Congruence				
Face (all congruent vs. prosodic incongruent)	0.85	3.19	0.27 (319)	.993
Face (all congruent vs. semantic incongruent)	2.38	3.19	0.75 (319)	.879
Face (all congruent vs. facial incongruent)	2.71	3.19	0.85 (319)	.830
Prosody (all congruent vs. prosodic incongruent)	14.28	3.19	4.48 (319)	< .001
Prosody (all congruent vs. semantic incongruent)	3.14	3.19	0.98 (319)	.759
Prosody (all congruent vs. facial incongruent)	4.65	3.19	1.46 (319)	.462
Semantics (all congruent vs. prosodic incongruent)	15.83	3.19	4.97 (319)	< .001
Semantics (all congruent vs. semantic incongruent)	17.80	3.19	5.59 (319)	< .001
Semantics (all congruent vs. facial incongruent)	2.28	3.19	0.72 (319)	.891

Note. The facial task and the congruent condition were used as the default level of task and congruence, respectively. When conducting pairwise comparison between prosodic and semantic tasks, prosody was set as the baseline level. *df* = degrees of freedom.

Experiment 1, participants remained to be more sensitive in processing emotional prosodic information than the verbal semantic information, giving rise to slower reaction time when attending to the tone of voice in the speech material. The reaction time was also increased only when involving semantic incongruency, but not when involving prosodic and facial incongruency.

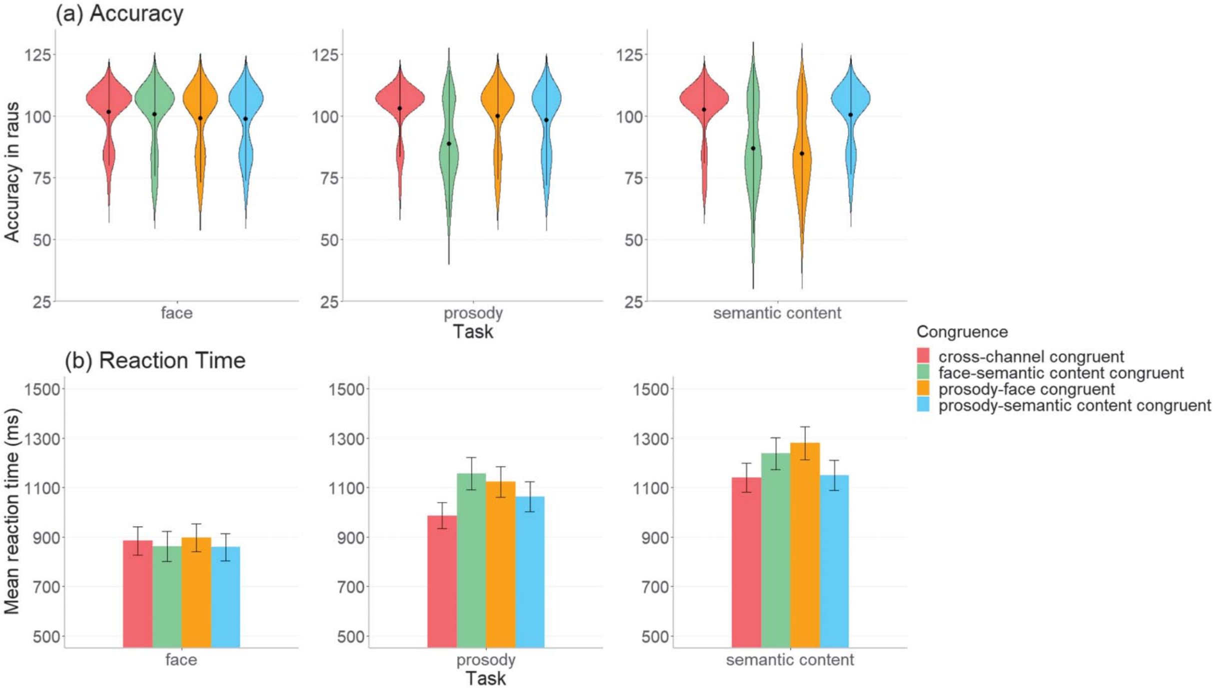
The accuracy and reaction time data of Experiment 2 converge with and extend the findings in Experiment 1. Despite the increased task demand with simultaneous presentation of semantic, prosodic, and facial cues for emotion identification, prosody continues to be more perceptually influential than semantic content. Prosody not only gains advantages over semantic content in reaction time when it is the attended channel, but it interferes with the identification of semantic emotions even when it is an unattended channel. Compared with the auditory informational channels, faces serve as the most perceptually salient and unambiguous channel in multimodal emotional processing, as we observed visual salience in mean accuracy and

Table 5. Linear mixed-effects model with task and congruence as the fixed effects and the logarithm of reaction time as the dependent variable in Experiment 2 (pairwise contrasts are indented).

