ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Attention Modulates the Role of Speakers' Voice Identity and Linguistic Information in Spoken Word Processing: Evidence From Event-Related Potentials

Yunxiao Ma,[a] Keke Yu,[a] Shuqi Yin,[a] Li Li,[b] Ping Li,[c] and Ruiming Wang[a]

[a] Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents, Ministry of Education, & Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China [b] The Key Laboratory of Chinese Learning and International Promotion, and College of International Culture, South China Normal University, Guangzhou, China [c] Department of Chinese and Bilingual Studies, Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong SAR, China

ARTICLE INFO

Article History:
Received July 19, 2022
Revision received December 4, 2022
Accepted January 20, 2023

Editor-in-Chief: Stephen M. Camarata
Editor: Susan Nittrouer

https://doi.org/10.1044/2023_JSLHR-22-00420

ABSTRACT

Purpose: The human voice usually contains two types of information: linguistic and identity information. However, whether and how linguistic information interacts with identity information remains controversial. This study aimed to explore the processing of identity and linguistic information during spoken word processing by considering the modulation of attention.
Method: We conducted two event-related potentials (ERPs) experiments in the study. Different speakers (self, friend, and unfamiliar speakers) and emotional words (positive, negative, and neutral words) were used to manipulate the identity and linguistic information. With the manipulation, Experiment 1 explored the identity and linguistic information processing with a word decision task that requires participants' explicit attention to linguistic information. Experiment 2 further investigated the issue with a passive oddball paradigm that requires rare attention to either the identity or linguistic information.
Results: Experiment 1 revealed an interaction among speaker, word type, and hemisphere in N400 amplitudes but not in N100 and P200, which suggests that identity information interacted with linguistic information at the later stage of spoken word processing. The mismatch negativity results of Experiment 2 showed no significant interaction between speaker and word pair, which indicates that identity and linguistic information were processed independently.
Conclusions: The identity information would interact with linguistic information during spoken word processing. However, the interaction was modulated by the task demands on attention involvement. We propose an attention-modulated explanation to explain the mechanism underlying identity and linguistic information processing. Implications of our findings are discussed in light of the integration and independence theories.

Speakers' voice is indispensable to everyday communication. Listeners could recognize speakers' identity and comprehend what speakers say via speakers' voice. In this regard, the voice consists of two important types of information: linguistic information and identity information (Scott, 2019). The linguistic information refers to speech acoustic features related to the phonological and semantic information the speaker expresses. The identity information consists of physical acoustic features related to speakers' identities, such as the fundamental frequency (f_o), formants, and formant transitions. Researchers can measure and visualize the two types of information via speech processing software like Praat (https://www.fon.hum.uva.nl/praat/). Many studies have discussed how listeners process identity and linguistic information (e.g., Belin et al., 2004; Feng et al., 2018; Seydell-Greenwald et al., 2020; Yu et al., 2019). However, how the two types of information interact with each other remains controversial.

Correspondence to Ruiming Wang: wangrm@scnu.edu.cn, and Ping Li: ping2.li@polyu.edu.hk. Yunxiao Ma and Keke Yu contributed equally to this paper and should be considered as co-first authors.
Disclosure: *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

The Independence Versus Integration Theories

Two competing theories, the independence versus integration theories, in the literature have explained the issue. The independence theory considered that the speakers' voice identity information plays a minor role in listeners' phonological and semantic processing of speakers' voices (Belin et al., 2004; Halle & Mohanan, 1985; Samuel, 2011). It is discarded or standardized during speech perception and does not play a role in processing the key linguistic information.

Much of the theory's evidence came from neurocognitive studies (Aglieri et al., 2021; Roswandowitz et al., 2018; Schirmer, 2018; Seydell-Greenwald et al., 2020; Whitehead & Armony, 2018). The studies suggested different brain regions involved in the identity and linguistic information processing. Specifically, according to these studies, the linguistic information processing mainly activated the left anterior/posterior superior temporal sulcus (STS) and inferior prefrontal area, while the identity information processing mainly involved the transverse temporal gyrus, temporal plane, anterior/mid/posterior superior temporal gyrus (STG)/sulcus, and part of the middle temporal gyrus (Aglieri et al., 2021; Belin et al., 2004; Blank et al., 2014; Roswandowitz et al., 2018; Schelinski et al., 2016).

The integration theory, unlike the independence theory, hypothesized that identity information would affect or directly participate in semantic processing (Holmes et al., 2018, 2021; Kapnoula & Samuel, 2019; Zhang & Chen, 2016). Listeners do not need a separate system for standardizing identity information. It is stored with the linguistic information together and helps listeners extract semantic information. For example, Kapnoula and Samuel (2019) required the participants to learn pseudowords associated with identity indices. The results showed that the participants performed better in words associated with fixed identity indices, suggesting that identity information could be encoded in lexical representation to help learners learn words better. Moreover, some neuroscience studies also supported the theory. Identity and linguistic information processing may recruit overlapped brain regions, such as the bilateral STG and the left inferior parietal lobule (Feng et al., 2018; Zhang & Chen, 2016).

Different Levels of Attention Engagement During Spoken Words Processing

The distinctions between independence and integration theories may be due to differences in experimental paradigms, specifically, the level of attention engagement. Previous studies used several tasks to explore the identity and linguistic information processing, such as vocal discrimination, voice recognition, word recall, and word–picture matching tasks (Aglieri et al., 2021; Creel et al., 2008; Holmes et al., 2018; Kapnoula & Samuel, 2019; McGettigan & Scott, 2012; Seydell-Greenwald et al., 2020). In these tasks, the word recall and word–picture matching tasks usually need participants to pay more attention to the speech materials' linguistic information to complete the task requirements (Creel et al., 2008; Holmes et al., 2018; Kapnoula & Samuel, 2019). However, the passive auditory, vocal discrimination, and voice recognition tasks tend to focus on the physical auditory features of materials and do not require participants' attention to the linguistic information (Aglieri et al., 2021; Feng et al., 2018; McGettigan & Scott, 2012; Roswandowitz et al., 2018; Seydell-Greenwald et al., 2020). Studies that supported the independence theory used adopted tasks with less attention to linguistic information. For example, Roswandowitz et al. (2018) recruited patients with unilateral focal brain lesions to complete a voice recognition task through functional magnetic resonance imaging (fMRI). In the task, participants needed to discriminate whether the words are spoken by their familiar speaker. After controlling for processing of acoustic voice features by entering vocal pitch and vocal timbre, hearing level, and lesion volume data as covariates, the fMRI results still showed that the right posterior/mid temporal lobe was crucially involved in voice-identity processing. The findings suggested that voice recognition had an independent neural basis, at least at the early stage of speech processing. Aglieri et al. (2021) also used the voice recognition task and found that voice identity classification primarily involved the STS/STG and the left inferior frontal regions. The results indicate that voice identity classification engages a unique brain activity area, which also supports the independence theory.

However, studies that support the integration theory usually used tasks with more attention to linguistic information. Holmes et al. (2018) examined whether the familiar voice facilitates word comprehension in two tasks. One was a word recognition task, participants listened to two sentences (one target and the other nontarget) each time, and then reported the words of the target sentences. The task demanded more attention to the sentences' linguistic information. The other was a voice identification task. Participants needed to identify the voices' identity, which requires little attention to the linguistic information. Their results showed that the participants could not recognize the familiar speaker's voice. However, the familiar voice could facilitate recalling words. The results suggested that identity information could play a role in processing linguistic information. Kapnoula and Samuel (2019) asked participants to learn novel words and then completed a word–picture matching task to choose corresponding pictures to the novel words they listened to. The results also

showed that the identity information could facilitate participants' learning of novel words. In conclusion, different levels of attention engagement in tasks may influence the identity and linguistic information processing during spoken words processing, both in behavioral and neural response patterns.

Listeners' Familiarity With Speakers' Voice Identity (Voice Familiarity)

Several recent studies manipulated speakers' voice familiarity to detect the influence of identity information on linguistic information processing. They have revealed that listeners' familiarity with speakers' voice identity affected their processing of linguistic information (Domingo et al., 2020; Holmes et al., 2018, 2021; Johnsrude et al., 2013; Schall et al., 2014). Firstly, familiar voice processing occurs in brain regions more anterior to the right STS, whereas unfamiliar voice processing occurs in brain regions more posterior to the right STS (Qin et al., 2020; Schall et al., 2014). More importantly, the familiar voices may aid in processing linguistic information (Domingo et al., 2020; Johnsrude et al., 2013). For example, Holmes et al. (2018) found that the participants recalled the words of the target sentences spoken by their familiar (friends) speakers better than their unfamiliar speakers. Domingo et al. (2020) suggested that the participants recalled more words from the sentences spoken by their familiar (friends and spouses) speakers than their unfamiliar speakers.

