Semantics Outperforms Prosody in Emotional Speech Processing: Evidence from a Complex Stroop Experiment

Jing Qi, Kaile Zhang, Gang Peng

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Semantics Outperforms Prosody in Emotional Speech Processing: Evidence from a Complex Stroop Experiment

Jing Qi[1]**, Kaile Zhang**[1]**, Gang Peng**[1]
[1] Research Centre for Language, Cognition, and Neuroscience,
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
jing.qi@connect.polyu.hk
kaile-keller.zhang@polyu.edu.hk
gang.peng@polyu.edu.hk

Abstract

Semantic and prosodic cues both play crucial roles in conveying feelings and emotions in speech communication. Previous studies on the salience effects in emotional speech processing have shown inconsistent results. Most past research has focused on two simple categories of emotion. In this study, we investigated the perceptual saliency of the two cues in Mandarin using semantics-prosody Stroop tasks involving seven basic emotions: happiness, sadness, anger, fear, disgust, surprise, and neutrality. The results, based on 36 normal Chinese adults, demonstrated a semantic salience effect. This suggests that individuals may rely more on semantic cues when integrating emotional speech across different channels in more complex and challenging situations.

1 Introduction

Emotion is an integral part of human language communication. To understand the emotion of speakers, various cues are integrated. In the auditory modality, semantics and prosody are two crucial channels. Semantic cues refer to the emotional meanings inherent in the speech contents, while prosodic cues include phonetic features such as duration, pitch, and intensity. Both cues play a role in emotion processing, but they can express the same or different states at the same time. Saying "I'm very happy!" with an angry tone of voice is one example. In this instance, there is a disagreement between the two information channels. People might rely more on one of the two cues for verbal emotion processing.

The inequality among channels of information mentioned above, commonly referred to as the sensory dominance or salience effect (Colavita, 1974). Stroop task is frequently used to investigate the salience effect of different channels. A typical Stroop test utilizes color words and word colors as two perceptually congruent (e.g. word “red” in red) or incongruent (e.g. word “red” in blue) dimensions. Participants are instructed to identify one dimension while ignoring the other (Stroop, 1935). Congruency effects indicate the semantic correspondence between the two dimensions, while task effects reveal the asymmetry of the two channels. The presence, magnitude, and direction of Stroop effects are modulated by both dimensional relatedness and imbalance (Melara & Algom, 2003).

The Stroop-like paradigm has been adapted to investigated channels in emotion processing. Participants are often asked to focus on the emotion of one channel while disregarding information from the other. Such research has gained consensus regarding the congruency effect in emotion processing, that is, congruent stimuli elicit faster and more accurate responses. (Barnhart et al., 2018; Lin et al., 2020; Pell, 2005; Schirmer et al., 2005; Schwartz & Pell, 2012;). However, findings regarding the sensory dominance effect of communication channels are mixed. Some researchers found a processing saliency of the semantic meaning over prosody(Kitayama & Ishii, 2002; Pell et al., 2011), while others claimed the predominance of prosodic cues (Ben-David et al., 2016; Filippi et al., 2017; Kim & Sumner, 2017; Lin et al., 2020).

The discrepancies regarding the perceptual salience of prosodic and semantic channels may stem from cultural backgrounds and experimental settings (Lin et al., 2020). Studies have reported a greater emphasis on semantic salience in Western cultures (Grimshaw, 1998; Kitayama & Ishii, 2002; Pell et al., 2011), while prosody appears to take precedence among participants from Asian countries (Ishii et al., 2003; Lin et al., 2020; Liu et al., 2015). Additionally, experimental settings including stimulus，number of choices, and task difficulty also affects the channel and modality salience effect (Lin et al., 2020).

The inconsistencies in previous studies highlight the need for further investigation into the dominance effects of prosody and semantics. Firstly, more studies have focused on subjects from Western cultural contexts, with only a few studies have addressing tonal languages such as Chinese (Lin et al., 2020; Lin et al., 2021; Xiao &Liu, 2024). Secondly, the stimulus settings in many studies might be too simple. Some studies utilized binary choices of positive and negative emotions, employing positive or negative prosody to express corresponding words (Schirmer and Kotz, 2003; Sutton et al., 2007). Others have used two discrete emotion categories, such as happy and sad prosody to convey synonymous words of "happy" and "sad" (Lin et al., 2020; Filippi et al., 2017). The simplicity might create the imbalance of difficulty between semantic and prosodic tasks, as binary judgments based on semantic information from sound are often more challenging than those based on prosody. In natural conversation, people normally make decisions from a much richer array of emotional categories.

