IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2018, accepted February 27, 2018, date of publication March 8, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2813358

Investigating Duration Effects of Emotional Speech Stimuli in a Tonal Language by Using Event-Related Potentials

JIANG CHANG, XUEYING ZHANG, QIPING ZHANG, AND YING SUN

College of Information Engineering, Taiyuan University of Technology, Taiyuan 030024, China

Corresponding authors: Xueying Zhang (tyzhangxy@163.com) and Qiping Zhang (qiping.zhang@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61371193.

ABSTRACT Studying event-related potentials (ERPs) is considered as an effective method for investigating cerebral mechanisms of processing emotional speech. It has been shown that the amplitudes of ERP components in the cognitive processing of emotional speech are modulated by acoustic characteristics, such as valence and arousal. However, whether the duration of emotional speech stimuli impacts emotion-related cognitive processing remains unclear. To better understand the effect of emotional speech stimulus duration on emotion-related cognitive processing, we explored whether emotional speech ERPs were influenced by the duration of stimuli presented. Specifically, this paper focused on the ERP investigation of different durations (short: 0.50–1.00 s; medium: 1.50–2.00 s; and long: 2.50–3.00 s) of Chinese emotional speech stimuli. Chinese is a typical tonal language, and the stimuli were excerpted from radio plays in order to make emotions more obvious and easier to distinguish. We investigated the three different stages of the emotional speech processing: sensory processing, salience detection, and cognition. During the experiment, participants passively listened to emotional utterances matched for semantics and prosody with four emotions (sadness, anger, happiness, and surprise). Our results showed significant differences in the amplitudes of ERP components for different emotions during short-duration emotional speech stimuli. These findings suggest that shorter duration emotional speech stimuli may be more effective for separating the ERP components representing different emotions (N100, P200, and N300).

INDEX TERMS Emotional speech, event-related potentials (ERPs), speech duration, speech perception, auditory emotion.

I. INTRODUCTION

Emotional speech refers to a special voice signal that has evolved through human development and is an indispensable information carrier for interpersonal communication [1]. Perceiving and identifying emotional information from speech stimuli play important roles in social life [2] and also have enormous adaptive value [3]. Emotional information in speech is usually expressed by semantic content [4] and prosody [5]. In general, speech stimuli are presented in chronological order. The presented semantic content is also different at different times for emotional speech. This difference is mainly manifested in the specific speaker of the voice, pronunciation, tone, and so on, where tone is the use of pitch contour in language to express emotional and other paralinguistic information (e.g., emphasis, contrast), as well as to distinguish lexical or grammatical meaning. Those languages that use tones to distinguish words or their inflections are tonal languages, such as Chinese, and those that do not are non-tonal languages, such as English, in which tone indicates nothing about the meaning of the word [6]. Mandarin is the standard Chinese language and has four tones and one neutral tone, the pronunciation of which is usually transcribed by using Hanyu/Chinese pinyin (henceforth simply pinyin). For example, the Chinese syllable ''ma'' can mean ''mother'' (pinyin: mā) with a high level tone, ''hemp'' (pinyin: má) with a high rising tone, ''horse'' (pinyin: ma) with a low dipping tone, or ''scold'' (pinyin: mà) with a high falling tone, or it can act as an interrogative particle (pinyin: ma) with no specific tone (neutral tone).

Chinese is a typical tonal language. Some researchers found that ''disgust'' or ''anger'' can be expressed by a falling successive addition tone, and ''happiness'' or ''surprise''

by a rising successive addition tone [7]. Chinese and English speakers use three regions in the left hemisphere: the inferior frontal gyrus, the anterior superior temporal gyrus, and the posterior middle gyrus [6]. Compared with English speakers, Chinese speakers perform better in frequency-modulated identification tasks [8]. Furthermore, Chinese speakers showed additional activation for tone perception in the frontal, parietal, and parietal regions of the left hemisphere, while English speakers showed activation in the lower right frontal cortex that is consistent with the role of the right hemisphere in pitch perception [9]. In addition, Chinese and English speakers have different styles of expression. English speakers are used to putting the result in front, followed by content and explanation, while in Chinese, the opposite is true. For example, the Chinese utterance ''我非常喜欢他 (Pinyin: wǒ fēi cháng xǐhuān tā) is translated into English as ''I like it very much''. In Chinese, if you do not hear the last word, then you will not know what it wants to express. Furthermore, the level of understanding of listeners in the moment also increases with the progress of time and the completeness of utterances, especially for Chinese language. Therefore, the emotional information conveyed by emotional speech should vary at different times based on the the completeness of the utterance. For example, a speech sample of 500 ms may consist entirely of emotional information. In contrast, a speech sample of 1500 ms that contains the same 500 ms of emotional content would have a lower concentration of emotional information (only one-third of its total duration). Thus, there should be a difference in emotional cognition for emotional utterances of different durations at the same moment. It remains unknown whether this difference is reflected in different ERP components. ERP measures are very time sensitive and can accurately extract signals related to cognitive processing on the scale of milliseconds. If the duration of an emotional utterance conveys its meaning, can ERP components be influenced by the duration of the emotional speech? Investigating the effect of ERP component duration on emotional speech processing is thought to be important for better understanding emotional perception.

Previous studies have proposed that processing emotional speech consists of three stages: sensory processing, meaning-related processing, and integration and identification of emotional information [10]. Most behavioral and electrophysiological researchers now agree that early processing of emotional information (first stage: N100/N1, second stage: P200/P2) is influenced by acoustic characteristics (e.g., fundamental frequency, F0, and intensity) [10], [11], and the third stage (including N300/N3 [12], [13], P300 [14]) and the late positive component (LPC) [15], [16]) are influenced by emotions, semantics and prosody [14] as well as by arousal (exciting or not exciting, active or inactive) [16]. Furthermore, previous studies have also reported that the left hemisphere responds more to linguistic prosody, while the right brain responds more to effective prosody [17]–[19]. In contrast to the right hemisphere, the left hemisphere seems to process intelligible speech [10], [20]. In a study of emotional speech duration, Dmitrieva *et al.* [21] found that recognition efficiency and response time were the best in cases of neutral and negative emotional speech for different durations, and the shortest stimulus duration (above 0.5 s) could be perceived by 14–17-year-olds. In addition, prior studies have demonstrated differences in the time required for identifying different emotions. Pell and Kotz [22] found that recognition rates for different emotions were associated with different stages of speech. For example, recognition rates for sadness, anger, fear and neutral were faster than those for happiness and disgust, and recognition results for happiness quickly increased following the end of the speech. However, to date, only behavioral studies have been performed to explore whether the duration of related speech stimuli impacts emotional speech recognition. Previous studies have indicated that the cognitive processing of emotional speech progressively unfolds over time [23]. Thus, the current investigation attempts to fill this gap in the literature by investigating whether duration impacts emotional speech processing.

In previous ERP studies of emotional speech, the prosody and semantic content of emotional speech were studied separately. For prosody studies, pseudo-sentences and pseudo-utterances (nonsense speech) with no semantic information [24], [25] or neutral semantic words were selected as emotional speech stimuli [13] in order to avoid interference of semantics on emotion recognition. For example, when meaningful utterances were included in the stimuli, researchers generally attempted to demonstrate the dynamic integration of emotional prosody and semantics [26]. These previous studies shared two common features. First, most emotional speech materials were recorded by professional actors. According to the results, however, emotion recognition was not influenced by speaker voice [16] or experimental tasks [4], [27]. Second, numerous emotional speech samples were created with minimal semantic content by choosing neutral text or pseudo-sentences. In addition, these previous studies differed in various aspects of the speech stimuli: the number of emotion categories (e.g., 3 types [13], [28], [29]; 5 types [23]; 7 types [24], [30]), the types of emotion investigated (e.g., happiness, and anger; anger, disgust, fear, and happiness; anger, disgust, fear, sadness, and happiness; anger, disgust, fear, happiness, surprise, and sadness), the content (e.g., words or sentences) and the language (e.g., English, German or Chinese) of the emotional speech. These previous studies consistently demonstrated that there are some differences between neutral and emotional speech and that the speech stimuli of different emotional categories (for example, anger and happiness, neutral and emotional) are associated with different P200 response amplitudes [16], [17], [30]. However, ERP results for emotional speech remain inconsistent. For example, in one report, angry speech had a larger P200 amplitude than happy speech [28], whereas other studies showed no difference between these two emotions [13], [16]. Overall, because of the different stimulus materials used in the literature, it remains challenging to determine which

factor(s) will affect ERP components of emotional speech stimuli. One possible explanation is that the durations of these emotional stimuli were different, varying from 0.12–3.50 s on average [16], [22], [24]. Consequently, whether duration will affect the ERP components during cognitive processing of emotional speech remains an open question.

To date, few studies have directly examined the effect of emotional speech stimulus duration on emotional perception. Determining how the duration of emotional speech stimuli impacts ERP components such as the N100, P200, and N300 could help provide a more complete understanding of the cognitive processing of emotional speech.

In this study, we took advantage of the temporal resolution of ERP methodology to explore whether ERP components were influenced by the duration of stimuli presented in emotional speech. We addressed the following three questions: (1) Do ERPs show group differences for emotional stimuli of different durations? (2) Do emotional stimuli of different durations impact different ERP components (N100, P200, and N300)? and (3) Do short-duration emotional stimuli significantly impact N100 amplitude? In addition, we asked whether the ERP N100 and P200 components showed greater amplitudes in the left hemisphere.

II. METHOD

A. PARTICIPANTS

A total of twenty right-handed native speakers of Chinese (10 females and 10 males, average age: 27 years, range: 22–35 years) participated in the experiment. All participants were graduate students with no reported hearing or neurological disorders and had normal or corrected-to-normal vision. The research protocol was approved by the Institutional Review Board of the Institute of Psychology, Chinese Academy of Sciences (H16012) and was conducted according to the principles expressed in the Declaration of Helsinki. All participants provided written informed consent and were given a well-prepared gift for their contribution.