Parameter	Estimate	*SE*	Test (*df*)	*p*
Task				
Face vs. prosody	−0.24	0.02	−9.90 (92.1)	< .001
Face vs. semantics	−0.35	0.02	−14.64 (92.6)	< .001
Prosody vs. semantics	−0.11	0.02	−4.73 (94.1)	< .001
Congruence				
All congruent vs. prosodic incongruent	−0.07	0.03	−2.59 (89.8)	.054
All congruent vs. semantic incongruent	−0.08	0.03	−3.23 (89.0)	.009
All congruent vs. facial incongruent	−0.02	0.03	−0.69 (87.6)	.903
Task × Congruence				
Face (all congruent vs. prosodic incongruent)	0.04	0.04	0.87 (80.4)	.819
Face (all congruent vs. semantic incongruent)	−0.02	0.04	−0.35 (81.2)	.985
Face (all congruent vs. facial incongruent)	0.03	0.04	0.58 (81.3)	.937
Prosody (all congruent vs. prosodic incongruent)	−0.16	0.04	−3.70 (85.7)	.002
Prosody (all congruent vs. semantic incongruent)	−0.13	0.04	−3.08 (81.8)	.015
Prosody (all congruent vs. facial incongruent)	−0.08	0.04	−1.77 (81.7)	.297
Semantics (all congruent vs. prosodic incongruent)	−0.08	0.04	−1.83 (85.6)	.266
Semantics (all congruent vs. semantic incongruent)	−0.10	0.04	−2.34 (86.4)	.097
Semantics (all congruent vs. facial incongruent)	−0.002	0.04	−0.04 (81.9)	1.000

Note. The facial task and the congruent condition were used as the default level of task and congruence, respectively. When conducting pairwise comparison between prosodic and semantic tasks, prosody was set as the baseline level. *df* = degrees of freedom.

Figure 4. Accuracy in (a) rationalized arcsine units (raus) and (b) mean reaction time in different tasks and congruency conditions averaged across participants in Experiment 2. In (a), accuracy in raus is displayed in the violin plots, with data distribution shape indicated by the density plots, mean values represented by the black dots, and 95% confidence intervals shown by the error bars. In (b), mean reaction time is displayed in the bar charts with error bars showing 95% confidence intervals.



reaction time. As in Experiment 1, we also found the facilitation effects posed by cross-channel congruency. Interestingly, such a congruence-induced facilitation effect holds true only when the emotion conveyed by semantic content is incongruent with the other two channels, but not when prosody or face serves as the incongruent channel, which also indicates the perceptual salience of paralinguistic cues in emotion processing.

General Discussion

Findings of the Current Study

In this study, we investigated the relative perceptual saliency of linguistic and paralinguistic cues in emotional processing. We directly contrasted prosodic with semantic signals of emotion in a two-dimensional cross-channel auditory Stroop task in Experiment 1, and increased task difficulty and modality by including visual facial expression as a third dimension in Experiment 2. As hypothesized, prosody is intrinsically more salient than lexical semantic content in emotion identification. Our data showed that prosody played a predominant role in eliciting more rapid responses in both experiments. Prosody also interfered with correct identification of emotion words even when the attended channel was semantic content and when facial emotional expression, a more perceptually transparent channel, was concurrently presented. However, prosody failed to take precedence over facial expression in the audiovisual task, which is somewhat contrary to our second hypothesis of auditory dominance in emotion processing. Furthermore, cross-channel and cross-modal congruence enhanced emotion identification performance with a smaller facilitation effect in facial and prosodic domains than the semantic domain. These findings clearly demonstrate the important role of paralinguistic cues in multimodal emotion identification.