In addition to the familiar versus unfamiliar voices, researchers also explored how speakers process their own voices (e.g., Johnson et al., 2021; Pinheiro, Rezaii, Nestor, et al., 2016; Pinheiro, Rezaii, Rauber, & Niznikiewicz, 2016). Johnson et al. (2021) used a voice perception task to investigate how individuals monitor and process their own voices. The researchers first asked participants to record two vowels (i.e., /a/ and /o/) and then modified the recorded vowels. They assigned the recording materials into three conditions: participant's own, participant's own modified, and unfamiliar voice. The fMRI results showed that the right anterior STG and the right inferior frontal gyrus (IFG) were explicitly sensitive to the participants' own voices but not their own modified and unfamiliar voices. The findings suggested that a network involving the right anterior STG and IFG was involved to monitor self-voice. Moreover, the self-voice processing may have a unique pattern. It is distinct from familiar speakers' like friends' voices (e.g., Graux et al., 2015; Johnson et al., 2021).

Compared with familiar voices, individuals process their own voice at a much earlier stage, as earlier as around 100 ms. Pinheiro, Rezaii, Nestor, et al. (2016) used emotional words to detect how self-voice affects spoken word processing via the voice recognition task. After hearing a spoken word, participants were asked to judge who said it. The results showed that the self-voice's emotional words elicited larger P2 and late positive potentials (LPP) amplitudes than the unfamiliar speaker's words. The results revealed the role of speaker identity in emotional word processing. The study also suggested that emotional words could be effective speech materials for detecting the identity and linguistic information processing.

As see from these studies, listeners processed speakers' voices differently in terms of whether the speakers' voices are familiar or unfamiliar or whether the voice is their own voice. It needs to be noted here that the speakers' own voices are also familiar but they are distinct from their familiar speakers' voice. More importantly, the studies implied that listeners' familiarity with speakers' voice may affect their processing of linguistic information.

This Study

Whether and how identity and linguistic information interact remains unclear. Based on the above analyses, the attention requirements in tasks may modulate these two types of information processing. Moreover, listeners' familiarity with speakers' voices may affect their processing of linguistic information. Taken together, the identity information may interact with the linguistic information during spoken word processing. However, the attention requirements to the linguistic information in different tasks may modulate their interaction. To demonstrate it, we conducted two event-related potential (ERP) experiments in this study to explore the identity and linguistic information processing by considering the speakers' voice familiarity and attention involvement.

In Experiment 1, we used a word decision task, which usually requires participants to decide whether the words they hear are real words or not (Vitale et al., 2018). It is a typical task that requires participants' attention to linguistic information. We used the emotional words spoken by the participants themselves, participants' friends, and unfamiliar speakers (participants did not know the unfamiliar speaker) as the materials (Pinheiro, Rezaii, Nestor, et al., 2016). Previous studies suggested that the identity information, especially the participants' own identity information, could be processed as early as 100 ms after stimuli onset (Graux et al., 2013; Pinheiro, Rezaii, Nestor, et al., 2016). It may be difficult to capture the interaction between identity and linguistic information at such an early stage. However, the emotional words could help to resolve the potential issue because the processing of emotional linguistic information is rapid and also begins at the early stage (Bernat et al., 2001; Sass et al., 2010). Thus, we could explore the potential interaction during the time course of spoken word processing with the emotional words.

We focused on three ERPs components in the experiment, including N100, P200, and N400. These components are usually elicited in the word decision task and emotional word processing (Pinheiro, Rezaii, Nestor, et al., 2016). The N100 is usually distributed in the frontal-central brain area and reaches the peak around 100 ms after stimuli onset (Ford et al., 2001; Keenan, 2001). It could reflect the phonemic or emotional information processing at a prelexical stage (Bernat et al., 2001; Marslen-Wilson, 1987; Sass et al., 2010). It is also related to self-generated voice and emotional words processing (Knolle et al., 2013; Pinheiro, Rezaii, Nestor, et al., 2016). The P200 is usually distributed at the central-parietal brain area and reaches the peak at 200–300 ms after stimuli onset (Graux et al., 2013). It is associated with the attention bias and self-relative information detection (Crowley & Colrain, 2004; Knolle et al., 2013). The N400 is usually distributed in the central-parietal brain area and continuously present at 300–600 ms after stimuli onset (Coronel & Federmeier, 2016; Hagoort et al., 2004). It is a classical component that related to semantic processing at a later stage (Hagoort et al., 2004). As previous studies implied different hemispheric lateralization patterns during spoken word processing (e.g., Pinheiro, Rezaii, Nestor, et al., 2016; Yu, Chen, Yin, et al., 2022), we would examine the N100, P200, and N400 amplitudes between the left, midline, and right hemispheres. The amplitude differences between different hemispheres could indicate the different processing extent or the different number of neural resources involved in processing (Duncan et al., 2009). Moreover, examining the three ERP components could also be helpful for detecting the time course of processing.

In Experiment 2, we used a passive oddball paradigm, which usually requires the participants to see a silent movie but ignores the spoken words during the experiment. It usually explores the pre-attentive processing of spoken words and does not require participants' explicit attention to the linguistic information (e.g., Näätänen et al., 2007; Yu, Chen, Wang, et al., 2022). We also used the emotional words spoken by the participants themselves, the participants' friends, and unfamiliar speakers as the materials. The oddball paradigm usually elicits the mismatch negativity (MMN). It is usually distributed in the frontal-parietal area and reaches a peak around 100–350 ms after stimuli onset (Chandrasekaran et al., 2007; Korpilahti, 2001; Yu et al., 2020). We would examine the MMNs elicited by different words in the left, midline, and right hemispheres.

Based on previous studies, we hypothesized that the identity information would interact with the linguistic information during spoken word processing. Moreover, their interaction would be modulated by attention involvement. In terms of the hypotheses, we expected to observe an interaction among speaker (self, friend, and unfamiliar), word type (negative, positive, and neutral), and hemisphere (left, midline, and right) in Experiment 1. However, in Experiment 2, the interaction among speaker (self, friend, and unfamiliar), word contrast (negative–positive, negative–neutral, and positive–neutral), and hemisphere (left, midline, and right) would not be significant due to the pre-attentive task during information processing.

Experiment 1

In Experiment 1, we used the ERP technique and word decision task to examine how the identity information interacts with the linguistic information during emotional word processing.

Participants

In total, we recruited 24 female undergraduate students (M_{age} = 19.761, SD = 1.611) from South China Normal University. We performed a statistical power analysis using G-power 3.1 software (Faul et al., 2007). Taking the suggested effective size (η_p^2 = .25) in G-power Manual, 12 participants are needed (power = 0.8, α = .05). We recruited the participants in pairs (12 pairs), and each pair of participants were familiar friends. The participants were native Chinese speakers. They had normal hearing and (corrected-normal) vision. They were all right-handed according to the Edinburgh handedness test (Oldfield, 1971) and reported no history of speech, language, neurological disorders, head damage, or mental illness. The study was approved by the Ethics Review Board of South China Normal University. The participants all signed a consent form before they took part in the experiment and received monetary compensation after the experiment.

Materials

We used 108 real Chinese adjective words and 108 pseudowords as our experimental material. The real words were chosen from Pinheiro, Rezaii, Nestor, et al. (2016). We first asked a proficient Mandarin–English speaker to translate the words into Chinese. Then we asked two English major doctoral students to evaluate the degree of concordance between English and Mandarin words. Their evaluations confirmed that the concordance between English and Mandarin words was high. The materials included 36 negative words (e.g., 丑陋 /chou3-lou4/, means ugly; see Appendix A), 36 positive words (e.g., 美丽 /mei3-li4/, means beautiful; see Appendix B), and 36 neutral words (e.g., 基础 /ji1-chu3/, means basic; see Appendix C). The

pseudowords (e.g., 敢几 /gan3-zi3/; see Appendix D) consisted of Chinese syllables but did not have meanings in Chinese.

Before the experiment, we recruited a group of 15 undergraduate students who did not take part in the ERP experiment to rate the pleasure, arousal, and familiarity of the real words using the 9-point Likert scale (see Table 1). The analysis of variance (ANOVA) results showed that the main effect of pleasure was significant, positive words > neural words > negative words, $F(2, 70) = 668.782$, $p < .001$, $\eta_p^2 = .950$. The main effect of arousal was significant, positive words = negative words > neural words, $F(2, 70) = 44.502$, $p < .001$, $\eta_p^2 = .560$. However, the main effect of familiarity were not significant, $F(2, 70) = .570$, $p = .457$, $\eta_p^2 = .016$. For the real words and pseudowords, we asked another group of 12 undergraduate students who did not participate in the ERP experiment to rate the words' intelligibility using the 9-point Likert scale ($M = 1.142$, $SD = 0.114$; $M = 7.438$, $SD = 0.461$; see Table 1). The t-test results showed that the intelligibility of real words was significantly higher than the pseudowords, $t(214) = 138.115$, $p < .001$, $d = 19.300$. The rating results confirmed the materials' effectiveness in the experiment.