B. STIMULUS MATERIAL

1) SPEECH SELECTION

Stimulus materials were selected by our lab from the emotional speech database of Taiyuan University of Technology (TYUT2.0) [31] for their clear, strong and easily distinguishable characteristics of emotional expression. Specifically, four types of emotional utterances, expressing sadness, anger, happiness, and surprise, were selected. The emotional speech database can be downloaded from the following website: http://www.tyut.edu.cn/cie/szysp/News_View.asp?NewsID=624. These utterances were excerpted from a Chinese radio play (drama, without background music). Very different from the usual proscenium theatre or films, radio plays are a "blind medium" that is totally dependent on sound [32], [33], using various artistic means (e.g., dialogue, music, and sound effects) to create audio images and depict scenarios. Sometimes, necessary comments may be added to help the audience better understand the scenarios. We chose radio play as the experimental material because such emotional utterances are richer in content and have more distinct emotions, which aided participants in quick and accurate identification of emotions.

In addition, it must be noted that the type and number of emotional utterances in TYUT2.0 are still constantly improving and increasing. The database currently contains 160 sadness, 180 anger, 111 happiness, 305 surprise, and 153 neutral samples. It contains more distinct speakers than speech databases recorded by professionals and almost equal numbers of male and female voices, all of which sound like young people between 20 and 40 years old. However, some studies have noted that the speaker's voice and gender have nothing to do with the ERP findings [16], [24]. Therefore, in our study, ignoring the speaker's age and gender, we selected 90 samples of each emotion (sadness, anger, happiness, and surprise) that met our duration requirements, for a total of 360 samples used in the experiments.

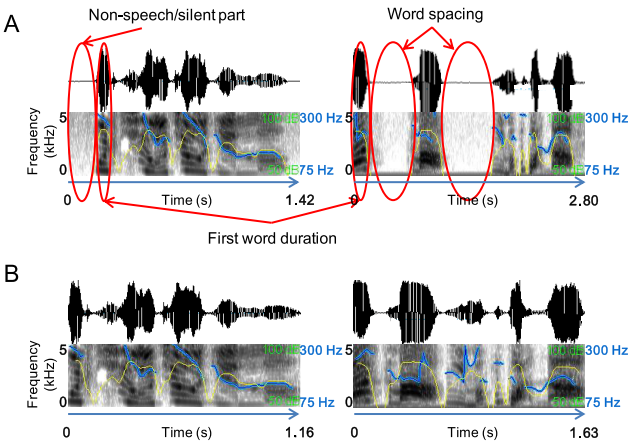

FIGURE 1. Speech stimulus before and after preprocessing. Oscillograms (top panel) and spectrograms (bottom panel) of the speech were produced by using Praat software. The dark shading indicates the energy of frequencies up to 5 kHz. The superimposed blue lines represent the fundamental frequency contour, which is perceived as speech melody. The superimposed yellow lines represent the intensity, which is up to 100 dB. Top (A): Invalid original speech stimulus for the superposition of ERP waveforms. The left utterance is sadness, "wǒ qiú qiú nǐ lā!" ("I beg you! "). The right utterance is surprise, "ā ? nǐ zěn me bú guǐčn ā ?" ("what? why do not you care? "). Bottom (B): Preprocessed speech signal by endpoint detection algorithm.

2) PREPROCESSING

It is well known that EEG has a high temporal resolution and can reflect a participant's cognitive processing in real time. An ERP is induced by a specific stimulus, and its acquisition usually requires multiple EEG signals representing the same event to be superimposed. However, speech stimuli are presented in chronological order; for example, for the speech stimulus shown in Fig. 1 (A), it is difficult to effectively superimpose the ERP waveform due to the inconsistency of the starting position and the spacing of the words in the speech materials. In particular, the starting position has a direct

impact on the extraction of the ERP waveform, requiring that the beginning of the speech signal not be silent—there must be a voice. That is, the time tolerance of waiting for speech playback is zero for a speech signal.

To ensure the effective superposition of ERP waves, signals were preprocessed by performing endpoint detection on all stimuli. An example of a preprocessed speech signal is shown in Fig. 1 (B). This process does not affect the acoustic characteristics such as fundamental frequency, energy, and intensity, and it only addresses the starting non-speech portion, word spacing and the duration of the first word of speech signals. First, removing the non-speech/silent part at the beginning of the speech signal ensured that all speeches started at the same time [34], i.e., speech stimuli were presented beginning at time zero. Second, adjusting the word spacing was necessary to reduce differences between the speech samples. Based on previously used experimental methods in speech signal processing, if the wording space duration was greater than 50 ms, then it was set to 50 ms; if it was less than 50 ms, then no change was made. Finally, adjusting the first word in each speech sample was also necessary to reduce between-sample differences. If the first word duration was longer than 400 ms, then no change was made because this duration comprised two words. If the first word duration was longer than 150 ms and less than 250 ms, then it was adjusted and compressed into 150 ms. It should be noted that very few utterances require this kind of processing. An endpoint detection algorithm was applied to emotional speech samples in MATLAB 7.0. The specific algorithm was as follows:

A double-threshold, two-stage detection method based on energy and zero-crossing rate was used to detect speech endpoints.

Step 1: A speech signal x was divided into equal frames to keep characteristics approximately constant in a short time interval. Each frame was recorded as $s_i(n)$, $n = 1, 2, \ldots, N$, where n is the discrete time series of the speech signal, N is the frame length, and i is the frame number.

Step 2: The short-term energy of each frame was as follows:

$$E_i = \sum_{n=1}^{N} s_i^2(n) \tag{1}$$

Step 3: The zero-crossing rate of each frame was as follows:

$$\mathrm{Z}_i = \sum_{n=1}^{N} |sgn[s_i(n)] - \mathrm{sgn}[s_i(n-1)]| \tag{2}$$

Where $s_i(n) \geq 0$, $\mathrm{sgn}[s_i(n)] = 1$, $s_i(n) \leq 0$, and $\mathrm{sgn}[s_i(n)] = 0$.

Step 4: Double-threshold two-stage detection method: for determining the speech start time, it was necessary to set a higher average energy threshold T_1 and a slightly lower threshold T_2, such that $T_2 = \alpha_1 E_N$, where E_N is the average energy of the noise segment, α_1 is the empirical parameters, and T_2 is used to determine the end of the speech. At this point, first-stage detection is complete. Based on the zero-crossing rate of the noise, Z_N, the second-stage detection also requires the experimenter to set a threshold T_3 for determining the onset and time in the speech. Adjusting threshold A enables detection of the speech endpoints and extraction of the first words. Start time detection of the speech sample and single word extraction x'_m (where m is the number of words) can be achieved by adjusting the threshold.

Step 5: Judging the duration of adjacent words: if the duration was greater than 50 ms, then it was set to 50 ms; if it was less than 50 ms, then no change was made.

Step 6: Judging the duration of the first word: if the first word was less than 150 ms in length, it was not processed. If it was longer than 150 ms, it was necessary to adjust the duration. Without affecting the acoustics, rhythm, integrity or fluency characteristics of the speech sample, one frame of the speech signal was randomly deleted from the first word. If it was still longer than 150 ms, then the procedure was repeated; otherwise, the algorithm exited the operation.

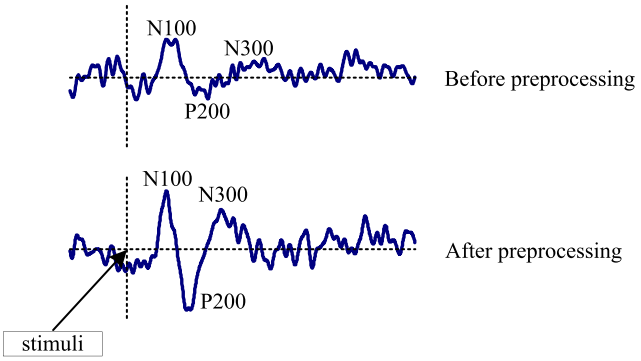

FIGURE 2. Observed ERP waveforms before and after preprocessing. For N100 and P200 components, although the latency is the same, the amplitudes of the components obtained are significantly larger after preprocessing. For the N300 component, the waveform's latency is advanced, and the peak of the waveform is more easily observed.

To ensure speech quality, all emotional speech was subject to a validity test after preprocessing. As seen in Fig. 2, the observed ERP waveform components (N100, P200, and N300) were more obvious after preprocessing.

3) EXPLANATION AND ANALYSIS

For the selected emotional speech sentences, the emotional language was required to be accurate and unambiguous. In addition, there were no significant differences in the intensity of emotional speech: intensity was maintained at approximately 80 dB for all speech durations.

Based on the duration of emotional speech, emotional speech samples were classified into three groups: short (0.50–1.00 s), medium (1.50–2.00 s), and long duration (2.50–3.00 s). Emotional speeches were an even mix of male and female voices for different durations. There were 360 emotional speech samples (90 for each emotion, divided equally among the three duration conditions). Neutral stimuli were a pure tone signal with a frequency of 250 Hz.

TABLE 1. Acoustic analysis of emotional speech: mean (standard deviation).