Perceptual Salience of Paralinguistic Information

In daily speech communication, interlocutors not only utter the words but also convey a certain thought, attitude, or intention. The successful interpretation of both what is said (literal meaning) and what is implicated (intended meaning; Grice, 1975) requires the listeners to pay special attention to the paralinguistic messages in an utterance, that is, how meaning is conveyed beyond literal interpretation. These paralinguistic cues, provided either vocally (e.g., prosody) or nonvocally (e.g., facial expressions, gestures, movements) by the speakers, disclose their emotional states as well. Previous developmental studies on perception of tone of voice (Filippi et al., 2017; Gil et al., 2014; Morningstar et al., 2018; Sauter et al., 2010; Scherer, 2003) and facial expressions (Adolphs, 2002; Ekman & Cordaro, 2011; Hoehl & Striano, 2010) have demonstrated that the ability to efficiently decode emotions based on the paralinguistic information shows its presence in early adulthood. Our results confirm and extend previous work to cross-channel and cross-modal experiments (Beall & Herbert, 2008; Kitayama & Ishii, 2002), revealing that paralinguistic signals in both auditory and visual modalities are perceptually more salient than purely linguistic indicators of emotive meaning for the Chinese participants. Notably, paralinguistic messages are so perceptually salient that they can even counteract the congruence-induced facilitation effect, which functions only when incongruent information is provided in the semantic channel rather than through the prosodic or facial channel. This is plausible due to the interconnections between different paralinguistic signals, which jointly provide a conceptual space that biases the processing of linguistic content. There has been evidence demonstrating that emotional prosody is interrelated with its corresponding facial expressions, and they both can activate a prototypical understanding of discrete emotions (Gil et al., 2014; Pell, 2005; Rigoulot & Pell, 2014). Other potential factors accounting for the perceptual disadvantage of semantic content might be attributed to the unique features in Chinese language and culture. For instance, Mandarin Chinese is a homonym-dense language without the direct grapheme–phoneme correspondences as found in alphabetic languages (C. Williams & Bever, 2010), which may require the listeners to spare more cognitive efforts in resolving potentially ambiguous words. In addition, Chinese participants, who were born and bred in an East Asian country with high-context culture, might show greater propensity to rely on implicit emotive mechanisms through contextual and nonverbal cues rather than by explicit verbal content (Hall, 1976).

Predominance of Prosody Over Semantic Content

As an indispensable paralinguistic carrier of the speech signal, emotional prosody perceptually dominates over semantic content in our two experiments, which replicates the findings in a series of previous research (Ben-David et al., 2016; Filippi et al., 2017; Kim & Sumner, 2017; Schirmer & Kotz, 2003; Schwartz & Pell, 2012). It is likely that stimulus presentation pattern in the current research engenders faster responses in prosodic tasks than semantic ones. That is, neither of our experiments can rule out the possibility that the prosodic cue may be identified at an earlier point in the stimulus than the lexical content. Since we adopted disyllabic words as targets of emotion processing, the semantic processing would require waiting until the second syllable is heard and processed before making a semantically based decision of the word. However, the prosody is integrated constantly as time unfolds and therefore may be available to the listener earlier than the presentation of the second syllable. Thus, it remains unclear whether the reaction time differences emerge due to the cross-channel/modal differences in the time at which the stimuli can be identified or processed.

Another possible explanation for prosodic salience is that, as Mandarin Chinese is a tonal language in which pitch variations can denote lexical meaning differences, Chinese speakers may have transferred their sensitivity to pitch patterns from lexical identification to emotion processing (Liu et al., 2015). Crucially, such prosodic salience persists despite an increase in task difficulty. Collectively, our behavioral findings provide a coherent picture with previous neurophysiological investigations. These studies have documented that early emotion decoding, as indexed by P200 component (Paulmann et al., 2013) with middle superior temporal sulcus activation (Wildgruber et al., 2006) for suprasegmental acoustic processing, is largely independent of task demand.

Predominance of Facial Expression Over Prosody

Although emotional prosody dominates over and interferes with semantic processing in emotion identification, its function in emotion identification does not appear to be as strong in the presence of visual facial expression in Experiment 2. This result corroborates a visual salience/dominance hypothesis, which is somewhat unexpected due to a long-held assumption that participants in East Asian countries, such as China and Japan, would be more attuned to vocal cues in audiovisual emotion processing (Ishii et al., 2003; Liu et al., 2015; Tanaka et al., 2010).