This study aims to investigate the salience of prosody or semantics in Emotional Speech Processing. In this study, we used a Stroop paradigm featuring a broader range of emotions to investigate the emotional speech perception of Mandarin native speakers based on semantic and prosodic cues. Referring to Ekman's basic emotion categories (Ekman, 1992), we selected seven emotion categories (neutral, happy, sad, angry, fearful, disgust, surprise) for the stimuli. 36 subjects were required to choose from seven options during the prosodic and semantic tasks. The accuracy rates and response times of the two tasks in both conditions will be recorded and compared. According to a previous study (Lin et al., 2020), Mandarin native speakers rely more on prosodic cue.

It is expected that the present study will contribute to the research objectives from two perspectives. Firstly, the study of Mandarin speakers may add information to explorations in high-context culture. Secondly, the complex experimental design can examine the process of emotion integration in scenarios that are closer to real-life situations.

2 Method

2.1 Participates

Thirty-six subjects (18 women and 18 men) completed all experimental tasks. Women had a mean age of 25.3 years (*SD* = 1.7), and men were also on average 25.3 years old (*SD* = 1.5). All participants were all native Mandarin speakers and postgraduate students. They all had normal or corrected-to-normal vision and were without any history of speech, language, hearing impairment or any neurological problem. The experiment was approved by the Institutional Review Board (IRB). Subjects completed written informed consent prior to inclusion in the experiment and were financially compensated for their time. The PolyU Institutional Review Board (IRB) approved of the ethics for this study (HSEARS20240818003).

2.2 Stimuli

The stimuli comprised 328 different sounds of disyllabic spoken words in Mandarin Chinese, representing seven types of emotions (including happy, sad, angry, fearful, disgust, surprise and neutral) in semantic contents and prosody simultaneously. Specifically, the stimuli were selected from a sound set consisting of 84 different semantic words across the seven emotional categories, each spoken with seven types of emotional prosody (see Table A1 in appendix A for examples of words corresponding to the seven emotions). Thus, the emotions of the two auditory channels in each stimulus could be congruent or incongruent. There were 76 congruent stimuli and 252 incongruent stimuli. (see Table A2 in appendix A for details of stimulus types) Each stimulus differed from the others in at least one of the two channels.

Semantic channel: The disyllabic words were sourced from the Affective Lexicon Ontology (Xu et al., 2008), which is an opensource database that categorizes words according to Ekman's 6 basic emotion categories (Ekman, 1992) and rates their emotional intensity. Words of 7 emotional groups were matched based on word frequency, utilizing the SUBTLEX-CH-WF (Cai & Brysbaert, 2010), which is derived from movie subtitles and thus reflects everyday spoken language effectively. Additionally, the semantics of the words were tested and evaluated by 15 native Mandarin speakers through an online task. In a forced-choice task with seven emotional categories, each category reached an accuracy rate exceeding 90%. In a word familiarity rating task, each word received an average familiarity score greater than 4 on a five-point scale (1=not familiar, 5=very familiar).

Prosodic channel: The words were produced in 7 types of emotional prosody by a professional broadcaster (male, age:30) who achieved the highest level on the Standard Mandarin Chinese Test. All sounds were tested and screened by five native Mandarin speakers who did not participate in Stroop experiments. For the 328 stimuli involved in this experiment, the accuracy of the prosody for each emotional category was above 92% in the forced-choice task with seven choices (ignoring the meaning of words). The average confidence level of the emotion for each stimulus was above 5.9 in 7 points scales.

To ensure the naturalness of the stimuli, we did not alter any of the original properties of the sounds. The statistics of the acoustic properties of the different emotions can be found in Table A3 in appendix A. Although there were differences in acoustic parameters between emotional types, all categories were covered in the experimental task in both the congruent and incongruent conditions to control the interference.