We then recorded all the words by the participants, who know each other, and one female native Chinese speaker that the participants had never met. Hence, we included three groups of speakers in the experiment: self, friend, and unfamiliar. The words were recorded at a sampling rate of 44.1 kHz via Cool Edit Pro (Adobe Systems Incorporated) firstly and then were normalized to 70 dB via Praat software (http://www.fon.hum.uva.nl/praat/). The duration and mean f_o for the three groups of speakers' recording materials were as shown in Table 2.

Procedure

The participants attended to the ERP experiment 1 month after recording. We used a word decision task in the experiment. The materials included 108 real words, 108 pseudowords, and 216 filters spoken by the participants, their friends, and unfamiliar speakers. The whole experiment consisted of six blocks (each being 108 trials). Each block consisted of different types of words spoken by different speakers with a same number of trials. The words were presented to the participants randomly. In each trial, the participants would see a red fixation on the screen for 500 ms and then hear a word. They need to judge whether the word is a real word or not and press a corresponding button on the keyboard (real word: "F"; pseudoword: "J") in 2,000 ms. Once the participants press a button, they would see a blank screen in a random interval (600/800/1,000/1,200 ms) and continue to the next trial. The experiment lasted for about 90 min. The participants could rest for 3–5 min between blocks.

Table 1. Rating of negative, neutral, and positive words in Experiment 1 (mean, standard deviation in parentheses).

Word type	Negative words	Neutral words	Positive words
Pleasure	3.589 (0.520)	6.223 (0.577)	8.186 (0.465)
Arousal	5.929 (0.920)	4.214 (0.733)	5.523 (0.778)
Familiarity	7.619 (0.387)	7.479 (0.666)	7.629 (0.627)
Intelligibility	7.520 (0.408)	7.368 (0.434)	7.425 (0.532)

Electroencephalogram Recording

The electroencephalogram (EEG) was recorded with 32 electrodes mounted on a standard cap (NeuroScan), according to the expanded 10–20 system (American Electroencephalographic Society, 1991), using Scan 4.5 software to record the EEG data (NeuroScan). The online reference electrode was placed at the tip of the nose. The ground electrode was placed at FPz, supra- and infra-orbitally from the left eye is recorded as the vertical electro-oculogram (EOG). The left versus right orbital rim was recorded as the horizontal EOG. The EEG was acquired at a sampling rate of 1000 Hz, with a bandpass filter of 0.05–100 Hz, and the impedance of each electrode was kept below 5 kΩ.

Data Analyses

Off-line signal processing was carried out using Curry 7.0 (NeuroScan). The reference electrode was converted to the average signal of the two mastoids (M1 and M2). The interference from the horizontal and vertical eye movements was then automatically detected and corrected. Then, the data were segmented with a 1,100-ms epoch window, including a 100-ms prestimulus baseline. After baseline correction was performed, any trials with artifact activities beyond the range of −100 to 100 mV were excluded. After rejecting bad trials, at least 80% of trials (more than 28 trials) in each condition for each participant remained. The remained trials were matched among experimental conditions. The trials that were included were then filtered at 0.1–30 Hz with a finite impulse response filter. The averaged ERPs elicited by different types of words were then obtained.

We mainly focused on three ERP components, N100, P200, and N400, in the experiment. Based on the scalp distribution of N100 (frontal-central area), P200 (central-parietal area), and N400 (central-parietal area; Coronel & Federmeier, 2016; Ford et al., 2001; Graux et al., 2013; Hagoort et al., 2004; Keenan et al., 2001) and the waveforms obtained in our study, we selected nine

Table 2. Duration and mean fundamental frequency (f_o) for the different types of words recorded by the participants (self), friends, and unfamiliar speaker in Experiment 1 (mean, standard deviation in parentheses).

Variable	Speaker	Negative words	Neutral words	Positive words
Duration (ms)	Self	104.309 (13.889)	104.485 (13.633)	105.046 (13.400)
	Friend	103.426 (12.139)	105.105 (12.905)	105.157 (13.167)
	Unfamiliar speaker	99.713 (11.115)	100.217 (9.916)	97.328 (13.299)
Mean f_o (Hz)	Self	236.066 (34.193)	232.890 (34.448)	233.523 (33.526)
	Friend	247.476 (23.132)	244.739 (19.782)	243.508 (20.096)
	Unfamiliar speaker	217.869 (35.488)	221.709 (23.836)	215.159 (28.176)

electrodes, F3, Fz, F4, FC3, FCz, FC4, C3, Cz, and C4, to further analyze the three components. According to the waveforms in each condition, we chose 70–120 ms as the time window of N100, 120–250 ms as the time window of P200, and 270–370 ms as the time window of N400. Then, we extracted the three components' amplitudes in each electrode. Considering the potential hemisphere effect, we calculated the averaged amplitudes of the three components in the left (F3, FC3, C3), midline (Fz, FCz, Cz), and right hemispheres (F4, FC4, C4) for each condition.

Finally, we conducted three-way repeated-measures ANOVAs with speaker (self, friend, and unfamiliar), word type (negative, positive, and neutral), and hemisphere (left, midline, and right) as within-subject factors for the N100, P200, and N400 amplitudes, respectively. The data's normality was tested with the Mauchly's test of sphericity. Detailed results were reported on https://osf.io/e9zny/?view_only=92d0f5bec23942d88419ec544c0223fd. When the data did not have a normal distribution, we used the Greenhouse–Geisser method to adjust the degrees of freedom. For multiple comparisons, we used the Bonferroni correction method.

Results

Three participants' data were excluded because they did not complete the experiment or the accuracy rate was less than the random level (50%). The accuracy in each condition was above 85% for the remained participants, indicating that the remained participants completed the experiment seriously. Figure 1 showed the waveforms and topographic maps for each condition. Table 3 showed the N100, P200, and N400 mean amplitudes in each condition.

N100

The ANOVA results showed that the main effect of hemisphere was significant, $F(2, 40) = 4.764$, $p = .014$, $\eta_p^2 = .192$. The N100 amplitude in the midline was significantly smaller than the left hemisphere ($p = .016$), but there was no significant difference between the left and right hemispheres ($p = .267$) or the midline and right hemispheres ($p = .930$). The main effect of speaker and word type were not significant, $F(2, 40) = 1.860$, $p = .169$, $\eta_p^2 = .085$; $F(2, 40) = 2.786$, $p = .074$, $\eta_p^2 = .122$. The interaction between speaker and word type, $F(4, 80) = 1.444$, $p = .148$, $\eta_p^2 = .080$; speaker and hemisphere, $F(2.130, 42.597) = 1.355$, $p = .269$, $\eta_p^2 = .063$; and word type and hemisphere, $F(4, 80) = 0.877$, $p = .482$, $\eta_p^2 = .042$, was not significant. The interaction among speaker, word type, and hemisphere was not significant either, $F(4.999, 99.971) = 1.955$, $p = .092$, $\eta_p^2 = .089$.