One potential account for this inconsistency could be that, while previous studies generally contrasted prosody and face in a bimodal emotion recognition task, we adopted a relatively novel Stroop-like protocol with three communication channels concurrently presented, thereby increasing task complexity and demand. In other words, few previous research has considered the influence of emotional semantic content when examining audiovisual salience effects, as the stimuli in previous studies usually contained neutral semantic content or were even devoid of linguistic meaning (e.g., pseudosentences; Liu et al., 2015; Tanaka et al., 2010). However, in the three-dimensional semantics–prosody–face Stroop task in Experiment 2, it is very likely that emotional prosody processing is encumbered by the co-occurring semantic information in the vocal channel with the complexity of cross-channel auditory processing of emotion outweighed by the simplicity and speed of concurrent and unambiguous visual signal processing. Additionally, there are some built-in asymmetries between the two modalities, with facial expressions presented as static images and emotional prosody embedded in the dynamic audio signal, which may have also contributed to the aforementioned differences in auditory and visual processing of emotion. Thus, future studies can employ more ecologically valid stimulus design such as video input synchronized with the audio signal.

On the other hand, the discrepancy between the current study and the previous ones reporting higher reliance on auditory signals in East Asians could arise from the differences in the arousal and valence of emotional stimuli. A larger cross-modal congruence effect in facial emotion identification has been consistently observed in the study conducted by Tanaka et al. (2010) using stimuli with high emotional arousal (i.e., happiness and anger) and by Liu et al. (2015) using stimuli with congruent hedonic valence (i.e., fear and sadness as unpleasant emotions). By contrast, the discrete emotional categories we selected for the current study (i.e., happiness and sadness) are incongruent in terms of both arousal and valence, thus triggering different processing patterns from the previous reports.

Limitations

There are several limitations in the current study when interpreting the findings. First, though the stimuli were displayed concurrently through different communication channels in both Stroop tasks, the behavioral measures do not reveal online processing of cross-channel/model information. It is unclear whether the observed trend in reaction time arises from differences in the speed with which the cues are processed or from the time at which the cue becomes available. As stated above, it might be the case that the asymmetries across channels (e.g., two-syllable words processed until the end vs. prosodic signals occurring from the beginning) and modalities (e.g., static photos of facial expression vs. dynamic auditory prosodic cues) could bias emotion processing. Second, though we transformed the accuracy data into raus to handle the ceiling effects statistically, the near-perfect scores obtained by the participants might limit how we interpret the findings. Furthermore, the evaluation of processing sensitivity, speed, and accuracy of emotional speech needs to take into account multiple factors, including other types of basic emotions (e.g., fear, anger, disgust; Herbener et al., 2007; Paulmann et al., 2013), forms of stimuli (e.g., syllables, nonwords, sentences, music, videos; Rao et al., 2013), representation of emotional information (e.g., emotion-laden vs. emotion-labeled stimuli; Altarriba & Basnight-Brown, 2010), participant characteristics (e.g., second or foreign language learners, mentally impaired patients, young and old people; Agustí et al., 2017; Eilola & Havelka, 2010; Lin et al., 2018; Sutton et al., 2007; L. E. Williams et al., 2010), and language background including dialectal influences (e.g., Cantonese; Cheang & Pell, 2011).

In future studies, it will be beneficial for us to include electrophysiological techniques, which allows detailed examination of the online temporal dynamics of different sensory modalities for emotion processing. For instance, event-related potential measures can track the sensory and cognitive stages of cross-modal emotion processing millisecond by millisecond, thus enriching our understanding of the neural correlates of the paralinguistic salience effect and congruence-induced facilitation effect (Zhu et al., 2010). We also expect to see more investigations extending from laboratory settings to real-world emotional communication settings, involving a wider range of task difficulty levels and covering more types of emotion, forms of stimuli, modes of emotion representation, groups of participants, and diversity of language, which can reflect evolutionary reality and ecological validity (Mitchell & Rossell, 2014).

Implications

Traditional models of lexical access have scarcely considered the role of paralinguistic cues in language comprehension (Marslen-Wilson, 1987; McClelland & Elman, 1986), while most emotion-related theories in psychology and psychiatry have been supported by evidence of nonverbal behavioral or physiological signals and viewed language as an independent conceptual process (Barrett et al., 2007; Brosch et al., 2010; Hinojosa et al., 2009; Marslen-Wilson, 1987; McClelland & Elman, 1986). Thus, linguistic and paralinguistic messages have largely been disassociated on previous theoretical accounts, which could be refined to some extent by the comparison of relative salience of semantic, prosodic, and facial cues in the current research. One crucial implication of our study is that emotion word processing is interwoven with and even shaped by paralinguistic dimension of communication, and congruence-induced facilitation effects reflect how these dimensions converge and achieve cross-channel correspondences (Nygaard et al., 2009).