Duration	Parameter	Sadness	Anger	Happiness	Surprise
Short	Speech speed (words/s)	4.98(1.02)	6.99(1.82)	6.22(1.21)	6.27(1.47)
	F0 mean (Hz)	178.05(15.10)	185.76(19.33)	188.03(27.96)	165.82(22.44)
	F1 mean (Hz)	846.13(151.95)	899.91(139.89)	807.47(197.52)	893.30(236.17)
	F2 mean (Hz)	1746.36(225.64)	1823.86(191.04)	1841.58(160.43)	1875.45(224.74)
	Intensity mean (dB)	76.93(5.72)	81.12(3.21)	80.26(5.17)	76.05(3.98)
	Duration (s)	0.76(0.16)	0.70(0.20)	0.73(0.19)	0.73(0.17)
Medium	Speech speed (words/s)	3.43(1.29)	5.44(1.45)	4.28(1.31)	4.71(1.58)
	F0 mean (Hz)	186.29(17.29)	185.76(19.72)	171.81(22.03)	176.37(16.45)
	F1 mean (Hz)	820.73(173.23)	910.25(113.47)	823.37(171.74)	824.71(181.39)
	F2 mean (Hz)	1829.8(197.65)	1895.86(226.77)	1798.23(226.84)	1862.64(210.38)
	Intensity mean (dB)	78.38(4.89)	79.43(2.85)	77.97(4.23)	78.01(5.32)
	Duration (s)	1.76(0.17)	1.67(0.14)	1.82(0.17)	1.67(0.25)
Long	Speech speed (words/s)	3.03(0.93)	4.43(1.48)	3.75(0.86)	4.32(1.40)
	F0 mean (Hz)	195.97(25.21)	216.41(27.32)	185.05(27.02)	210.36(19.00)
	F1 mean (Hz)	787.12(125.61)	942.84(140.16)	819.97(123.84)	884.74(277.02)
	F2 mean (Hz)	1758.43(214.23)	1923.10(198.17)	1823.87(212.54)	1791.39(206.54)
	Intensity mean (dB)	82.22(3.06)	78.86(3.71)	77.44(5.07)	84.35(6.60)
	Duration (s)	2.63(0.36)	2.50(0.40)	2.63(0.37)	2.61(0.39)

The purpose was to make emotions easy to distinguish and to make the experimental task simple. Only those whose recognition rate exceeded 90% were selected for the official experiment. All stimuli were acoustically analyzed using Praat software (32-bit, version 5.4.22). Praat is an open-source, cross-platform, and multi-functional phonetic professional program for the analysis, labeling, processing and synthesis of speech in phonetics; it can be downloaded from this page: http://www.fon.hum.uva.nl/praat/. Table 1 presents the emotional acoustic parameters of speech speed, fundamental frequency F0, formant F1, formant F2, intensity, and duration for three durations. Among them, fundamental frequency F0, intensity and duration were obtained directly by Praat software, and speech speed = number of words/duration (s). However, formants F1 and F2 were obtained by setting the maximum formant frequency value of Praat. The selected parameters refer to the standards of 5000 Hz for men and 5500 Hz for women. Table 2 lists Chinese example sentences expressed in pinyin (Chinese phonetic alphabet), and literal English translations are provided in parentheses. There were 5 example sentences for each emotion at each duration. In Table 1, short duration refers to 3 to 6 words, medium durations to 5 to 10 words, and long durations to 7 to 12 words.

4) EXPERIMENTAL DESIGN AND DATA PROCESSING

The current study aimed to explore whether ERP waveforms elicited by emotional speech were influenced by the duration of presented stimuli. The experimental design is shown in Fig. 3(A). This is a 3 × 4 within-subjects design with factors of duration (short, medium, long) and emotion (sadness, anger, happiness, and surprise). In addition, ROIs were defined for ERP analysis based on scalp regions of interest. Six ROIs were defined for critical scalp sites: left frontocentral (LFC) FC1 and FC3; left central (LC) C3 and C1; left centro-parietal (LCP) CP3 and CP1; right frontal (RFC) FC2 and FC4; right central (RC) C2 and C4; and right centro-parietal (RCP) CP2 and CP4.

Based on existing ERP literature on emotional prosody processing [13], [24], three ERP components (N100, P200, and N300) were analyzed for the three stages of emotional speech processing.

5) PROCEDURE

Each participant was seated comfortably in an electrically shielded chamber in front of a computer monitor at a distance of 110 cm. The experiment was divided into three blocks (block 1: short duration; block 2: medium duration; block 3: long duration). Each block was composed of 150 (trials) stimuli, including 30 pure-tone signals and 120 emotional utterances. To avoid continuous presentation of the same type of emotion, stimuli were pseudo-randomized. Different orders of stimuli were presented to different participant in blocks of 150 trials each [35] During the experiment, each stimulus was only presented once. Participants were instructed to press a button as quickly as possible when the pure tone was presented and were not required to take any action when other sounds were heard. This design was intended to make the experimental task simpler and easier without identifying emotions. Participants were more likely to concentrate on tasks and did not feel tired. In addition, muscle artifacts can be reduced via this method.

Stimuli were pseudo-randomized and presented to participants in blocks of 150 trials. The order of the three blocks was counterbalanced. In total, 450 trials were presented across the entire experiment, and participants rested after each block. Each experimental trial began with a 500 ms fixation cross displayed centrally on the screen, followed by a 500 ms blank display. The sound stimulus (including emotion and pure tones) was subsequently presented for 3000 ms, as shown in Fig. 3(A). The entire experiment was completed in approximately 40 minutes.

6) PROCEDURE

The electroencephalogram (EEG) was recorded by using a 64-channel EEG Quick-cap (Neuroscan, USA). Bipolar horizontal and vertical electrooculograms were recorded for

TABLE 2. Example sentences.

Duration	Sadness	Anger	Happiness	Surprise
Short	wǒ zhī dào cuò le (I knew I had made a mistake)	nǐ xiū xiǎng (Over my dead body)	tài bàng le (That is fantastic)	nán dào shì tā (Could he)
	wǒ yuān ā (I was wronged)	bàn bú dào (That is beyond my power)	wǒ chéng gōng le (I made it)	nǐ gàn ma ne (What are you doing)
	bào yīng ā (This is karma)	wǒ gēn nǐ pīn le (I'll fight it out with you)	wǒ yào jié hūn le (I'm going to get married)	hái xián shǎo ā (It's too soon, ah)
	qiú qiú nǐ men (I beg you)	wǒ piān huí qù (I simply must go)	wǒ men yíng le (We won it)	zhè shì shí me (What is this)
	wǒ hèn nǐ (I hate you)	bié dé yì tài zǎo (Don't count your chickens before they are hatched)	zhēn shì guò yǐn ā (It was really terrific)	nǐ zěn me lái le (Why are you here)
Medium	wǒ kě lián de hái zǐ (My poor child)	zhè shì bàn bú dào (I really can't make it)	nǐ kě shì lì dà gōng le (You did a great job)	nǐ zěn me hái bú chī fàn (Why don't you have a meal)
	wǒ duì bú qǐ nǐ ā (I am so sorry for you)	wǒ hái yào dǎ nǐ ne (I also want to beat you)	wǒ hái yào xiè xiè nǐ men ne (Must thank you)	zěn me zhè me kuài jiù kàn wán le (How can you read so quickly)
	duì wǒ de dǎ jī tài dà le (It was a great shock to me)	hái gǎn shàng wǒ men jiā de mén (You dare to come to my home)	zhè jué zhāo hái zhēn de guǎn yòng ā (This trick really works.)	nǐ zěn me shí me dōu zhī dào (How do you know everything)
	dōu shì wǒ de cuò (It's all my fault)	mā de hún dàn jìng zhǎo wǒmá fán (Shit, shit, shit, dare find fault with me in trouble)	tài hǎo le tài hǎo le (Wonderful! Wonderful!)	nán dào tā men yǒu tóng huǒ (Do they have partners)
	wǒ de mìng hǎo kǔ ā (What a miserable destiny I've got)	bié shuō le nǐ yán sù diǎn (Stop! You must be serious)	yě duì yán zhī yǒu lǐ (The story sounded perfectly plausible.)	yě méi yǒu qīn qī hǎo yǒu ma (Are there no relatives or friends)
Long	cāng tiān bú zhǎ yǎn ā (The god is unfair)	wǒ xiǎng shā le nǐ wǒ qiā sǐ nǐ (I want to kill you by strangling you)	wǒ jiù zhī dào nǐ shì ài wǒ de (I know you love me)	nǐ bú shì yǐ jīng bèi zhuā dào le ma (Have you already been arrested)
	wǒ duì bú qǐ ér zǐ ā (I am sorry for my son)	nǐ men kuài gěi wǒ fàng kāi tā (Leave her alone)	méi cuò ā zhè gè rén jiù shì wǒ (Right, it is me)	zhè hái zǐ gāi bú huì chū shí me shì ba (Will this child be OK)
	zhè kě jiào wǒ zěn me huó ā (How can I live alone)	chòu xiǎo zǐ nǐ yòu chū mài le wǒ (You have betrayed me, brat)	wǒ yě duì nǐ guā mù xiàng kàn ā (I also view you differently)	zěn me kě néng yǒu 30wàn mǎi fáng zǐ ne (How can we have 300,000 for the house)
	wǒ de mìng wéi shí me zhè me kǔ (How poor my life is)	nǐ zhè gè qín shòu bú rú de dōng xī (You are worse than a beast)	wǒ yuàn yì wǒ dāng rán yuàn yì le (I am certainly willing)	nín jìn jiān láo qù zuò shí me (What are you doing in prison)
	nǐ zěn me sǐ de zhè me cǎn ā (How miserably you died)	tiān hái méi tā xià lái wǒ hái méi sǐ ne (The heaven will not fall down because I did not die)	hǎo de hǎo de zhēn shì xiè xiè nǐ men le (Ok thank you very much)	nǐ zěn me bǎ chē tíng zài xuán yá biān shàng (Why do you stop the car beside the cliff)