2.3 Procedure

In the experiment, subjects were asked to finish two Stroop tasks (a prosodic task and a semantic task) separated by 12 to 60 hours. The order of two tasks was balanced between subjects to avoid familiarity effect. In prosodic task, they needed to choose the emotion conveyed by prosody of the sounds from seven choices as quickly and accurately as possible while ignoring the semantics. In semantic tasks, the requirements reversed.

Each task included a practice session with 49 trials of unrepeated stimuli (7×7) before the formal test session. Participants were required to achieve 80% accuracy within a 5-second reaction time to ensure they could understand and follow the instructions. During the task, the location of the keys for the options remained constant but was randomized between subjects. The practice session also helped them become familiar with the key locations. There was no significant difference in the number of practice sessions for the two tasks (prosodic vs. semantic: 1.58 vs. 1.22). In the formal session, there were 420 trials divided equally into 14 blocks. 210 stimuli with incongruent emotional prosody and semantics were played once, while 70 congruent stimuli were repeated 3 times to equalize the number of stimuli in two conditions. The order of stimuli was completely randomized.

Each trial began with a fixation cross for 1000 ms, followed by a visual notice “Listen carefully!” (in Mandarin and English) displayed for 1000 ms to attract subjects’ attention. Stimulus would then be presented binaurally over headphones, with the options and requirements displayed on the screen simultaneously. Subjects were required to push the keys on a keyboard (fixed position for each emotion) as quickly as possible while maintaining accuracy to select the emotion conveyed in the attended channel. We recorded accuracy and response time from stimulus onset.

The experiment was conducted in a sound-insulated room with subjects seated in a comfortable chair approximately 70 cm from the monitor. The experimental program was written by E-Prime (version 3.0.3.80; Psychology Software Tools, 2012). Auditory stimuli were presented binaurally at 70 dB SPL through Audio-Technica headphones. Detailed instructions were included in the program before practice and formal sessions.

2.4 Statistical Analyses

Linear mixed-effects models were performed to analyze the data using R (Version 4.3.3; R Core Team, 2024) with the lme4 package (Bates et al., 2015). We focus on three variables separately: Accuracy (ACC), Response Time (RT), and Speed-Accuracy Tradeoff (SAT). Among the various methods of calculating the SAT, we used the Balanced Integration Score (BIS) proposed by Liesefeld et al. (2015), which integrates speed and accuracy with equal weights (Liesefeld & Janczyk, 2019).

Considering that RT data exhibits positive skewness, we performed a log transformation to RT data. The BIS data showed a clear left-skewed distribution, so we used Box-Cox transformation (Box & Cox, 1964; Sakia, 1992) to achieve a normally distributed BIS data.

In the linear mixed-effects models, ACC, the logarithm of RT, and transformed BIS were respectively entered as dependent variables. Congruency (congruent vs. incongruent) and task (semantic vs. prosodic) were entered as fixed factors, with congruent condition in prosodic task set as the default level. Subjects and items were entered as random intercepts. Tukey’s post hoc

Parameter	Any effect	Estimate	SE	*t*-value	*p*-value
Task	No	-0.007	0.004	-1.755	0.079
Congruency	Yes	-0.029	0.011	-2.600	0.0098**
Task× Congruency	Yes	0.019	0.006	3.371	<0.001***
Prosody - Semantics (Congruent)		0.007	0.004	1.662	0.344
Prosody - Semantics (Incongruent)		-0.012	0.004	-2.853	0.023*
Congruent - Incongruent (Prosody)		0.029	0.004	6.960	<0.001***
Congruent - Incongruent (Semantics)		0.010	0.004	2.445	0.069

Table 1: Linear mixed-effects model with accuracy as the dependent variable. (Significant codes: *p* < 0.05: '*'; *p* < 0.01 '**'; *p*<0.001 '***')

Parameter	Any effect	Estimate	SE	t-value	p-value
Task	Yes	0.038	0.005	8.017	<0.001***
Congruency	Yes	0.084	0.014	6.215	<0.001***
Task × Congruency	Yes	-0.024	0.007	-3.501	<0.001***
Prosody - Semantics (Congruent)		-0.039	0.006	-6.743	<0.001***
Prosody - Semantics (Incongruent)		-0.015	0.006	-2.516	0.057
Congruent - Incongruent (Prosody)		-0.085	0.006	-14.584	<0.001***
Congruent - Incongruent (Semantics)		-0.061	0.006	-10.469	<0.001***