The current study provides important insights for multimodal communication practices in real-life settings. The salient role of paralinguistic cues in emotion understanding as previously identified in studies conducted in western countries speaking Indo-European languages is now shown to generalize to a syllable-timed tonal language (i.e., Mandarin Chinese) with a typical high-context East Asian Chinese culture that relies heavily on contextual cues and implicit messages (Hall, 1976). This sheds light on how interlocutors in a high-context culture resolve discrepancies and ambiguity among multiple sources of information by manipulating their selective attention. In addition to basic research on healthy participants, more applied work with cross-channel/modal emotion processing can be conducted to find characteristic patterns to be utilized to facilitate the diagnosis and therapeutic intervention for people who show disadvantages in emotion processing, language comprehension, and social adaptation. For instance, individuals with hearing loss have degraded auditory input and can benefit more from combining visual information in emotion perception and speech processing. However, much more work is needed to discover how to optimize audiovisual training to mitigate the negative effects of hearing impairment (Picou et al., 2018; Yu et al., 2017). If successful, clinical applications have the potential to shape the intervention trajectory of emotion cognition and advance speech communication and social life for a large number of special populations such as patients with psychotic disorders (e.g., schizophrenia, autism, Alzheimer's dementia), people with hearing impairments (e.g., severe–profound hearing loss, recipients of cochlear implants), the elderly people, and children with learning disabilities (e.g., dyslexia; Agustí et al., 2017; de Jong et al., 2009; Diamond & Zhang, 2016; Irwin & DiBlasi, 2017).

Conclusion

In summary, the current research examined the relative perceptual salience of multisensory emotional information with a cross-channel auditory-alone task and a cross-modal audiovisual emotion Stroop task. Results indicated that paralinguistic signals, including tone of voice and facial expression, are more perceptually salient than linguistic messages (i.e., lexical semantic content) in multimodal emotion processing. Congruent emotions across channels and modalities lead to more rapid and accurate identification performance to a great extent, but this facilitation effect is attenuated by the perceptual saliency of paralinguistic signals (i.e., face and prosody). Our findings from the Chinese speakers support the paralinguistic salience effect and congruence-induced facilitation effect in cross-modal emotion integration, which provides the cross-linguistic justification for further investigations to reveal its neural basis and temporal dynamics with potential clinical applications.

Acknowledgments

This study was supported by grants (awarded to Ding and Zhang) from the major project of National Social Science Foundation of China (18ZDA293). Zhang additionally received support for international exchange and research from the University of Minnesota's Grand Challenges Exploratory Research Grant.

References

Adolphs, R. (2002). Neural systems for recognizing emotion. *Current Opinion in Neurobiology, 12*(2), 169–177. https://doi.org/10.1016/S0959-4388(02)00301-X

Agustí, A. I., Satorres, E., Pitarque, A., & Meléndez, J. C. (2017). An emotional Stroop task with faces and words. A comparison of young and older adults. *Consciousness and Cognition, 53,* 99–104. https://doi.org/10.1016/j.concog.2017.06.010

Altarriba, J., & Basnight-Brown, D. M. (2010). The representation of emotion vs. emotion-laden words in English and Spanish in the affective Simon task. *International Journal of Bilingualism, 15*(3), 310–328. https://doi.org/10.1177/1367006910379261

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

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

Barrett, L. F., Lindquist, K. A., & Gendron, M. (2007). Language as context for the perception of emotion. *Trends in Cognitive Sciences, 11*(8), 327–332. https://doi.org/10.1016/j.tics.2007.06.003

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

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

Ben-David, B. M., Multani, N., Shakuf, V., Rudzicz, F., & van Lieshout, P. H. (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

Brosch, T., Pourtois, G., & Sander, D. (2010). The perception and categorisation of emotional stimuli: A review. *Cognition and Emotion, 24*(3), 377–400. https://doi.org/10.1080/02699930902975754

Cheang, H. S., & Pell, M. D. (2011). Recognizing sarcasm without language: A cross-linguistic study of English and Cantonese. *Pragmatics & Cognition, 19*(2), 203–223. https://doi.org/10.1075/pc.19.2.02che