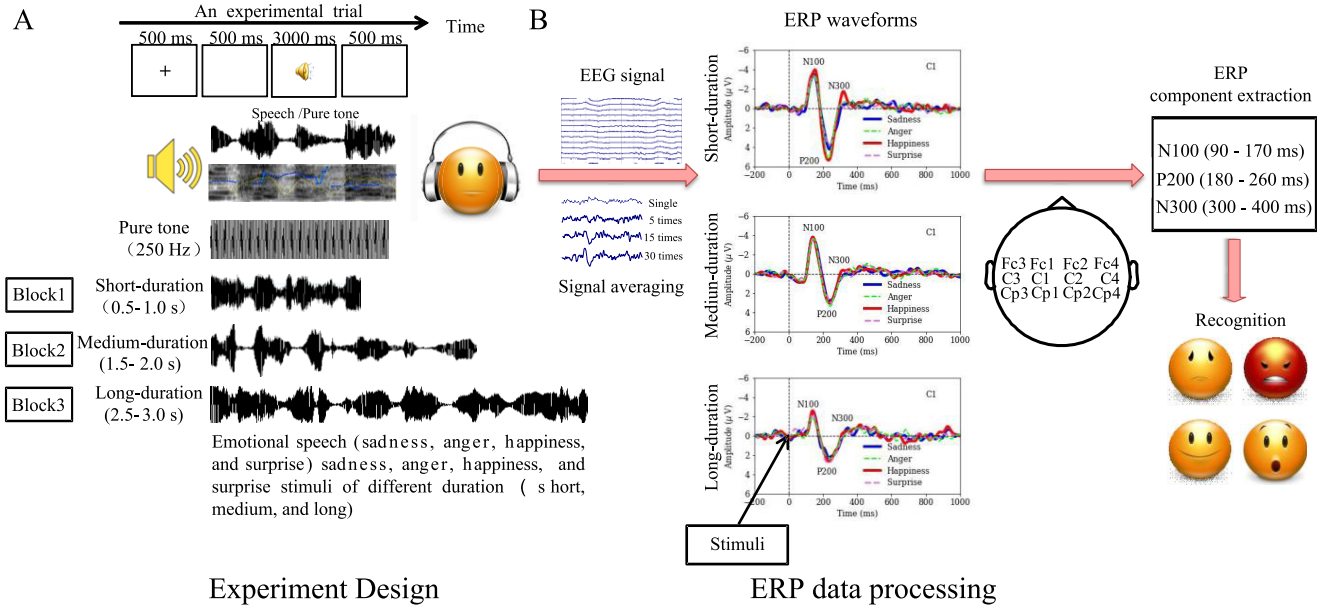

FIGURE 3. Experimental design and data processing. Left (A): experimental design flow chart for three durations (short, medium, and long) of emotional (sadness, anger, happiness, and surprise) stimuli. Right (B): data processing from EEG signal to the ERP signal.

purposes of artifact rejection. Signals were continuously recorded with a bandpass filter for 0–200 Hz and digitized at a sampling rate of 1000 Hz. Additionally, electrode resistance was kept below 5 kΩ. The EEG reference was placed on the tip of the nose and grounded by the cap electrode.

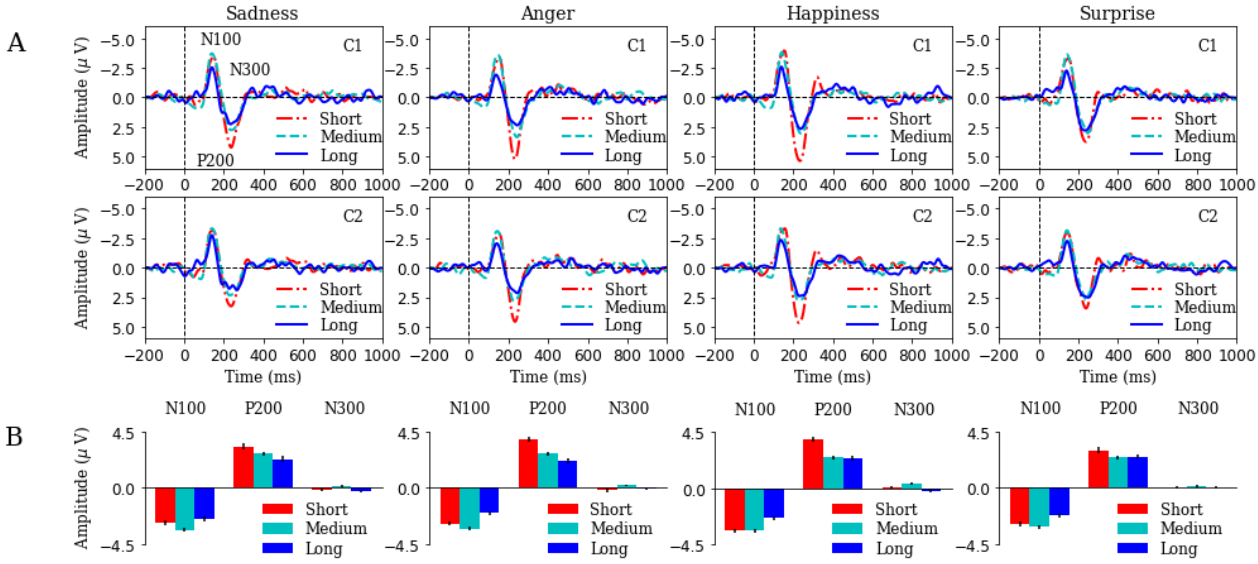

FIGURE 4. Average waveforms and amplitudes for different emotional conditions at different durations. Top (A): grand average waveforms at C1 and C2 electrodes for emotional (sadness, anger, happiness, and surprise) stimuli at different durations (short, medium, and long). Bottom (B): average amplitudes (N100, P200, and N300) for emotional (sadness, anger, happiness, and surprise) stimuli at different durations (short, medium, and long)

For offline analysis (see Fig. 3(B)), data were resorted based on the reference average and filtered offline with a 0.5–30 Hz bandpass filter. Trials with movement and electrooculogram artifacts exceeding ± 100 μV were omitted from the average. ERP waveforms were time-locked from stimulus onset, and waveforms in the time range between 200 ms before stimulus onset and 1,000 ms after stimulus offset were averaged. Additionally, the 200 ms pre-stimulus period was used in the baseline correction.

7) DATA ANALYSIS

Behavioral responses were not reported because the experimental task was used to ensure that participants were able to complete the experiment carefully. Peak amplitude was calculated between 90 and 170 ms (N100 component) and 180 and 260 ms (P200 component). Mean amplitude was calculated between 300 and 400 ms (N300 following negativity) after stimulus onset.

This study used the IBM SPSS Statistics 19.0 software package. Amplitudes were entered into a repeated-measures ANOVA based on the within-subject factors of speech duration (short, medium, long) and emotion (sadness, anger, happiness, surprise). Six regions of interest (ROIs) were used for ERP analysis. Analyses were corrected for nonsphericity using the Greenhouse-Geisser method. All significant tests were two-tailed at the preset significance alpha level of $p<0.05$. In addition, post hoc tests were conducted as appropriate, using a Bonferroni correction for multiple comparisons.

The present work first reports the main effects of duration and emotion as well as their interaction effect. Next, ROI analysis with follow-up comparisons on the lateral plane (left hemisphere (LH): LFC, LC, LCP; right hemisphere (RH): RFC, RC, RCP) and sagittal plane (fronto-central (FC): LFC, RFC; central (CN): LC, RC; centro-parietal (CP): LCP, RCP) is presented. The results are presented as mean amplitudes and their standard deviations.

III. RESULTS

A. ERP DATA

The analysis of ERP data revealed differences in how the human brain processed different durations of emotional speech stimuli at the level of the N100, P200, and N300. Fig. 4 shows grand average waveforms at C1 and C2 electrodes as well as amplitudes (N100, P200, and N300) for different emotional conditions at different durations. Fig. 5 shows the ERP responses to different emotions for short-duration stimuli. Fig. 6 shows the ERP responses to different emotions for medium-duration stimuli. Fig. 7 shows the ERP responses to different emotions for long-duration stimuli.

B. ERP COMPONENT ANALYSIS

1) N100 PEAK AMPLITUDE (90-170 MS)

The main effect of duration was significant ($F(2, 57) = 22.64$, $p<0.001$). As depicted in Fig. 4, the speech stimuli with medium duration (-3.30 (0.18)) showed stronger N100 amplitudes than short- (-3.01 (0.19)) and long-duration (-2.29 (0.15)) stimuli. The interaction of duration and emotion was also significant ($F(6,171) = 3.70$, $p<0.01$). The post hoc Bonferroni-corrected tests showed that the N100 amplitude was more negative in the happiness emotion condition (-3.44 (0.18)) than for the other three emotions ($p<0.05$), but only in the short-duration condition (Fig. 5). There were no differences in N100 amplitudes between the 4 emotions