Table 2: Linear mixed-effects model with the logarithm of RT as the dependent variable.

tests using the lsmeans package (Lenth, 2016) were conducted when there was a significant effect. The full models for ACC, RT and BIS analyses are represented as follows:

$$\begin{aligned} \mathrm{ACC} = {} & \beta_0 + \beta_1 \times \mathrm{Task} + \beta_2 \times \mathrm{Congruency} \\ & + \beta_3 \times \mathrm{Task} \times \mathrm{Congruency} \\ & + b_{\mathrm{items}} + b_{\mathrm{subjects}} + \varepsilon_{ij} \end{aligned} \tag{1}$$

$$\begin{aligned} \mathrm{RT}_{(\log)} = {} & \beta_0 + \beta_1 \times \mathrm{Task} + \beta_2 \times \mathrm{Congruency} \\ & + \beta_3 \times \mathrm{Task} \times \mathrm{Congruency} \\ & + b_{\mathrm{items}} + b_{\mathrm{subjects}} + \varepsilon_{ij} \end{aligned} \tag{2}$$

$$\begin{aligned} \mathrm{BIS}_{(\mathrm{transformed})} = {} & \beta_0 + \beta_1 \times \mathrm{Task} \\ & + \beta_2 \times \mathrm{Congruency} \\ & + \beta_3 \times \mathrm{Task} \times \mathrm{Congruency} \\ & + b_{\mathrm{subjects}} + \varepsilon_{ij} \end{aligned} \tag{3}$$

3 Results

The results of the linear mixed-effects models for ACC, RT and BIS are shown in Table1, Table 2 and Table 3 respectively.

3.1 Accuracy

Overall, the participants responded with high accuracy (*M* = 92.7%, *SD* = 2.7%). Figure 1 illustrates the accuracy data for the congruent and incongruent conditions in the two tasks, which are normally distributed according to the *Kolmogorov-Smirnov* test.

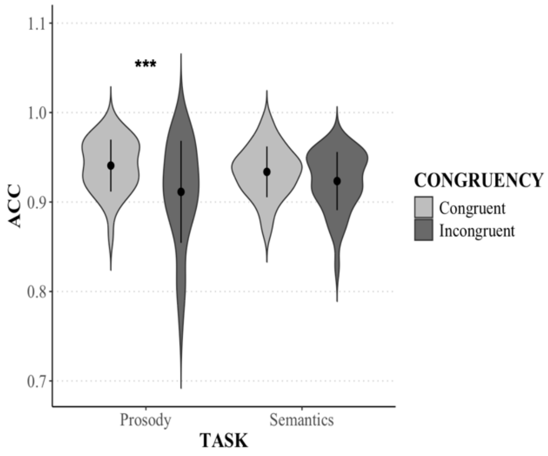

Figure 1: Accuracy in the two tasks and two congruence conditions**.**

Linear mixed-effects analyses showed no main effect of task, $\chi^2(1) = 3.08$, p >0.05. The main effect of congruency condition ($\chi^2(1)$ = 6.76, *p* =0.009) was significant. Congruent stimuli elicited more (2.0% ± 2.8%) accurate responses than incongruent ones (β_2 = -0.029, *SE* = 0.011, *t* = -2.6, *p* <0.01). There was an interaction effect between task and congruency($\chi^2(1)$ = 11.36, β_3 = 0.019, *SE* = 0.006, *t* =3.371, *p* <0.001).

Post hoc tests showed significantly higher accuracy for the semantic task in the incongruent condition (*t*=-2.853, *p*=0.023). The difference between the two conditions was significant in the prosodic task (*t*=6.96, *p* <0.001) but did not arise in

Parameter	Any effect	Estimate	SE	t-value	p-value
Task	No	-2.901	1.565	-1.853	0.067
Congruency	Yes	-7.301	1.565	-4.664	<0.001***
Task× Congruency	No	3.594	2.214	1.610	0.110
Prosody - Semantics (Congruent)		2.901	2.38	1.219	0.616
Prosody - Semantics (Incongruent)		-0.663	2.38	-0.279	0.992
Congruent - Incongruent (Prosody)		7.301	2.38	3.069	0.014*
Congruent - Incongruent (Semantics)		3.373	2.38	1.571	0.399

Table 3: Linear mixed-effects model with the transformed BIS as the dependent variable.

the semantic task (t = 2.445, p =0.069). This suggested that the conflicting information from semantics reduces correctness more than prosody.