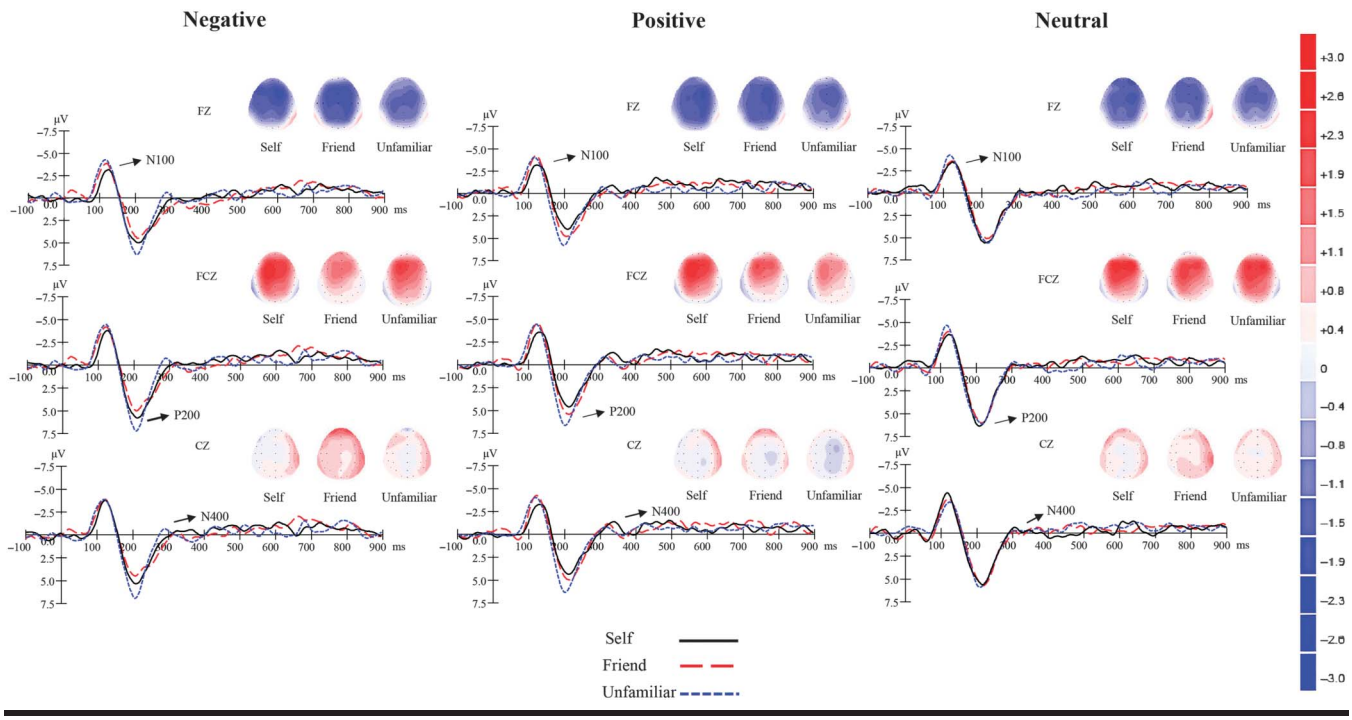
P200

The main effect of speaker was significant, $F(2, 40) = 4.572$ $p = .016$, $\eta_p^2 = .186$. The P200 amplitude of unfamiliar speaker was significantly larger than participants themselves ($p = .049$), but there was no significant difference between participants themselves and friends ($p = 1.000$) or friends and unfamiliar speaker ($p = .054$). The main effect of hemisphere was significant, $F(1.450, 29.01) = 33.743$, $p < .001$, $\eta_p^2 = .628$. The P200 amplitude in the right hemisphere was significantly larger than the midline ($p < .001$) and the left hemisphere ($p < .001$). Besides, the P200 amplitude in the left hemisphere was significantly larger than in the midline hemisphere ($p = .039$). The main effect of word type was not significant, $F(2, 40) = 0.296$, $p = .746$, $\eta_p^2 = .015$. The interactions between speaker and word type, $F(4, 80) = 1.060$, $p = .382$, $\eta_p^2 = .050$; speaker and hemisphere, $F(2.474, 49.479) = 0.665$, $p = .618$, $\eta_p^2 = .032$; and word type and hemisphere, $F(2.258, 45.158) = 1.046$, $p = .389$, $\eta_p^2 = .050$, were not significant. The interaction among speaker, word type, and hemisphere was not significant, $F(4.062, 81.231) = 1.299$, $p = .277$, $\eta_p^2 = 0.061$.

N400

The main effect of hemisphere was significant, $F(1.548, 30.962) = 6.892$, $p = .006$, $\eta_p^2 = .256$. The N400 amplitude in the midline was significantly smaller than the right hemisphere ($p = .001$), but there was no significant difference between the left and right hemispheres ($p = .292$)

Figure 1. Waveforms and topographic maps for N100, P200, and N400 in negative, positive, and neural words spoken by the participants (self), friends, and unfamiliar speaker in Experiment 1.



or the left and midline hemispheres (p = .275). The main effects of speaker, $F(2, 40) = 0.653$, p = .526, η_p^2 = .032, and word type, $F(2, 40) = 0.718$, p = .494, η_p^2 = .035, were not significant. The interactions between speaker and word type, $F(4, 80) = 0.115$, p = .977, η_p^2 = .006; speaker and hemisphere, $F(2.542, 50.846) = 0.520$, p = .641, η_p^2 = .025; and word type and hemisphere, $F(2.301, 46.014) = 0.186$, p = .859, η_p^2 = .009, were all not significant.

However, the interaction among speaker, word type, and hemisphere was significant, $F(8, 160) = 3.137$, p = .003, η_p^2 = .136. We further conducted two-way ANOVAs to the different speakers with word type (negative, positive, and neutral) and hemisphere (left, midline, and right) as the within-subject factors. The results showed that for words spoken by the unfamiliar speaker, the main effect of hemisphere, $F(2, 40) = 10.258$, $p < .001$, η_p^2 = .339, was significant. The N400 amplitudes in the left hemisphere and midline were significantly smaller than the right hemisphere (p = .020; $p < .001$). There was no significant difference between the left hemisphere and midline (p = .990). The main effect of word type was not significant, $F(2, 40) = 0.094$, p = .911, η_p^2 = .005. However, the interaction between hemisphere and word type, $F(2.207, 44.136) = 3.263$, p = .043, η_p^2 = .140, was significant. For the positive and negative words, the N400 in the left hemisphere and midline was significantly smaller than the right hemisphere (p = .029, p = .012; p = .027, p = .002). There were no significant differences between other hemispheres ($ps > .05$). For the neutral words, the N400 in the midline was significantly smaller than the left hemisphere (p = .048). There were no significant differences between other hemispheres ($ps > .05$).

For words spoken by participants' friends, the main effects of hemisphere, $F(1.330, 26.603) = 2.542$, p = .114, η_p^2 = .113, and word type, $F(1.497, 29.945) = 0.664$, p = .480, η_p^2 = .032, were not significant. The interaction between hemisphere and word type, $F(2.347, 46.933) = 1.741$, p = .182, η_p^2 = .080, was not significant.

For words spoken by the participants themselves, the main effects of hemisphere, $F(1.461, 29.227) = 2.873$, p = .087, η_p^2 = .126, and word type, $F(2, 40) = 0.030$, p = .970, η_p^2 = .002, were not significant. The interaction between hemisphere and word type, $F(4, 80) = 0.411$, p = .800, η_p^2 = .020, was not significant.

Discussion

Experiment 1 examined whether identity information interacts with linguistic information during the emotional word processing with a word decision task. We found an interaction among speaker, word type, and hemisphere in N400 amplitudes but not in N100 and P200. The results showed that in the task that requires explicit attention to linguistic information, identity information interacted with

Table 3. Amplitudes (uv) for N100, P200, and N400 in negative, positive, and neural words spoken by the participants (self), friends, and unfamiliar speaker in Experiment 1 (mean, standard deviation in parentheses).

	N100			P200			N400		
Word pair	**Left**	**Midline**	**Right**	**Left**	**Midline**	**Right**	**Left**	**Midline**	**Right**
Self-negative	–0.746 (0.845)	–0.847 (0.880)	–0.521 (0.851)	0.727 (0.718)	0.796 (0.886)	1.322 (0.881)	–0.246 (0.757)	–0.152 (1.045)	0.321 (0.829)
Self-positive	–0.392 (0.854)	–0.309 (0.550)	–0.535 (0.646)	0.903 (0.912)	0.721 (0.734)	1.381 (0.992)	–0.272 (0.575)	–0.251 (0.602)	0.260 (0.741)
Self-neutral	–0.201 (1.179)	–0.217 (0.507)	–0.605 (0.710)	1.105 (1.172)	0.609 (0.549)	1.412 (1.046)	0.076 (0.880)	–0.311 (0.947)	–0.062 (0.630)
Friend-negative	–0.644 (0.844)	–1.001 (0.833)	–0.890 (1.029)	0.616 (0.776)	0.296 (0.811)	0.886 (1.074)	0.078 (0.876)	–0.024 (0.937)	0.243 (0.848)
Friend-positive	–0.361 (0.806)	–0.618 (1.057)	–0.545 (0.756)	0.613 (0.829)	0.599 (0.700)	1.161 (1.096)	0.011 (0.481)	–0.252 (0.608)	–0.051 (0.773)
Friend-neutral	–0.666 (0.617)	–0.733 (0.722)	–0.609 (0.703)	0.608 (0.797)	0.554 (0.586)	1.057 (0.783)	–0.192 (0.830)	–0.226 (0.917)	–0.269 (0.864)
Unfamiliar speaker-negative	–0.437 (0.707)	–0.652 (0.800)	–0.615 (0.611)	0.777 (0.783)	0.420 (0.709)	1.111 (0.809)	–0.059 (0.809)	–0.341 (0.883)	–0.010 (0.941)
Unfamiliar speaker-positive	–0.553 (0.823)	–0.930 (0.795)	–0.646 (0.697)	0.477 (1.023)	0.354 (0.816)	0.743 (1.097)	–0.112 (0.982)	–0.376 (0.939)	–0.048 (0.716)
Unfamiliar speaker-neutral	–0.418 (0.711)	–0.828 (0.812)	–0.701 (0.760)	0.767 (0.868)	0.333 (0.819)	1.001 (1.120)	–0.196 (0.721)	–0.317 (0.917)	0.045 (0.772)

linguistic information at the later stage of spoken word processing.

Specifically, the N400 results showed that the identity information interacted with linguistic information with regard to hemispheric lateralization patterns. For the unfamiliar speakers, the neutral words evoked larger N400 amplitudes in the midline than the left hemisphere, while the negative and positive words evoked larger N400 amplitudes in the left and midline hemispheres than in the right hemisphere (N400 is a negative component and its values were negative; see Figure 1). However, we did not find such patterns for the words spoken by participants themselves and their friends. The results first supported the integration theory. More importantly, our findings suggested that the interaction occurs at the later stage of processing.