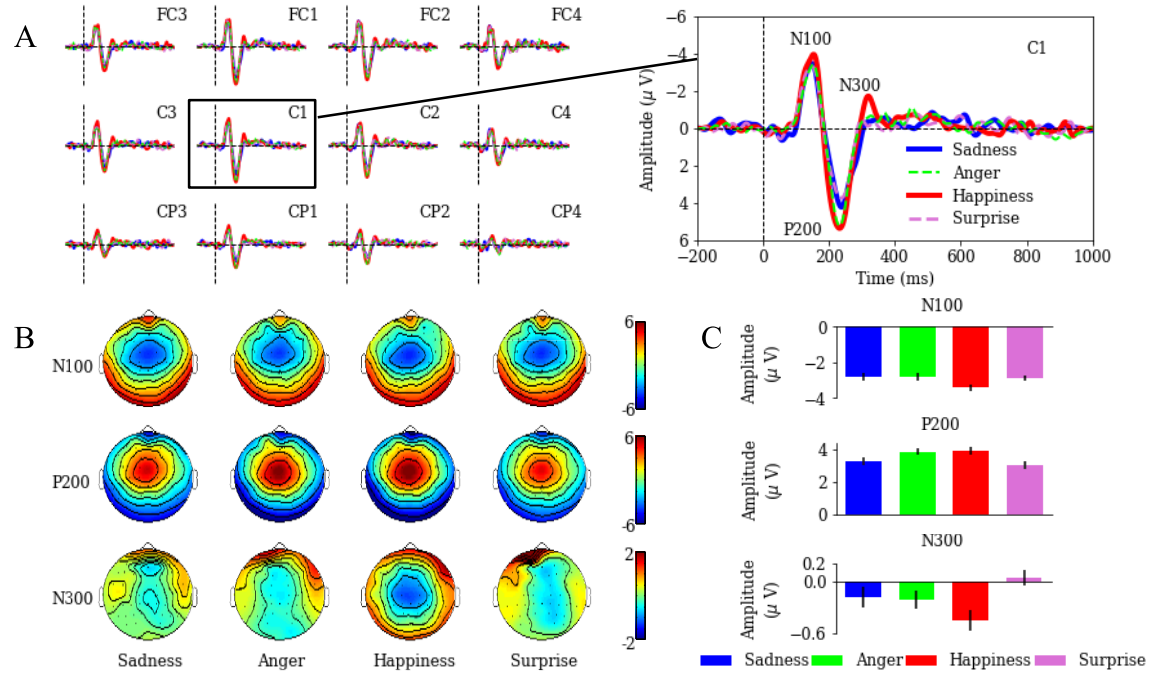

FIGURE 5. ERP responses to different emotions for short-duration stimuli. Top (A): ERP waveforms are shown at FC1, FC2, FC3, FC4, C1, C2, C3, C4, CP1, CP2, CP3 and CP4 for different emotions (sadness, anger, happiness, and surprise). Lower left (B): voltage topographies of the N100, P200 and N300 components on the scalp are shown for each emotion (sadness, anger, happiness, and surprise). Lower right (C): average amplitudes (N100, P200, and N300) for emotional (sadness, anger, happiness, and surprise) stimuli in the short-duration condition.

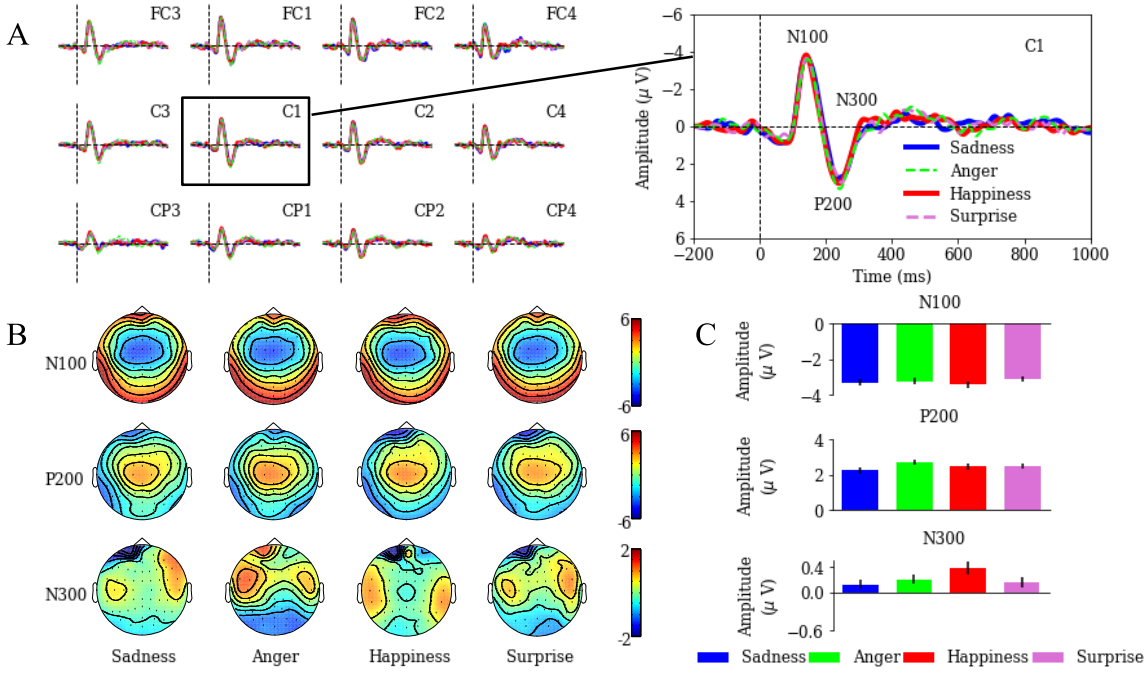

FIGURE 6. ERP responses to different emotions for medium-duration stimuli. Top (A): ERP waveforms are shown at FC1, FC2, FC3, FC4, C1, C2, C3, C4, CP1, CP2, CP3, and CP4 for different emotions (sadness, anger, happiness, and surprise). Lower left (B): voltage topographies of the N100, P200, and N300 components on the scalp are shown for each emotion (sadness, angry, happiness, and surprised). Lower right (C): average amplitudes (N100, P200, and N300) for emotional (sadness, anger, happiness, and surprise) stimuli in the medium-duration condition.

(sadness: −3.35 (0.18), anger: −3.25 (0.18), happiness: −3.44 (0.18), and surprise: −3.14 (0.16)) for the medium-duration condition (Fig. 6). For the long-duration condition, angry speech stimuli (−2.03 (0.15)) triggered lower N100 amplitudes than sad (−2.51 (0.18)) or happy (−2.41 (0.23)) stimuli ($p<0.01$).

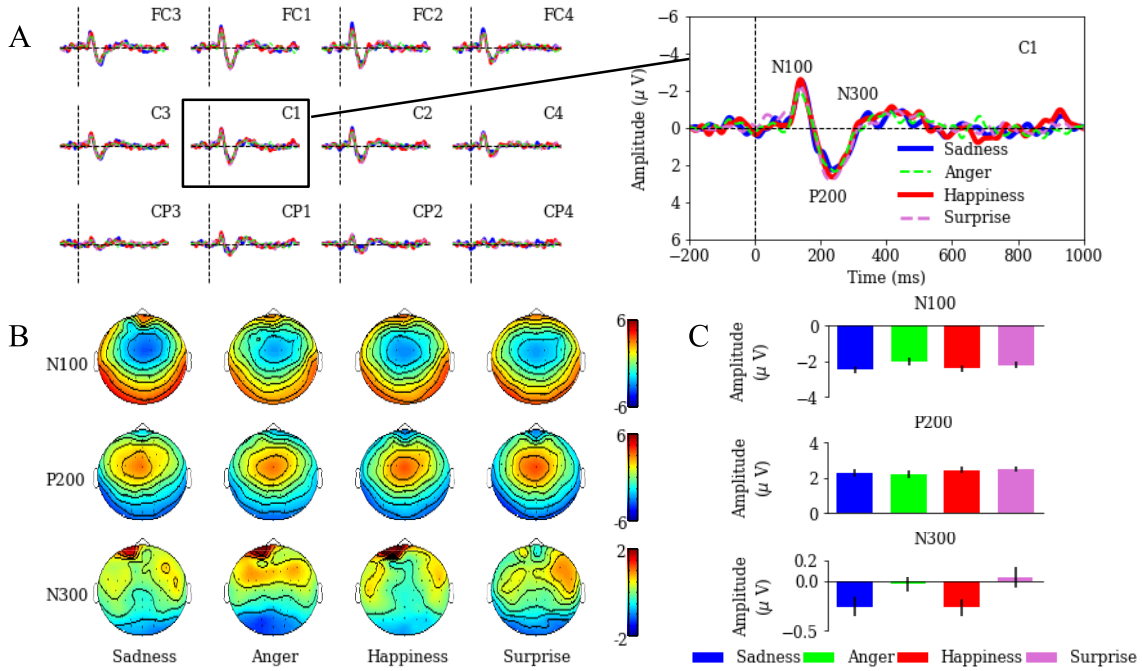

FIGURE 7. ERP responses to different emotions for long-duration stimuli. Top (A): ERP waveforms are shown at FC1, FC2, FC3, FC4, C1, C2, C3, C4, CP1, CP2, CP3 and CP4 for different emotions (sadness, anger, happiness, and surprise). Lower left (B): voltage topographies of the N100, P200 and N300 components on the scalp are shown for each emotion (sadness, anger, happiness, and surprise). Lower right (C): average amplitudes (N100, P200, and N300) for emotional (sadness, anger, happiness, and surprise) stimuli in the long-duration condition.

ROI analysis demonstrated a significant main effect of region ($F(5, 285) = 202.67$, $p<0.001$). Lateral plane comparisons revealed significantly larger N100 amplitudes in the left hemisphere (-3.49 (0.52)) than in the right hemisphere (-3.09 (0.68)) ($t(19) = 4.148$, $p<0.01$) in the medium-duration condition, and no significant differences were found in the other two duration conditions ($p>0.05$). Sagittal plane comparisons showed significant differences among three regions. The N100 amplitudes were consistently highest in FC (short: -3.63(0.13); medium: -4.33 (0.15); and long: -2.91 (0.18)), followed by CN (short: -3.27 (0.16); medium: -3.62 (0.13); and long: -2.46 (0.08)) and CP (short: -2.10 (0.10); medium: -1.92 (0.13); and long: -1.54 (0.08)) across all three duration conditions. The results are shown in Fig. 8.