3.2 Reaction Time

In analysis of RT data, incorrect responses and responses over 2 SDs from the mean were excluded (Baayen & Milin, 2010; Lin et al., 2020), which respectively accounted for 7.3% and 4% of the overall data set. Reaction time data in the two tasks and conditions are displayed in Figure 2. Reported mixed-effects analyses in Table 2 were conducted based on the logarithm of RT.

Analyses on the logarithm transformed reaction time showed main effects of task ($\chi^2(1) = 64.28$, $p < 0.001$), congruency (χ^2 (1) = 38.63, $p < 0.001$), and a significant interaction (χ^2 (1) = 12.26, $p < 0.001$). Participants responded 49 ± 228 ms faster to the prosody task than to the semantic task (β_1 = 0.038, SE = 0.005, t = 8.017, $p < 0.001$), and 132 ± 72 ms faster to the congruent stimuli (β_2 = 0.084, SE = 0.014, t = 6.215, $p < 0.001$).

Post hoc tests found that the response was faster in congruent condition for any task (prosodic task: t = -14.58, $p < 0.001$; semantic task: t = -10.47, p <0.001). Significant difference between tasks existed in the congruent condition (t = -6.74, $p < 0.001$), but disappeared in the incongruent condition (t = -2.52, p = 0.057). This suggests that the difference between the tasks narrowed after being disturbed by inconsistent messages. Further, it is likely that there was greater negative interference from semantics than prosody in the incongruent condition

3.3 Speed-Accuracy Tradeoff

We used the BIS as a parameter for the SAT, whose larger value indicates that the subject did better (Liesefeld & Janczyk, 2019). Reported analyses in Table 3 were based on the transformed BIS using Box-Cox transformation (Box & Cox, 1964).

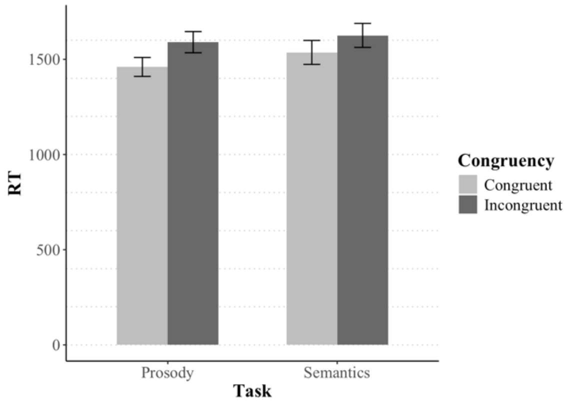

Figure 2: Reaction time in the two tasks and two congruence conditions**.**

Analyses only showed a main effect of congruency (χ^2 (1) = 21.75, $p < 0.001$). Participants responded better in congruent condition (β_2 = -7.3, SE = 1.57, t = -4.66, $p < 0.001$). Task effects were not significant (χ^2 (1) = 3.43, p = 0.064) and there was no interaction (χ^2 (1) = 2.59, p = 0.107). In the posttest, worse performance in the incongruent condition than congruent only occurred in the prosodic task (t = -3.07, p = 0.014), suggesting a significant role for semantic interference.

4 Discussion

To be clear, we summarize the results of the statistical analysis in Table 4. Combining the three parameters, the main effect of congruency remained significant, indicating that two channels with congruent information perform faster and more effectively than incongruent ones. This is not surprising, as people are more frequently exposed to congruent affective information conveyed by both channels in daily life (Nygaard & Queen, 2008).

We were particularly interested in the predominance effects of prosody and semantics. Our results showed that only the RT exhibited a main effect of the task, with prosody being

Parameter	Significant effect in LMM	*Post hoc test*	
		Prosody vs Semantics	**Congruent vs Incongruent**
ACC	Congruency, Task× Congruency	Better semantic task in incongruent condition	Better in congruent condition only in prosodic task
RT	Task, Congruency, Task× Congruency	Faster prosodic task in congruent condition	Faster in congruent condition
SAT(BIS)	Congruency	NS differences	Better in congruent condition only in prosodic task

Table 4: Summary for results of statistical analysis.

recognized significantly faster than semantics. However, this alone might be insufficient to prove that prosody plays a more important role in the identification process. This result might stem from the fact that prosodic signals are acquired earlier than semantic signals. Subjects need to finish listening to a disyllabic word before making a judgment about its semantics. This possibility was also reported in the study by Lin et al.(2020). Additionally, we found an interaction effect where the prosodic task was significantly faster only in the congruent condition. This might suggest a difference in the interference caused by inconsistent information from various channels. Further, semantic passages may cause greater latency, thereby negating the advantage of the faster speed of the prosodic task.