However, the N400 results were inconsistent with Holmes et al. (2018) and Kapnoula and Samuel (2019). Holmes et al. (2018) and Kapnoula and Samuel found that the trained familiar voice could facilitate linguistic information processing, but we did not find the similar results. Participants were able to obtain sufficient identity information from the tasks in their studies. Participants in the work of Holmes et al. (2018) completed a word recall task after listening to the sentences. Compared with words in our study, the sentences have a longer duration which may help participants acquire more identity information. In the work of Kapnoula and Samuel, participants received a training phase before completing the word–picture task. As with Holmes et al. (2018), the training phase provided participants with much identity information. Furthermore, we found no difference between the self-voice and the friend's voice. However, previous research suggested that the self-voice processing might have a distinct processing pattern compared to that of familiar friends' voices (e.g., Johnson et al., 2021; Pinheiro, Rezaii, Nestor, et al., 2016; Pinheiro, Rezaii, Rauber, & Niznikiewicz, 2016). We speculate that this is because the word decision task in Experiment 1 required more attention to the linguistic information than to the identity information. The linguistic information processing may cover up the potential difference between the self-voice and the friend's voice. Therefore, we did not observe their differences explicitly in our experiment.

At the early processing stage, we found the main effects of hemisphere in N100 and P200, and the main effect of speaker in P200. The N100 amplitude in the midline was larger than the left hemisphere (N100 is a negative component and its values were negative). The P200 amplitude was significantly larger in the unfamiliar speaker than the self-voice. It was larger in the bilateral hemispheres than the midline. Previous studies suggested that N100 and P200 could reflect the extraction of acoustic cues (unbiased lateralization of the N100), self-related information detection (right hemisphere of the P200), and integration of acoustic characteristics cues (left hemisphere of the P200), respectively (Bernat et al., 2001; Chang et al., 2018; Conde et al., 2018; Crowley & Colrain, 2004; Knolle et al., 2013; Kotz & Paulmann, 2011; Marslen-Wilson, 1987; Sass et al., 2010). Considering the implications of N100 and P200, our results implied that individuals process the acoustic cues firstly and then focus on speakers' identity information at P200 time window.

In general, Experiment 1 indicated the time course of identity and linguistic information processing in the task that requires explicit attention to the linguistic information. Listeners first processed the acoustic cues and identity information at the early stage of emotional word processing (N100 and P200 time window). At the later stage (N400 time window), listeners integrated the identity with linguistic information and the two types of information interacted during the processing. However, previous studies implied that the task would modulate identity and linguistic information processing (Domingo et al., 2020; Holmes et al., 2018; Kapnoula & Samuel, 2019). To test the task's role, we used the passive oddball paradigm in Experiment 2. The paradigm usually requires participants to watch a silent movie and ignore the speech stimuli. It requires no explicit attention to either identity or linguistic information.

Experiment 2

We used a passive oddball paradigm with the ERP technique to further detect how the identity information interacts with the linguistic information during preattentive emotional word processing and whether the processing pattern differs from Experiment 1.

Participants

We recruited another group of 24 female undergraduate students (M_{age} = 20.261, SD = 1.451) from South China Normal University in Experiment 2. The recruitment requirements were the same as in Experiment 1. The study was approved by the Ethics Review Board of South China Normal University. The participants all signed a consent form before they took part in the experiment and received monetary compensation after the experiment. No participants took part in Experiment 1.

Materials

We first chose one negative word (贫穷 /pin2-qiong2/, "poor"), one positive word (富有 /fu4-you3/, "rich"), and one neutral word (接近 /jie1-jin4/, "being close to") from

Experiment 1. The three words differed in the rating of pleasure and valence (贫穷 [poor], pleasure: 3.407; arousal: 5.721, familiarity: 8.100; 富有 [rich], pleasure: 8.193; arousal: 6.300, familiarity: 7.907; 接近 [being close to], pleasure: 6.300; arousal: 4.436, familiarity: 8.036). Similar to Experiment 1, the words were recorded by the participants, participants' friends, and one unfamiliar female native Chinese speaker. The detailed procedure of recording was identical to Experiment 1. We used the words to construct nine types of word pairs, that is, self-negative (the negative word spoken by the participants themselves, 贫穷 [poor]) versus self-positive (the positive word spoken by the participants themselves, 富有 [rich]), self-negative versus self-neutral (the neutral word spoken by the participants themselves, 接近 [being close to]), self-positive versus self-negative, friend-negative (the negative word spoken by the participants' friends, 贫穷 [poor]) versus friend-positive (the positive word spoken by the participants' friends, 富有 [rich]), friend-negative versus friend-neutral (the neutral word spoken by the participants' friends, 接近 [being close to]), friend-positive versus friend-neutral, unfamiliar-negative (the negative word spoken by the participants' unfamiliar speaker, 贫穷 [poor]) versus unfamiliar-positive (the positive word spoken by the participants' unfamiliar speaker, 富有 [rich]), unfamiliar-negative versus unfamiliar-neutral (the neutral word spoken by the participants' unfamiliar speaker, 接近 [being close to]), and unfamiliar-positive versus unfamiliar-neutral.

Procedure

We used the passive oddball paradigm in Experiment 2. The paradigm usually consists of one type of standard stimuli and one type of deviant stimuli (e.g., Näätänen et al., 2007). In the experiment, there were 18 blocks in total. Each word pair have two blocks (e.g., the self-negative word as the standard stimuli and the self-neutral word as the deviant stimuli, and vice versa). In each block, we would first present the participants with 10 standard stimuli and then present 82 standard stimuli and 18 deviant stimuli pseudorandomly. The order of the blocks was counterbalanced between participants. Each word was presented for 2,000 ms, and the interstimulus interval was 600 ms. The participants could have 3 min of rest between blocks. During the experiment, the participants were asked to watch a silent movie seriously and ignore the auditory stimuli. In order to ensure the participants watch the movie carefully, they need to answer questions about the movie's content after the experiment. The whole experiment lasted about 120 min.

EEG Recording

The EEG recording was the same as in Experiment 1.

Data Analyses

The EEG data were firstly preprocessed as the method in Experiment 1, after rejecting bad trials, at least 80% of trials (more than 168 trials) in each condition for each participant remained. However, the individual EEG epochs were processed with a 100-ms prestimulus baseline and a 500-ms poststimulus epoch. After we obtained the waveforms of the standard and deviant stimuli in each word pair condition, we further subtracted the waveform of the standard stimuli from that of the deviant stimuli to get the MMN waveform of each word pair condition.

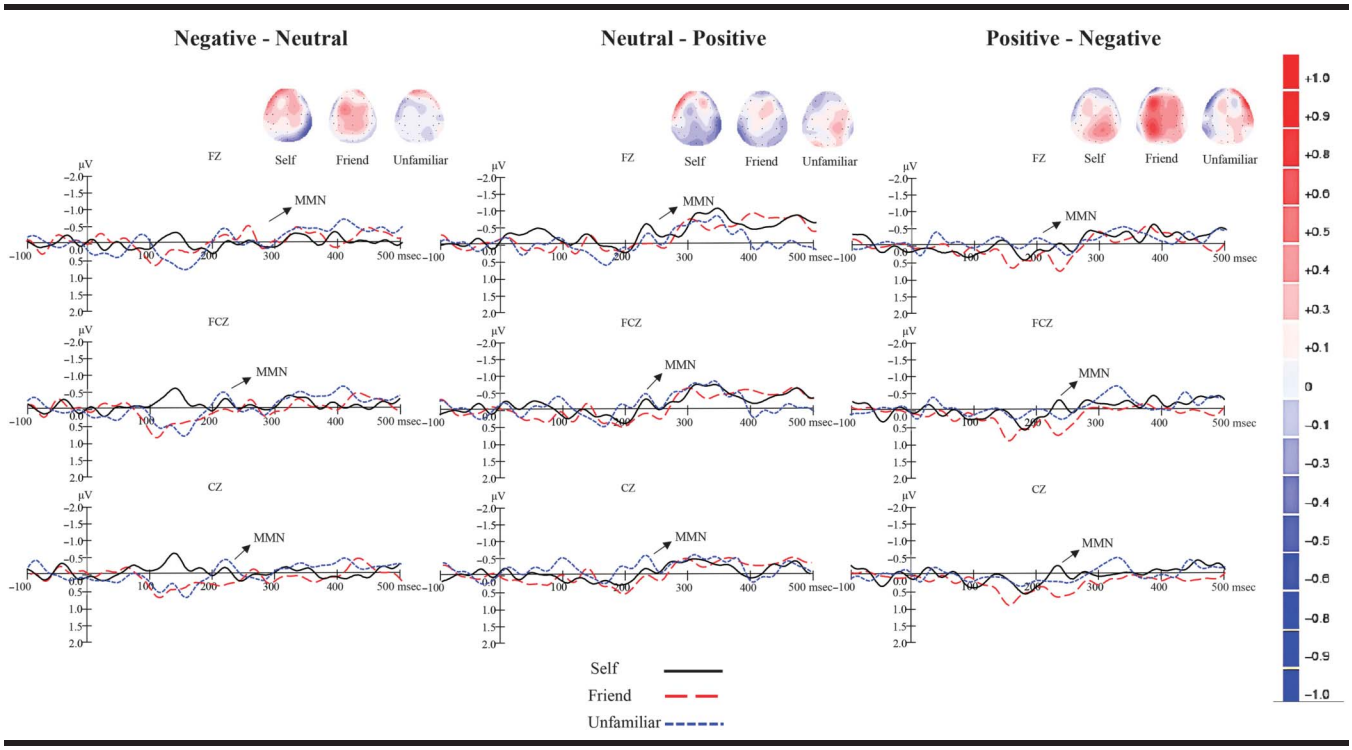
Based on the scalp distribution of MMN (frontal–parietal area; e.g., Näätänen et al., 2007; Yu et al., 2017) and the waveforms obtained in our study, we selected nine electrodes, F3, Fz, F4, FC3, FCz, FC4, C3, Cz, and C4, to analyze the MMNs. According to the obtained difference waves in each condition, we chose 100–300 ms as the time window. We first detected the MMN peak automatically with the software and then examined it manually within the corresponding time windows for each participant in each condition. The MMN amplitude was then calculated in a time window ranging from 20 ms before the detected peak in electrode Fz to 20 ms after that peak. We further calculated the averaged MMN amplitudes in the left (F3, FC3, C3), midline (Fz, FCz, Cz), and right hemispheres (F4, FC4, C4) for each condition. Finally, we conducted a three-way repeated-measures ANOVA with the speaker (self, friend, and unfamiliar), word pair (negative-neutral, neutral-positive, and positive-negative), and hemisphere (left, midline, and right) as within-subject factors for the MMN amplitudes. The data's normality was tested with the Mauchly's test of sphericity. Detailed results were reported on https://osf.io/e9zny/?view_only=92d0f5bec23942d88419ec544c0223fd. When the data did not have a normal distribution, we used the Greenhouse–Geisser method to adjust the degrees of freedom. For multiple comparisons, we used the Bonferroni correction method.