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

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception & Psychophysics, 25*(4), 345–347. https://doi.org/10.3758/BF03198814

de Gelder, B., Vroomen, J., de Jong, S. J., Masthoff, E. D., Trompenaars, F. J., & Hodiamont, P. (2005). Multisensory integration of emotional faces and voices in schizophrenics. *Schizophrenia Research, 72*(2–3), 203. https://doi.org/10.1016/j.schres.2004.02.013

de Jong, J. J., Hodiamont, P. P. G., & de Gelder, B. (2010). Modality-specific attention and multisensory integration of emotions in schizophrenia: Reduced regulatory effects. *Schizophrenia Research, 122*(1–3), 143. https://doi.org/10.1016/j.schres.2010.04.010

de Jong, J. J., Hodiamont, P. P. G., Van den Stock, J., & de Gelder, B. (2009). Audiovisual emotion recognition in schizophrenia: Reduced integration of facial and vocal affect. *Schizophrenia Research, 107*(2–3), 286–293. https://doi.org/10.1016/j.schres.2008.10.001

de Silva, L. C., Miyasato, T., & Nakatsu, R. (1997). Facial emotion recognition using multi-modal information. In*Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications* (Vol. 1, pp. 397–401). Institute of Electrical and Electronics Engineers.

Diamond, E., & Zhang, Y. (2016). Cortical processing of phonetic and emotional information in speech: A cross-modal priming study. *Neuropsychologia, 82,* 110–122. https://doi.org/10.1016/j.neuropsychologia.2016.01.019

Egeth, H. E., & Sager, L. C. (1977). On the locus of visual dominance. *Perception and Psychophysics, 22*(1), 77–86. https://doi.org/10.3758/BF03206083

Eilola, T. M., & Havelka, J. (2010). Behavioural and physiological responses to the emotional and taboo Stroop tasks in native and non-native speakers of English. *International Journal of Bilingualism, 15*(3), 353–369. https://doi.org/10.1177/1367006910379263

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

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

Gil, S., Aguert, M., Bigot, L. L., Lacroix, A., & Laval, V. (2014). Children's understanding of others' emotional states. *International Journal of Behavioral Development, 38*(6), 539–549. https://doi.org/10.1177/0165025414535123

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). Academic Press.

Hall, E. T. (1976). *Beyond culture*. Anchor Press.

Herbener, E. S., Rosen, C., Khine, T., & Sweeney, J. A. (2007). Failure of positive but not negative emotional valence to enhance memory in schizophrenia. *Journal of Abnormal Psychology, 116*(1), 43–55. https://doi.org/10.1037/0021-843X.116.1.43

Hinojosa, J. A., Carretié, L., Valcárcel, M. A., Méndez-Bértolo, C., & Pozo, M. A. (2009). Electrophysiological differences in the processing of affective information in words and pictures. *Cognitive, Affective & Behavioral Neuroscience, 9*(2), 173–189. https://doi.org/10.3758/CABN.9.2.173

Hoehl, S., & Striano, T. (2010). The development of emotional face and eye gaze processing. *Developmental Science, 13*(6), 813–825. https://doi.org/10.1111/j.1467-7687.2009.00944.x

Irwin, J., & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. *Language and Linguistics Compass, 11*(3), 77–91. https://doi.org/10.1111/lnc3.12237

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

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

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

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

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

Liu, P., Rigoulot, S., & Pell, M. D. (2015). Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia, 67,* 1–13. https://doi.org/10.1016/j.neuropsychologia.2014.11.034

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Mitchell, R. L. C., & Rossell, S. L. (2014). Perception of emotion-related conflict in human communications: What are the effects of schizophrenia. *Psychiatry Research, 220*(1), 135–144. https://doi.org/10.1016/j.psychres.2014.07.077

Morningstar, M., Nelson, E. E., & Dirks, M. A. (2018). Maturation of vocal emotion recognition: Insights from the developmental and neuroimaging literature. *Neuroscience & Biobehavioral Reviews, 90,* 221–230. https://doi.org/10.1016/j.neubiorev.2018.04.019

Niedenthal, P. M. (2007). Embodying emotion. *Science, 316*(5827), 1002–1005. https://doi.org/10.1126/science.1136930

Nygaard, L. C., Cook, A. E., & Namy, L. L. (2009). Sound to meaning correspondences facilitate word learning. *Cognition, 112*(1), 181–186. https://doi.org/10.1016/j.cognition.2009.04.001

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

Ovaysikia, S., Tahir, K. A., Chan, J. L., & DeSouza, J. F. (2011). Word wins over face: Emotional Stroop effect activates the frontal cortical network. *Frontiers in Human Neuroscience, 4,* 234. https://doi.org/10.3389/fnhum.2010.00234