2) P200 PEAK AMPLITUDE (180-260 MS)

The main effect of duration was significant ($F(2, 57) = 34.95$, $p<0.001$). As shown in Fig. 4, this effect consisted of higher P200 amplitudes in the short-duration condition (3.53 (0.23) than those in both the medium- (2.47 (0.16)) and long-duration (2.35 (0.19)) conditions. The interaction effect of duration and emotion was also significant ($F(6,171) = 8.27$, $p<0.001$). In the short-duration condition alone, the 4 emotions showed significant differences in P200 amplitudes, with happiness (3.94 (0.23)) and anger (3.88 (0.21)) amplitudes higher than those for sadness (3.29 (0.25)) and surprise (3.01(0.23)) ($p<0.01$). There were no emotion-related differences for the medium- and long-duration conditions.

We found a significant effect for ROI ($F(2, 57)=3.79$, $p<0.05$). For lateral plane comparisons, the P200 amplitude was larger in the left hemisphere than that in the right hemisphere in both the medium-duration ($p<0.001$) and long-duration ($p<0.05$) conditions, but no significant difference was found in the short-duration condition. For sagittal plane comparisons, P200 amplitudes at CP were lowest in all three duration conditions (short: 2.48 (.10); medium: 2.01 (0.10); long: 1.61 (0.10)). The comparison of FC and CN showed mixed results. At short durations, there was no significant difference between these regions (FC: 4.12 (0.20); CN: 3.99 (0.12)). At medium and long durations, significant differences were observed between them (medium condition, FC: 2.59 (0.10), CN: 2.83 (0.11); long condition, FC: 2.92 (0.18), CN: 2.52 (0.13)). The results are shown in Fig. 9.

3) N300 MEAN AMPLITUDES (300-400 MS)

The main effect of duration was significant ($F(2, 57) = 3.79$, $p<0.05$). Post hoc comparisons indicated that the N300 amplitude at short durations (-0.20 (0.10)) was significantly higher than those at medium (-0.19 (0.09)) and long durations (-1.35 (0.09)). In addition, there was no significant difference between N300 amplitudes for medium and long durations. However, there was no interaction between the emotion and duration ($F(6,171) = 1.83$, $p>0.05$).

The main effect of ROI was significant ($F(5, 285) = 43.72$, $p<0.001$). For lateral plane comparisons, there was no significant difference between the left and right hemispheres in any of the three duration conditions ($p>0.05$). For sagittal

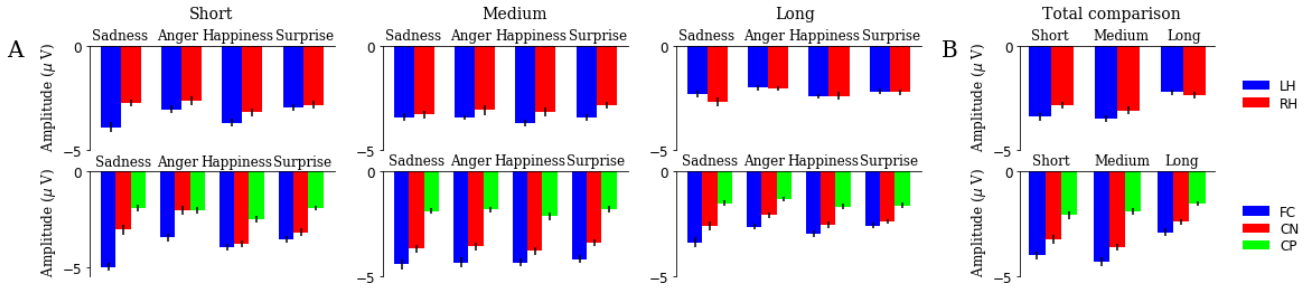

FIGURE 8. Average amplitude of N100 by duration, emotion, lateral region (LH, RH), and sagittal region (FC vs. CN vs. CP). Left (A): average amplitude (N100) of lateral regions (left: LH, right: RH) or sagittal regions (fronto-central: FC, central: CN, centro-parietal: CP) for the 4 emotions (sadness, anger, happiness and surprise) and 3 durations (short, medium, and long). Right (B): Total N100 average amplitudes for all emotions at different durations for lateral regions (LH, RH) and sagittal regions (FC vs. CN vs. CP).

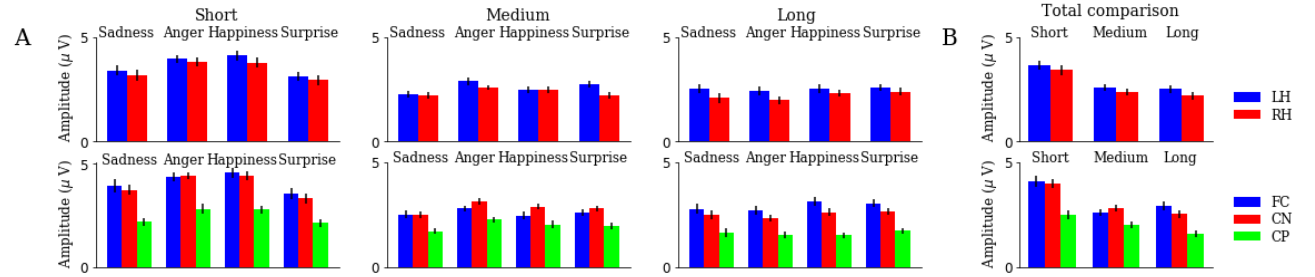

FIGURE 9. Average amplitude of P200 by duration, emotion, lateral region (LH, RH), and sagittal region (FC vs. CN vs. CP). Left (A): average amplitude (P200) of lateral regions (left: LH, right: RH) or sagittal regions (fronto-central: FC, central: CN, centro-parietal: CP) for the 4 emotions (sadness, anger, happiness and surprise) and 3 durations (short, medium, and long). Right (B): Total P200 average amplitudes for all emotions at different durations for lateral regions (LH, RH) and sagittal regions (FC vs. CN vs. CP).

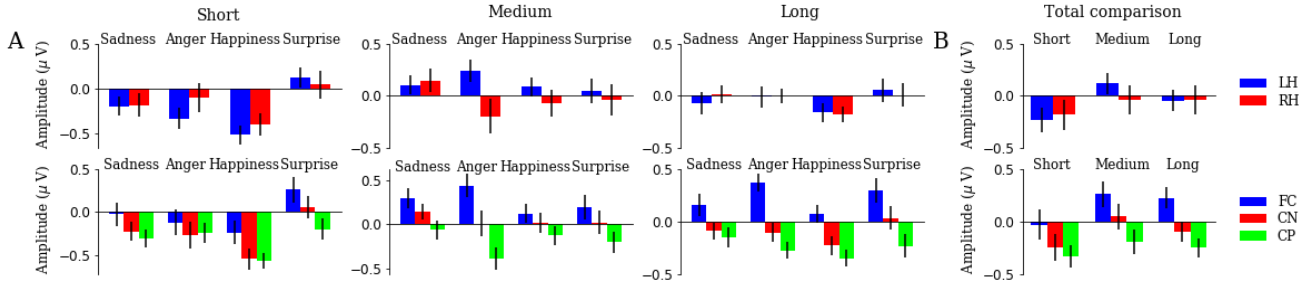

FIGURE 10. Average amplitude of N300 by duration, emotion, lateral region (LH, RH), and sagittal region (FC vs. CN vs. CP). Left (A): average amplitude (N300) of lateral regions (left: LH, right: RH) or sagittal regions (fronto-central: FC, central: CN, centro-parietal: CP) for the 4 emotions (sadness, anger, happiness and surprise) and 3 durations (short, medium, and long). Right (B): Total N300 average amplitudes for all emotions at different durations for lateral regions (LH, RH) and sagittal regions (FC vs. CN vs. CP).

plane comparisons, there were significant differences among the three regions. The N100 amplitudes were highest across all three duration conditions in CP (−3.33 (0.06)), followed by CN (−0.25 (0.09)) and FC (−0.03 (0.11)). The results are shown in Fig. 10.

C. CORRELATIVE ANALYSES

The association between duration of emotional speech and ERP amplitude was evaluated using Pearson's correlation analysis, and the results are shown in Table 3. Duration was positively correlated with N100 and P200 amplitudes: an emotional utterance of longer duration was associated with N100 and P200 components of smaller amplitude. No significant associations were observed for N300 amplitude except at FC locations.

TABLE 3. Correlation analysis between duration of emotional speech and ERP amplitude.

Brain regions	N100		P200		N300	
	r	p	r	p	r	p
LH	-0.513**	<0.001	-0.591**	<0.001	0.234	0.072
RH	-0.299*	<0.05	-0.625**	<0.001	0.126	0.197
FC	-0.368*	<0.05	-0.505**	<0.001	0.132*	0.026
CN	-0.459**	<0.001	-0.735**	<0.001	0.069	0.147
PC	-0.448**	<0.001	-0.617**	<0.001	0.083	0.295

**. p<0.001, *. p<0.05

IV. DISCUSSION

In the current study, we investigated the influence of emotional stimulus duration on ERP components for emotional speech in an effort to investigate cognitive processing of

emotional speech perception. To this end, we presented Chinese emotional speech stimuli of different durations (short: 0.50–1.00 s, medium: 1.50–2.00 s, and long: 2.50–3.00 s). Participants passively listened to emotional utterances matched for semantic content and prosody with four emotions (sadness, anger, happiness, and surprise). Our results showed that ERP effects were observed at all three durations, which further indicated the sensitivity of the N100, P200, and N300 components to prosodic manipulations in emotional speech [18], [24], [29], [36], [37]. The amplitudes of the N100 and P200 components were associated with the duration of emotional speech. A short-duration emotional speech was more likely to induce a larger ERP waveform, which made it easier to observe differences between emotions for ERP components (i.e., the N100, P200, and N300). In the ROI analysis, N100 and P200 amplitudes were highest in the FC region, and N300 amplitudes were the highest in the CP region. In addition, semantic processing occurred to a greater degree in the left hemisphere than the right hemisphere at medium (N100, P200) and long durations (P200). The differences between emotions observed for the N100, P200, and N300 suggests that the duration of emotional speech plays a critical role in emotional cognitive processing. These findings also suggest a hemispheric lateralization in early semantic processing of emotional speech.