Results

One participant did not complete the whole experiment because of physical discomfort. Thus, we only included 23 participants' data in the statistical analysis. Figure 2 shows the MMN waveforms and topographic maps in each word pair condition. Table 4 showed the MMN amplitudes in each condition.

The ANOVA results showed that the main effect of hemisphere was significant, $F(2, 44) = 7.918$, $p = .003$, $\eta_p^2 = .265$. The MMN amplitude in the midline was significantly smaller than the right hemispheres ($p = .001$). The interaction between speaker and hemisphere was

Figure 2. Waveforms and topographic maps for mismatch negativities (MMNs) in different word pairs spoken by the participants (self), friends, and unfamiliar speaker in Experiment 2.



significant, $F(2.459, 54.109) = 4.935$, $p = .007$, $\eta_p^2 = .183$. To further detect the potential differences between hemispheres in each type of speaker, we conducted simple effect analysis. The results showed that for the participants themselves, the MMNs in the left and midline hemispheres were significantly smaller than the right hemisphere ($p = .030$; $p < .001$). However, there was no significant difference between the left and midline hemispheres ($p = .99$). For the unfamiliar speaker, the MMN in the midline was significantly smaller than the left hemisphere ($p < .001$). However, there was no significant difference between the left and right hemispheres ($p = .303$) or the midline and right hemispheres ($p = .152$). For the participants' friends, there were no significant differences between hemispheres ($ps > .05$).

The main effects of speaker and word pair were not significant, $F(2, 44) = 1.174$, $p = .319$, $\eta_p^2 = .051$; $F(1.234, 27.152) = 0.996$, $p = .345$, $\eta_p^2 = .043$). The interaction between speaker and word pair, $F(1.625, 35.750) = 0.628$, $p = .507$, $\eta_p^2 = .028$; word pair and hemisphere, $F(2.255, 49.611) = 0.297$, $p = .770$, $\eta_p^2 = .013$, was not significant. The interaction among speaker, word pair, and hemisphere was not significant, $F(3.682, 80.999) = 0.192$, $p = .738$, $\eta_p^2 = .021$.

Table 4. Amplitudes (uv) for mismatch negativities (MMNs) in different word pairs spoken by the participants (self), friends, and unfamiliar speaker in Experiment 2 (mean, standard deviation in parentheses).

Word pair	MMNs		
Self-negative vs. self-neutral	−0.363 (0.718)	−0.682 (0.676)	−0.506 (0.770)
Self-neutral vs. self-positive	−0.519 (1.719)	−1.047 (1.624)	−0.877 (1.703)
Self-positive vs. self-negative	−0.445 (1.031)	−0.612 (1.086)	−0.462 (0.871)
Friend-negative vs. friend-neutral	−0.513 (.813)	−0.376 (0.824)	−0.376 (0.824)
Friend-neutral vs. friend-positive	−0.941 (2.326)	−1.029 (2.062)	−0.942 (2.448)
Friend-positive vs. friend-negative	−0.258 (1.167)	−0.450 (1.150)	−0.196 (0.995)
Unfamiliar speaker-negative vs. unfamiliar speaker	−0.953 (1.460)	−0.969 (1.304)	−0.505 (1.339)
Unfamiliar speaker-neutral vs. unfamiliar speaker-positive	−0.996 (1.019)	−0.970 (1.058)	−0.410 (1.256)
Unfamiliar speaker-positive vs. unfamiliar speaker-negative	−0.764 (.816)	−0.840 (0.929)	−0.338 (0.881)

Discussion

Experiment 2 further investigated whether identity information interacts with linguistic information using the passive oddball paradigm that does not demand much attention to linguistic information. The MMN results showed that there was no significant interaction between the speaker and word pair, which indicates that the identity and linguistic information may be processed independently in the task. Although we did not find the interaction between the linguistic and identity information, the speaker identity interacts with the hemispheric region. The results indicated that the self-voice processing is different from the unfamiliar voice processing with regard to hemispheric pattern. Conde et al. (2018) also found differences between self and unfamiliar voices. Their results showed that the words spoken by participants themselves elicited increased P3 amplitudes compared to an unfamiliar speaker. As the P3 component could reflect the mobilization of attention resources to task-relevant events (Kok, 2001; Polich, 2007; Spencer et al., 2001), the results indicated that the self-voice recruited more attention resources than the unfamiliar voice during the processing. Their study used an active oddball paradigm, which required participants to pay attention to each stimulus and count the number of deviant stimuli. However, our experiment did not require participants' attention to the auditory stimuli. The MMN differences between self-voice and unfamiliar speaker further indicated that when attention was not required in the task, processing the self-voice and unfamiliar speaker differed in the hemispheric pattern. Moreover, we also observed hemispheric pattern differences between the self-voice and friend's voice. Although they were both familiar to listeners, the results indicated that listeners process the two types of familiar voices differently. Lastly, the hemispheric patterns between the friend and unfamiliar speaker were also distinct. Taken together, the hemispheric pattern of identity information processing would be varied in gradient from self, friend, to unfamiliar speakers.

General Discussion

Previous studies proposed distinct views (the independence vs. integration theories) to explain identity and linguistic information processing. In our study, Experiment 1 indicated that identity information interacted with linguistic information at the later stage of spoken word processing, which supports the integration theory (Holmes et al., 2018, 2021; Kapnoula & Samuel, 2019; Zhang & Chen, 2016). However, Experiment 2 did not find an interaction between identity and linguistic information, which is consistent with the independence theory (Aglieri et al., 2021; Feng et al., 2018; McGettigan & Scott, 2012; Seydell-Greenwald et al., 2020). A crucial difference between our two experiments was whether the task required explicit attention to the linguistic information. Experiment 1 required it, while Experiment 2 did not. Therefore, identity information would interact with linguistic information during spoken word processing. Importantly, the interaction would be modulated by the task demands.

As reviewed before, most evidence of the integration theory came from the studies that adopted tasks with a greater emphasis on linguistic information (Holmes et al., 2018, 2021; Kapnoula & Samuel, 2019; Zhang & Chen, 2016). By contrast, studies that supported the independence theory mainly used tasks that do not require much attention to linguistic information (Aglieri et al., 2021; Feng et al., 2018; McGettigan & Scott, 2012; Roswandowitz et al., 2018; Seydell-Greenwald et al., 2020). Our research used both types of tasks in the same study and provided direct evidence for the view that the attention requirements to the linguistic information affected the interaction between identity and linguistic information.

Moreover, based on previous studies and our findings, we tried to improve the independence and integration theories. We propose an attention-modulated explanation for voice perception. Initially, there were two potential patterns for processing the identity and linguistic information, the integrated versus independent patterns. We considered an attention classifier to determine whether the listener adopts an integrated or independent pattern to process the speech stimuli. When listeners receive a speech stimulus, if they need to extract the linguistic information explicitly, they would adopt the integration pattern to process the two types of information. Here, we need to note that the interaction mainly occurs at the later stage of semantic processing. However, if they do not need to explicitly consider the linguistic information, they will turn to the independent pattern.