Paulmann, S., Bleichner, M., & Kotz, S. A. (2013). Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology, 4*(345). https://doi.org/10.3389/fpsyg.2013.00345

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. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion, 35*(2), 192–201. https://doi.org/10.1007/s11031-011-9206-0

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

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

Qualtrics. (2018). *Qualtrics*. https://www.qualtrics.com

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology, 16*(2), 143–160. https://doi.org/10.1007/s10772-012-9172-2

Rigoulot, S., & Pell, M. D. (2014). Emotion in the voice influences the way we scan emotional faces. *Speech Communication, 65,* 36–49. https://doi.org/10.1016/j.specom.2014.05.006

Sander, D., Grandjean, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., & Vuilleumier, P. (2005). Emotion and attention interactions in social cognition: Brain regions involved in processing anger prosody. *NeuroImage, 28*(4), 848–858. https://doi.org/10.1016/j.neuroimage.2005.06.023

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

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1–2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

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

Schmid, C., Büchel, C., & Rose, M. (2011). The neural basis of visual dominance in the context of audio-visual object processing.

NeuroImage, 55(1), 304–311. https://doi.org/10.1016/j.neuroimage.2010.11.051

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

Spence, C. (2009). Explaining the Colavita visual dominance effect. *Progress in Brain Research, 176,* 245–258. https://doi.org/10.1016/S0079-6123(09)17615-X

Spence, C., Parise, C., & Chen, Y. C. (2012). The Colavita visual dominance effect. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes*. CRC Press/Taylor & Francis. https://doi.org/10.1201/b11092-34

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

Studebaker, G. A. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research, 28*(3), 455–462. https://doi.org/10.1044/jshr.2803.455

Sutton, T. M., Altarriba, J., Gianico, J. L., & Basnight-Brown, D. M. (2007). The automatic access of emotion: Emotional Stroop effects in Spanish–English bilingual speakers. *Cognition and Emotion, 21*(5), 1077–1090. https://doi.org/10.1080/02699930601054133

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I feel your voice: Cultural differences in the multisensory perception of emotion. *Psychological Science, 21*(9), 1259–1262. https://doi.org/10.1177/0956797610380698

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.

Wildgruber, D., Ackermann, H., Kreifelts, B., & Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: fMRI studies. *Progress in Brain Research, 156,* 249–268. https://doi.org/10.1016/S0079-6123(06)56013-3

Williams, C., & Bever, T. (2010). Chinese character decoding: A semantic bias? *Reading and Writing, 23*(5), 589–605. https://doi.org/10.1007/s11145-010-9228-0

Williams, L. E., Light, G. A., Braff, D. L., & Ramachandran, V. S. (2010). Reduced multisensory integration in patients with schizophrenia on a target detection task. *Neuropsychologia, 48*(10), 3128–3136. https://doi.org/10.1016/j.neuropsychologia.2010.06.028

Yow, W. Q., & Markman, E. M. (2011). Bilingualism and children’s use of paralinguistic cues to interpret emotion in speech. *Bilingualism: Language and Cognition, 14*(4), 562–569. https://doi.org/10.1017/S1366728910000404

Yu, L., Rao, A., Zhang, Y., Burton, P. C., Rishiq, D., & Abrams, H. (2017). Neuromodulatory effects of auditory training and hearing aid use on audiovisual speech perception in elderly individuals. *Frontiers in Aging Neuroscience, 9,* 30. https://doi.org/10.3389/fnagi.2017.00030

Zhu, X. R., Zhang, H. J., Wu, T. T., Luo, W. B., & Luo, Y. J. (2010). Emotional conflict occurs at an early stage: Evidence from the emotional face-word Stroop task. *Neuroscience Letters, 478*(1), 1–4. https://doi.org/10.1016/j.neulet.2010.04.036

Appendix A

List of Word Stimuli

Table A1. Spoken stimuli adopted in Experiments 1 and 2.