A. N100: EFFECTS OF DIFFERENT DURATIONS ON SENSORY PROCESSING OF EMOTIONAL SPEECH

Evidence exists in the literature that the N100/N1 is the first stage at which emotional prosody is processed [38] and is likely to be influenced by physical characteristics of stimuli or automatic attention allocation [39]. For example, it has been reported that N100 amplitude increases with pitch salience [40] and is increased by attended stimuli [19], [41]. In addition, one study reported that the N100 amplitude was more negative in happiness and anger conditions relative to neutral prosody [13]. The results of the current study are consistent with this claim. In our results, the N100 amplitude was significantly different for different durations. The N100 amplitude was more negative in the medium-duration condition than at the other two durations. A significant main effect of emotion was found in the short-duration condition. In accordance with previous literature [13], [28], the N100 was more negative in the happiness condition than in the anger, sadness, and surprise conditions. Overall, emotions were significantly different between duration conditions. In addition, the N100 amplitude varied by duration condition for different emotions. The N100 amplitude was the most negative with happiness prosody at short durations and with sadness prosody at long durations. This result may be due to acoustic parameters or other factors such as attention, motivation, or arousal at different durations for the same emotion.

One study reported hemisphere differences in the sensitivity to speech information, whereby the left hemisphere responded to intelligible semantic information [20]. According to the results of ROI analyses, N100 amplitudes were highest at the FC region in all three duration conditions, and the advantage of the left hemisphere over the right hemisphere only occurred in he medium-duration condition. This result departed from our hypothesis for the short-duration condition. Considering that the left and right hemispheres have different sensitivities to semantic and prosodic processing, one plausible explanation is that semantic activation of the N100 by short-speech stimuli was not strong enough to appear in the left hemisphere, while semantic processing of long-speech stimuli was beyond the brain's capacity in such a short time. Therefore, only speech stimuli in the medium-duration condition showed stronger N100 amplitudes in the left hemisphere.

B. P200: EFFECTS OF DURATION ON SALIENCE DETECTION OF EMOTIONAL INFORMATION

The P200 component is the second stage of emotional speech processing. This stage reflects the initial encoding and salience detection of the emotion, and this pathway projects from the superior temporal gyrus (STG) to the anterior superior temporal sulcus (STS) [10]. Previous studies have reported that different P200 amplitudes can reflect different emotional states, especially between emotional and neutral stimuli [13], [23]. This initial emotional encoding appears to be influenced by stimulus pitch [42], valence, and arousal [16], as well as female and male voices [24]. In this study, a significant difference in emotion was found in ROIs for the short-duration condition. Relative to surprise prosody, the P200 amplitude was more positive for happiness and anger. Through comparison of the three durations, this study found that the short-duration condition exhibited the most highly significant differences between emotions. Previous studies showed a significant difference for emotional speech of approximately 3 seconds in duration [13], [28]. Thus, the present paper conjectured that emotional speech may integrate more emotional information at short durations (200 ms) relative to medium and long durations, including acoustic features and semantic content. It may be easier to observe differences in P200 amplitude for different emotions. Conversely, it may be more difficult to observe differences for medium and long-duration stimuli because of the lower integration of emotional information. In conclusion, P200 amplitudes for three different durations exhibit significant emotion-related differences from each other. However, more significant emotion-related differences were observed for emotional speech of shorter-duration than for the other duration conditions.

The results of the ROI analysis of P200 amplitudes were similar to those of N100 amplitudes. P200 amplitudes were the lowest at the CP electrode site in comparison with FC and CN electrode sites. However, in contrast to N100 amplitudes, for P200 amplitudes, the advantage of the left hemisphere over the right hemisphere occurred not only in the medium-duration condition but also in the long-duration condition. In accordance with hemispheric lateralization, this finding

also indicated that semantic processing of speech stimuli was more evident in the second stage of processing emotional speech.

C. N300: EFFECTS OF DIFFERENT DURATIONS ON INTEGRATION AND IDENTIFICATION OF EMOTIONAL CUES

The N300 occurs at the third stage of processing emotional speech, the stage of integration and identification of semantics and prosody. Different waveforms can be obtained from different experimental methods and conditions. Based on the time the peak appears and on the positive or negative trend of the waveform, the observed ERP components at this stage include the N300/N3 [12], [13], P300 [14], N400 [43], the late positive component (LPC) [15], [16] [44] and other negative components(280–480 ms) [23], as mentioned in the literature. At this stage, significant differences between emotions, between neutral and emotional stimuli, between semantics and prosody [14], or between high and low arousal [16] can be distinguished.

In this study, the N300 amplitude was clearly observed, although this component was relatively small compared to the N100 and P200. This was particularly true at CP electrode sites, which exhibited significantly greater responses than those at FC and CN electrode sites. The main effect of emotion in the short-duration condition was more significant than those of the medium- and long-duration conditions. However, the interaction of duration and emotion was not significant. Overall, the N300 component observed in the current study may further elucidate the multi-stage process of emotional speech. Differences in emotional speech of approximately 3 seconds in duration have been reported previously [16]. However, in the current study, significant differences between emotions were only found in the short-duration condition due to the smaller N300 amplitudes. Nevertheless, as can be inferred from this study, the duration of speech stimuli can affect the N300 amplitude, and emotion-related differences in the N300 may decrease as duration increases.

D. LIMITATIONS

We used emotional speech materials from an existing database of Chinese emotional speech (TYUT2.0); these materials were excerpted from Chinese radio plays, and the semantic content of emotional speech stimuli was intelligible to participants. Future studies should explore the relative contributions of other emotional speech databases with unintelligible semantic content, such as a database of German emotional speech [45], the Persian emotional speech database (Persian ESD) [46], whose semantics would be unintelligible for Chinese participants, during processing of emotional speech with different durations. In addition, the real natural emotional speech (not the acted emotions) should also be valued. Moreover, the findings might only apply to tonal languages. In future studies, this experiment will be repeated with non-tonal languages. Furthermore, behavioral data regarding emotion recognition over a wider range of emotions should be collected in future studies to facilitate a more comprehensive understanding of the temporal course of distinct types of emotional speech and its relationship with duration. In addition, the current study cannot assert that the relationship between speech duration and ERP amplitude was reversed. Since the associations observed between N300 amplitude and speech duration did not reach significance except in FC regions, future studies are needed to confirm these associations. Moreover, it is possible that other variables, such as attention, motivation, or task parameters, can explain the observed findings. More studies are needed to address these possibilities.

V. CONCLUSION

To our knowledge, this is the first study to explore the influence of duration and emotion in tonal language on ERP components at different stages of emotional speech processing. Our findings suggest that the duration of emotional speech may impact ERP components associated with various stages of vocal emotional processing, particularly the extraction of meaning-related information from an emotional speech signal. This study suggested that short emotional speech stimuli lead to facilitated emotion as distinguished by the P200 component and the subsequent negative N300 component. Shorter-duration emotional speech stimuli were associated with more significant differences in emotional stimulation. Our findings may provide an experimental reference and partial support for the functional model of vocal emotional perception.

REFERENCES

[1] X. Y. Zhang, Y. Sun, W. Zhang, and J. Chang, "Key technologies in speech emotion recognition," *Taiyuan Li Gong Da Xue Xue Bao*, vol. 46, no. 6, pp. 629–636, Nov. 2015.
[2] H. Thönnessen, F. Boers, J. Dammers, Y. H. Chen, C. Norra, and K. Mathiak, "Early sensory encoding of affective prosody: Neuromagnetic tomography of emotional category changes," *Neuroimage*, vol. 50, no. 1, pp. 250–259, Jan. 2010.
[3] L. Jiménezortega, J. Espuny, P. H. De Tejada, C. Vargasrivero, and M. Martínloeches, "Subliminal emotional words impact syntactic processing: Evidence from performance and event-related brain potentials," *Frontiers Human Neurosci.*, vol. 11, no. 4, p. 192, Apr. 2017.
[4] S. A. Kotz and S. Paulmann, "When emotional prosody and semantics dance cheek to cheek: ERP evidence," *Brain Res.*, vol. 151, no. 1, pp. 107–118, Jun. 2007.
[5] K. Lakshminarayanan, S. D. Ben, V. van Wassenhove, D. Orbelo, J. Houde, and D. Poeppel, "The effect of spectral manipulations on the identification of affective and linguistic prosody," *Brain Lang.*, vol. 84, no. 2, pp. 250–263, Feb. 2003.
[6] D. Grasu. (Dec. 22, 2015). *Tonal vs. Non-Tonal Languages: Chinese vs. English*. [Online]. Available: http://www.lexington.ro/en/blog/itemlist/user/47-dianagrasu.html?start=5
[7] A. Li, Q. Fang, and J. Dang, "Emotional intonation in a tone language: Experimental evidence from Chinese," in *Proc. ICPHS*, 2011, pp. 17–21.
[8] H. Luo, A. Boemio, M. Gordon, and D. Poeppel, "The perception of FM sweeps by Chinese and English listeners," *Hear Res.*, vol. 224, nos. 1–2, pp. 75–83, Feb. 2007.
[9] D. Klein, R. J. Zatorre, B. Milner, and V. Zhao, "A Cross-linguistic PET study of tone perception in mandarin Chinese and English speakers," *Neuroimage*, vol. 13, no. 4, pp. 646–653, Apr. 2001.
[10] S. Annett and S. A. Kotz, "Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing," *Trends Cognit. Sci.*, vol. 10, no. 1, pp. 24–30, Jan. 2006.
[11] S. A. Kotz and S. Paulmann, "Emotion, language, and the brain," *Lang. Linguist Compass*, vol. 5, no. 3, pp. 108–125, Mar. 2011.