We found the interaction between identity and linguistic information during emotional word processing in their hemispheric patterns. Previous studies have indicated that the processing of identity and linguistic information depends on the different neural basis (identity information: transverse temporal gyrus, temporal plane, anterior/mid/posterior STG; linguistic information: left anterior/posterior STS and inferior prefrontal area; Aglieri et al., 2021; Belin et al., 2004; Blank et al., 2014; Roswandowitz et al., 2018; Schelinski et al., 2016). However, how different brain regions interact or how the brain network supports the integrated processing of identity and linguistic information remains to be investigated in future studies. Additionally, researchers can also explore how brain regions associated

with attention (e.g., Johnson et al., 2021) interact with voice perception regions. Exploration of the issue with source localization analysis or fMRI technique could also provide evidence for the modulation of attention on identity and linguistic information processing.

One limitation of this study is that we recruited only females as the participants in the study. Previous studies suggested that the listeners' gender may affect their voice perception (e.g., García-García et al., 2016). Future studies should further examine how male listeners process the two types of information during emotional word processing. Another potential issue is that we adopted the participants themselves, their friends, and unfamiliar speakers to manipulate the identity information. The three types of speakers' voices' familiarity may be continuous, but not categorical (familiar vs. unfamiliar) to listeners. Therefore, our findings could not directly indicate that voice familiarity interacts with linguistic information in spoken words processing. Future studies could explore the issue with a fined control of voices' familiarity.

Conclusions

The overall findings indicated that the identity information would interact with linguistic information during spoken word processing. However, the interaction was modulated by the task demands. An attention-modulated account was proposed to explain the mechanism underlying identity and linguistic information processing by considering the integration and independence theories.

Author Contributions

Yunxiao Ma: Conceptualization (Equal), Formal analysis (Lead), Investigation (Equal), Writing – original draft (Equal). **Keke Yu:** Conceptualization (Equal), Methodology (Lead), Writing – original draft (Equal). **Shuqi Yin:** Investigation (Equal). **Li Li:** Resources (Equal), Writing – original draft (Equal). **Ping Li:** Conceptualization (Equal), Supervision (Equal), Writing – review & editing (Equal). **Ruiming Wang:** Funding acquisition (Lead), Supervision (Equal), Resources (Equal), Writing – review & editing (Equal).

Data Availability Statement

Original data files presented in this article can be found at https://osf.io/e9zny/?view_only=92d0f5bec23942d88419ec544c0223fd

Acknowledgments

This research was funded by 2022 Guangdong-Hong Kong-Macao Greater Bay Area Exchange Programs of South China Normal University (Ruiming Wang), the Key Project of National Social Science Foundation of China (19ZDA360; Ruiming Wang), and Key Laboratory for Social Sciences of Guangdong Province (2015WSY009; Li Li). The first author was also supported in part by a scholarship from the Hong Kong Polytechnic University (Joint Supervision Scheme #A0034784).

References

Aglieri, V., Cagna, B., Velly, L., Takerkart, S., & Belin, P. (2021). Virtual reality for teaching ESP vocabulary: A myth or a possibility. *International Journal of English Language Education, 5*(2), 1–13. https://doi.org/10.5296/ijele.v5i2.11993

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Bernat, E., Bunce, S., & Shevrin, H. (2001). Event-related brain potentials differentiate positive and negative mood adjectives during both supraliminal and subliminal visual processing. *International Journal of Psychophysiology, 42*(1), 11–34. https://doi.org/10.1016/S0167-8760(01)00133-7

Blank, H., Wieland, N., & von Kriegstein, K. (2014). Person recognition and the brain: Merging evidence from patients and healthy individuals. *Neuroscience & Biobehavioral Reviews, 47,* 717–734. https://doi.org/10.1016/j.neubiorev.2014.10.022

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2007). Mismatch negativity to pitch contours is influenced by language experience. *Brain Research, 1128*(1), 148–156. https://doi.org/10.1016/j.brainres.2006.10.064

Chang, J., Zhang, X., Zhang, Q., & Sun, Y. (2018). Investigating duration effects of emotional speech stimuli in a tonal language by using event-related potentials. *IEEE Access, 6,* 13541–13554. https://doi.org/10.1109/ACCESS.2018.2813358

Conde, T., Gonçalves, Ó. F., & Pinheiro, A. P. (2018). Stimulus complexity matters when you hear your own voice: Attention effects on self-generated voice processing. *International Journal of Psychophysiology, 133,* 66–78. https://doi.org/10.1016/j.ijpsycho.2018.08.007

Coronel, J. C., & Federmeier, K. D. (2016). The N400 reveals how personal semantics is processed: Insights into the nature and organization of self-knowledge. *Neuropsychologia, 84,* 36–43. https://doi.org/10.1016/j.neuropsychologia.2016.01.029

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013

Crowley, K. E., & Colrain, I. M. (2004). A review of the evidence for P2 being an independent component process: Age, sleep and modality. *Clinical Neurophysiology, 115*(4), 732–744. https://doi.org/10.1016/j.clinph.2003.11.021

Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied, 26*(2), 236–247. https://doi.org/10.1037/xap0000247

Duncan, C. C., Barry, R. J., Connolly, J. F., Fischer, C., Michie, P. T., Näätänen, R., Polich, J., Reinvang, I., & Van Petten, C. (2009). Event-related potentials in clinical research: Guidelines

for eliciting, recording, and quantifying mismatch negativity, P300, and N400. *Clinical Neurophysiology, 120*(11), 1883–1908. https://doi.org/10.1016/j.clinph.2009.07.045

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Feng, G., Gan, Z., Wang, S., Wong, P. C., & Chandrasekaran, B. (2018). Task-general and acoustic-invariant neural representation of speech categories in the human brain. *Cerebral Cortex, 28*(9), 3241–3254. https://doi.org/10.1093/cercor/bhx195

Ford, J. M., Mathalon, D. H., Heinks, T., Kalba, S., Faustman, W. O., & Roth, W. T. (2001). Neurophysiological evidence of corollary discharge dysfunction in schizophrenia. *American Journal of Psychiatry, 158*(12), 2069–2071. https://doi.org/10.1176/appi.ajp.158.12.2069

García-García, I., Kube, J., Gaebler, M., Horstmann, A., Villringer, A., & Neumann, J. (2016). Neural processing of negative emotional stimuli and the influence of age, sex and task-related characteristics. *Neuroscience & Biobehavioral Reviews, 68,* 773–793. https://doi.org/10.1016/j.neubiorev.2016.04.020

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., & Bruneau, N. (2015). Is my voice just a familiar voice? An electrophysiological study. *Social Cognitive and Affective Neuroscience, 10*(1), 101–105.

Graux, J., Gomot, M., Roux, S., Bonnet-Brilhault, F., Camus, V., & Bruneau, N. (2013). My voice or yours? An electrophysiological study. *Brain Topography, 26,* 72–82. https://doi.org/10.1007/s10548-012-0233-2

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438–441. https://doi.org/10.1126/science.1095455

Halle, M., & Mohanan, K. P. (1985). Segmental phonology of modern English. *Linguistic Inquiry, 16*(1), 57–116.

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science, 29*(10), 1575–1583. https://doi.org/10.1177/0956797618779083

Holmes, E., To, G., & Johnsrude, I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to voice training. *Psychological Science, 32*(6), 903–915. https://doi.org/10.1177/0956797621991137

Johnson, J. F., Belyk, M., Schwartze, M., Pinheiro, A. P., & Kotz, S. A. (2021). Expectancy changes the self-monitoring of voice identity. *European Journal of Neuroscience, 53*(8), 2681–2695. https://doi.org/10.1111/ejn.15162

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science, 24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language, 107,* 111–127. https://doi.org/10.1016/j.jml.2019.05.001

Keenan, J. P., Nelson, A., O'connor, M., & Pascual-Leone, A. (2001). Self-recognition and the right hemisphere. *Nature, 409*(6818), 305–305. https://doi.org/10.1038/35053167

Knolle, F., Schröger, E., & Kotz, S. A. (2013). Prediction errors in self- and externally generated deviants. *Biological Psychology, 92*(2), 410–416. https://doi.org/10.1016/j.biopsycho.2012.11.017

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology, 38*(3), 557–577. https://doi.org/10.1017/S0048577201990559

Korpilahti, P., Krause, C. M., Holopainen, I., & Lang, A. H. (2001). Early and late mismatch negativity elicited by words and speech-like stimuli in children. *Brain and Language, 76*(3), 332–339. https://doi.org/10.1006/brln.2000.2426

Kotz, S. A., & Paulmann, S. (2011). Emotion, language, and the brain. *Language and Linguistics Compass, 5*(3), 108–125. https://doi.org/10.1111/j.1749-818X.2010.00267.x

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25*(1–2), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: What's wrong, what's right and what's left? *Trends in Cognitive Sciences, 16*(5), 269–276. https://doi.org/10.1016/j.tics.2012.04.006