In the conflict condition, semantic interference was greater than that of prosodic interference. Results from both the ACC and SAT analyses support this conclusion. A significantly poorer performance in the conflict condition was observed only in the prosodic task. When judging semantics, no significant difference was found between the congruent and incongruent conditions. We might therefore conclude that there is a semantic salience effect in the emotional word processing.

We also confirmed the predominance of semantics by analyzing the incorrect responses. We counted the incorrect options that matched the emotion of the misleading channel in both tasks for each participant (i.e., in prosodic task, participants selected angry for a word “angry” spoken sadly). The proportions of matched incorrect options (P_{MIO}) among all incorrect options were calculated. A larger P_{MIO} indicates a stronger salience effect from the misleading channel. Using a paired t-test, we found the P_{MIO} in the prosodic task (P_{MIO_P} = 17.4%± 8.8%) was significantly larger than that in semantic task (P_{MIO_S} =11.4% ± 6.8%), with *t (35)* = 2.675, *p* = 0.011. Incorrect responses in the prosodic tasks are more influenced by semantics than vice versa.

Our results did not show prosody salience effect during emotional speech processing in Mandarin speakers, which differs from the findings of Lin et al. (2020). They used only two emotions, and prosody performed better than semantics. This suggests that changes in complexity due to the number of emotion categories might influence the strategies people use for emotional speech processing, leading to a shift in cue dominance. For example, in the study by Grimshaw (1998), stimuli involved words “mad” “sad” “glad”, “fad” expressed by four emotions (angry, sad, happy, and neutral). Increased reaction time and decreased accuracy in inconsistent conditions were observed only in the prosodic task. In other words, this indicated a semantic dominance effect.

The study suggests that the contrast in task difficulty might have an influence on the Stroop effect which cannot be ignored. In the two-choice task, participants could quickly establish the relationship between the acoustic features (e.g., pitch) of prosody and emotions, enabling them to respond without fully listening to the stimulus. The acoustic features of the semantic cues were more complex, and there were fewer repeated features to help subjects establish patterns. This imbalance in difficulty might affect their strategies in the task, leading to faster responses in prosodic task. In our study, the complexity of the options increased the difficulty level of the prosodic task, thereby equalizing the difficulties of the two tasks. After the experiment, subjects were asked to rate the difficulty of the two tasks using a 5-point scale (1=very simple, 5=very hard). There was no significant difference between the prosodic (2.78) and semantic (2.47) tasks. The increased difficulty also weakened the ceiling effect. In this context, the results of the study may be more robust.

This study has some limitations. Firstly, it has not clearly delineated how variations in complexity specifically influence the cues relied upon by native Mandarin speakers. Complexity can be affected both by the number of emotional categories involved in the task and by controlling the number of channels. Future research will explore this issue by changing more variables. Secondly, this study only utilized male speakers. Future investigations will include female speakers to provide a more comprehensive examination of gender differences and identity recognition in emotional perception processing. Lastly, as this study is purely behavioral, it has certain limitations. We will consider employing more brain imaging techniques for further exploration.

5 Conclusion

This study explored the salience effect of prosody and semantics in speech emotion processing through a complex stroop experiment. It was found that semantic information was more salient than prosody cues, evidenced by the greater influence of semantic information on prosodic judgments. Task difficulty was better controlled in this study, which may have yielded more robust results. Complex tasks are more relevant to real life than previous studies, therefore this study informs natural emotional speech processing and provides reference for exploring neural basis of emotional recognition with potential clinical applications.

Acknowledgments

This research was partially supported by a fellowship award from the Research Grants Council of the Hong Kong SAR, China (Project No. PolyU/RFS2122-5H01), and a research grant from the National Natural Science Foundation of China (NSFC: 12304526).

References

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12–28.