	Emotional prosody			
	Happy		Sad	
Emotional semantic content	Male speaker	Female speaker	Male speaker	Female speaker
Happy	有趣 (interesting)	开心 (happy)	安康 (sound)	欢喜 (joyful)
	融洽 (harmonious)	快乐 (pleased)	得意 (elated)	知足 (contented)
	喜悦 (delightful)	高兴 (glad)	舒适 (comfortable)	安宁 (peaceful)
	舒畅 (carefree)	愉快 (cheerful)	热切 (cordial)	如愿 (gratified)
Sad	伤心 (sad)	伤感 (melancholic)	悲痛 (sorrowful)	阴暗 (gloomy)
	压抑 (repressed)	失落 (disappointed)	消沉 (low-spirited)	难过 (upset)
	悲哀 (mournful)	哀愁 (grieved)	苦涩 (bitter)	黯淡 (downcast)
	凄凉 (dismal)	心痛 (heart-broken)	苦闷 (sulky)	抑郁 (depressive)

Appendix B

Procedures for Developing Stimuli Used in Experiments 1 and 2

The preparation of the materials used in the two experiments involved three basic steps. Stimuli were first developed respectively for three communication channels, namely, semantic content, prosody, and facial expression. The semantic content set contained 48 Chinese adjectives, most of which were selected from the Chinese Affective Words System (Wang et al., 2008), a widely used database providing established norms for valence, arousal, dominance rating, and frequency of words in the Chinese language. One half of the items were synonyms of "happy," whereas the other half were synonyms of "sad." In order to produce the stimuli in a prosodic channel, these emotional words were recorded from three male and three female student amateur actors who were native speakers of Mandarin Chinese. The speakers invited for recording were instructed to enunciate the emotional words clearly at a comfortable rate in both happy and sad prosody and not allow the word meaning to affect their tone of voice. The spoken words were recorded in a quiet laboratory setting and digitized into a Macbook Pro computer with AVID Mbox Mini at a sampling rate of 44.100 kHz. Faces consisted of 48 highly validated black and white photographs of actors (half of them were men and the other half were women) showing happy or sad expressions. They were chosen from another well-established emotional database, the Chinese Affective Picture System (Bai et al., 2005).

Afterward, a norming study was conducted with the aim of selecting the materials that were relatively familiarized, correctly identified as for emotional valence and similarly rated in terms of intensity by most of the participants. A group of 24 native speakers of Mandarin Chinese (13 women and 11 men; $M_{\text{age}} = 24.1$, $SD = 2.4$) were invited to complete an online questionnaire designed and distributed through Qualtrics using their own computers (Qualtrics, 2018). Participants were first instructed to judge whether they were familiar with the emotional adjectives, which were expressed in neutral prosody (0 = *not familiar*, 6 = *very familiar*). Then, it was followed with three blocks requiring them to identify the valence in a two-choice task ("happy" or "sad") and rate the intensity of the emotion on a 7-point Likert scale (0 = *not intense*, 6 = *very intense*) expressed by three communication channels: (a) the semantic content displayed in written text, (b) the prosody of the emotional words played as auditory stimuli, and (c) the facial expression illustrated in visual materials. The presentation order of three experimental blocks focusing on emotional valence and intensity expressed in different sensory channels was counterbalanced across participants.

Finally, the validated materials were adopted as stimuli in both experiments based on three criteria: (a) The spoken words in neutral prosody should have received an averaged rating of > 3 for familiarity; (b) 90% (corresponding to 22 of 24 participants) or more of participants should have correctly recognized the expressed emotion as "happy" or "sad"; and (c) the displays should have received an averaged rating of > 3 for intensity. Males and females did not differ in their identification accuracy and intensity rating ($p > .05$). The collected data for the finally included stimuli are shown in Tables B1 and B2.

Table B1. Familiarity rating for the emotional words used in Experiment 1 and 2.

Emotion type	*M*	*SD*
Happy	5.3	0.4
Sad	5.1	0.3

Note. Participants rated the familiarity of the emotional words on a 7-point scale (0 = *not familiar*, 6 = *very familiar*).

Table B2. Identification accuracy of emotional valence and rating of emotional intensity for stimuli adopted in Experiment 1 and 2.

		Identification accuracy of emotional valence		Rating of emotional intensity	
Stimulus type	Emotion type	*M*	*SD*	*M*	*SD*
Face	Happy	99.5%	1.4%	4.2	0.6
	Sad	99.2%	2.2%	3.7	0.5
Prosody	Happy	96.3%	3.2%	4.3	0.4
	Sad	95.6%	3.7%	3.9	0.5
Semantic content	Happy	99.0%	1.8%	4.2	0.7
	Sad	99.5%	1.4%	4.5	0.6

Note. Participants identified the emotional valence of the materials in a two-choice task ("happy" or "sad") and rated the intensity on a 7-point scale (0 = *not intense*, 6 = *very intense*).