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Pinheiro, A. P., Rezaii, N., Nestor, P. G., Rauber, A., Spencer, K. M., & Niznikiewicz, M. (2016). Did you or I say pretty, rude or brief? An ERP study of the effects of speaker's identity on emotional word processing. *Brain and Language, 153-154,* 38–49. https://doi.org/10.1016/j.bandl.2015.12.003

Pinheiro, A. P., Rezaii, N., Rauber, A., & Niznikiewicz, M. (2016). Is this my voice or yours? The role of emotion and acoustic quality in self-other voice discrimination in schizophrenia. *Cognitive Neuropsychiatry, 21*(4), 335–353. https://doi.org/10.1080/13546805.2016.1208611

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Qin, P., Wang, M., & Northoff, G. (2020). Linking bodily, environmental and mental states in the self—A three-level model based on a meta-analysis. *Neuroscience & Biobehavioral Reviews, 115,* 77–95. https://doi.org/10.1016/j.neubiorev.2020.05.004

Roswandowitz, C., Kappes, C., Obrig, H., & von Kriegstein, K. (2018). Obligatory and facultative brain regions for voice-identity recognition. *Brain, 141*(1), 234–247. https://doi.org/10.1093/brain/awx313

Samuel, A. G. (2011). Speech perception. *Annual Review of Psychology, 62,* 49–72. https://doi.org/10.1146/annurev.psych.121208.131643

Sass, S. M., Heller, W., Stewart, J. L., Silton, R. L., Edgar, J. C., Fisher, J. E., & Miller, G. A. (2010). Time course of attentional bias in anxiety: Emotion and gender specificity. *Psychophysiology, 47*(2), 247–259. https://doi.org/10.1111/j.1469-8986.2009.00926.x

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience, 27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707

Schelinski, S., Borowiak, K., & von Kriegstein, K. (2016). Temporal voice areas exist in autism spectrum disorder but are dysfunctional for voice identity recognition. *Social Cognitive and Affective Neuroscience, 11*(11), 1812–1822. https://doi.org/10.1093/scan/nsw089

Schirmer, A. (2018). Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing.

Social Cognitive and Affective Neuroscience, 13(1), 1–13. https://doi.org/10.1093/scan/nsx142

Scott, S. K. (2019). From speech and talkers to the social world: The neural processing of human spoken language. *Science, 366*(6461), 58–62. https://doi.org/10.1126/science.aax0288

Seydell-Greenwald, A., Chambers, C. E., Ferrara, K., & Newport, E. L. (2020). What you say versus how you say it: Comparing sentence comprehension and emotional prosody processing using fMRI. *NeuroImage, 209*. https://doi.org/10.1016/j.neuroimage.2019.116509

Spencer, K. M., Dien, J., & Donchin, E. (2001). Spatiotemporal analysis of the late ERP responses to deviant stimuli. *Psychophysiology, 38*(2), 343–358. https://doi.org/10.1111/1469-8986.3820343

Vitale, J., Kosson, D. S., Resch, Z., & Newman, J. P. (2018). Speed-accuracy tradeoffs on an affective lexical decision task: Implications for the affect regulation theory of psychopathy. *Journal of Psychopathology and Behavioral Assessment, 40*(3), 412–418. https://doi.org/10.1007/s10862-018-9652-z

Whitehead, J. C., & Armony, J. L. (2018). Singing in the brain: Neural representation of music and voice as revealed by fMRI. *Human Brain Mapping, 39*(12), 4913–4924. https://doi.org/10.1002/hbm.24333

Yu, K., Chen, Y., Wang, M., Wang, R., & Li, L. (2022). Distinct but integrated processing of lexical tones, vowels, and consonants in tonal language speech perception: Evidence from mismatch negativity. *Journal of Neurolinguistics, 61*. https://doi.org/10.1016/j.jneuroling.2021.101039

Yu, K., Chen, Y., Yin, S., Li, L., & Wang, R. (2022). The roles of pitch type and lexicality in the hemispheric lateralization for lexical tone processing: An ERP study. *International Journal of Psychophysiology, 177,* 83–91. https://doi.org/10.1016/j.ijpsycho.2022.05.001

Yu, K., Li, L., Chen, Y., Zhou, Y., Wang, R., Zhang, Y., & Li, P. (2019). Effects of native language experience on Mandarin lexical tone processing in proficient second language learners. *Psychophysiology, 56*(11). https://doi.org/10.1111/psyp.13448

Yu, K., Wang, R., & Li, P. (2020). Native and nonnative processing of acoustic and phonological information of lexical tones in Chinese: Behavioral and neural correlates. In H.-M. Liu, F-M. Tsao, & P. Li (Eds.), *Speech perception, production and acquisition: Multidisciplinary approaches in Chinese languages* (pp. 79–99). Springer. https://doi.org/10.1007/978-981-15-7606-5_5

Yu, K., Zhou, Y., Li, L., Su, J. A., Wang, R., & Li, P. (2017). The interaction between phonological information and pitch type at pre-attentive stage: An ERP study of lexical tones. *Language, Cognition and Neuroscience, 32*(9), 1164–1175. https://doi.org/10.1080/23273798.2017.1310909

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1252. https://doi.org/10.1037/xhp0000216

Appendix A

Negative Words

不适 (Sick)	嫉妒 (Jealous)	愚蠢 (Stupid)	痛苦 (Painful)
丑陋 (Ugly)	干燥 (Dry)	懒惰 (Lazy)	破旧 (Shabby)
内疚 (Sinful)	心烦 (Upset)	挫败 (Failed)	空白 (Blank)
冷淡 (Aloof)	恐怖 (Horrid)	无助 (Helpless)	粗鲁 (Rude)
反常 (Abnormal)	恐惧 (Fearful)	无用 (Useless)	糟糕 (Terrible)
发疯 (Mad)	恶意 (Wicked)	暴力 (Violent)	羞耻 (Shamed)
可怜 (Pathetic)	悲剧 (Tragic)	暴怒 (Enraged)	虚弱 (Weak)
困惑 (Confused)	惊吓 (Scared)	残酷 (Cruel)	贫穷 (Poor)
失去 (Lost)	愚笨 (Stupid)	潮湿 (Wet)	错误 (Wrong)

Appendix B

Positive Words

亲切 (Kind)	宝贵 (Precious)	有望 (Hopeful)	纯净 (Pure)
优雅 (Lovely)	富有 (Wealthy)	有用 (Useful)	绝妙 (Incredible)
伶俐 (Bright)	干净 (Clean)	杰出 (Brilliant)	苗条 (Slender)
充足 (Satisfied)	平静 (Calm)	极好 (Fabulous)	英俊 (Handsome)
关心 (Caring)	强壮 (Strong)	活泼 (Alive)	被爱 (Loved)
周到 (Thoughtful)	忠实 (Honest)	深厚 (Deep)	超级 (Super)
和蔼 (Friendly)	忠诚 (Loyal)	温和 (Gentle)	辉煌 (Brilliant)
嬉戏 (Joyful)	文雅 (Elegant)	神圣 (Divine)	迷人 (Charming)
完整 (Full)	明智 (Wise)	神奇 (Magical)	高级 (High)

Appendix C

Neutral Words

不变 (Constant)	基础 (Basic)	当地 (Local)	私人 (Private)
中心 (Central)	大量 (Plural)	接近 (Near)	简短 (Brief)
中立 (Neutral)	奇异 (Fantastic)	收集 (Collected)	自动 (Automatic)
主要 (Main)	实际 (Actual)	敞开 (Open)	著名 (Famous)
仔细 (Careful)	宽阔 (Broad)	日常 (Daily)	规律 (Regular)
公民 (Civil)	平均 (Average)	明显 (Plain)	通常 (Usual)
共同 (Common)	平坦 (Flat)	普通 (Plain)	锋利 (Sharp)
华丽 (Gorgeous)	平方 (Square)	有关 (Related)	随便 (Informal)
圆形 (Round)	年度 (Annual)	浓密 (Thick)	随意 (Casual)

Appendix D

Pseudowords

世留	后政	忙两	线造
业屋	吗送	快先	者没
个开	吧并	忽质	背低
中使	呀打	情口	脑吃
举则	呢击	慢志	船度
了围	呼院	成更	药应
于深	品功	抗首	行突
从区	响孩	护严	被别
位跳	哥太	改把	规即
住米	啊坐	整影	视校
倒书	啦技	文马	评且
光金	嘴西	显天	读亮
八棉	器黑	晴咱	走六
兴找	回只	朋种	赶己
具层	处听	本决	身推
农南	复都	母眼	轻精
准叫	够条	油强	这续
出故	头全	海少	速姐
刚机	娘失	满友	采冲
动沉	容但	火布	问刻
助片	对密	灯放	阳在
化府	将最	由结	集专
单帝	床认	界在	雨卖
及旁	式给	看百	青钱
反早	弟律	确家	静哪
发特	张况	积钢	音团
句红	很调	系坏	须晚