[12] V. Bostanov and B. Kotchoubey, "Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations," *Psychophysiology*, vol. 41, no. 2, pp. 259–268, Mar. 2004.
[13] J. M. Iredale, J. A. Rushby, S. McDonald, A. Dimoska-Di Marco, and J. Swift, "Emotion in voice matters: Neural correlates of emotional prosody perception," *Int. J. Psychophysiol.*, vol. 89, no. 3, pp. 483–490, Sep. 2013.
[14] I. J. Wambacq and J. F. Jerger, "Processing of affective prosody and lexical-semantics in spoken utterances as differentiated by event-related potentials," *Brain Res. Cognit. Brain Res.*, vol. 20, no. 3, pp. 427–437, Mar. 2004.
[15] A. Schirmer, C. B. Chen, A. Ching, L. Tan, and R. Y. Hong, "Vocal emotions influence verbal memory: Neural correlates and interindividual differences," *Cognit. Affect. Behav. Neurosci.*, vol. 13, no. 1, pp. 80–93, Jan. 2013.
[16] S. Paulmann, M. Bleichner, and S. A. Kotz, "Valence, arousal, and task effects in emotional prosody processing," *Frontiers Psychol.*, vol. 4, no. 4, p. 345, Jun. 2013.
[17] A. Schirmer, "Timing speech: A review of lesion and neuroimaging findings," *Brain Res. Cognit. Brain Res.*, vol. 21, no. 2, pp. 269–287, Oct. 2004.
[18] A. P. Pinheiro, S. Galdo-Álvarez, A. Rauber, A. Sampaio, M. Niznikiewicz, and O. F. Gonçalves, "Abnormal processing of emotional prosody in Williams syndrome: An event-related potentials study," *Res. Device Disabil.*, vol. 32, no. 1, pp. 47–133, Jan. 2010.
[19] G. Pourtois, D. B. Gelder, J. Vroomen, B. Rossion, and M. Crommelinck, "The time-course of intermodal binding between seeing and hearing affective information," *Neuroreport*, vol. 11, no. 6, pp. 1329–1333, Apr. 2000.
[20] S. K. Scott, C. C. Blank, S. Rosen, and R. J. Wise, "Identification of a pathway for intelligible speech in the left temporal lobe," *Brain*, vol. 123, no. 12, pp. 2400–2406, 2000.
[21] E. S. Dmitrieva and V. Y. Gel, "Man, K. A. Zaitseva, and A. M. Orlov, "Perception of the emotional component of vocal signals at different durations of the stimulus," *Fiziol Cheloveka*, vol. 32, no. 5, pp. 36–40, May 2005.
[22] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS ONE*, vol. 6, no. 11, p. e27256, Nov. 2011.
[23] S. Paulmann, D. V. M. Ott, and S. A. Kotz, "Emotional speech perception unfolding in time: The role of the basal ganglia," *PLoS ONE*, vol. 6, no. 3, pp. 1451–1454, Mar. 2011.
[24] S. Paulmann and S. A. Kotz, "Early emotional prosody perception based on different speaker voices," *Neuroreport*, vol. 19, no. 2, pp. 209–213, Jan. 2008.
[25] L. Sokka *et al.*, "Alterations in attention capture to auditory emotional stimuli in job burnout: An event-related potential study," *Int. J. Psychophysiol.*, vol. 94, no. 3, pp. 427–436, Dec. 2014.
[26] S. Paulmann and S. A. Kotz, "An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context," *Brain Lang.*, vol. 105, no. 1, pp. 59–69, Apr. 2008.
[27] I. J. Wambacq, K. J. Shea-Miller, and A. Abuhuziefa, "Non-voluntary and voluntary processing of emotional prosody: An event-related potentials study," *Neuroreport*, vol. 15, no. 3, pp. 555–559, Apr. 2004.
[28] A. P. Pinheiro, M. Vasconcelos, M. Dias, N. Arrais, and O. F. Goncalves, "The music of language: An ERP investigation of the effects of musical training on emotional prosody processing," *Brain Lang.*, vol. 140, no. 1, pp. 24–34, Jan. 2015.
[29] D. Agrawal *et al.*, "Electrophysiological responses to emotional prosody perception in cochlear implant users," *Neuroimage Clin.*, vol. 2, no. 1, pp. 229–238, Jan. 2013.
[30] S. Paulmann, M. D. Pell, and S. A. Kotz, "How aging affects the recognition of emotional speech," *Brain Lang.*, vol. 104, no. 3, pp. 262–269, Mar. 2008.
[31] J. Song, X. Y. Zhang, Y. Sun, and J. Chang, "Establishment of emotional speech database based onfuzzy comprehensive evaluation method," *Modern Electron. Techn.*, vol. 39, no. 13, pp. 51–54, Jul. 2016.
[32] T. Crook, *Radio Drama Theory and Practice*. London, U.K.: Routledge, 1999.
[33] D. M. P. Das, "Radio Play: The kolkata story," *Global Media J. Indian Ed.*, vol. 4, no. 2, p. 1, Dec. 2013.
[34] J. Chang, X. Y. Zhang, Q. P. Zhang, H. T. Chen, Y. Sun, and F. Y. Hu, "ERP research on the emotional voice for different languages and non-speech utterances," *J. Tsinghuauniv*, vol. 12, pp. 1131–1136, Dec. 2016.
[35] S. Pakarinen, L. Sokka, M. Leinikka, A. Henelius, J. Korpela, and M. Huotilainen, "Fast determination of MMN and P3a responses to linguistically and emotionally relevant changes in pseudoword stimuli," *Neurosci. Lett.*, vol. 577, pp. 28–33, Aug. 2014.
[36] A. P. Pinheiro, M. E. Dajer, A. Hachiya, A. N. Montagnoli, and D. Tsuji, "Graphical evaluation of vocal fold vibratory patterns by high-speed video laryngoscopy," *J. Voice*, vol. 28, no. 1, pp. 106–111, Jan. 2014.
[37] A. P. Pinheiro *et al.*, "Sensory-based and higher-order operations contribute to abnormal emotional prosody processing in schizophrenia: An electrophysiological investigation," *Psychol. Med.*, vol. 43, no. 3, pp. 603–618, Mar. 2013.
[38] A. Schirmer and S. Kotz, "ERP evidence for a sex-specific stroop effect in emotional speech," *J. Cognit. Neurosci.*, vol. 15, no. 8, pp. 1135–1148, Nov. 2003.
[39] T. Rosburg, N. N. Boutros, and J. M. Ford, "Reduced auditory evoked potential component N100 in schizophrenia—A critical review," *Psychiatry Res*, vol. 161, no. 3, pp. 259–274, Oct. 2008.
[40] A. Seither-Preisler, R. Patterson, K. Krumbholz, S. Seither, and B. Lütkenhöner, "Evidence of pitch processing in the N100m component of the auditory evoked field," *Hearing Res.*, vol. 213, nos. 1–2, pp. 88–98, Mar. 2006.
[41] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, "Electrical signs of selective attention in the human brain," *Science*, vol. 182, no. 4108, pp. 80–177, Oct. 1973.
[42] C. Pantev, L. T. Roberts, B. Ross, and C. Wienbruch, "Tonotopic organization of the sources of human auditory steady-state responses," *Hearing Res.*, vol. 101, nos. 1–2, pp. 62–74, Nov. 1996.
[43] L. Rohr and R. A. Rahman, "Affective responses to emotional words are boosted in communicative situations," *NeuroImage*, vol. 109, pp. 273–282, Apr. 2015.
[44] J. J. Stekelenburg and J. Vroomen, "Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events," *Frontiers Integr. Neurosci.*, vol. 6, no. 6, p. 25, May 2012.
[45] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 1517–1520.
[46] N. Keshtiari, M. Kuhlmann, M. Eslami, and G. Klanndelius, "Recognizing emotional speech in Persian: A validated database of Persian emotional speech (Persian ESD)," *Behav. Res. Methods*, vol. 47, no. 1, p. 295, Jan. 2015.

JIANG CHANG is currently pursuing the Ph.D. degree with the College of Information Engineering, Taiyuan University of Technology, China. Her current research interests include speech signal processing and event-related potential analysis.

XUEYING ZHANG received the Ph.D. degree in underwater acoustic engineering from Harbin Engineering University in 1998. She is currently a Professor with the Taiyuan University of Technology. She supervised a lot of national, province, or ministry research projects. She has published over 100 papers in the periodicals of national level and international conferences and about 40 papers were indexed by SCI, EI, or ISTP. Her research interests include auditory model, emotional speech recognition, and machine learning.

QIPING ZHANG received the Ph.D. degree in information science from the University of Michigan in 2012. She is currently a Professor with the Hundred Talent Program, Taiyuan University of Technology, and an Associate Professor with the Palmer School of Library and Information Science, Long Island University. Her research interests include cognitive psychology and human–computer interaction. She is a member of the ACM.

YING SUN received the Ph.D. degree in engineering from the Taiyuan University of Technology, Taiyuan, China. She is currently a Lecturer with the Taiyuan University of Technology. She supervised two of the province research projects. She has published over 10 papers in the periodicals of national level and international conferences and about four papers were indexed by SCI, EI, or ISTP. Her research interests include speech signal processing and speech emotion recognition.