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

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

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211-243.

Cai Q, & Brysbaert M.(2010) SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*. 2010 Jun 2;5(6):e10729.

Colavita, F.B.(1974). Human sensory dominance. *Perception & Psychophysics*, 16(2), 409–412. https://doi.org/10.3758/ BF03203962

Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3), 550–553. https://doi.org/10.1037/0033-295X.99.3.550

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.

Grimshaw, G. M. (1998). Integration and interference in the cerebral hemispheres: Relations with hemispheric specialization. *Brain and Cognition*, *36*(2), 108-127.

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), 49–55.

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

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

Liesefeld, H. R., Fu, X., & Zimmer, H. D. (2015). Fast and careless or careful and slow? Apparent holistic processing in mental rotation is explained by speed-accuracy trade-offs. *Journal of Experimental Psychology Learning, Memory, and Cognition,* 41(4), 1140–1151

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods,* 51(1), 40–60.

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.

Lin Y, Ding H, Zhang Y. (2021). Unisensory and Multisensory Stroop Effects Modulate Gender Differences in Verbal and Nonverbal Emotion Perception. *Journal of Speech Language and Hearing Research*, 64(11):4439-4457.

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

Melara, R. D., & Algom, D. (2003). Driven by information: A tectonic theory of Stroop effects. *Psychological Review*, 110(3), 422–471.

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

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.

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.

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

R Core Team (2024). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The Statistician*, 41(2), 169-178.

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.

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.

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

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

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

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.

Xiao, C., & Liu, J. (2024). Semantic effects on the perception of emotional prosody in native and non-native Chinese speakers. *Cognition and Emotion*, 1-11.

Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the Affective Lexicon Ontology. *Journal of the China Society for Scientific and Technical Information*, 27(2), 180–185.

Appendices

Appendix A: Experimental Materials

Emotions	Examples of Words
Happy	愉快，快乐，开心，……
Sad	悲凉，忧伤，难过，……
Angry	恼火，发火，恼怒，……
Fearful	害怕，慌乱，吓人，……
Disgust	厌恶，厌倦，讨厌，……
Surprise	惊讶，奇妙，惊叹，……
Neutral	冷静，先生，开始，……

Table A1: Examples of words in 7 emotions

Note: There are at least 10 different disyllabic words in each emotion category. Except for the neutral category, words in other types are synonyms of emotion words.

		Prosody						
	Emotions	Neutral	Happy	Angry	Sad	Fearful	Disgust	Surprise
Semantics	Neutral	11	6	6	6	6	6	6
	Happy	6	11	6	6	6	6	6
	Angry	6	6	11	6	6	6	6
	Sad	6	6	6	11	6	6	6
	Fearful	6	6	6	6	11	6	6
	Disgust	6	6	6	6	6	11	6
	Surprise	6	6	6	6	6	6	10

Table A2: Counting of stimuli of different stimulus types

Note: The numbers in the table represent the number of different semantic words in each semantics*prosody category. There were no stimuli with the same word and prosody at the same time. 49 stimuli (containing seven semantic categories with seven prosodic categories) were used in practice session, in which only one (surprise word in surprise prosody) was repeated in the formal session.

Emotion Category	Grouped by Prosody			Grouped by Semantics		
	f0(Hz)	Intensity(dB)	Duration(ms)	f0(Hz)	Intensity(dB)	Duration(ms)
Happy	213(±40)	68(±2.5)	868(±115)	182(±49)	65(±4.3)	778(±110)
Sad	119(±32)	60(±3.2)	1053(±128)	192(±69)	66(±4.3)	825(±194)
Angry	235(±63)	68(±2.3)	705(±68)	186(±60)	66(±4.0)	741(±177)
Fearful	189(±34)	68(±3.5)	745(±109)	189(±53)	65(±3.9)	791(±181)
Disgust	168(±38)	65(±3.8)	612(±72)	182(±55)	65(±4.2)	725(±157)
Surprise	237(±43)	68(±2.4)	654(±80)	204(±70)	66(±4.3)	783(±192)
Neutral	150(±31)	64(±2.8)	773(±70)	178(±46)	66(±3.7)	770(±131)

Table A3: Mean (± SD) of acoustic parameters for different emotional subgroups

Note: N = 47 for most emotional categories except for the surprise category (N=46).